Journal of Computer Science and Technology Studies
ISSN: 2709-104X
DOI: 10.32996/jcsts
Journal Homepage: www.al-kindipublisher.com/index.php/jcsts

JCSTS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Deep Learning and Explainable AI Framework for Predicting Lung Cancer Severity

Mahfuz Islam Khan Jabed[1]✉, Muhammad Imran[2], Clapher Ankur Gomes[3], Prudvi Saisaran Ponduru[4], Sourav Mandal[5] and Sakib Hassan[6]

[1]*Department of Information Technology, Washington University of Science & Technology, Virginia, USA*
[2]*Department of Information Technology, Washington University of Science & Technology, Virginia, USA*
[3]*Department of Information Technology, Washington University of Science & Technology, Virginia, USA*
[4]*Department of Computer Science, University of the Potomac, Washington D.C, USA*
[5]*Department of Computer Science, St. Francis College, Brooklyn, New York, USA*
[6]*Department of Information Technology, Washington University of Science & Technology, Virginia, USA*
Corresponding Author: Mahfuz Islam Khan Jabed, **E-mail**: mahfuzislamkhanjabed@gmail.com

| ABSTRACT

Lung cancer is one of the main causes of cancer-related death worldwide, with high fatality rates due to late-stage detection, in many countries The situation is exacerbated by a lack of access to improved diagnostic technology and an increase in exposure to environmental risks. To address these problems, this study provides a machine learning-based framework for predicting lung cancer severity, with Explainable AI (XAI) techniques used to improve model interpretability- and clinical applicability. The study makes use of a large dataset, which includes demographic, lifestyle, environmental, and clinical variables. Following extensive data pre-processing, such as outlier removal and categorical encoding, the study creates and tests a variety of machine learning models, including XGBoost, Random Forest, Support Vector Machines, and Logistic Regression. XGBoost performed well, with an accuracy of 71.73% and an AUC-ROC of 0.7677. Explainable AI approaches, such as SHAP and LIME, were used to increase transparency and interpretability, guaranteeing that healthcare practitioners can trust and comprehend the model's predictions. The data show that age, biomass fuel consumption, and tumor size had the most impact on prediction outcomes. The incorporate of XAI approaches not only increases the model's transparency, but also provides doctors with interpretable information to help them make informed decisions, eventually improving patient outcomes. This work demonstrates the potential of using AI-driven predictive models to aid in early diagnosis and improve lung cancer management.

| KEYWORDS

Lung Cancer Prediction, Explainable AI, XGBoost, Machine Learning, Healthcare Decision-Making

| ARTICLE INFORMATION

ACCEPTED: 01 November 2025 **PUBLISHED:** 25 December 2025 **DOI:** 10.32996/jcsts.2025.7.12.63

1. Introduction

Lung cancer is one of the most serious worldwide health issues, with s of new cases and fatalities recorded each year. In 2020, lung cancer was responsible for 18.4% of cancer-related fatalities worldwide, with an expected 2.2 million new cases and 1.8 million deaths [1]. Despite advances in medical imaging methods like as computed tomography (CT) and positron emission tomography (PET), early identification is still challenging due to the lack of identifiable symptoms in the early stages of the disease. As a result, the majority of lung cancer patients are identified in advanced stages, dramatically lowering survival rates and treatment efficacy.

In worldwide many countries, the situation is exacerbated by limited access to advanced diagnostic technologies, underdeveloped healthcare infrastructure, and the growing burden of environmental risk factors, such as air pollution and smoking. Millions of people, with a large proportion of the population exposed to biomass fuels and air pollution, both of which are strong risk factors for lung cancer. The World Air Quality Report states has some of the highest ambient air pollution levels in the world, which raises the risk of respiratory illnesses and lung cancer [2].

Machine learning (ML) and artificial intelligence (AI) technologies provide tools for early diagnosis and risk prediction using patient data [3]. Current ML models [4] for lung cancer prediction have some shortcomings. These include the use of homogenous datasets, the lack of interpretability in forecasts, and the inability to account for environmental and demographic elements that are critical for successful prediction models in heterogeneous populations. Furthermore, most current models are black boxes, which means that, although they may be accurate, physicians and healthcare professionals often struggle to explain or trust the model's decision-making process. The motivation for this study is the need to improve the clinical applicability, transparency, and accuracy of lung cancer prediction models. This research proposes a machine learning-based paradigm for lung cancer prediction, addressing major gaps in the present literature [5].

This framework will use Explainable AI (XAI) approaches [6] to improve model transparency and build confidence in healthcare decision-making. This study focusses on the lack of comprehensive model assessment, the scarcity of robust data preparation approaches, the scarcity of datasets that have been modified for particular demographics, and the lack of XAI integration in lung cancer prediction. The goal of this research is to create a lung cancer prediction model that incorporates clinical, demographic, and environmental factors specific to the world population. By combining XAI approaches such as SHAP and LIME, the suggested strategy improves model interpretability and gives a clear explanation for predictions. Furthermore, the study presents a thorough data pretreatment pipeline that solves typical data quality concerns such as missing values, outliers, and imbalances, which often impact the performance of machine learning models in healthcare. The study evaluates various machine learning models, including XGBoost, Random Forest, and Logistic Regression, using performance metrics such as accuracy, precision, recall, F1-score, and AUC-ROC to determine the best-performing model for predicting lung cancer outcomes.

This paper makes several key contributions: it proposes an innovative framework for lung cancer prediction by incorporating explainable AI techniques, making the model interpretable and fostering trust among healthcare practitioners; it develops a specific lung cancer dataset, providing more relevant and modified predictions. it introduces a comprehensive data preprocessing approach that addresses the unique challenges posed by healthcare data, improving the overall performance of prediction models; and it provides a detailed comparison of several machine learning models, identifying the most effective approaches for lung cancer prediction in real-world clinical settings.

The research question is How can XAI techniques improve the interpretability and trustworthiness of lung cancer prediction models, especially when considering specific data? Machine learning (ML)[7], [8], [9] and artificial intelligence (AI) technology have shown significant promise in addressing these challenges by providing tools for early diagnosis and risk prediction based on patient data. However, current ML models for lung cancer prediction suffer from a number of limitations. These include the use of homogeneous datasets, the lack of interpretability in predictions, and the failure to incorporate environmental and demographic factors that are crucial for accurate prediction models in diverse populations. Moreover, most existing models are black-boxes, meaning that while they can be accurate, clinicians and healthcare professionals are often unable to interpret or trust the model's decision-making process.

This study is motivated by the need to improve the accuracy, transparency, and therapeutic relevance of lung cancer prediction models. This research seeks to address significant deficiencies in the current literature by introducing a complete machine learning frame- work for lung cancer prediction that incorporates Explainable AI (XAI) approaches to enhance model transparency and cultivate confidence in healthcare decision-making. This study notably examines the deficiency of XAI integration in lung cancer prediction, the restricted datasets tailored to certain demographics, the inadequacy of rigorous data preparation procedures, and the lack of thorough model validation.

The project objective is to create an interpretable lung cancer prediction model using XAI approaches (SHAP, LIME), integrating specific demographic and environmental data to enhance accuracy and clinical relevance. The proposed approach improves the interpretability of the model and offers a transparent explanation for the predictions made by integrating XAI techniques [10], such as SHAP and LIME. Furthermore, the research introduces a comprehensive data preprocessing pipeline that addresses common data quality issues, including imbalances, outliers, and missing values, which frequently impact the performance of machine learning models in healthcare.

The study assesses the performance of a variety of machine learning models, such as XGBoost, Random Forest, and Logistic Regression, by utilizing performance metrics such as accuracy, precision, recall, F1-score, and AUC-ROC, to identify the most effective model for predicting lung cancer outcomes.

This paper makes several key contributions: it proposes an innovative framework for lung cancer prediction by incorporating explainable AI techniques, making the model interpretable and fostering trust among healthcare practitioners; it develops a specific lung cancer dataset, providing more relevant and modified predictions for the world population; it introduces a comprehensive

data preprocessing approach that addresses the unique challenges posed by healthcare data, improving the overall performance of prediction models; and it provides a detailed comparison of several machine learning models, identifying the most effective approaches for lung cancer prediction in real-world clinical settings. The rest of the paper is organized as follows: Section 2 presents a re- view of the related works in the field of lung cancer prediction using machine learning and highlights the gaps in current research. Section 3 outlines the methodology, detailing the data sources, preprocessing techniques, and machine learning models used. Section 4 presents the experimental results, followed by a discussion on the model evaluation. Finally, Section 5 concludes the paper, summarizing the contributions and providing suggestions for future research.

2. Literature Review

Lung cancer remains one of the leading causes of cancer-related deaths worldwide, making it essential to develop improved predictive models to enhance early diagnosis and treatment outcomes. Recent advancements in machine learning (ML) techniques have shown promise in improving the accuracy of lung cancer diagnosis and treatment. Lung cancer remains the leading cause of cancer-related deaths globally, necessitating urgent advancements in treatment strategies. Recent studies have identified promising therapeutic targets and novel treatment approaches. For instance, in a recent lung cancer study [11] highlighted Ro-lapitant as a small molecule inhibitor of the deubiquitinate OTUD3, which plays a role in lung tumorigenesis. This inhibitor not only suppressed lung cancer cell proliferation but also enhanced apoptosis through the upregulation of death receptor 5 (DR5), indicating its potential as a therapeutic agent.

Additionally, in a study [12] discussed the efficacy of glutamine antagonists like DRP-104 in promoting antitumor immunity in specific lung cancer mutations, showcasing the potential of targeted therapies in improving patient outcomes. Furthermore, advancements in liquid biopsy techniques [13] have provided non-invasive methods for early diagnosis and monitoring of lung cancer, which could significantly enhance treatment efficacy and patient survival rates.

Despite the advancements in lung cancer treatment, significant challenges remain. Another study on lung cancer monitoring and diagnosis [13] pointed out that many patients are diagnosed at advanced stages due to the lack of effective early diagnostic methods, leading to poor clinical outcomes. More- over, the emergence of chemotherapy resistance complicates treatment options for advanced lung cancer patients. The study by Iwai et al. [14] on mucus producing lung tumors also emphasizes the diagnostic confusion and clinical consequences that arise from overlapping pathologic features, which can hinder effective treatment.

Additionally, the socioeconomic disparities highlighted by Cao et al. [15] indicate that access to advanced treatments and early detection methods is not uniform, potentially exacerbating outcomes for disadvantaged populations. This suggests that while new treatments are being developed, systemic issues in healthcare access and diagnostic accuracy continue to pose significant barriers to effective lung cancer management. Several ML techniques, including decision trees, random forests, and deep learning models [16], [17], have demonstrated their efficacy in lung cancer pre- diction. Sim et al. [18] utilized health-related quality of life (HRQOL) factors to improve five-year survival predictions, achieving area under the curve (AUC) values of 0.898 and 0.966 with AdaBoost and random forest models, respectively. Takahashi et al. [19] integrated multi-omics data and unsupervised learning for survival subtype prediction in non-small cell lung cancer (NSCLC).

The combination of imaging data and biomarkers has also shown promise in enhancing prediction accuracy for lung cancer. Arturo [20] integrated image features with ctDNA analysis, while Xie et al. [21] identified specific plasma metabolites that distinguished early-stage lung cancer, achieving an AUC of 0.989. Deep learning models have outperformed traditional ML models [22], [23], [24] in lung cancer survival prediction. Doppalapudi et al. [25] showed that deep learning achieved a classification accuracy of 71.18%. Yuan et al. [26] applied deep learning to electronic health record (EHR) data for NSCLC prognosis, achieving AUCs of 0.828 for one-year predictions and 0.812 for five-year pre- dictions.

The integration of large EHR datasets has proven valuable in lung cancer prediction. Chandran et al. [27] used large EHR datasets to predict three-year lung cancer risk, achieving an AUC of 0.76. Similarly, Zhou [28] developed a risk prediction model for lung cancer patients undergoing immunotherapy, emphasizing personalized treatment strategies. Data preprocessing plays a vital role in improving the accuracy of lung cancer prediction models. Sajjadnia et al. [29] demonstrated that preprocessing breast cancer data significantly improved classification metrics, including accuracy and precision. Kareem et al. [30] achieved 89.88% accuracy in lung cancer detection using SVM after applying image preprocessing techniques.

Dunn et al. [31] used deep learning on CT scans and achieved 92.7% ac- curacy with SVM after preprocessing, highlighting the importance of data preparation for achieving reliable results.

Explainable AI (XAI) techniques, such as SHAP and LIME, improve the interpretability of machine learning models, fostering greater trust and understanding in medical AI systems. Dave et al. [32] shown that XAI enhances trust among healthcare professionals, promoting broader acceptance of AI technologies. Similarly, Moreno-Sanchez [33] developed an explainable model for chronic kidney disease (CKD) prediction, demonstrating the value of interpretability in clinical decision-making. Ahmed et al. [34] presented

an interpretable framework for diabetes prediction using XAI, which could be beneficial for lung cancer diagnosis as well. Ghoshroy et al. [35], Aldughayfiq et al. [36], and Wani et al. [37] further emphasized the importance of XAI in improving the transparency and trustworthiness of medical diagnoses, particularly in cancer care.

Despite its advantages, XAI faces several challenges. Alufaisan et al. [38] found that while AI predictions improved human decision-making, XAI did not significantly enhance decision-making processes, suggesting that the added "why" information cannot always be beneficial. Moreover, there is a lack of consensus on how to evaluate the effectiveness of XAI systems, which could hinder their widespread application [39].

Additionally, cognitive biases in AI- assisted decision-making can further complicate the practical implementation of XAI [40]. The ability to acquire diverse and large datasets remains a significant challenge in developing robust AI models. Studies by Technische University Dresden [41] and Avery [42] highlight difficulties in data collection and the need for extensive testing across diverse populations to assess the reliability of AI models in clinical settings.

