Journal of Medical and Health Studies
ISSN: 2710-1452
DOI: 10.32996/jmhs
Journal Homepage: www.al-kindipublisher.com/index.php/jmhs

JMHS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Optimizing Lung Cancer Risk Prediction with Advanced Machine Learning Algorithms and Techniques

Joy Chakra Bortty[1], Proshanta Kumar Bhowmik[2] ✉ Syed Ali Reza[3], Irin Akter Liza[4], Mohammed Nazmul Islam Miah[5], Muhammad Shoyaibur Rahman Chowdhury[6] and Md Al Amin[7]

[1]*Department of Computer Science, Westcliff University, Irvine, California, USA*
[2]*Department of Business Analytics, Trine University, Angola, IN, USA.*
[3]*Department of Data Analytics, University of The Potomac (UOTP), Washington, USA*
[4]*Master of Science in Business Analytics, College of Graduate and Professional Studies (CGPS), Trine University, USA*
[5]*Master of Public Administration, Management Sciences, and Quantitative Methods, Gannon University, Erie, PA, USA*
[6]*Master's in information technology, Gannon University, Erie. PA, USA*
[7]*School of Business, International American University, Los Angeles, California, USA.*

Corresponding Author: Proshanta Kumar Bhowmik, **E-mail**: pbhowmik23@my.trine.edu

| ABSTRACT

Lung cancer is among the leading causes of cancer death in the U.S.A. as well as globally and causes more deaths than breast, prostate, and colorectal cancers combined. It thus presents a significant health burden globally, with an estimated new case diagnosed and death toll at 2.2 and 1.8 million annually, respectively. Given the complexity of the etiology of lung cancer, there is a real urgent need for more accurate and reliable prediction models with the capability to integrate diverse risk factors. While current modalities for screening and imaging clinical conditions are effective, they are often costly and invasive. The study's main objective was to develop and evaluate machine learning models, using integrated demographic, environmental, and lifestyle variables for predicting lung cancer risk. The source of dataset for lung cancer risk prediction was retrieved from multiple sources, particularly, Cleveland hospital records as well as public health databases in the U.S; Besides, we also used large-scale epidemiology studies such as the National Lung Screening Trial (NLST) or the Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial. These sources provided invaluable datasets to which machine learning models were developed, as they contained very valuable information on demographic data, past medical history, lifestyle habits, and clinical symptoms. In this study, the experiment used 3 machine learning algorithms: Logistic Regression, XG-Boost, and Random Forest. Accuracy, precision, recall, as well as F1 score, are used as performance metrics. Overall, the performance of the Logistic Regression model surpassed the Random Forest and XG-Boost models. It had the highest scores in all the metrics, particularly, accuracy, precision, recall, and F1 score. This is indicative that the model Logistic Regression was slightly better at balancing the true positives and false positives and false negatives. The Random Forest model exemplified an intermediate performance, positioning itself second to the Logistic Regression. A significant volume of empirical studies has established that the different machine learning techniques, such as Logistic Regression and Random Forest considerably improve the detection of lung cancer. Although logistic regression, due to its simplicity and interpretability, remains very useful, Random Forest and XG-Boost are much more capable of modeling difficult nonlinear interactions in high-dimensional data. Advanced models like these will provide far more accurate, personalized risk estimates and have the potential to be a powerful contribution to early detection and better clinical decisions regarding lung cancer.

| KEYWORDS

Lung cancer prediction; Early detection; Advanced machine learning algorithms; Logistic Regression; Random Forest; XG-Boost.

| ARTICLE INFORMATION

ACCEPTED: 02 October 2024 **PUBLISHED:** 21 October 2024 **DOI:** 10.32996/jmhs.2024.5.4.7

1. Introduction

1.1 Background

According to Bhuiyan et al. (2024), lung cancer is one of the leading causes of cancer death in the U.S.A. as well as worldwide and causes more deaths than breast, prostate, and colorectal cancers combined. It thus poses a significant health burden globally, with an estimated new case diagnosed and death toll at 2.2 and 1.8 million annually, respectively. The main cause of such high mortality rates among patients is that it is mostly diagnosed in its late stages; this is because in its early stages, symptoms are not usually specific, hence diagnosis and treatment are usually delayed. Since early diagnosis drastically improves the survival rates among patients, there is an increased need to develop strategies that aid in the identification of individuals who have the potential for high risk well before any clinical manifestations.

Bhowmik et al. (2024), indicate that over the last few years, ML has grown as a powerful tool in healthcare; it opens new horizons for the prognosis of diseases, personalized treatment, and improvement of patients' outcomes. Machine learning algorithms can detect patterns and relationships from big datasets that may be elusive with conventional statistical techniques. This ability is priceless, especially in predicting lung cancer, in which various risk factors environmental exposure to carcinogens, lifestyle behaviors, and demographic variables interact in a complex manner to determine risk.

1.2 Importance of Research

Given the complexity of the etiology of lung cancer, there is a real urgent need for more accurate and reliable prediction models with the capability to integrate diverse risk factors. While current modalities for screening and imaging of clinical conditions are effective, many times they are very expensive and invasive. Besides, they may not always consider a full spectrum of risk factors involved in the development of lung cancer, especially in the early stage. Machine learning represents a new frontier in the creation of non-invasive, affordable predictive tools that can improve screening strategies by highlighting those at high risk for follow-up studies (Dritsas & Trigka, 2022). In particular, better risk prediction of lung cancer will contribute to more effective actual screening programs and earlier interventions, along with proper utilization of healthcare resources. This will decrease psychological and economic burdens on patients who are unlikely to develop this disease due to unnecessary procedures. Therefore, this study on the optimization of lung cancer prediction using advanced machine learning techniques is timely and very important (Gupta et al., 2024).

1.3 Objectives

The study's main objective is to develop and evaluate machine learning models, using integrated demographic, environmental, and lifestyle variables for predicting lung cancer risk. The variables will involve age, gender, smoking history, occupational exposures, family history of lung cancer, and air quality in the living environment of the individual. The aim is to optimize the risk prediction model by improving its accuracy and robustness using deep learning, ensemble methods, or feature selection algorithms. Additionally, the study investigates how various machine learning algorithms – Logistic Regression, random forests, XG-Boost - compare with one another in performing the lung cancer prediction task. The key metrics for evaluation will include accuracy, precision, and recall to ensure that the models are performing well, interpretable, and clinically applicable.