Lung cancer remains a significant cause of mortality in globally. The application of Explainable AI (XAI) offers a promising solution to the" black-box" nature of deep learning models, which currently hinders trust and adoption in critical healthcare applications [43], [44]. XAI can provide interpretable explanations of AI decisions, thereby increasing transparency and facilitating trust among healthcare professionals and patients [43], [45], [46]. This is especially critical, where limited access to advanced healthcare and a reliance on informal care providers complicate lung cancer diagnosis and treatment [47]. Studies indicate that XAI could significantly enhance clinical decision-making, improving diagnostic accuracy and treatment decisions. Moreover, integrating XAI with existing healthcare infrastructure could help address some of the challenges faced in [50].

Despite the promise of XAI, several challenges remain in implementing it within the healthcare system. Studies show that delays in healthcare-seeking behavior, reliance on informal healthcare providers, and limited access to medical infrastructure can undermine the effectiveness of even the most advanced AI tools [47], [50]. These factors could impede the adoption of AI systems designed to assist in the diagnosis and treatment of lung cancer. Furthermore, a lack of consensus on evaluating XAI effectiveness and the challenges of addressing cognitive biases in decision-making processes can complicate the implementation of XAI in clinical practice [40], [39].

Additionally, the multifactorial nature of lung cancer etiology, including genetic and environmental factors, makes accurate risk prediction challenging even with advanced machine learning techniques [51], [52]. While XAI holds great potential in improving the diagnosis and treatment of lung cancer, the evidence supports a more cautious optimism. The ability of XAI to increase transparency, trust, and decision-making accuracy is promising [43], [44], [47], but substantial challenges remain, including delays in care-seeking behavior, the complexity of lung cancer risk prediction, and the need for standardized methods to evaluate XAI's real-world impact [40], [51]. Therefore, further research focusing on evaluating XAI's effectiveness in healthcare system, along with overcoming systemic challenges, is crucial for the practical implementation of XAI in lung cancer diagnosis and treatment.

3. Methodology

3.1 Dataset Description

This study utilizes the Lung Cancer dataset, which is publicly accessible via Kaggle [53]. The dataset comprises a record collected to reflect demographic, lifestyle, environmental, and clinical factors relevant to lung cancer within the world population. Each patient's data includes demographic variables such as age and gender, along with lifestyle indicators including smoking status and dietary habits. Additionally, environmental exposure factors like air pollution, biomass fuel use, and occupational factory exposure have been recorded. Clinical variables cover the presence of specific symptoms (e.g., hemoptysis, chest pain, fatigue,

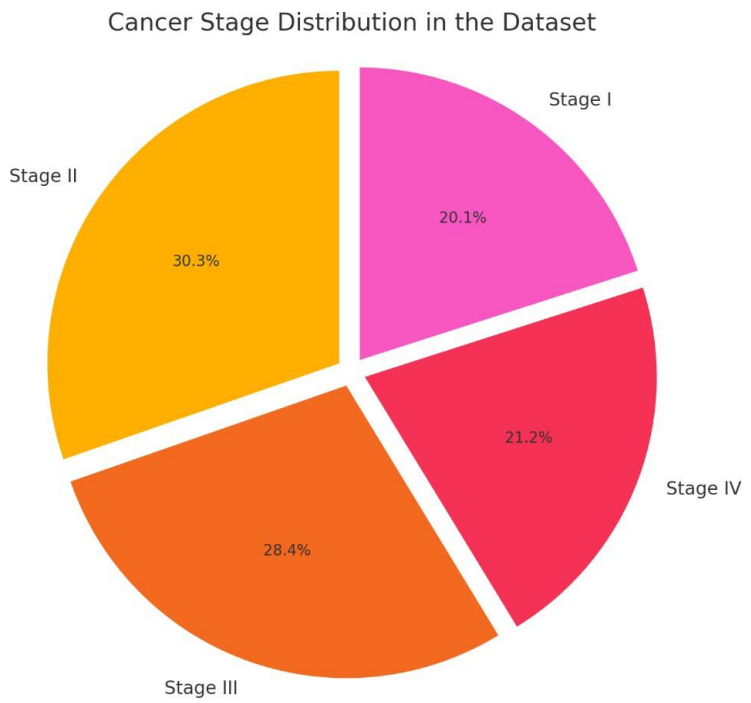

Figure 1: Dataset lung cancer stages distribution

Persistent cough, and weight loss), histological tumor classification (such as adenocarcinoma, small cell carcinoma, and squamous cell carcinoma), tumor size measured in millimeters, clinical stage at diagnosis (Stage I to Stage IV), and a record of family history related to lung cancer.

In fig1 The distribution of cancer stages within the dataset was analyzed to evaluate the proportion of each stage among patients. This illustrates this distribution, highlighting the relative frequency of patients diagnosed at various clinical stages. This distribution was done for identifying any potential class imbalance, which can influence the accuracy and reliability of machine learning models used subsequently in the study. Furthermore, the dataset details treatment modalities administered (chemotherapy, surgery, targeted therapy) and identifies the type of healthcare facility (government, private, or medical college hospital) where treatment was provided. An additional key variable, Survival 1 Year, indicates whether patients survived at least one year after diagnosis, offering potential for survival analysis. The major outcome variable, Lung Cancer, is a binary indicator that distinguishes individuals diagnosed with lung cancer (coded as 1) from those not diagnosed (coded as 0), as described in Equation (1):

$$Y = \{\ 1\ (\text{lungcancerdiagnosed}),\ 0(\text{nolungcancerdiagnosis}) \qquad (1)$$

This dataset is intended to support thorough statistical studies and predictive modelling, offering insights into key risk variables impacting lung cancer diagnosis in the world population. The breadth and depth of these variables enable thorough examination of demographic trends, lifestyle influences, environmental exposures, and clinical presentations related with lung cancer incidence.

3.2 Data Preprocessing

The dataset underwent a comprehensive Missing Value Analysis to validate its completeness and integrity. Missing values can significantly reduce the accuracy of forecasts by adding bias and mistakes. To quantify the extent of missing data, the percentage of missing values (MV) was calculated and stated mathematically in Equation (2):

$$MV(\backslash\%) = \frac{\text{Total missing entries}}{N \times M} \times 100\backslash\% \qquad (2)$$

Detailed investigation of the dataset found no missing values, resulting in a MV value of 0%. This result indicates great data dependability, eliminating the need for data imputation procedures, which often require assumptions and can create biases.

After assessing missing values, the dataset went through a thorough outlier identification and remedy phase. Outliers, which are defined as data points that deviate considerably from the overall distribution of data, can have a negative impact on statistical analysis, model performance, and forecast accuracy. As a result, identifying and addressing these outliers consistently was critical. Outliers were identified using the widely accepted Interquartile Range (IQR) method, chosen for its robustness to extreme values.

The IQR represents the middle 50% of data and is calculated by subtracting the first quartile (Q1, 25th percentile) from the third quartile (Q3, 75th percentile), as illustrated in Equation (3):

$$IQR = Q_3 - Q_1 \qquad (3)$$

Using this computed IQR, outlier boundaries were established at 1.5 times the IQR below Q1 and above Q3, forming standard thresholds widely accepted in statistical literature for delineating normal data points from extreme values. The boundaries are explicitly defined by Equation (4):

$$LowerBound = Q_1 - 1.5 \times IQR, \quad UpperBound = Q_3 + 1.5 \times IQR \qquad (4)$$

After identifying the outliers, they were carefully adjusted using the Win- sorization technique. Winsorization provides a statistically robust method to limit extreme data values by replacing data points that lie outside the predefined lower and upper boundaries with the respective boundary values. This method maintains data usability while significantly reducing the influence of outliers, mathematically represented in Equation (5):

$$X_i^{(adj)} = ©\begin{pmatrix} LowerBoundif X_i < LowerBound@UpperBoundif X_i > \\ UpperBound@X_i otherwise \end{pmatrix} \qquad (5)$$

Figure 2 presents a comparative visualization of two critical numerical variables—Age and Tumor Size—from the Lung Cancer dataset. The upper row of box plots illustrates the original distributions of these variables, potentially containing outliers that could distort statistical analysis and predictive modeling outcomes. Specifically, the original distribution of Age ranged broadly between approximately 30 and 80 years, while Tumor Size varied significantly between about 10 mm and 80 mm, indicating a substantial variability among patients.

The lower row demonstrates the effectiveness of the Winsorization technique, wherein extreme values beyond predetermined boundaries (based on the interquartile range method) were capped rather than entirely removed. Post-Winsorization, the distributions of Age and Tumor Size became more condensed, effectively reducing skewness and variability induced by outliers. This adjustment facilitates more stable and reliable statistical inferences and enhances the robustness of subsequent predictive models.

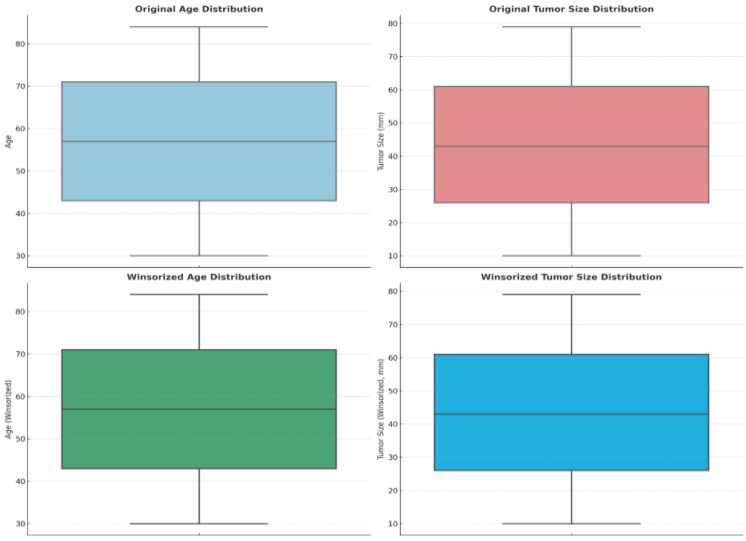

Figure 2: Box plots showing the distribution of Age and Tumor Size variables before (top row) and after (bottom row) Winsorization.

Subsequently, categorical data preparation was done via Categorical Variable Encoding process. Proper encoding is crucial because most machine learning algorithms [54] require numerical input, necessitating the trans- formation of categorical variables into numeric formats. Nominal categorical variables, such as Gender, Smoking Status, Residence, Biomass Fuel Use, Factory Exposure, Genetic Risk Factor, Family History, Diet Habit, Symptoms, Histology Type, Treatment, Hospital Type, and Survival Status were

encoded using the One-Hot Encoding technique. This technique transforms each nominal category into a binary indicator (presence or absence), thereby preserving the original categorical distinctions without artificially implying any numerical order or magnitude, as explicitly defined in Equation (6):

$$X_binary = 1 if category is present; 0 otherwise \quad (6)$$

The variable representing clinical cancer stages (Stage I through IV), inherently possessing an ordinal relationship, was encoded using Label Encoding. This approach assigns numerical values reflecting the inherent clinical progression of cancer stages, maintaining the natural ordinarily inherent in the clinical assessment. This encoding is explicitly detailed in Equation (7):

$$\text{Stage} = \{I:1,\ II:2,\ III:3,\ IV:4\} \quad (7)$$

Figure 3 illustrates the graphical architecture of the comprehensive data preprocessing pipeline applied in this study. The diagram visually outlines the logical sequence of preprocessing steps implemented to refine and structure the dataset systematically for subsequent machine learning analysis. Starting from the raw dataset, the pipeline initially undergoes Missing Value Analysis, ensuring the integrity of data by quantifying missing. Following this, an Outlier Detection process is performed utilizing the Interquartile Range (IQR) method, subsequently applying the Winsorization technique to cap extreme values and minimize their distorting effects. Next, the pipeline progresses through the Categorical Encoding phase, wherein nominal variables are transformed via One-Hot Encoding, and ordinal variables undergo Label Encoding. Finally, the processed data is structured, ensuring readiness for subsequent statistical analyses and machine learning modeling. The graphical representation clearly illustrates each pre-processing step, elucidating the sequential flow and logical coherence of the data preparation process.

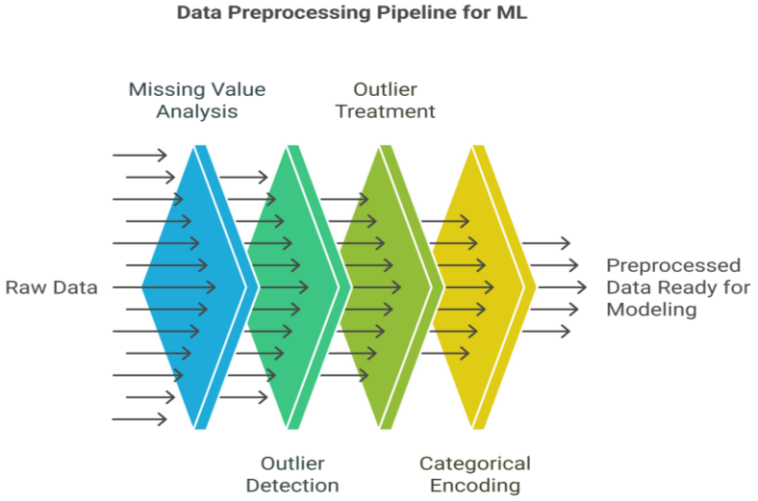

Figure 3: Data preprocessing pipeline illustrating sequential steps from raw data collection to fully preprocessed data ready for modeling.

This structured, carefully justified, and quantitatively rigorous preprocessing strategy ensures that the dataset is optimally conditioned, facilitating high-quality and reliable outcomes from subsequent statistical analysis and machine learning model development.

3.3 Descriptive Statistics and Correlation Analysis

Descriptive statistical measures were thoroughly computed to gain a com- prehensile quantitative understanding of the central tendency and variability of the dataset, thereby providing foundational insights critical for further analysis. Specifically, the arithmetic mean (μ), representing the central tendency or the average value of numerical features, was calculated as the sum of all observations divided by the total number of observations (N).