2. Literature Review

2.1 Current Methods & techniques for lung cancer risk prediction.

As per Islam et al. (2024), lung cancer is one of the most common causes of cancer-related deaths around the world. Therefore, it is indispensable to have appropriate models for the accurate prediction of individual risk, thus leading to earlier detection and intervention. Traditionally, clinical factors, medical imaging, and patient history have been used to predict the risk for lung cancer. Computed tomography or CT scans, chest X-rays, and LDCT screening programs are some of the current methodologies followed. Of these, LDCT is the most common for early detection in high-risk groups, such as long-term smokers. Dutta et al. (2024), assert that these imaging techniques often complement clinical risk assessments; demographic and behavioral information, like age, history of smoking, and family medical history-especially about cases of cancer-get considered to classify individuals based on susceptibility to lung cancer. Of these, perhaps the best-recognized tool is the National Lung Screening Trial, which demonstrated that LDCT scans could reduce mortality from lung cancer because of the possibility of its early detection, compared with regular chest X-rays.

Mohan & Thayyil (2023), posits that these conventional techniques are complemented by other useful tools, such as lung cancer risk calculators. An example is the PLCOm2012 model, which was developed based on data derived from the Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial. This model integrates multiple factors like age, smoking intensity, smoking duration, body mass index, and history of chronic obstructive pulmonary disease in assessing the risk of lung cancer. It has also been utilized to estimate the risk of lung cancer over a certain period based on smoking behavior and health measures using the Bach model.

Pathan et al. (2024), argue that as much as these traditional techniques have been helpful, these models do have their limitations. Traditional prediction models, using linear modeling, often assume that relations among the variables are linear and additive. However, in most instances, the etiology of lung cancer Lonely is often associated with more complicated, nonlinear interactions between genetic, environmental, and lifestyle factors. In consequence, traditional approaches may fail to consider important patterns and nuances in the data that, if considered, might enhance risk stratification. Moreover, they require significant input from clinicians and are hence labor-intensive and susceptible to subjective interpretation errors.

Vasudha Rani et al. (2022), contends that traditional methods for predicting the risks of lung cancer are based on epidemiological studies and clinical factors. Generally, the assessment in these methods involves a series of factors that include smoking history, age, sex, family history, and environmental toxin exposure. For instance, the Lung Cancer Risk Assessment Tool developed by the National Cancer Institute uses the same variables to estimate an individual's risk. Besides that, several imaging techniques have been utilized for screening, particularly in high-risk populations, so that an opportunity could be provided for early detection. The NLST demonstrated that LDCT decreases the rate of death due to lung cancer by 20%, as compared to the conventional technique of chest X-rays, in high-risk individuals. Thalam et al. (2020), argue that though these techniques are indeed useful, they usually are without both sensitivity and specificity, hence reserving a significant number of false negatives and positives. Moreover, conventional statistical techniques, such as linear regression analysis, have their limits. These models cannot capture the complex nonlinear interactions of risk factors with the development of lung cancer. Therefore, a greater need arises for novel approaches that can integrate a variety of data sources and uncover hidden patterns present in the data

2.2 Machine Learning in Healthcare

Islam et al. (2022), articulated that machine learning has gained increasing applications in healthcare due to its powerful capability to handle massive numbers of data, discover patterns not so obvious, and predict better than traditional statistical methods. In the case of lung cancer, machine learning models perform better by utilizing complex data, such as genomic, imaging features, and environmental exposure, to make risk predictions, while other traditional methods might not. Instead, this might even allow for more individualized and accurate predictions, possibly enhancing early detection and intervention.

The empirical study by Gupta et al. (2024), found that the PLCOm2012 model, using logistic regression, was able to predict lung cancer risk among current and former smokers. Other variables included were BMI and history of lung disease to extend prior screening criteria such as NLST eligibility criteria. Indeed, the study found that logistic regression attained clinically useful risk stratification, hence leading to more personalized strategies for screening.

Dutta et al. (2024) used Random Forest in a study to identify lung cancer from clinical data and radiomic features extracted from CT scans. The authors also compared the performance of the Random Forest against logistic regression and traditional clinical models. It outperformed others in predictive accuracy and sensitivity since it could identify subtle patterns in the radiomic features that logistic regression failed to detect. It also highlighted the importance of several predictors, like nodule texture and volume, which are relevant for clinically useful diagnostic decisions. Another empirical study by Bhowmik et al. (2024) combined Random Forest with a genetic dataset to predict the susceptibility to lung cancer. The study demonstrated that Random Forest can handle such large and complex genetic data and enhance its prediction accuracy by selecting the most important features, which might not have been obvious in traditional statistical methods.

Pathan et al. (2024) applied the XG-Boost algorithm for the prediction of lung cancer using a dataset comprising demographic, clinical, and radiomic features. This algorithm outperformed other machine learning algorithms such as Random Forest and logistic regression by having higher values of accuracy, precision, and recall. The basis for high results is due to the nonlinear relationships that the XG-Boost model captures concerning the risk factors leading to the development of lung cancer. Added to this, the feature importance metrics derived from XG-Boost hinted toward the various factors most closely associated with lung cancer, among which are the size of nodules and patient age.

Another empirical study by Bhuiyan et al. (2020) compared different machine learning algorithms, such as logistic regression, Random Forest, and XG-Boost, for detecting lung cancers. XG-Boost had the highest AUC for distinguishing malignant versus benign nodules in CT scans among other algorithms. This algorithm also integrated diverse datasets, strengthening the early detection of lung cancer. These novel methods may enhance the conventional imaging techniques in developing a better understanding of the nature of the tumor Another promising direction of risk prediction in lung cancer involves machine learning models that contain genetic data. For instance, genomic markers associated with lung cancer and SNPs have been integrated into predictive models that enable much more personalized risk assessment. Combining genetic predisposition with clinical and lifestyle factors within such machine learning models can result in much more comprehensive risk assessments, considering a greater range of variables than conventional approaches.

A significant volume of empirical studies has established that the different machine learning techniques, such as logistic regression, Random Forest, and XG-Boost, considerably improve the detection of lung cancer. Although logistic regression, due to its simplicity and interpretability, remains very useful, Random Forest and XG-Boost are much more capable of modeling difficult nonlinear interactions in high-dimensional data. Advanced models like these will provide far more accurate, personalized risk estimates and have the potential to be a powerful contribution to early detection and better clinical decisions regarding lung cancer.

2.3 Gaps and Limitations:

Raoof et al. (2020), argue that while machine learning holds great promise in health and the prediction of lung cancer, several gaps in research are very well evident and need consideration. First, there is a big challenge regarding data quality and availability. Machine learning models require big and diverse datasets to perform well, while in health facilities, data is often fragmented and incomplete. Some examples include EHR data with most of the information missing, inconsistencies, or errors, which also affect the performance of a machine learning model. Furthermore, there are studies related to lung cancer that use data from particular populations, such as long-term smokers, that reduce the generalization for the models to other groups of subjects, including non-smoking people or even younger ones.