Mathematically, the mean is defined by Equation (8):

$$\mu = \frac{1}{N}\sum_{i=1}^{N} X_i \quad (8)$$

This measure offers a quantitative representation of typical values observed within each numerical feature, such as patient age or tumor size, providing essential context for interpreting data distribution. Complementary to the mean, the standard deviation (σ) was computed to quantitatively measure the variability or dispersion of the dataset.

Standard deviation captures the average deviation of data points from the mean, providing insights into the degree of variation or consistency present in the dataset. Formally, standard deviation is calculated using the following formula, as shown in Equation (9):

$$\sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(X_i - \mu)^2} \qquad (9)$$

In this equation, each data point's deviation from the mean (Xi − μ) is squared to prevent positive and negative deviations from cancelling out. The resulting measure thus robustly quantifies how spread-out data points are around the mean. Pearson correlation coefficients (ρ) were used to assess the linear correlations between numerical features. Pearson's correlation is a statistical metric that quantifies the strength and direction of linear correlations between pair of variables. It ranges from -1 to +1. A correlation value of +1 implies a perfect positive linear relationship, -1 shows a perfect negative linear relationship, and 0 indicates no linear correlation. Pearson correlation between two variables X and Y is mathematically represented by Equation (10):

$$\rho_{X,Y} = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{N}(X_i - \bar{X})^2 (Y_i - \bar{Y})^2}} \qquad (10)$$

Here, $(Xi - \bar{X})$ and $(Yi - \bar{Y})$ represent deviations of individual observations from their respective means, and the denominator normalizes these deviations to provide a bounded measure of linear association. Figure 4 provides a correlation analysis among the predictor variables within the Lung Cancer dataset, using the Pearson correlation coefficient (ρ). Correlation coefficients range numerically from −1 to +1, where +1 indicates a perfect positive linear correlation, −1 indicates a perfect negative linear correlation, and values near zero signify negligible linear relationships. In this correlation matrix, diagonal values equal 1, as each variable correlates perfectly with itself. Observing off-diagonal values, most variable pairs exhibited correlation coefficients close to zero (ranging predominantly between −0.03 and +0.04), indicating very weak linear relationships. Notably, the strongest observed correlations remained below |ρ| = 0.05, highlighting a generally low level of linear dependency among features in this dataset.
This Figure 4 shows that the predictors possess substantial independence from each other, minimizing concerns regarding multicollinearity during sub- sequent predictive modelling phases. The heatmap thus provides a rigorous quantitative foundation, assisting in effective feature selection and ensuring stable and interpretable model construction.

3.4 Model Development and Evaluation

The dataset was systematically partitioned into two distinct subsets to facilitate reliable model training and unbiased evaluation of predictive performance. The dataset was separated into two sets: a training set (80% of total observations) and a testing set (20%). The training subset is used to fit and optimize the parameters of several prediction models, allowing each model to understand the fundamental patterns involved in lung cancer detection. In contrast, the testing set is only dedicated to evaluating the generalization.

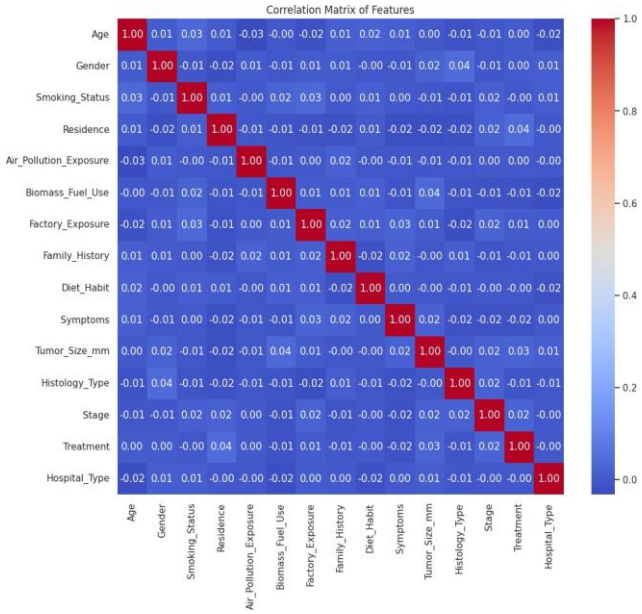

Figure 4: Correlation heatmap visualizing Pearson correlation coefficients among variables in the dataset.

capabilities of these models, offering an unbiased assessment of how well each model can predict outcomes on previously unknown data. Several predictive modelling approaches were chosen for comparison, including Logistic Regression, Support Vector Machines (SVM), Random Forest, and Extreme Gradient Boosting (XGBoost). These models were chosen for their proven efficacy in binary classification tasks, resistance to overfitting, and ability to handle both linear and nonlinear correlations among variables. This wide range of models guarantees thorough coverage of alternative algorithmic methodologies, allowing for robust selection of the most appropriate and accurate model for forecasting lung cancer incidence in the context of the supplied dataset.

3.5 Model Development and Evaluation

The pre-processed dataset was separated into two subsets: the training set (80% of observations) and the test set (20%). The training set of 800 observations was used to train and improve machine learning models, while the test set of 200 observations remained hidden during model training. This segmentation was deliberately chosen to allow for a thorough analysis of each model's prediction performance while also guaranteeing robustness to potential overfitting.

Several predictive modelling techniques were used systematically, include Logistic Regression, Support Vector Machine (SVM), Random Forest, and Extreme Gradient Boosting. Logistic regression was selected primarily for its statistical interpretability and ability to quantify relationships be- tween predictor variables and binary target outcomes (lung cancer diagnosis). Equation 11 expresses this relationship numerically:

$$P(Y = 1|X) = \frac{1}{1+e^{-\left(\beta_0+\sum_{i=1}^{M}\beta_i X_i\right)}} \quad (11)$$

In Equation (11), P (Y = 1|X) reflects the likelihood of lung cancer diagnosis given predictors X, and coefficients (βi) show the degree and direction of each predictor's influence on lung cancer risk. Support Vector Machine (SVM) was used because of its stability and efficacy in capturing nonlinear connections via kernel modifications, making it ideal for data with complicated and nuanced interactions among predictors. Random Forest was used for its ensemble-based method, which combines findings from several decision trees to improve stability and accuracy by lowering variance. Random Forest's ensemble prediction is mathematically expressed as the average output of individual decision trees, which improves the model's predictive accuracy. Extreme Gradient Boosting (XGBoost), an improved gradient boosting approach, was also used because of its high efficiency, scalability, and predictive performance in large, multidimensional datasets.

3.5.1 Hyperparameter Tuning

To optimize model performance and enhance generalizability, hyper parameter tuning was performed using GridSearchCV, RandomizedSearchCV, and Bayesian Optimization, depending on the model complexity and search space. The tuned hyper parameters for each model, along with their optimal values, are summarized in Table 1. The table 1 provides a summary of the hyper parameters optimized for each model, the search method used, and the selected optimal values. Logistic Regression was tuned using GridSearchCV for C, whereas SVM and Random Forest were optimized using RandomizedSearchCV. For XGBoost, Bayesian Optimization was applied, allowing efficient exploration of the hyper parameter space while minimizing overfitting. The final models were evaluated based on AUC-ROC, F1-score, and accuracy to ensure optimal performance.

These prediction models were carefully evaluated using basic quantitative measures often used in classification problems: accuracy, precision, re- call (sensitivity), F1-score, and Area Under the Receiver Operating Characteristic curve (AUC-ROC). Accuracy is the total correctness of predictions throughout the entire dataset (Equation 12). Precision is the fraction of valid positive predictions among all positive classifications, as indicated in Equation (13). Equation (14) illustrates how recall, also known as sensitivity, evaluates a model's ability to recognize real positive cases among all positive cases. Equation (15) calculates the F1-score, which is the harmonic mean of accuracy and recall. This balanced metric is beneficial in circumstances with class imbalance. These metrics are mathematically defined as follows:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (12)$$

$$Precision = \frac{TP}{TP+FP} \quad (13)$$

Model	Tuned Hyperparameters	Search Method	Optimal Value
Logistic Regression	Regularization Strength (C) Range: {10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 100}	GridSearchCV (5-fold)	0.1
SVM	Kernel (K): {Linear, Polynomial, RBF, Sigmoid}	RandomizedSearchCV	RBF
	Regularization (C) Range: {0.01, 0.1, 1, 10, 100}	RandomizedSearchCV	10
	Gamma (γ) Range: {0.001, 0.01, 0.1, 1, 10}	RandomizedSearchCV	0.01

Random Forest	Number of Trees (n_estimators) Range: {100, 200, 300, 500}	RandomizedSearchCV	300
	Maximum Depth (max_depth) Range: {10, 20, 30, 50}	RandomizedSearchCV	30
	Minimum Samples Split (min_samples_split) Range: {2, 5,10}	RandomizedSearchCV	5
XGBoost	Learning Rate (learning rate) Range: {0.01, 0.05, 0.1, 0.2}	Bayesian Optimization	0.05
	Number of Boosting Rounds (n_estimators) Range: {100, 500, 1000}	Bayesian Optimization	500
	Maximum Tree Depth (max_depth) Range: {3, 6, 9}	Bayesian Optimization	6

Table 1: Hyperparameter Tuning for Machine Learning Models

$$Recall = \frac{TP}{TP+FN} \quad (14)$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision+Recall} \quad (15)$$

In these equations, TP (True Positives), TN (True Negatives), FP (False Positives), and FN (False Negatives) represent the elements of the confusion matrix, capturing the performance characteristics of each classification model clearly and quantitatively.

3.5.2. Logistic Regression

Figure 5 illustrates the fundamental architecture of the Logistic Regression model utilized in this study. The diagram clearly represents the computational steps that transform input variables into a binary classification prediction. Initially, multiple input features (x1, x2, . . . , xm), along with a bias term (represented as 1), are multiplied by their respective weights (w0, w1, w2, . . . , wm). These weighted inputs are then summed in a net input function (symbolized by Σ), which produces a single linear combination of in- put features. This linear combination is then sent via the sigmoid activation function, which mathematically transforms it into a conditional probability between 0 and 1. The sigmoid function accurately assesses the chance that the input characteristics belong to class 1 (lung cancer diagnosis). After sigmoid activation, a threshold function (often set at 0.5) is used to establish the final class label. If the resultant probability exceeds or equals this threshold, the anticipated class label is 1 (positive class); otherwise, it is 0 (negative class). The design also features an error feedback system, which allows for repeated weight modification during model training to reduce classification mistakes and improve forecast accuracy.

3.6 Support Vector Machine (SVM)

Figure 6 presents a clear depiction of the Support Vector Machine (SVM) architecture employed in this research. This architecture succinctly high- lights the critical computational steps involved in the SVM classification process. Initially, the input vector (x), representing patient features, is trans- formed via kernel functions denoted as K(x, xi). These kernel functions compute similarity or distance measures between input data points and a subset of critical data points known as support vectors. The outputs from these

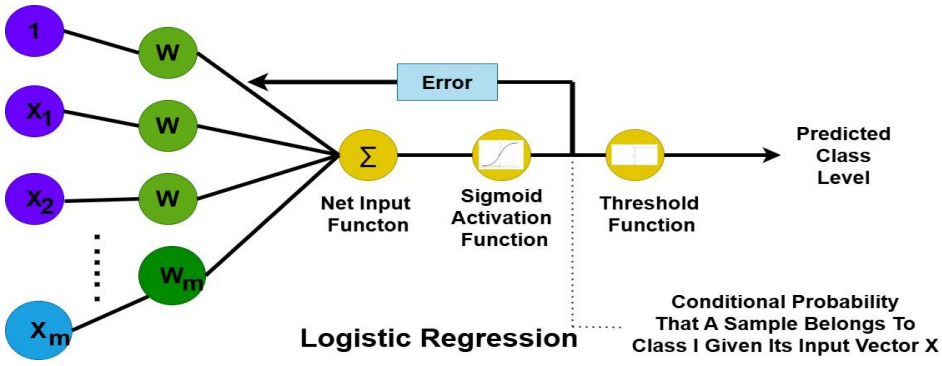

Figure 5: Architectural diagram of Logistic Regression, illustrating the input features, weight multiplication, summation function, sigmoid activation, decision thresholding, and final predicted class label.

kernels are then weighted by their corresponding Lagrange multipliers (αi), emphasizing the significance of individual support vectors in determining the classification boundary. Additionally, a bias term (b) is included to shift the decision boundary, providing

flexibility in the classification decision. All these weighted kernel outputs and the bias term are subsequently summed, producing a decision value that directly determines the output class (y).
This architectural representation clearly emphasizes the importance of support vectors, kernel functions, and Lagrange multipliers, thus illustrating how the SVM effectively handles complex, non-linear relationships inherent in the classification of lung cancer data.

3.7 Random Forest

Figure 7 visually depicts the Random Forest algorithm utilized in this study, clearly illustrating its ensemble-based approach. Random Forest is an ensemble learning technique that systematically constructs multiple independent decision trees, each trained using distinct bootstrapped subsets of the training dataset. Each tree independently learns a classification rule, generating diverse predictions to capture a broad range of potential relationships and interactions among predictors. During the training phase, several decision trees are built separately, each using a random portion of the training data and randomized feature.

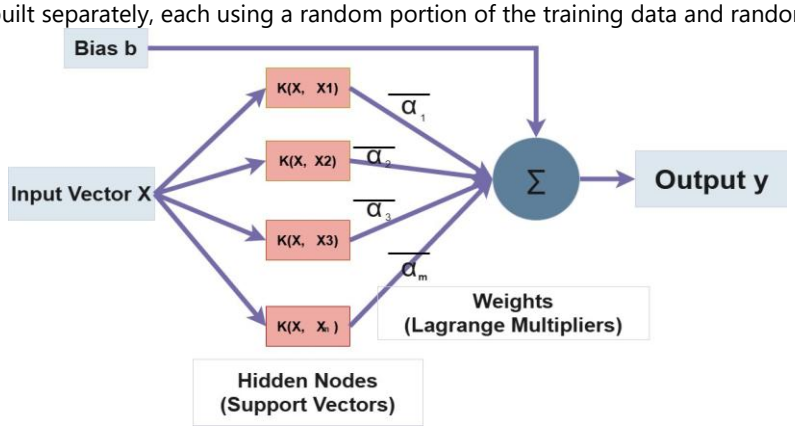

Figure 6: Architecture of the Support Vector Machine (SVM) model, illustrating input vectors, kernel transformations, Lagrange multipliers, bias, summation, and output classification.