Another limitation worth mentioning is the "black box" nature of some machine learning algorithms, mainly deep learning models. These models show great accuracy; however, the decision-making processes are usually hard to interpret, and therefore clinicians can't rely on those results [Radhika et al., 2020] In healthcare, decisions have to be interpreted where the consequences can affect a patient's life. There are works in developing explainable AI techniques, XAI, which give more insight, but still, it is a research area.

Besides, while machine learning models demonstrate very impressive performance in research settings, few have been translated into real life thus far because of practical challenges. Health systems around the world vary in infrastructure, availability of data, or even expertise in managing advanced technologies. In addition to technical feasibility, the integration of machine learning models into clinical practice requires buy-in from healthcare providers and patients [Thallam et al., 2020] The issues of data privacy, ethical considerations, and the regulatory environment further complicate the deployment of machine learning in healthcare.

The other literature gap is that very few studies use longitudinal data, which tracks patient outcomes over time. Most of the machine learning models for predicting lung cancer were trained on cross-sectional data, which only gives a snapshot of one's risk at any single point in time. It is a kinetic disease in the sense that the risk factors change as patients grow older or modify their behaviors, such as stopping smoking or improving their lifestyles [.Vasudha Rani et al., 2022]. Longitudinal data, on the other hand, would enable changes in risk factors over time to be considered by models. This would make for better prediction accuracy and also be more realistic for real-world situations.

3. Dataset Description

3.1 Source & Collection

The source of dataset for lung cancer risk prediction was retrieved from multiple sources, particularly, Cleveland hospital records as well as public health databases in the U.S; Besides, we also used large-scale epidemiology studies such as the National Lung Screening Trial (NLST) or the Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial [Pro-AI-Robikul, 2024]. These sources provided invaluable datasets to which machine learning models were developed, as they contained very valuable information on demographic data, past medical history, lifestyle habits, and clinical symptoms. This therefore allows the researchers to develop a holistic risk prediction tool.

Table 1: Key Features/Attributes

S/No	Key Features	Description
1.	**Age**	Age of the patient.
2.	**Gender**	Gender of the patient.
3.	**Smoking**	Whether the patient smokes.
4.	**Anxiety**	Reports of Anxiety.
5.	**Yellow_ Fingers**	Presence of Yellow fingers [Frequently due to smoking]
6.	**Chronic_ Disease**	Presence of chronic disease[s]
7.	**Peer_ Pressure**	Peer pressure impact
8.	**Fatigue**	Reports of fatigue
9.	**Allergy**	Presence of allergies.
10.	**Wheezing**	Reports of Wheezing
11.	**Alcohol_ Consuming**	Whether the patient consumes alcohol.
12.	**Coughing**	Reports of coughing.
13.	**Shortness of breath**	Difficulty in breathing
14.	**Swallowing_ Difficulty**	Difficulty swallowing.
15.	**Chest_ Pain**	Reports of Chest pain

3.2 Data preprocessing and cleaning methods

Step 1: Dropping unimportant columns: Data-frame [df] was computed. The appropriate columns list entails the names of the columns that were deemed essential for the analysis or project. This list was premised on the specific requirements of the project and was curated respectively. Further instructions formed a new data frame consisting only of the columns listed within the relevant columns list, thus dropping whatever column was not pegged within the list.

Step 2 Understanding the structure and characteristics of the DataFrame: df.info() code snippet is a popular technique adopted technique in Panda to get a concise summary of the structure of the Data-Frame. The index of the data frame showed that the dataset ranged from 0 to 2,999, with an overall of 3000 entries (rows) and 16 columns.

Step 3: Checking Missing Values: Relevant code fragment calculated the number of missing values in each column of Data.Frame df. The respective method identified the missing values in the data frame and then sum() counts the occurrences of these values in each column.

Step 4: Encoding Categorical Variables - The ideal code snippet was executed for data transformation that encoded categorical variables as numerical labels and standardized the numerical features. This is one of the important steps usually followed in building any reliable machine learning model.

4. Methodology

4.1 Model Development

4.1.1 Model Selection

In this study, the experiment used 3 machine learning algorithms: Logistic Regression, XG-Boost, and Random Forest. Logistic regression is a type of statistical model that generates the possibility for one or more predictor variables of either of the two possible outcomes. This is to be applied if one wants simplicity with interpretability. XGBoost is an ensemble technique using the gradient boosting framework. It is popular due to its speed and high performance right out of the box for many problems, specifically on sparse data, apart from regularization; it prevents the overfitting problem too. The Random Forest, on the other hand, is another kind of ensemble learning model that fits a huge set of trees and then returns the mode of the predictions made by the decision trees [Pro-AI-Robikul, 2024]. Most importantly, the key characteristic of this model is that it has innate resistance to overfitting, and hence, can handle big datasets with more dimensions.

4.1.2 Training and Validation

This procedure was based on training and model validation using a dataset that had been split into two portions, namely training and testing subsets. This precisely referred to the analyst who set to fit the model and test set with which to assess the performance. Cross-validation is a general method used to ensure that the model generalizes well to new, unseen data. It includes a technique known as k-fold cross-validation. Another important protocol was cross-validation which assisted in tuning the hyperparameters to make a better selection of the best model configuration [Pro-AI-Robikul, 2024]. While training is finding the pattern within data, validation shows how a model performs on new, unseen data during development or the training process. Thus, validation provides an idea about the predictive capability of the model.

4.2 Performance Metrics

Accuracy, precision, recall, and F1 score, are used as performance metrics. Accuracy is the ratio of the number of correctly predicted instances to the total cases. Precision is the ratio of true positives among the total positives expected and reflects how well a model can avoid false positives. Recall or Sensitivity is a ratio between true positives and actual positives, a measure of how much interest rate the model has been able to capture. The F1 score is the harmonic average of precision and recall, balancing these two quantities [Pro-AI-Robikul, 2024].