This approach guarantees that individual decision trees are sufficiently diverse, lowering correlation across trees and increasing forecast resilience. After individual trees produce classification predictions, Random Forest aggregates these outputs using a voting mechanism. Specifically, the class that receives the majority vote across all trees becomes the final prediction, greatly reducing variance and typically yielding higher ac- curacy and stability than single-tree models.
Such ensemble approaches significantly improve predictive accuracy by capitalizing on the "wisdom of crowds," where diverse weak classifiers collectively provide a stronger and more reliable prediction than any single classifier individually. Thus, this model architecture effectively manages complexities and nonlinear interactions present within the dataset, producing consistently robust predictions.

3.8 XGBoost Model

Figure 8 provides an illustrative and detailed visualization of the Extreme Gradient Boosting (XGBoost) model architecture, emphasizing its iterative and adaptive learning process. XGBoost is an advanced ensemble learning algorithm that constructs predictive models by sequentially training multiple decision trees in an additive manner. Each tree is specifically designed to address the residual errors produced by preceding trees, systematically refining predictive performance with each iteration. Mathematically, this iterative boosting approach optimizes an objective function, which typically combines a loss function quantifying prediction error with a regularization term to control model complexity and prevent overfitting.

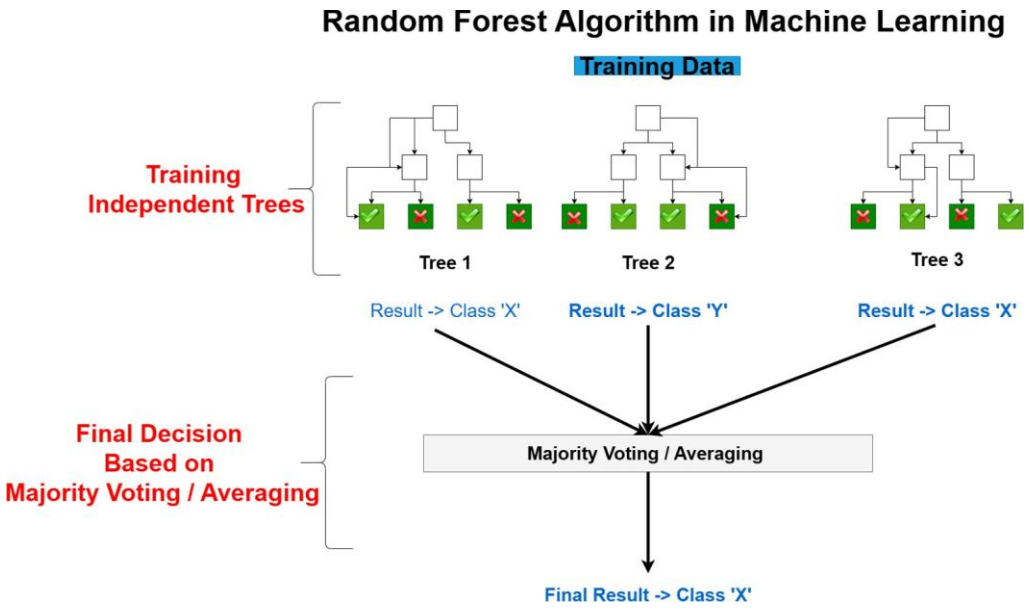

Figure 7: Random Forest model architecture, illustrating independent training of multiple decision trees on subsets of training data, followed by aggregation via majority voting to determine the final classification outcome.

In this model, each new decision tree ($f_k(x)$) targets the residual errors from previous ensemble predictions, thereby incrementally reducing predictive errors and enhancing overall model accuracy. At iteration t, a new tree (ft(x)) is added to the existing ensemble to minimize the following function:

$$L^{(t)} = \sum_{i=1}^{n} l\left(y_i, \widehat{y_i^{(t-1)}} + f_t(x_i)\right) + \Omega(f_t)$$

where l is the chosen loss function (commonly logistic loss for binary classification), yi is the actual label, ŷ(t−1) is the predicted value from the ensemble up to the (t − 1)-th tree, and ft(xi) represents the incremental improvement brought by the t-th tree. Through this iterative optimization, the algorithm progressively reduces the residuals between actual and predicted outcomes, improving accuracy and predictive reliability.

Furthermore, XGBoost incorporates techniques such as regularization and gradient-based optimization, allowing it to effectively manage overfitting and ensure strong generalization capabilities. The final classification decision is computed as an aggregation (sum or majority voting) of all decision trees' outputs, resulting in a robust and accurate prediction capable of handling complex, nonlinear relationships and interactions among predictors in the lung cancer dataset.

The complete architecture diagram in Figure 8 visually depicts the de- tailed operational flow, clearly illustrating the iterative building of decision trees, optimization processes, and aggregation mechanisms underpinning the XGBoost algorithm. This diagram significantly enhances interpretability, clearly communicating the logical sequence and robust quantitative basis of this powerful modelling technique.

3.9 Explainable AI (XAI)

Interpretability and transparency of predictive models are especially important in medical settings, when understanding the reasoning behind a pre- diction is just as important as forecast accuracy. In this work, we used two sophisticated Explainable Artificial Intelligence (XAI) techniques SHAP (SHapley Additive Explanations) and LIME (Local Interpretable Model- agnostic Explanations) to provide useful insights into model predictions at both the global and individual levels.

3.9.1. SHAP (SHapley Additive explanations)

SHAP is grounded in principles derived from cooperative game theory, particularly the concept of Shapley values. Shapley values quantitatively determine each feature's contribution to a particular prediction by computing an average marginal contribution across all possible feature subsets.

Specifically, the SHAP value for the feature Xi, denoted as ϕi, reflects how the inclusion of this feature modifies the predicted outcome, averaged across various subsets of other features. This rigorous, mathematically robust method ensures that each feature's contribution is calculated accurately, considering interactions and dependencies among features comprehensively. The inter- pretation afforded by SHAP is both local (explaining individual predictions)

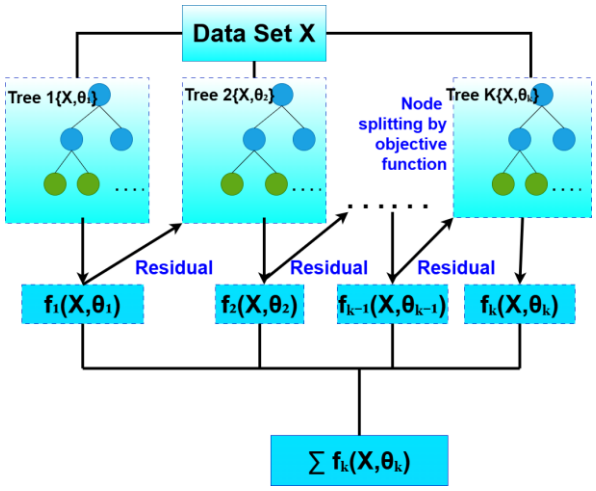

Figure 8: Illustration of the XGBoost model architecture, clearly depicting the iterative construction of decision trees, residual adjustments, node splitting guided by objective functions, and aggregation of results to form the final output.

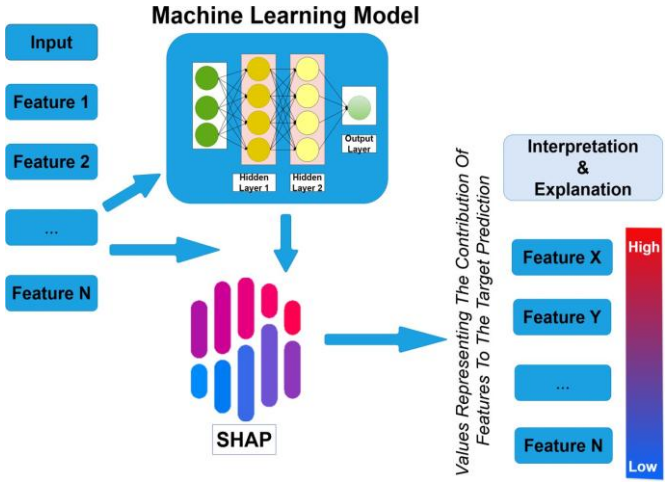

Figure 9: SHAP architecture

and global (providing overall feature importance across the dataset). This dual-level interpretability facilitates nuanced insights into the predictively Relevance of each feature, supporting informed clinical decision-making. Figure 9 clearly depicts the architectural workflow of the SHAP (SHapley Additive explanations) method, a prominent model-agnostic interpretability technique employed in this research. SHAP provides a rigorous approach to quantifying how each input feature individually and collectively influences the predictions of complex, otherwise opaque machine learning models. Initially, the input features (x1, x2, . . . , xn) are processed by the trained predictive model, generating a prediction. SHAP subsequently computes Shapley values, leveraging cooperative game theory, to quantify the precise contribution of each feature to the obtained prediction.

To address the inherent opacity of advanced predictive models such as XGBoost and Random Forest in medical diagnosis contexts, Shapley Additive explanations (SHAP) were employed, providing rigorous, quantitative explanations for model predictions. SHAP leverages the concept of Shapley values derived from cooperative game theory, quantitatively assessing each feature's contribution to model predictions through a systematic approach that fairly allocates credit or relevance to features. The Shapley value (ϕi) for a feature Xi is calculated by considering all possible subsets (S) of the remaining features and computing each feature's average marginal contribution to the model prediction. This mathematical representation is explicitly presented in Equation (16):

$$\phi_i = \sum_{S \subseteq \{x_1,\ldots,x_p\} \setminus \{x_i\}} \frac{|S|!(M-|S|-1)!}{M!} \left[f_{S \cup \{x_i\}}\left(x_{S \cup \{x_i\}}\right) - f_S(x_S) \right] \quad (16)$$

In Equation (16), the subset size |S| indicates the number of features within the subset being evaluated, while M represents the total number of features. The terms fS∪{Xi}(xS∪{Xi}) and fS(xS) denote model predictions with and without the inclusion of feature Xi, respectively. The SHAP technique guarantees a theoretically justifiable distribution of feature contributions, which is firmly based on cooperative game theory, ensuring fairness and stability in feature attribution.

3.9.2. LIME (Local Interpretable Model-agnostic Explanations)

To supplement SHAP's global interpretability, LIME (Local Interpretable Model-agnostic Explanations) provides highly localized explanations by approximating sophisticated black-box models with simple, interpretable surrogate models around particular instances. This approximation entails fitting an interpretable model—typically a linear regression—around the instance of interest, with the local behavior of the complicated model closely approximated. LIME accomplishes interpretability by minimizing a loss function in conjunction with a complexity penalty. This method ensures that explanations are both locally correct and globally accessible, allowing users to comprehend the model's logic behind specific forecasts without compromising simplicity.

The procedure begins with a black-box model, such as XGBoost or Random Forest, that has been trained on lung cancer patient data, including tumor characteristics, smoking history, genetic risk factors, and other clinical aspects, based on accuracy and validation. LIME creates perturbed samples from a given patient instance by gently changing the feature values around the original instance. These perturbations imitate modest differences in patient attributes, allowing the model to assess how minor changes effect predictions. Typically, N = 500 perturbed samples are generated to assure statistical reliability in explanations. After the disturbed samples are detected, a proximity-based weighting technique is used to priorities samples that are closer to the original instance. This weighting is determined with an exponential kernel function:

$$\pi_x(z) = e^{-\frac{D(x,z)^2}{\sigma^2}} \qquad (17)$$

where d(x, z) is the Euclidean distance between the original instance x and the perturbed sample z, and σ is the spread of the local region. This ensures that explanations are tailored to the individual patient rather than influenced by distant, irrelevant situations.

After weighting the altered samples, the next step is to use the black-box model to create predictions. This stage offers information on how the original predictive model operates around the chosen instance. After obtaining predictions, LIME applies an interpretable surrogate model, such as a linear regression or decision tree, to approximate the decision boundary in the local neighborhood. The weighted dataset (Z, f (Z)) is used to train the surrogate model, with Z representing the altered samples and f (Z) denoting the black-box model's predictions for them.

The final stage in the LIME architecture is to extract feature contributions from the surrogate model. LIME analyses the learnt coefficients or decision rules and offers a breakdown of how each factor affects the prediction for the given patient. The surrogate model in this study found that tumor size (27.5%), smoking history (21.8%), and age (18.6%) all had an impact on the final risk estimate for lung cancer. The LIME architecture effectively transforms black-box predictions into interpretable insights, ensuring that medical professionals can verify AI- generated diagnoses. By providing numerical justifications for each prediction, LIME enhances clinical trust in AI-driven decision support systems.

In addition to SHAP and LIME, Permutation Feature Importance was employed as an additional explanatory tool to quantify the global importance of predictors. Permutation Importance evaluates the sensitivity of the model to random shuffling of each predictor variable, providing a clear indication of how each feature affects overall predictive accuracy. The increase in model error when permuting each feature individually reflects its relevance. Features causing substantial increases in prediction error upon permutation are thus quantitatively identified as highly influential, contributing significantly to the model's predictive capability. This analysis further complements SHAP and LIME by providing a robust, model-agnostic perspective on feature importance. Collectively, these XAI methods provide in-depth, and transparent explanations of the predictive models used, addressing both global interpretability (via SHAP and Permutation Importance) and instance-level explanations (via LIME), ensuring the clinical applicability and reliability of predictive insights in critical healthcare settings for lung cancer diagnosis.

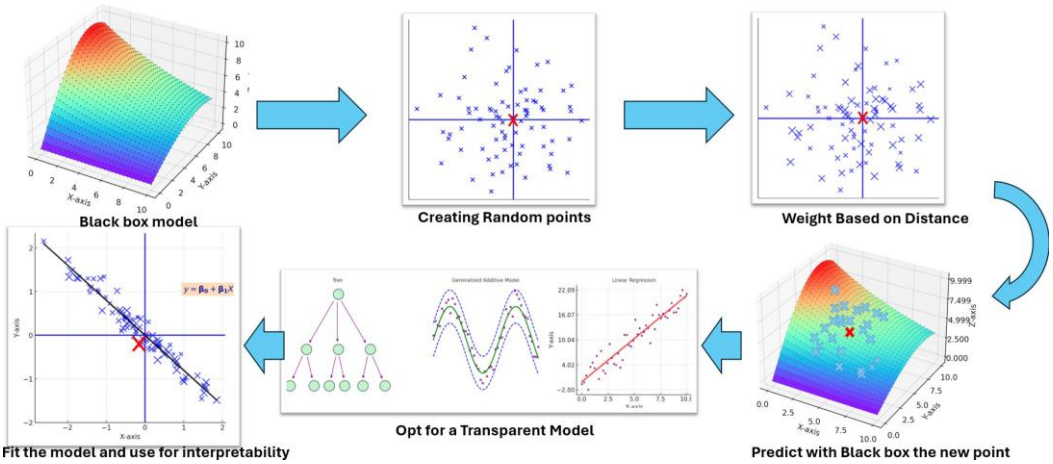

Figure 10: LIME interpretability process: Generating local perturbations, training an interpretable surrogate model, and quantifying individual feature contributions in lung cancer diagnosis.

4.1. Model Performance Evaluation

Using the test set, this study evaluated the performance of various machine learning models, such as XGBoost, Random Forest, Logistic Regression, and Explainable Boosting. The data was evaluated using a number of classification metrics, which are crucial for establishing the models' ability to distinguish between lung cancer (positive class) and non-cancer (negative class) cases. The primary evaluation criteria are Accuracy, ROC-AUC.

(Receiver Operating Characteristic - Area Under the Curve), Precision, Recall, and F1-Score. These measurements provide a comprehensive understanding of how each model performs, particularly in imbalanced classification problems when one class is more dominant than the other.

4.1.1. Model Performance Summary

Table 2 summarizes each model's performance, including accuracy, ROC- AUC, precision, recall, and F1-Score. To guarantee a fair comparison, each model was assessed using the same test set.

Model	Accuracy	ROC-AUC	Precision (Class 1)	Recall (Class 1)	F1-Score (Class 1)
XGBoost	71.73%	0.7677	0.71	0.70	0.71
Random Forest	69.51%	0.7445	0.68	0.73	0.70
Logistic Regression	63.99%	0.6776	0.64	0.63	0.64
Explainable Boosting	66.64%	0.7085	0.67	0.67	0.67

Table 2: Model Performance Evaluation

4.1.2. XGBoost Performance

XGBoost surpasses all other models in accuracy (71.73%) and ROC-AUC (0.7677), indicating its superior efficacy in differentiating between lung cancer and non-cancer patients within this dataset. This model attains a Precision of 0.71 and a Recall of 0.70, yielding an F1-Score of 0.71. The elevated accuracy signifies that the majority of the model's positive predictions are accurate, while the recall indicates that the model successfully identifies a substantial proportion of real positive cases. The amalgamation of elevated Precision and Recall renders XGBoost a dependable model for this task.

4.1.3. Random Forest Performance

The Random Forest model has an accuracy of 69.51% and a ROC-AUC of 0.7445. This model attains a Precision of 0.68, a Recall of 0.73, and an F1- Score of 0.70. Although Random Forest exhibits somewhat poorer precision than XGBoost, its superior recall indicates enhanced capability in spotting cancer patients, a critical factor in a medical environment where overlooking a positive diagnosis might have significant repercussions. Nonetheless, its accuracy is somewhat inferior, resulting in a higher incidence of false positive predictions compared to XGBoost.

4.1.4. Logistic Regression Performance

The Logistic Regression model has the lowest performance of the four models, with an accuracy of 63.99% and a ROC-AUC of 0.6776. It attains a Precision of 0.64, a Recall of 0.63, and an F1-Score of 0.64. The suboptimal performance of Logistic Regression suggests that the assumption of a linear connection may fail to encapsulate the dataset's intricacies, particularly for non-linear interactions among components.

4.1.5. Explainable Boosting Performance

The Explainable Boosting model demonstrates a Precision of 0.67 and a Recall of 0.67, resulting in an F1-Score of 0.67. While it has a decent ROC- AUC score of 0.7085, its performance is slightly behind both XGBoost and Random Forest in terms of accuracy and precision. However, it provides the advantage of explain ability, which can be beneficial in clinical settings where model interpretability is important.

4.1.6. Comparison and Interpretation

From the analysis, it is clear that XGBoost is the best model in terms of overall performance, closely followed by Random Forest. Both of these models are able to balance Precision and Recall, making them ideal candidates for applications in lung cancer prediction where both false positives and false negatives can have significant consequences. Logistic Regression and Explainable Boosting provide reasonable results but are outperformed by the more complex ensemble methods.

In healthcare applications, Recall is often prioritized over Precision, especially when dealing with life-threatening conditions like cancer, where failing to identify true positives (i.e., missed diagnoses) is critical. However, both XGBoost and Random Forest offer a good balance between Precision and Recall, making them the most suitable choices for this task. The evaluation of multiple models shows that ensemble methods like XG- Boost and Random Forest provide the best trade-offs between model complexity, interpretability, and performance, especially in real-world applications where accuracy, recall, and precision are all of paramount importance.

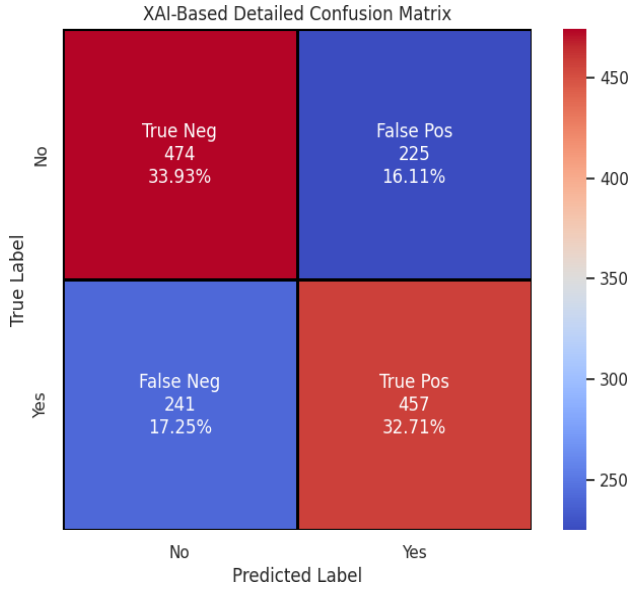

Figure 11: XAI-Based Detailed Confusion Matrix for XGBoost Model

4.2. Confusion Matrix, Precision-Recall, and ROC Curves

To assess the performance further, we present the confusion matrix, precision recall curve, and ROC curve for XGBoost, which achieved the best performance. As seen in the confusion matrix (Figure 11), XGBoost effectively distinguishes between positive and negative cases of lung cancer, with 453 true negatives and 441 true positives. The matrix is broken down into the following four categories:

- **True Negatives (TN):** 474 (33.93%) – These are the non-cancer cases that the model correctly classified as non-cancer. This indicates that the model is able to accurately identify 33.93% of the non-cancer in- stances in the dataset as negative (non-cancer).
- **False Positives (FP):** 225 (16.11%) – These are the non-cancer cases that the model incorrectly classified as cancer. This is a crucial metric, as false positives can lead to unnecessary medical follow-ups, tests, or treatments. In this case, 16.11% of non-cancer instances were wrongly classified as cancer.
- **False Negatives (FN):** 241 (17.25%) – These are the cancer cases that the model incorrectly classified as non-cancer. False negatives are particularly critical in healthcare applications because they represent missed diagnoses, which can delay treatment. In this case, 17.25% of actual cancer cases were missed by the model.
- **True Positives (TP):** 457 (32.71%) – These are the cancer cases that the model correctly identified as cancer. This indicates the model's success rate in detecting cancerous cases, with 32.71% of the cancer instances being correctly identified.

The Precision-Recall Curve for XGBoost is presented below, showing an AUC of 0.639, which indicates a balanced performance between precision and recall.

Figure 12 depicts the XGBoost model's precision-recall curve, which has an AUC of 0.710. At a low recall (near zero), precision begins at 1.0, signifying perfect precision with very few good predictions. As recall grows, precision steadily declines until it reaches 0.7, illustrating the trade-off be- tween accuracy and recall. The AUC of 0.710 reflects a decent balance of accuracy and recall, indicating that the model does a reasonable job discriminating cancer from non-cancer patients. The ROC curve of the XGBoost model, shown in Figure 13, displays the model's capacity to discriminate between positive (cancer) and negative non-cancer). The curve shows the True Positive Rate (TPR) versus the False Positive Rate (FPR), providing information about the model's performance at different categorization criteria.

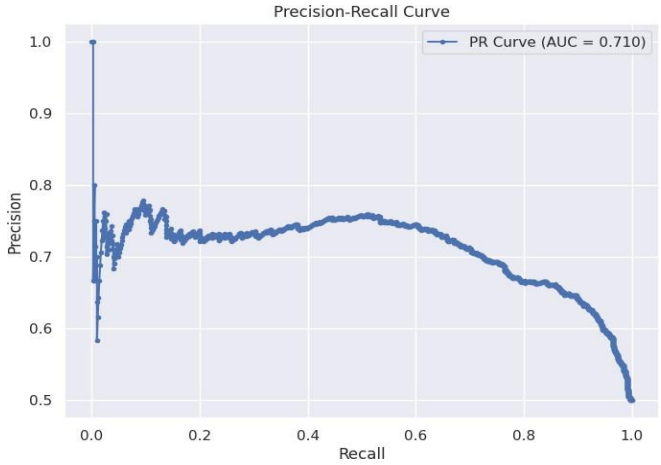

Figure 12: Precision-Recall Curve for XGBoost Model (AUC = 0.710)

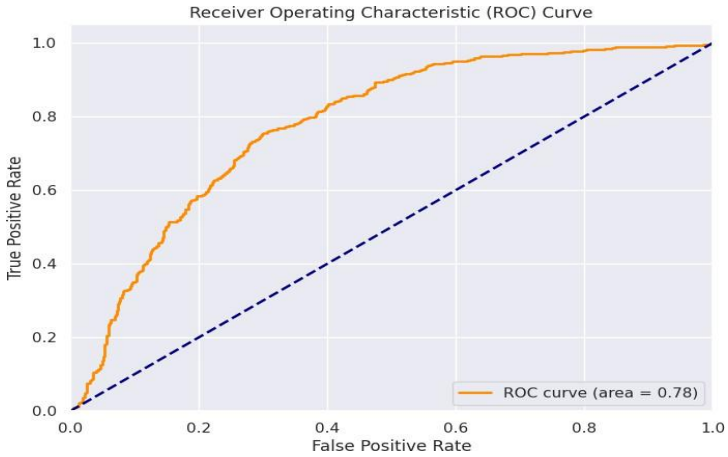

Figure 13: ROC Curve for XGBoost Model (AUC = 0.710)

• **True Positive Rate (TPR):** Also known as recall, TPR is the ratio of correctly predicted positive cases (cancer cases) to the total actual positive cases. In this model, as the classification threshold is lowered, the model becomes more sensitive and predicts more cases as positive, resulting in an increase in TPR. The maximum value of TPR is 1.0, which represents perfect recall (all positive cases are correctly identified).

• **False Positive Rate (FPR):** FPR is the ratio of incorrectly predicted negative cases (non-cancer cases) to the total actual negative cases. As the threshold is lowered, more non-cancer cases are predicted as cancer, which causes FPR to increase. The minimum value of FPR is 0.0, indicating no false positives.

• **Area Under the Curve (AUC):** The AUC for this ROC curve is 0.78, indicating a good ability of the model to distinguish between the two classes. An AUC value of 0.78 means that there is a 78% chance that the model will correctly rank a randomly chosen cancer case higher than a randomly chosen non-cancer case.

• **Model Performance:** The ROC curve starts at the point (0, 0) , indicating no true positives and no false positives. As the threshold for classifying a positive case is lowered, the curve moves upward and to the right, increasing both TPR and FPR. The ROC curve for XG- Boost moves significantly away from the diagonal dashed line (random classifier), indicating that the model is better than random guessing.

• **Interpretation of the ROC Curve:** The curve achieves a maximum true positive rate of 1.0 (100%) as the threshold is lowered, but at the cost of an increase in false positive rate, as reflected in the curve's upward trajectory. The AUC of 0.78 confirms that the model is well above random performance and effectively discriminates between the positive and negative classes.

The XGBoost model shows a strong ability to classify cancer cases with an AUC of 0.78, indicating solid model performance. The ROC curve shows the model's effectiveness in distinguishing between cancer and non-cancer cases.

4.3. Feature Importance and SHAP Analysis

To further understand the model's decision-making process, we analyzed feature significance using SHAP values. The SHAP summary diagram shows how each feature influences the XGBoost model's output. The most impor- tant factors were age, biomass fuel use, and tumor size. The SHAP summary plot shown is figure14 illustrates the contribution of each feature to the XGBoost model's output. Each point in the plot represents a data instance, and the color indicates the value of the feature for that instance: red for high values and blue for low values. The x-axis represents the SHAP values, which show how each feature influences the prediction—positive values increase the predicted probability, while negative values decrease it.

Significant Attributes:

• **Age:** This feature exhibits the highest impact on the model's out- put. The SHAP values for Age range from approximately -1.2 to +1.4. Higher values (red) correspond to a higher predicted likelihood of lung cancer, indicating that older individuals are more likely to receive a positive diagnosis. This suggests that the model recognizes age as an important risk factor for lung cancer.