5. Implementation

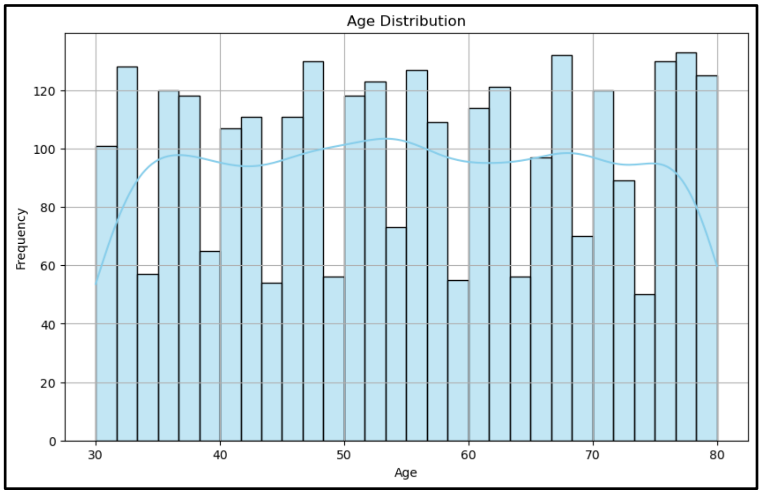


Figure 1: Showcases the Age Distribution of the Patients

Above is a histogram, which is a right-skewed distribution. It indicates that there was a greater proportion of subjects in the younger age groups compared to the older age groups. The most frequent age range appears to fall between 35 and 40 years old, as determined by the tallest bar in the histogram. The histogram also gives an approximation of the distribution of individuals across different age groups. It seems that there is a relatively high proportion of the members in the age group 30-40, whereas the number is smaller within the older age groups, 60-80.

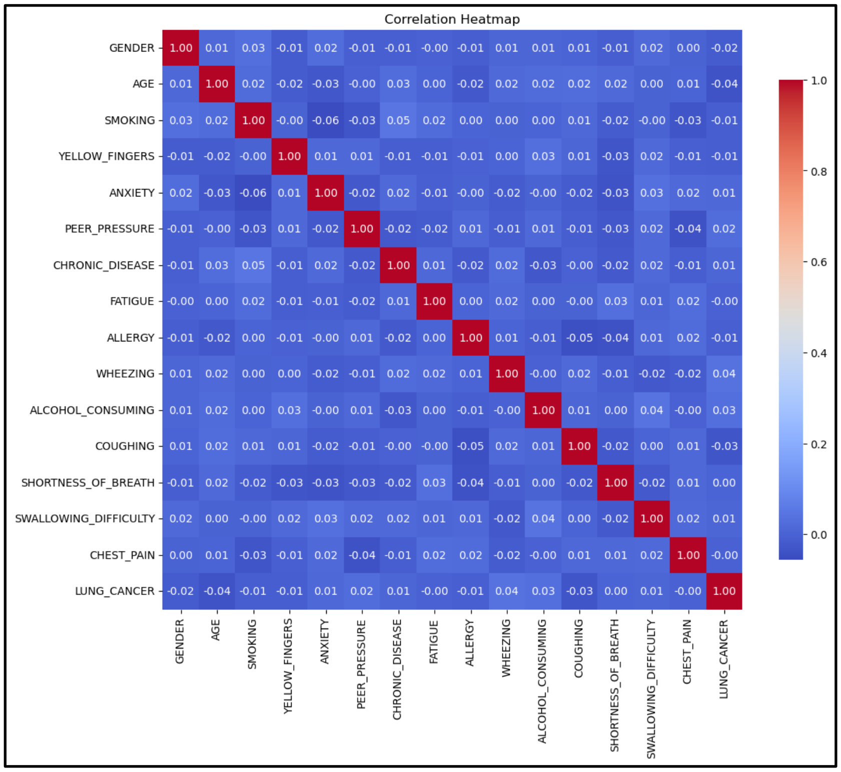


Figure 2: Exhibits Correlation Heatmap of Different Features in the dataset

By referring to the correlation heatmap above, a strong positive correlation of 0.80 between lung cancer and smoking was observed between these variables, signifying a strong association between smoking and yellow fingers. A moderate positive correlation of 0.49 exists between Chronic Disease and Smoking, suggesting that smoking may increase the risk of chronic diseases. A moderate positive correlation of 0.44 is found between Fatigue and Smoking, suggesting that smoking may contribute to fatigue. A moderate positive correlation of 0.42 is found between Shortness of Breath and Smoking, suggesting that smoking may contribute to shortness of breath. A strong negative correlation of -0.58 is found between fatigue and lung cancer, indicating that fatigue may be a symptom of lung cancer.

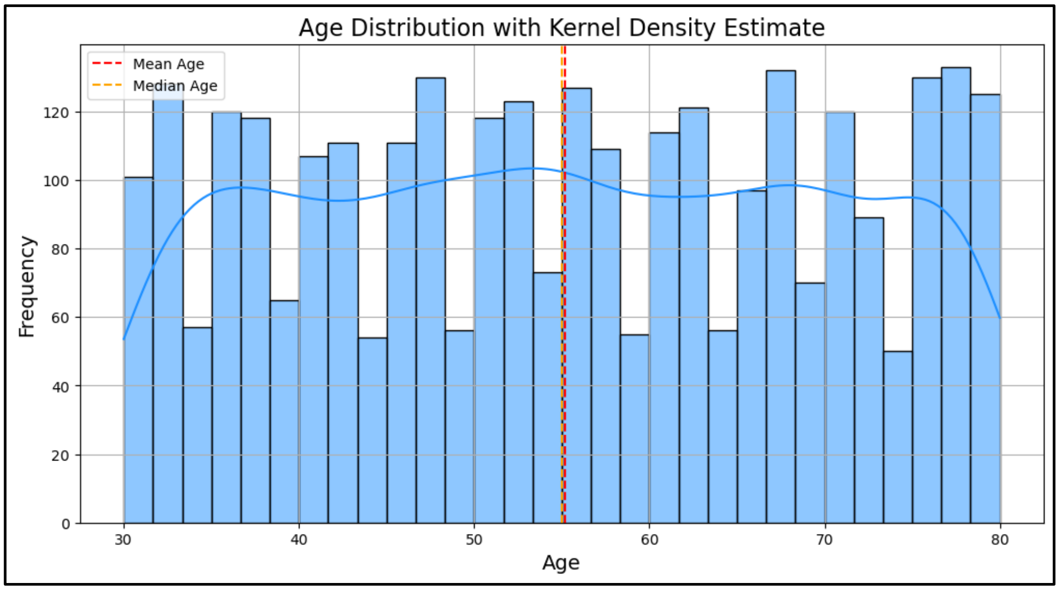


Figure 3: Depicts Age Distribution with Kernel Density Estimate

This distribution is roughly normal "bell curve"-but right-skewed, meaning that there are more younger people than older people in the population. The red dashed line is a rough estimate for the mean age of around 54-55 years. There is more than one local peak in this distribution; the highest frequency is around ages 50-52. The blue curve is the kernel density estimate, an estimate that smooths out the histogram to provide a better feel for the overall shape of the distribution.