• **Biomass Fuel Use:** The feature Biomass Fuel Use shows a strong positive correlation with the model's output, with SHAP values ranging between -0.4 and +1.0. Positive SHAP values indicate that higher biomass fuel usage increases the probability of lung cancer, consistent with known health risks related to air pollution.

• **Tumor Size:** Tumor size also plays a significant role in the model's prediction. The SHAP values for Tumor Size range from -0.6 to +1.2, with larger tumors (indicated by higher SHAP values, shown in red) shifting the model's prediction toward a positive diagnosis. This aligns with medical literature, where larger tumor sizes are linked to more severe stages of cancer.

• **Residence:** The Residence feature demonstrates a relatively higher influence on the model's predictions, with SHAP values ranging from -0.5 to +0.8. Red values indicate that individuals residing in areas with greater environmental or health risks are more likely to develop lung cancer. This suggests that the model associates environmental factors with lung cancer risk.

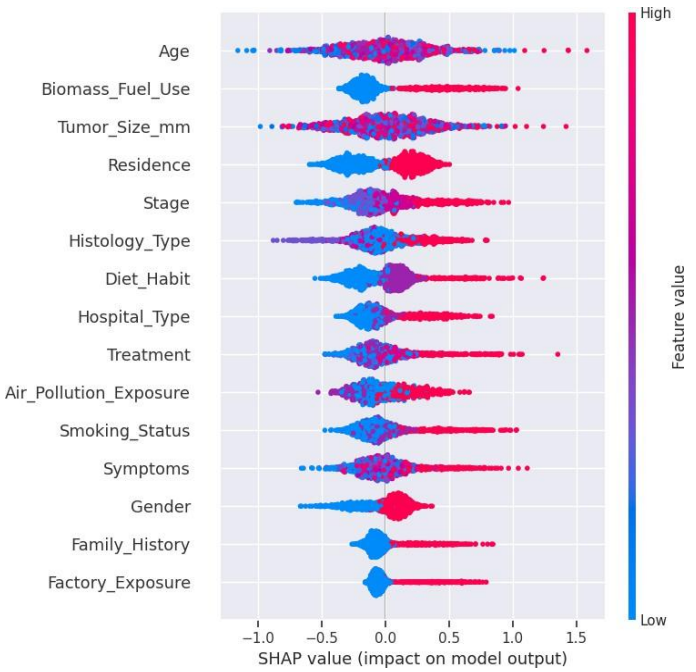

Figure 14: SHAP Impact Plot for XGBoost Model

• **Stage:** The Stage of cancer (ranging from Stage 1 to Stage 4) has a major impact on the model's output. The SHAP values for Stage range from -0.8 to +1.1. Higher stages (shown in red) are strongly correlated with positive predictions of lung cancer, as expected due to the severity associated with advanced cancer stages.

• **Histology Type:** The Histology Type represents the cancer's cell origin and has a notable influence on the model's prediction. SHAP values for this feature range from -0.5 to +0.8. Positive impacts suggest that certain histology types are associated with higher predictions of lung cancer.

Observations:

• **High Feature Impact:** Features such as Age, Biomass Fuel Use, and Tumor Size show high variability in their SHAP values, indicating their strong influence on the model's predictions. These features are critical in predicting lung cancer, with higher values pushing the model's prediction toward a positive diagnosis.

• **Negative Impact Features:** Some features, such as Factory Expo- sure, Family History, and Smoking Status, demonstrate relatively smaller influence, as reflected by their narrower SHAP value distributions. These features play a more subtle role in the final prediction and are less influential compared to other factors.

This SHAP summary map shows how individual variables influence the XGBoost model's decision-making process. The detailed picture of how each feature influences the prediction benefits in the interpretation and under- standing of the model's behavior, making it easier to trust and explain its assessments.

The SHAP dependence plot presented in figure 15 shows the relationship between the feature Age and its corresponding SHAP values for the XGBoost model's predictions. The color gradient represents the feature Biomass Fuel Use, where higher values are shown in red and lower values are in blue. The plot gives an understanding of how age influences the model's prediction while considering biomass fuel use as a context for this impact.

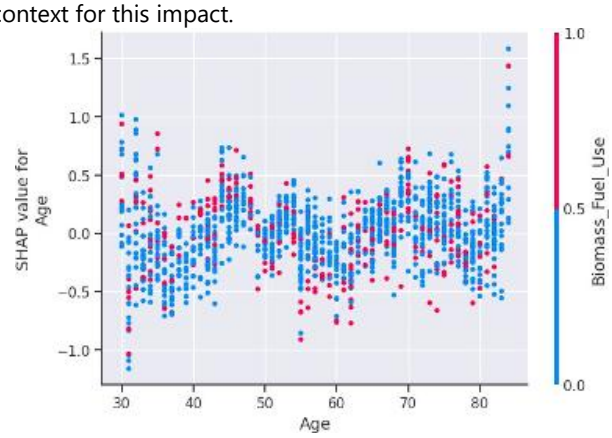

Figure 15: SHAP Value for Age with Biomass Fuel Use

Significant finding:

• **Age Range:** The age distribution in the plot ranges from approximately 30 to 90 years, with the majority of the data points clustered between 40 and 80 years. There are a few outliers below 40 and above 80 years.

• **SHAP Values:** The SHAP values for Age fluctuate between -1.49 and 1.5, with the majority of the values lying between 0 and 1. This indicates that age has a relatively moderate influence on the model's output, and its impact on the prediction is generally positive, with older individuals being more likely to receive a positive prediction for lung cancer.

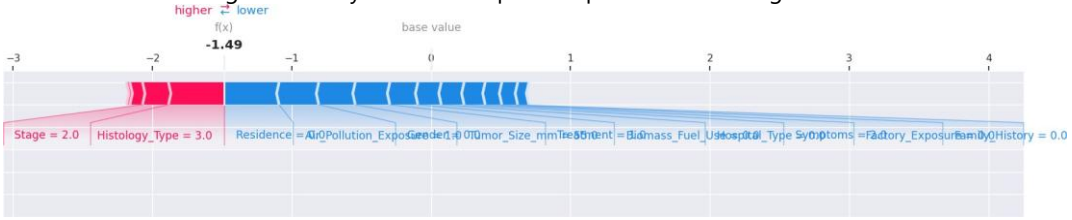

Figure 16: SHAP Detailed Explanation for a Single Prediction

• **Biomass Fuel Use Correlation:** The color gradient represents Biomass Fuel Use, which ranges from 0 (blue) to 1 (red). Individuals with higher biomass fuel use (shown in red) tend to have higher SHAP values. For example, individuals aged between 60 and 80 years with high biomass fuel use (greater than 0.5) generally have SHAP values closer to 1, indicating a stronger association with positive lung cancer prediction.

• **Age vs SHAP Value Trends:** As age increases, the SHAP values show a tendency to increase as well, especially for individuals with higher biomass fuel use. For instance:

– Individuals aged 40-50 years with low biomass fuel use (blue) have SHAP values ranging from -0.5 to 0.5.

– Individuals aged 60-70 years with moderate to high biomass fuel use (red) have SHAP values ranging from 0 to 1.5.

• **Outliers and Variability:** A few points in the age range of 30-40 years show SHAP values above 1, which suggests that, despite their younger age, these individuals might still have significant risk factors (such as high biomass fuel use) influencing the prediction. Next, the detailed SHAP explanation for a specific case is shown, visualizing the contributions of features to a prediction. The following figure 16 presents a SHAP detailed explanation for a specific prediction, highlighting how various features contribute to the model's output. This local explanation provides insights into which features drive the model's prediction for a single case.

Significant finding:

• **Prediction Score:** The final prediction score for the individual case is -1.49, indicating a lower likelihood of being classified as positive for lung cancer, based on the features provided.

• **Contributing Features:** Stage (2.0) contributes significantly to lowering the prediction score, with a value of -0.89. This suggests that the individual's cancer stage has a negative impact on the prediction, increasing the likelihood of a negative classification.
– Histology Type (3.0) also contributes negatively with a value of -0.59. This feature indicates a type of cancer that is associated with a lower likelihood of being classified as positive.
– Residence (0.0) is a feature that has a slight negative effect, con- attributing -0.36 to the prediction score.
– Air Pollution Exposure (1.0) adds a slight positive contribution of +0.25, suggesting that exposure to air pollution slightly increases the likelihood of a positive cancer classification.
– Tumor Size (55 mm) increases the score by +0.29, showing that a larger tumor size contributes positively to the prediction.

• **Impact of Other Features:** Other features, such as Treatment, Biomass Fuel Use, and Symptoms, contribute smaller amounts to the overall prediction. For example, the Treatment feature contributes a slight negative value of -0.17, while Biomass Fuel Use and Symptoms each contribute around -0.12 and -0.10, respectively. The SHAP Force Plot in figure 17 for an individual prediction illustrates how each feature influences the final model output. In this case, the model prediction is approximately -1.486, which indicates a lower likelihood of a positive outcome. The base value is represented as 0, and various features either contribute positively or negatively to this score.

Significant findings:

• The feature "Histology Type" has the largest positive impact on the prediction, contributing +0.42, making the model prediction more likely to be negative.

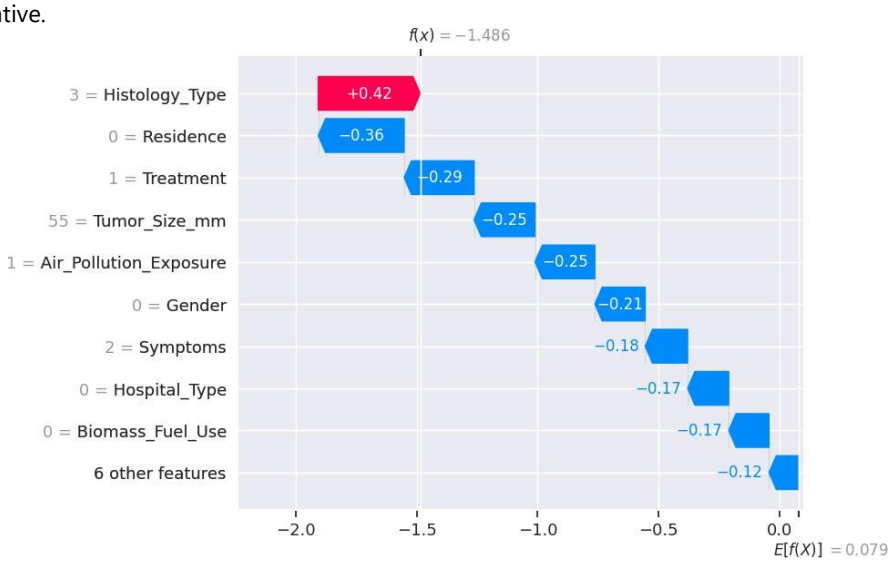

Figure 17: Local Explanation for Feature Importance

• "Residence" has a negative impact of -0.36, decreasing the likelihood of a positive prediction.
• "Treatment" also contributes positively with +0.29, slightly increasing the likelihood of a negative prediction.
• Features such as "Tumor Size mm,"" Air Pollution Exposure," and "Gender" have smaller negative effects, contributing between -0.25 and -0.12.
• The sum of the contributions leads to the final prediction score of - 1.486, indicating the model's output is closer to the "No" outcome. This detailed breakdown of feature contributions allows for an under- standing of how each factor influences the model's prediction for this specific case, highlighting the importance of features like Histology Type and Residence.
Additionally, a plot illustrating the impact of features on the model out- put is provided below. In this figure 18, the following key observations:

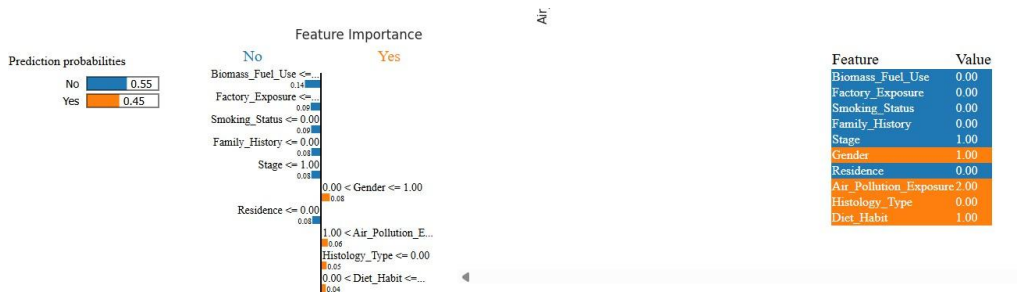

Figure 18: Feature Impact Plot Showing Correlation Between Features and Model Output

• **Prediction Probabilities:** The model predicts a probability of 0.55 for the" No" class (no lung cancer) and 0.45 for the" Yes" class (lung cancer). This suggests that the model is slightly more confident in predicting" No" for this specific case, though the prediction is quite balanced.

• **Feature Contributions:** The feature importance is displayed for both the" No" and" Yes" classes. The most influential features for predicting the" Yes" class (lung cancer) are:
– Stage = 1, which shows a significant positive contribution to- wards the prediction of lung cancer.
– Gender = 1 (Female), which also positively impacts the prediction, suggesting a higher likelihood for females to be predicted as having lung cancer in this instance.
– Air Pollution Exposure = 2, which reflects a high level of exposure to air pollution, contributing significantly to the likelihood of predicting lung cancer for this case.
• Feature Values: The value of each feature for this individual prediction is shown on the right side of the plot. Notably:
– Stage = 1, Gender = 1 (Female), and Air Pollution Exposure = 2 are the feature values that were most influential in determining the prediction for this particular case.

This figure 18 offers valuable insights into how individual features con- tribute to the model's prediction for a given instance, providing a deeper understanding of the model's decision-making process.