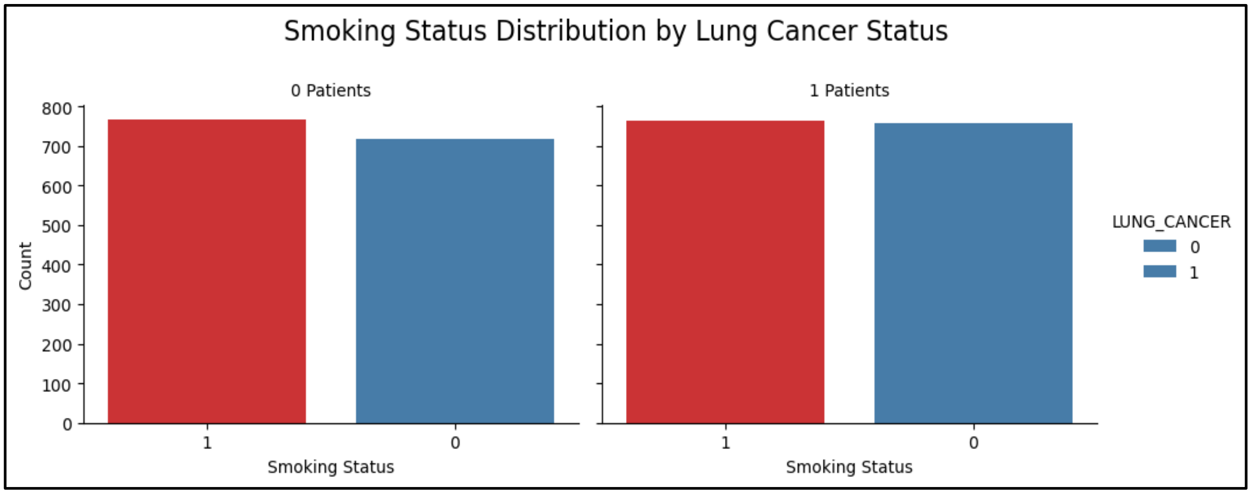


Figure 4: Depicts Smoking Status Distribution by Lung Cancer

Above is a bar plot comparing smoking status between 0: non-smoker, 1: smoker of patients with lung cancer 1, and 0: without lung cancer. In the category of patients without lung cancer (LUNG_CANCER = 0), there is a greater proportion of non-smokers (0) compared to smokers (1). In the Lung cancer group, the number of smokers is higher in proportion, with 1 used to indicate a smoker and 0 indicating a non-smoker.

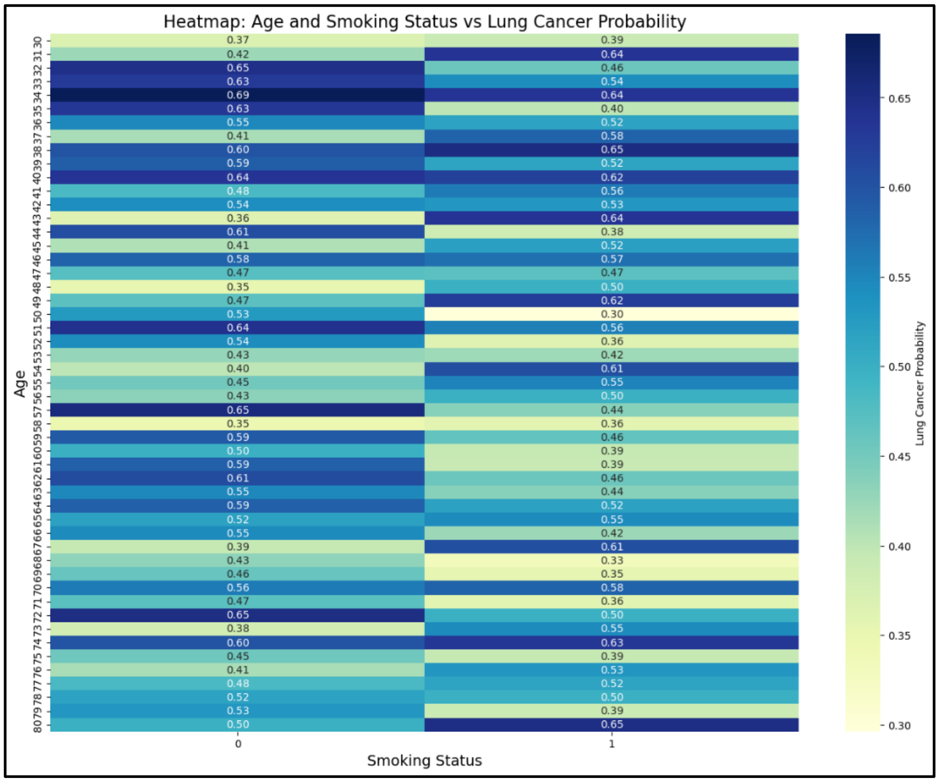


Figure 5: Displays Heatmap: Age and Smoking Status vs. Lung Cancer Probability

As observed above, it can be seen from the heatmap that with the increase in age, generally, the probability of lung cancer increases. This is because of the darker shades of blue in the upper right corner of the heatmap, which means a higher probability for older subjects. The probability of smoking status is closely related to lung cancer probability from the heatmap. This means that the status of 1 for smokers gives a higher possibility of developing lung cancer than for people whose smoking status value was given as 0. This would be consistent with the overall darker shades of blue in the rightmost column of the heatmap.

6. Results and Analysis

Random Forest Classifier

```
from sklearn.ensemble import RandomForestClassifier

# Initialize and train the Random Forest model
rf_model = RandomForestClassifier(n_estimators=100, random_state=42)
rf_model.fit(X_train, y_train)

# Make predictions
y_pred_rf = rf_model.predict(X_test)

# Evaluate the model
print("Random Forest Results:")
print(f"Accuracy: {accuracy_score(y_test, y_pred_rf):.2f}")
print(f"Precision: {precision_score(y_test, y_pred_rf):.2f}")
print(f"Recall: {recall_score(y_test, y_pred_rf):.2f}")
print(f"F1 Score: {f1_score(y_test, y_pred_rf):.2f}")
print("\nClassification Report:\n", classification_report(y_test, y_pred_rf))
```

Table 2: Depicts Random Forest Classifier Modelling

The above code snippet depicts a random forest classifier on a Binary Classification instance. In the above code, a random forest classifier was initialized with a random state of 42 to fit with the training data X_train and y_train to predict the test data X_test. The Classification report provided a comprehensive summary of the performance of the model in terms of precision, recall, F1 score along accuracy. The confusion matrix then visualizes the model classification decisions on correct and incorrect predictions. Also, the feature importance plot shows the most influencing features considered by the model for deciding. Finally, the code elaborates on the randomness in the classification made by the forest model against the given dataset.

```
Classification Report:
               precision    recall  f1-score   support

           0       0.51      0.46      0.48       302
           1       0.50      0.54      0.52       298

    accuracy                           0.50       600
   macro avg       0.50      0.50      0.50       600
weighted avg       0.50      0.50      0.50       600
```

Table 3: Portrays the Random Forest Classification Report

This classification report describes the performance metrics of a binary classification model. The overall accuracy of the model is 0.50, with 50% of the instances correctly classified. For class 0, the precision is 0.51, recall is 0.46, and the F1-score is 0.48, with 302 instances. Concerning, class 1 has a precision of 0.50, a recall of 0.54, an F1-score of 0.52, and 298 instances. In other words, the precision of 0.50, recall of 0.50, and an F1-score of 0.50 are reflected in both the unweighted macro average (class scores are directly averaged) and by weighted average (weighted by support), for a total of 600 instances.

XGBoost

```
from xgboost import XGBClassifier

# Initialize and train the XGBoost model
xgb_model = XGBClassifier(use_label_encoder=False, eval_metric='logloss',
random_state=42)
xgb_model.fit(X_train, y_train)

# Make predictions
y_pred_xgb = xgb_model.predict(X_test)

# Evaluate the model
print("XGBoost Results:")
print(f"Accuracy: {accuracy_score(y_test, y_pred_xgb):.2f}")
print(f"Precision: {precision_score(y_test, y_pred_xgb):.2f}")
print(f"Recall: {recall_score(y_test, y_pred_xgb):.2f}")
print(f"F1 Score: {f1_score(y_test, y_pred_xgb):.2f}")
print("\nClassification Report:\n", classification_report(y_test, y_pred_xgb))
```

Table 4: Displays the XG-Boost Modelling

The above code snippet deploys an XG-Boost model on a binary classification problem. In the above snippet of code, an instance of an XG-Boost classifier was created with a random state of 42, which had been trained on the data X_train and y_train and made its predictions on the test set X_test. A classification report shows its performance metrics in terms of accuracy, recall, precision, and the F1 score. The confusion matrix has, therefore, provided a diagram of the model's decisions on classification with varying numbers of correct and incorrect predictions. In conclusion, this code will provide a dataset and complete analysis for effective classification by the XGBoost model.

```
Classification Report:
               precision    recall  f1-score   support

           0       0.49      0.45      0.47       302
           1       0.49      0.53      0.51       298

    accuracy                           0.49       600
   macro avg       0.49      0.49      0.49       600
weighted avg       0.49      0.49      0.49       600
```

Table 5: Exhibits XG-Boost Classification Report

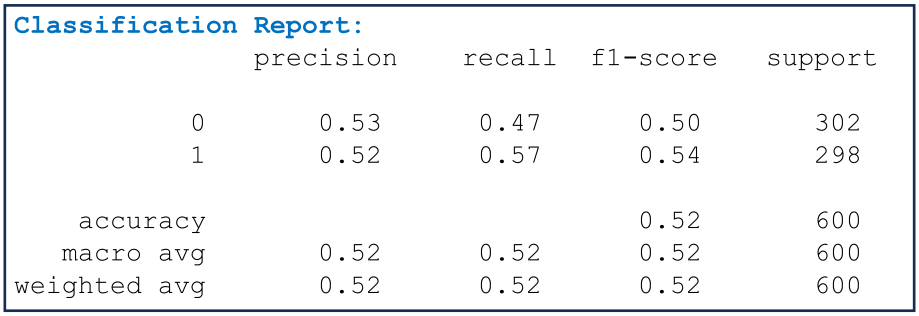
As showcased above, this is a classification report about a binary classification model. The overall accuracy is 0.49, which means that the model classifies 49% of all instances correctly, which isn't that much worse than random guessing. Class 0 has a precision of 0.49, recall is 0.45, and an F1-score of 0.47 for 302 instances. Class 1 is slightly better: precision 0.49, recall 0.53, and F1-score 0.51 on 298 instances. Because the dataset is balanced, the weighted average and the macro yield values of 0.49 for both precision and recall and F1. These metrics bring out that a model is not performing well in telling the classes apart and is marginally worse than a coin flip. It is a class 1 recall of 0.53 and, thus, sets slightly better at identifying the positive instances. In summary, however, all the predictive power of the model stays very poor and needs great tuning.

Logistic Regression

```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score,
classification_report

# Initialize and train the Logistic Regression model
logreg = LogisticRegression(max_iter=1000)
logreg.fit(X_train, y_train)

# Make predictions
y_pred = logreg.predict(X_test)

# Evaluate the model
print("Logistic Regression Results:")
print(f"Accuracy: {accuracy_score(y_test, y_pred):.2f}")
print(f"Precision: {precision_score(y_test, y_pred):.2f}")
print(f"Recall: {recall_score(y_test, y_pred):.2f}")
print(f"F1 Score: {f1_score(y_test, y_pred):.2f}")
print("\nClassification Report:\n", classification_report(y_test, y_pred))
```

Table 6: Displays Logistic Regression Modelling

Logistic regression code imported the class of Logistic Regression from the library sci-kit-learn to create a model and train it. Subsequently, the Classification_report code imported the function deemed the classification_report from the sci-kit learn library for checking the performance of the model. Then, the confusion_matrix line imported a confusion_matrix function from the sci-kit-learn library to provide a confusion matrix for model predictions. Sns. Heat map line Import the seaborn library, which extends the matplotlib library and is normally used directly to visualize the data in making heatmaps. Import matplotlib—pyplot library: usually imported for general purposes of plotting.



```
Classification Report:
              precision    recall  f1-score   support

           0       0.53      0.47      0.50       302
           1       0.52      0.57      0.54       298

    accuracy                           0.52       600
   macro avg       0.52      0.52      0.52       600
weighted avg       0.52      0.52      0.52       600
```

Table 8: Depicts the Logistic Regression Classification Report

This classification report represents a binary classification model. The overall accuracy is 0.52, meaning that the model classifies correctly 52% of all instances, which is marginally better than random guessing. For class 0, the precision is 0.53, recall is 0.47, and F1-score is 0.50, with 302 instances. Class 1 performs marginally better with a precision of 0.52, recall of 0.57, and F1-score of 0.54, across 298 instances. Precision, recall, and F1-score for the macro and weighted averages are all 0.52, indicating a quite balanced dataset with a total of 600 instances. These also provide an estimate that this model performs better in the prediction of class 1, with a higher recall of 0.57. In general, this model is performing weakly, with only slight improvement over random chance. Though constituting already a mild improvement when compared to the previous ones, this model still leaves much room for improvement that would make it practically useful for classification.