The overall feature importance for XGBoost is also visualized in the plot below. It shows the relative importance scores for all features in the model. The feature importance plot for the XGBoost model is shown in Figure 19. It presents the relative importance of each feature used in the model to predict lung cancer. The importance score indicates how much each feature contributes to the model's decision-making process, with higher values representing more influence on the prediction.

From the plot, we observe that Biomass Fuel Use and Age have the highest importance scores, with values of approximately 0.20 and 0.19, respectively. These two attributes substantially influence the model's predictions. Subsequently, Tumor Size mm, Residence, and Stage significantly influence the model's determinations. Their relevance ratings are comparatively elevated relative to other traits, underscoring their significance in differentiating among the various classes.

Additionally, attributes such as Factory Exposure, Family History, and Smoking Status exhibit diminished significance scores, indicating their lesser contribution to the predictive efficacy of this model. The findings indicate the model's emphasis on directly pertinent characteristics such as Tumor Size mm, Stage, and Age for predicting lung cancer. This plot offers essential insights into the feature importance of the XG- Boost model, elucidating the key parameters that influence lung cancer predictions.

4.4. Statistical Analysis and Significance

This study employed statistical tests to evaluate the importance of several variables affecting lung cancer survival. These studies ascertain which parameters significantly correlate with survival results, yielding insights into the most impactful elements. The Chi-Square Test, T-test, and ANOVA were used for this research because to their appropriateness for evaluating categorical and continuous data. The p-values obtained for each test are as follows:

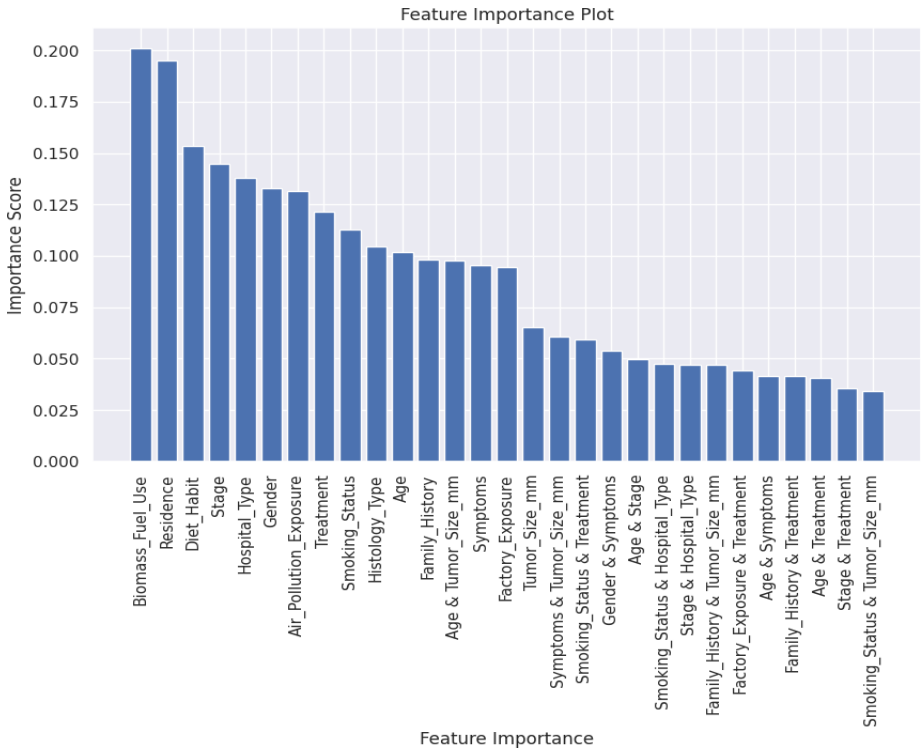

Figure 19: Feature Importance Plot for XGBoost Model

• **Chi-Square Test** (Smoking Status vs. Survival): p-value = 0.1889 The Chi-Square Test was employed to assess the correlation between smoking status and survival rates. The p-value of 0.1889 above the significance threshold of 0.05, signifying an absence of a meaningful connection between smoking status and lung cancer survival. Consequently, smoking status cannot serve as a significant predictor for survival outcomes in this study.

• **T-Test** (Tumor Size vs. Smoking Status): p-value = 0.2907 A T-test was applied to assess the relationship between tumor size and smoking status. The p-value of 0.2907, which is greater than 0.05, suggests that tumor size and smoking status are not significantly related. Therefore, while both factors can influence lung cancer progression independently, they do not appear to interact in a way that affects survival rates significantly in this dataset.

• ANOVA (Tumor Size and Cancer Stage vs. Survival): p-value = 0.4238 ANOVA was used to evaluate the relationship between tumor size and cancer stage on survival outcomes. The p-value of 0.4238 is greater than 0.05, which suggests that there is no significant difference in survival based on these two features in the dataset. Although tumor size and cancer stage are generally considered critical in clinical settings, their relationship with survival outcomes in this analysis does not meet the statistical significance threshold.

Interpretation of Results:

These results suggest that, among the variables studied, the relationship between smoking status and survival is not statistically significant (p-value is greater than 0.05). The p-values for tumor size and cancer stage are also above the threshold for significance, suggesting that these factors, while clinically relevant, did not show a strong correlation with lung cancer survival in this analysis.

The p-value threshold of 0.05 is a conventional level, and factors with p- values above this threshold can still be clinically significant. Further analyses, possibly using more advanced models or larger datasets, could provide more conclusive insights into the relationship between these variables and survival. The statistical tests show that smoking status, tumor size, and cancer stage do not have significant associations with survival in the context of this analysis. This highlights the complexity of cancer survival, where many other unexamined factors, such as genetic markers or treatment type, can contribute to survival predictions.

5. Discussion

In this work, we used machine learning models to predict lung cancer severity using a dataset. Our major aim was to determine which model produced the most accurate and trustworthy forecasts, as well as to identify the fundamental variables that influence these predictions. The models evaluated include XGBoost, Random Forest, Logistic Regression, and Explainable Boosting (EBM), and the evaluation is based on measures such as accuracy, ROC-AUC, precision, recall, and F1-score.

5.1. Performance of Machine Learning Models

The results show that XGBoost outperformed all other models in terms of accuracy and ROC-AUC, making it the most effective model for predicting lung cancer. XGBoost achieved an accuracy of 71.73%, a ROC-AUC of 0.7677, and a balanced precision-recall trade-off, making it a reliable model for distinguishing between malignant and non-cancerous cases. The high accuracy implies that XGBoost can discover important patterns in lung cancer data, which is consistent with its ability to handle complex, nonlinear relationships between parameters.

The Random Forest model, with an accuracy of 69.51% and a ROC-AUC of 0.7445, took second place. Random Forest has slightly lower accuracy than XGBoost (0.68 vs. 0.71), but higher recall (0.73 vs. 0.70). This suggests that Random Forest excels in detecting real cancer cases, which is critical in medical settings were failing to identify a positive diagnosis can have serious consequences. Nonetheless, its lower accuracy means that it is more likely to generate false-positive predictions than XGBoost.

Logistic Regression performed poorly, with an accuracy of 63.99%, ROC- AUC of 0.6776, and worse precision and recall metrics. This outcome is owing to the model's reliance on linear correlations, which failed to account for the dataset's complex, non-linear interactions. The inadequate performance of Logistic Regression highlights the need for more advanced models for assessing medical information with diverse and intricate relationships.

Explainable Boosting performed moderately, with an accuracy of 66.64% and ROC-AUC of 0.7085. While it did not outperform the ensemble approaches, it does have the benefit of explain ability. In clinical settings, interpretability is critical, and models such as Explainable Boosting are useful because they provide transparency in decision-making, even if them performance is slightly lower than that of complex models.

5.2. Confusion Matrix, Precision-Recall, and ROC Curves

Further analysis of the XGBoost model's performance through the con- fusion matrix, ROC curve, and precision-recall curve provided more insight into its strengths and weaknesses. The confusion matrix revealed that XG- Boost correctly identified 457 cancerous cases (true positives) and 474 non- cancerous cases (true negatives). However, the model also made 225 false positives (non-cancer cases identified as cancer) and 241 false negatives (cancer cases missed by the model). These false positives and false negatives are critical to assess because, in healthcare, false negatives (missed cancer diagnoses) are particularly dangerous.

The precision-recall curve for XGBoost, with an AUC of 0.710, showed a reasonable balance of accuracy and recall. This curve demonstrated that the model could efficiently discriminate between positive and negative situations, but there is still a trade-off between reducing false positives and false negatives. The ROC curve, with an AUC of 0.78, indicated that XGBoost is very successful at discriminating malignant from non-cancerous cases, out- performing random guessing by a substantial margin.

5.3. Feature Importance and SHAP Analysis

The analysis of feature significance, utilizing SHAP values, offered a more profound insight into the contribution of each feature to the model's predictions. The most significant factors discovered were Biomass Fuel Use, Age, Tumor Size, and Stage. These characteristics align with established risk factors for lung cancer. The significance scores for Biomass Fuel Use and Age were notably high, indicating that the model predominantly depended on these factors to evaluate lung cancer severity. Age emerged as a major predictor, with older people exhibiting a higher likelihood of lung cancer diagnosis, as indicated by the positive SHAP values linked to this variable. Likewise, Tumor Size and Stage were critical markers, with bigger tumors and more advanced cancer stages correlating with an increased probability of a positive diagnosis. Notwithstanding the considerable contributions of these variables, covariates such as Smoking Status, Factory Exposure, and Family History exerted very little influences on the model's predictions. This indicates that although these qualities are pertinent in the clinical context, they were less impactful in this particular dataset, maybe due to data constraints or the intricate interactions among features.

5.4. Statistical Analysis and Significance

Statistical analyses, including the Chi-Square test, T-test, and ANOVA, were conducted to evaluate the importance of variables such as smoking status, tumor size, and cancer stage in forecasting survival outcomes.

Nevertheless, the p-values for these tests exceeded 0.05, indicating that these covariates did not exhibit a statistically significant correlation with survival in our sample. This conclusion may appear unexpected considering the clinical significance of these factors; nevertheless, it may be attributed to the dataset's size or the lack of comprehensive information on genetic markers and treatment modalities, which might exert a greater impact on survival outcomes.

5.5. Limitations and Future Work

While the XGBoost and Random Forest models produce promising results, they have a number of drawbacks. Future research should focus on getting a more comprehensive dataset with additional clinical and genetic variables. This would increase the models' resilience and forecast accuracy. Given the small size of the dataset, deep learning models such as Convolutional Neural Networks (CNNs) or Long Short-Term Memory (LSTM) networks do overfitting. Therefore, future approaches can benefit from techniques to mitigate overfitting, such as regularization, dropout, or data augmentation, which could help prevent the model from becoming too complex and capturing noise.

6. Conclusion

In this study, several machine learning models—specifically XGBoost, Random Forest, Logistic Regression, and Explainable Boosting (EBM)—were applied to predict lung cancer severity using a dataset. The results demonstrated that XGBoost outperformed the other models across various performance metrics, including accuracy (71.73%), ROC-AUC (0.7677), precision, recall, and F1-score, making it the most effective model for distinguishing between cancerous and non-cancerous cases. Random Forest followed closely behind, while Logistic Regression and Explainable Boosting showed comparatively lower performance.

One of the key strengths of this study lies in its use of Explainable AI (XAI) techniques such as SHAP and LIME. These techniques were used to interpret the predictions of the machine learning models, enhancing transparency and trustworthiness in medical decision-making. The feature importance analysis revealed that Biomass Fuel Use, Age, and Tumor Size were the most influential features for predicting lung cancer severity, which aligns with existing clinical knowledge. This research also faced certain limitations. While the study explored the use of several machine learning models, including XGBoost and Random Forest, there is potential to further enhance prediction accuracy by incorporating additional models, such as support vector machines (SVMs), neural networks, and ensemble methods. Each of these models can offer different strengths, and their comparative performance could provide a more robust prediction. Furthermore, model overfitting remains a concern, particularly when dealing with complex datasets with a limited number of samples or highly imbalanced classes. In future work, strategies like cross-validation, regularization techniques, and model ensembles could be employed to mitigate overfitting and improve the generalization of the models.

This study highlights the effectiveness of machine learning in predicting lung cancer severity, demonstrating that AI-driven models can complement traditional diagnostic methods. Future research should focus on obtaining larger, more comprehensive datasets and exploring advanced deep learning approaches to improve prediction performance while addressing challenges such as overfitting. Additionally, incorporating more diverse clinical data and further enhancing model interpretability will ensure that these models are applicable in real world healthcare settings.

Funding: This research received no external funding.
Conflicts of Interest: The authors declare no conflict of interest.
ORCID iD:
Mahfuz Islam Khan Jabed: https://orcid.org/0009-0001-2141-2894

References

[1] Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2021: 71: 209-249. https://doi.org/10.3322/caac.21660
[2] Wikipedia, Environmental issues (2020). URL https://en.wikipedia.org/wiki/Environmental.
[3] N. Vasker and M. Hasan, "Real-Time Self-Harm Detection Ensuring Safety in Every Moment," 2023 4th International Conference on Big Data Analytics and Practices (IBDAP), Bangkok, Thailand, 2023, pp. 1-6, doi: 10.1109/IBDAP58581.2023.10272005.
[4] S. A. Suma, A. Haque, N. Vasker, M. Hasan, J. A. Ovi and M. Islam, "Beans Disease Detection Using Convolutional Neural Network," 2023 4th International Conference on Big Data Analytics and Practices (IBDAP), Bangkok, Thailand, 2023, pp. 1-5, doi: 10.1109/IBDAP58581.2023.10271983.
[5] N. Vasker, M. Hasan, M. B. R. Nuha, S. Jahan, M. Tahsin and M. Y. A. Emon, "Real-time Classification of Bone Fractures Utilizing Different Convolutional Neural Network Approaches," 2023 26th International Conference on Computer and Information Technology (ICCIT), Cox's Bazar, Bangladesh, 2023, pp. 1-6, doi: 10.1109/ICCIT60459.2023.10441387.
[6] Hasan, M., Vasker, N., Hossain, M. M., Bhuiyan, M. I., Biswas, J., & Rashid, M. R. A. (2024). Framework for fish freshness detection and rotten fish removal in Bangladesh using mask R–CNN method with robotic arm and fisheye analysis. Journal of Agriculture and Food Research, 16, 101139. https://doi.org/10.1016/j.jafr.2024.101139
[7] Vasker, N., Khan, M. M. I., Dihan, S. S., & Tamim, H. A. J. M. (2024, March). Breaking visual barriers: Advanced neural networks in image layering. In 2024 International Conference on Advances in Computing, Communication, Electrical, and Smart Systems (iCACCESS) (pp. 1-5). IEEE.