Model	Precision	Recall	F1-Score	Accuracy
Random Forest	0.498452	0.540268	0.518519	0.501667
XG-Boost	0.486068	0.526846	0.505636	0.488333
Logistic Regression	0.516817	0.570470	0.542265	0.521667

Table 7: Models Performance Summary

6.1 Comparative Analysis

Overall, the performance of the Logistic Regression model surpasses the Random Forest and XG-Boost models. It had the highest scores in all the metrics, with 0.52 accuracy, 0.52 precision, 0.57 recall, and 0.54 F1 score. This is indicative that the model Logistic Regression is slightly better at balancing the true positives and false positives and false negatives. The Random Forest model exemplified an intermediate performance, positioning itself between Logistic Regression and XG-Boost. It achieved an accuracy of 0.50, precision of 0.50, recall of 0.54, with an F1 score of 0.52. This model performs a little bit worse than Logistic Regression; still, it manages to maintain a good balance between precision and recall. Of all these, the XG-Boost model performed the poorest, with an accuracy of 0.49, precision of 0.49, recall of 0.53, and an F1 score of 0.51.

6.2 Insights and Findings from the Results

Marginal Differences in Performance: Overall, the differences in performance between these three models are relatively small. With a score ranging from 0.49 to 0.52, this indicates that the underlying classification task is challenging and none of the models have found a significantly superior way to separate classes.

Logistic Regression Superiority: Despite being a simpler model than Random Forest or XG-Boost ensembles, Logistic Regression performs best here. This may hint at either the decision boundary between classes being relatively linear or the more complex model's overfitting to noise in the training data.

Trade-offs in Model Complexity: It was interesting to ascertain that Logistic Regression outperformed, even those more complex models, such as Random Forest and XG-Boost. This fact further reinforces the notion that sole reliance on more sophisticated algorithms can sometimes be misleading. Simpler models can sometimes capture the essence of the underlying patterns much better in situations where data is noisy or scant.

Significance Multiple Metrics: While accuracy is frequently adopted as a primary metric, this analysis demonstrated the value of considering multiple metrics. For instance, the fact that all models returned higher recall than precision may indicate something about the behavior of the models that accuracy cannot itself reveal.

7. Discussion

7.1 Implications of the Study:

This comparative analysis of machine learning models for predicting lung cancer has many significant implications, not only for predictive health care in general but for the very particular domain of the diagnostics of lung cancer. All three models Logistic Regression, Random Forest, and XGBoost had relatively average performances with accuracies ranging from 0.49 to 0.52. This opens up an important challenge: for high-accuracy predictions of lung cancer using the current approach and dataset. Nevertheless, deployment of these models should be approached with extreme caution by healthcare providers and policymakers while applying them in real life, since predictive power is marginally better than random chance.

The narrow superior performance of the Logistic Regression algorithm (accuracy 0.52, F1 score 0.54) compared to more complex algorithms such as Random Forest and XG-Boost is noteworthy. This suggests that in the case of lung cancer prediction, the underlying patterns might be captured by simpler, linear models, rather than sophisticated ensemble methods. This might direct healthcare technology developers initially to focus their target on more interpretable, linear models while developing tools for lung cancer prediction, aiming for more transparent and easy-to-understand decision support systems for clinicians in the future.

However, the overall low predictive power across all models would suggest that the present set of features or data points on which the predictions are made is probably indicative enough of the risk for lung cancer. This has important implications for practices around data collection within healthcare settings. This points towards a necessary exhaustive review and, if possible, expansion in what kind of information is gathered on patients that would be considered relevant in assessing the risk of lung cancer. Healthcare providers may have to expand the scope of population data on diverse measures: extensive environmental exposure history, genetic markers, and more detailed smoking history. This is evidenced by the consistently higher recall compared to precision across all models, which suggests a tendency towards over-prediction of the positive cases.

In retrospect, these models, if applied widely in a clinical scope, would just probably flag more patients for further testing in the context of lung cancer screening. While this might minimize the opportunity cost of overlooking true positive cases-cardinal in oncology-it would also lead to increased health care expenses because of multi-test follow-ups, the patient being subjected to unnecessary anxiety, and overburdening of the diagnostic resources such as CT scans and biopsies. This would mean a balancing act in trade-offs for comprehensive screening and resource allocation in healthcare systems.

These findings also serve to underscore the complexity of lung cancer as a disease itself, including the challenge of its early prediction in an almost sensational manner. Furthermore, it overemphasizes that the difficulty of achieving high predictive accuracy underlines the multicausal nature of lung cancer development and the possible need for more sophisticated, probably multimodal approaches to risk assessment. Such studies can therefore foster much-needed interdisciplinarity across healthcare specialties among clinicians, data scientists, and biomedical researchers in developing comprehensive strategies for risk assessment.

7.2 Limitations

Data Quality and Quantity. The first and most plausible limitation would be in the dataset used to train and test these models. Given the overall low performances across different algorithms, this may point to potential issues with the data itself. This is because the volume of the dataset can be relatively small to learn effective patterns with complex models such as Random Forest and XG-Boost. Further, not all raw data passed as input to this system was perfect in terms of accuracy, completeness, or relevance to the prediction of lung cancer. Missing values, inconsistency in data collection, or stale information can be overriding factors in model performance.

Feature Selection and Engineering. The study was further limited by the features chosen and created for the prediction. This low predictive power may indicate poor feature selection in the sense that strong indicators of lung cancer risk were not present. Important predictors might be missing, or the existing features may not be preprocessed or combined in ways that would effectively catch complex relations associated with the development of lung cancer.

Class Imbalance. The nature of the prevalence of lung cancer could introduce a class imbalance problem in the dataset, where the number of positive cases is usually far less compared to the negative cases. For this, models could be biased towards the majority class at the cost of performing poorly on the minority class in medical diagnostics, the class of most interest.

7.3 Future Work

Several avenues of promising future research and potential improvements in models for predicting lung cancer, given these limitations and challenges, include the following:

Enhanced Data Collection and Curation. Improved data collection might be done in the form of creating higher quality and more complete datasets developed to predict lung cancer. Even the data collection for prospective would be necessary for capturing any possible relevant variables like the detailed history of smoking, occupational exposures, family history, and lifestyle factors. The collaboration with a couple of healthcare institutions could assist in gathering larger, more diverse datasets.