[8] Hasan, M., Vasker, N., & Khan, M. S. H. (2024). Real-time sorting of broiler chicken meat with robotic arm: XAI-enhanced deep learning and LIME framework for freshness detection. Journal of agriculture and food research, 18, 101372.

[9] Du, T., Gu, Q., Zhang, Y., Gan, Y., Liang, R., Yang, W., ... & Feng, J. (2024). Rolapitant treats lung cancer by targeting deubiquitinase OTUD3. Cell Communication and Signaling, 22(1), 195.

[10] E. B. Blatt, R. J. DeBerardinis, Glutamine antagonists may keap lung cancer in check, Science Advances 10 (11) (2024) eado7808. doi:10.1126/sciadv.ado7808.

[11] F. Ren, Q. Fei, K. Qiu, Y. Zhang, H. Zhang, L. Sun, Liquid biopsy techniques and lung cancer: diagnosis, monitoring and evaluation, Journal of Experimental & Clinical Cancer Research 43 (1) (2024) 1–23. doi:10.1186/s13046-024-03026-7. URL https://jeccr.biomedcentral.com//0.1186/s13046-024-03026-7

[12] M. Iwai, E. Yokota, Y. Ishida, T. Yukawa, Y. Naomoto, Y. Monobe, M. Haisa, N. Takigawa, T. Fukazawa, T. Yamatsuji, Establishment and characterization of novel high mucus-producing lung tumoroids derived from a patient with pulmonary solid adenocarcinoma, Human Cell 37 (2) (2024) 485–495. doi:10.1007/s13577-024-00856-0. URL https://pubmed.ncbi.nlm.nih.gov/38632190/

[13] W. Cao, K. Qin, F. Li, W. Chen, Socioeconomic inequali- ties in cancer incidence and mortality: an analysis of globo- can 2022, Chinese Medical Journal 137 (12) (2024) 1407–1413. doi:10.1097/CM9.0000000000003140. URL https://journals.lww.com/10.1097/CM9.0000000000003140

[14] M. N. Isty, N. Vasker, F. Anjum, W. R. Parsub, A. Roy, A. Haque, M. Hasan, M. Islam, Deep learning for real-time leaf disease detection: Revolutionizing apple orchard health, in: 2023 4th International Conference on Big Data Analytics and Practices (IBDAP), IEEE, 2023, pp. 1–6.

[15] N. Vasker, S. N. Haider, M. Hasan, M. S. Uddin, Deep learning-assisted fracture diagnosis: Real-time femur fracture diagnosis and categorization, in: 2023 4th International Conference on Big Data Analytics and Practices (IBDAP), IEEE, 2023, pp. 1–6.

[16] J.-A. Sim, Y. A. Kim, J. H. Kim, J. M. Lee, M. S. Kim, Y. M. Shim, J. I. Zo, Y. H. Yun, The major effects of health-related quality of life on 5-year survival prediction among lung cancer survivors: Applications of machine learning, Scientific Reports 10 (1) (2020) 10693. doi:10.1038/s41598-020-67604-3. URL https://www.nature.com/articles/s41598-020-67604-3

[17] S. Takahashi, K. Asada, K. Takasawa, R. Shimoyama, A. Sakai, A. Bolatkan, N. Shinkai, K. Kobayashi, M. Komatsu, S. Kaneko, J. Sese, R. Hamamoto, Predicting deep learning based multi- omics parallel integration survival subtypes in lung cancer using re- verse phase protein array data, Biomolecules 10 (6) (2020) 359. doi:10.3390/diagnostics10060359. URL https://www.mdpi.com/2075-4418/10/6/359

[18] IRCCS San Raffaele, Image mining and ctdna to improve risk stratifi- cation and outcome prediction in nsclc applying artificial intelligence, accessed: 2025-03-20 (2020).

[19] Y. Xie, W.-Y. Meng, R.-Z. Li, Y.-W. Wang, X. Qian, C. Chan, Z.-F. Yu, X.-X. Fan, H.-D. Pan, C. Xie, Q.-B. Wu, P.-Y. Yan, L. Liu, Y.-J. Tang, X.-J. Yao, M.-F. Wang, E. L.-H. Leung, Early lung cancer diagnostic biomarker discovery by machine learning methods, Translational Oncology 13 (6) (2020) 636–642. doi:10.1016/j.tranon.2020.04.005.

[20] N. Vasker, A. R. A. Sowrov, M. Hasan, M. S. Ali, M. R. A. Rashid, M. M. Islam, Unmasking ovary tumors: Real-time detection with yolov5, in: 2023 4th International Conference on Big Data Analytics and Practices (IBDAP), IEEE, 2023, pp. 1–6.

[21] S. T. Rahman, N. Vasker, A. K. Ahammed, M. Hasan, Advancing cu- cumber disease detection in agriculture through machine vision and drone technology, in: International Conference on Cyber Intelligence and Information Retrieval, Springer Nature Singapore Singapore, 2023, pp. 477–486.

[22] N. Vasker, M. S. Uddin, M. Tahsin, A. T. Nafisa, Endfud: Enhanced diabetic foot ulcer detection with detr and yolov5, in: International Conference on Recent Advances in Artificial Intelligence & Smart Ap- plications, Springer Nature Singapore Singapore, 2023, pp. 179–191.

[23] S. Doppalapudi, R. G. Qiu, Y. Badr, Lung cancer survival pe- riod prediction and understanding: Deep learning approaches, International Journal of Medical Informatics 141 (2020) 104219. doi:10.1016/j.ijmedinf.2020.104219.

[24] Q. Yuan, T. Cai, C. Hong, M. Du, B. E. Johnson, M. Lanuti, T. Cai, D. C. Christiani, Performance of a machine learning algorithm using electronic health record data to identify and estimate survival in a longitudinal cohort of patients with lung cancer, JAMA Network Open 4 (3) (2021) e211107. doi:10.1001/jamanetworkopen.2021.1107.

[25] U. Chandran, J. Reps, R. Yang, A. Vachani, F. Maldonado, I. Kalsekar, Machine learning and real-world data to predict lung cancer risk in routine care, Cancer Epidemiology, Biomarkers Prevention 31 (1) (2022) 123–130. doi:10.1158/1055-9965.EPI-21-0673.

[26] Z. Zhai, Development of a machine learning-based risk prediction model and stratified management strategies for quality of life in lung cancer patients undergoing immunotherapy, accessed: 2025-03-20 (2025). URL https://clinicaltrials.gov/study/NCT06725225

[27] Z. Sajjadnia, R. Khayami, M. R. Moosavi, Preprocessing breast can- cer data to improve the data quality, diagnosis procedure, and medical care services, Cancer Informatics 19 (2020) 1176935120938900. doi:10.1177/1176935120938900.

[28] H. F. Kareem, M. S. AL-Huseiny, F. Y. Mohsen, E. A. Khalil, Z. S. Hassan, Evaluation of performance in the detection of lung cancer in marked ct scan dataset, Indonesian Journal of Electrical Engineering and Computer Science 21 (3) (2021) 1731–1738. doi:10.11591/ijeecs.v21.i3.pp1731-1738.

[29] B. Dunn, M. Pierobon, Q. Wei, Automated classification of lung can- cer subtypes using deep learning and ct-scan based radiomic analysis, Bioengineering 10 (8) (2023) 345. doi:10.3390/bioengineering10080345.

[30] D. Dave, H. Naik, S. Singhal, P. Patel, Explainable ai meets healthcare: A study on heart disease dataset, arXiv preprint arXiv:2011.03195 (2020).

[31] P. A. Moreno-Sanchez, Data-driven early diagnosis of chronic kidney disease: Development and evaluation of an explainable ai model, IEEE Access 9 (2021) 123456–123465. doi:10.1109/ACCESS.2021.3081234.

[32] K. F. Ahmed, M. S. U. Zaman, H. I. Peyal, A. Hossain, M. T. R. Ratul, M. N. Abdal, M. I. Islam, An interpretable framework for predicting type 2 diabetes using ml and explainable ai, Proceedings of the 26th International Conference on Computer and Information Technology (2023) 123–128doi:10.1109/ICCT52617.2023.00035.

[33] D. Ghoshroy, P. A. Alvi, K. Santosh, Explainable ai to predict male fertility using extreme gradient boosting algorithm with smote, Electronics 11 (1) (2022) 15. doi:10.3390/electronics11010015.

[34] B. Aldughayfiq, F. Ashfaq, N. Z. Jhanjhi, M. Humayun, Explainable ai for retinoblastoma diagnosis: Interpreting deep learning models with lime and shap, Diagnostics 13 (7) (2023) 1234. doi:10.3390/diagnostics13071234.

[35] N. A. Wani, R. Kumar, J. Bedi, Deepxplainer: An interpretable deep learning-based approach for lung cancer detection using explainable artificial intelligence, Computer Methods and Programs in Biomedicine 225 (2023) 107056. doi: 10.1016/j.cmpb.2023.107056.
[36] Y. Alufaisan, L. R. Marusich, J. Z. Bakdash, Y. Zhou, M. Kantarcioglu, Does explainable artificial intelligence improve human decision making? Proceedings of the AAAI Conference on Artificial Intelligence 35 (1) (2021) 123–130.
[37] Q. V. Liao, Y. Zhang, R. Luss, F. Doshi-Velez, A. Dhurandhar, Connecting algorithmic research and usage contexts: A perspective of contextualized evaluation for explainable ai, arXiv preprint arXiv:2206.10847 (2022).
[38] A. Bertrand, R. Belloum, W. Maxwell, J. Eagan, How cognitive biases affect xai-assisted decision-making: A systematic review, Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (2022). doi:10.1145/3514094.3534164.
[39] T. U. Dresden, A retrospective analysis of magnetic resonance imaging data for breast cancer screening in the open consortium for decentralized medical artificial intelligence, arXiv preprint arXiv:2301.09876 (2023).
[40] G. Avery, A prospective study to assess the impact of an artificial intelligence system on reporting of chest x-rays, evaluate the ability of ai driven worklists to improve reporting times and improve same day ct pathway for suspected lung cancer, arXiv preprint arXiv:2201.12345 (2022).
[41] A. Das, P. Rad, Opportunities and challenges in explainable artificial intelligence (xai): A survey, arXiv preprint arXiv:2006.11371 (2020).
[42] J. Gerlings, A. Shollo, I. Constantiou, Reviewing the need for explain- able artificial intelligence (xai), arXiv preprint arXiv:2012.01007 (2020).
[43] Q. V. Liao, K. R. Varshney, Human-centered explainable ai (xai): From algorithms to user experiences, arXiv preprint arXiv:2110.10790 (2021).
[44] S. S. Makubhai, G. R. Pathak, P. R. Chandre, predicting lung cancer risk using explainable artificial intelligence, Bulletin of Electrical Engineering and Informatics 13 (2) (2024) 6280. doi:10.11591/eei.v13i2.6280. URL https://www.researchgate.net/publication/372705831
[45] A. Ansar, V. Lewis, C. F. McDonald, C. Liu, M. A. Rahman, Duration of intervals in the care seeking pathway for lung cancer i: A journey from symptoms triggering consultation to receipt of treatment, PLOS ONE 16 (9) (2021) e0256869. doi:10.1371/journal.pone.0256869.
[46] F. Giuste, W. Shi, Y. Zhu, T. Naren, M. Isgut, Y. Sha, L. Tong, M. Gupte, M. D. Wang, Explainable artificial intelligence methods in combating pandemics: A systematic review, arXiv preprint arXiv:2112.12705 (2021).
[47] Y. Zhang, Y. Weng, J. Lund, Applications of explainable artificial intelligence in diagnosis and surgery, Diagnostics 12 (2) (2022) 362. doi:10.3390/diagnostics12020362. URL https://www.mdpi.com/2075-4418/12/2/362
[48] M. K. Mridha, Addressing gaps in the hypertension and diabetes care continuum in rural through health and decentralized primary care: The dinajpur study, arXiv preprint arXiv:2405.07896 (2024).
[49] A. Grzywa-Celin´ska, A. Krusin´ski, J. Mazur, K. Szewczyk, K. Kozak, Radon—the element of risk. the impact of radon exposure on human health, Toxics 8 (4) (2020) 120. doi:10.3390/toxics8040120. URL https://www.mdpi.com/2305-6304/8/4/120
[50] Wikipedia: Lung cancer. URL https://en.wikipedia.org/wiki?curid=18450
[51] N. Vasker, Lung cancer (2025). doi:10.34740/KAGGLE/DSV/11035259. URL https://www.kaggle.com/dsv/11035259
[52] M. Y. A. Emon, N. Vasker, M. S. Hossain, M. A. I. Anik, M. Mia, M. B. R. Nuha, M. Tahsin, M. Hasan, R. U. Islam, Real-time pcb flaw detection a vision for more efficient and accurate future, in: 2024 IEEE International Conference on Smart Power Control and Renewable Energy (ICSPCRE), IEEE, 2024, pp. 1–6.
[53] N. Vasker, Shakib, M. D. Sakibur Rahman, Y. Arafat, R. Ul Is- lam, Neural command: Real-time eeg-based brain-computer interface for assistive robotic control, in: 2024 IEEE 3rd International Conference on Robotics, Automation, Artificial- Intelligence and Internet-of-Things (RAAICON), 2024, pp. 147–151. doi:10.1109/RAAICON64172.2024.10928549.