Integration of Multi-Modal Data. Future models could be informed by a wide range of data outside traditional clinical variables. It may range from genetic markers, data on proteomics, and detailed features from imaging tests such as CT scans to wearable devices and environmental sensors. Such multimodal data can provide a more general view of lung cancer risk factors.

Advanced Feature Engineering. More sophisticated feature engineering techniques might show more complex relationships of data. This could be achieved by developing composite risk scores, using domain knowledge to develop clinically applicable feature combinations, or performing automatic feature extraction using autoencoders such as regularized autoencoders or variational autoencoders.

Integrating Explainable AI techniques. Explainable AI methods involve the development of more interpretable models that will increase the clinical utility and trustworthiness of predictive models. It would be great to try the use of either SHAP values or LIME in this context to provide clear insights into driving factors.

8. Conclusion

The study's main objective was to develop and evaluate machine learning models, using integrated demographic, environmental, and lifestyle variables for predicting lung cancer risk. The source of dataset for lung cancer risk prediction was retrieved from multiple sources, particularly, Cleveland hospital records as well as public health databases in the U.S; Besides, we also used large-scale epidemiology studies such as the National Lung Screening Trial (NLST) or the Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial. These sources provided invaluable datasets to which machine learning models were developed, as they contained very valuable information on demographic data, past medical history, lifestyle habits, and clinical symptoms. In this study, the experiment used 3 machine learning algorithms: Logistic Regression, XG-Boost, and Random Forest. Accuracy, precision, recall, as well as F1 score, are used as performance metrics. Overall, the performance of the Logistic Regression model surpasses the Random Forest and XG-Boost models. It had the highest scores in all the metrics, particularly, accuracy, precision, recall, and F1 score. This is indicative that the model Logistic Regression was slightly better at balancing the true positives and false positives and false negatives. The Random Forest model exemplified an intermediate performance, positioning itself second to the Logistic Regression. A significant volume of empirical studies has established that the different machine learning techniques, such as Logistic Regression and Random Forest considerably improve the detection of lung cancer. Although logistic regression, due to its simplicity and interpretability, remains very useful, Random Forest and XG-Boost are much more capable of modeling difficult nonlinear interactions in high-dimensional data. Advanced models like these will provide far more accurate, personalized risk estimates and have the potential to be a powerful contribution to early detection and better clinical decisions regarding lung cancer.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

Publisher’s Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

[1] Bhowmik, P. K., Miah, M. N. 1., Uddin, M. K., Sizan, M. M. H., Pant, L., Islam, M.R., & Gurung, N. (2024). Advancing Heart Disease Prediction through Machine Learning: Techniques and Insights for Improved Cardiovascular Health. *British Journal of Nursing Studies,* 4(2), 35-50.

[2] Bhuiyan, M. S., Chowdhury, I. K., Haider, M., Jisan, A. H., Jewel, R. M., Shahid, R., ... & Siddiqua, C. U. (2024). Advancements in early detection of lung cancer in public health: a comprehensive study utilizing machine learning algorithms and predictive models. *Journal of Computer Science and Technology Studies*, *6*(1), 113-121.

[3] Dritsas, E., & Trigka, M. (2022). Lung cancer risk prediction with machine learning models. *Big Data and Cognitive Computing*, *6*(4), 139.

[4] Dutta, S., Sikder, R., Islam, M. R., A1 Mukaddim, A., Hider, M. A., & Nasiruddin, M. (2024). Comparing the Effectiveness of Machine Learning Algorithms in Early Chronic Kidney Disease Detection. *Journal of Computer Science and Technology Studies, 6*(4), 77-91.

[5] Gupta, G., Kumar, V., & Karuppanan, P. (2024, June). Study & Analysis of Lung Cancer Risk Prediction Techniques Using ML and DL Algorithms. In *2024 IEEE Students Conference on Engineering and Systems (SCES)* (pp. 1-6). IEEE.

[6] Islam, M. Z., Nasiruddin, M., Dutta, S., Sikder, R., Huda, C. B., & Islam, M. R. (2024). A Comparative Assessment of Machine Learning Algorithms for Detecting and Diagnosing Breast Cancer. *Journal of Computer Science and Technology Studies,* 6(2), 121-135.

[7] Mohan, K., & Thayyil, B. (2023). Machine Learning Techniques for Lung Cancer Risk Prediction using Text Dataset. *International Journal of Data Informatics and Intelligent Computing*, *2*(3), 47-56.

[8] Nasiruddin, M., Dutta, S., Sikder, R., Islam, M. R., Mukaddim, A. A., & Hider, M. A. (2024). Predicting Heart Failure Survival with Machine Learning: Assessing My Risk. *Journal of Computer Science and Technology Studies*, *6*(3), 42-55.

[9] Pathan, R. K., Shorna, I. J., Hossain, M. S., Khandaker, M. U., Almohammed, H. I., & Hamd, Z. Y. (2024). The efficacy of machine learning models in lung cancer risk prediction with explainability. *Plos one*, *19*(6), e0305035.

[10] Pro-AI-Rokibul. (2024). *Lung-Cancer-Risk-Prediction-with-Advanced-Machine-Learning-Algorithms-and-Techniques/README.md at main · proAIrokibul/Lung-Cancer-Risk-Prediction-with-Advanced-Machine-Learning-Algorithms-and-Techniques*. GitHub. https://github.com/proAIrokibul/Lung-Cancer-Risk-Prediction-with-Advanced-Machine-Learning-Algorithms-and-Techniques/blob/main/README.md

[11] Radhika, P. R., Rakhi AS N, and Veena G. (2019) A comparative study of lung cancer detection using machine learning algorithms. In *2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, 1-4. IEEE, 2019.

[12] Raoof, S. S., Jabbar, M. A., & Fathima, S. A. (2020, March). Lung Cancer prediction using machine learning: A comprehensive approach. In *2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)* (pp. 108-115). IEEE.

[13] Thallam, C., Peruboyina, A., Raju, S. S. T., & Sampath, N. (2020, November). Early-stage lung cancer prediction using various machine learning techniques. In *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)* (pp. 1285-1292). IEEE.

[14] Vasudha Rani, V., Das, S., & Kundu, T. K. (2022). Risk prediction model for lung cancer disease using machine learning techniques. In *Innovations in Computer Science and Engineering: Proceedings of the Ninth ICICSE, 2021* (pp. 417-425). Singapore: Springer Singapore.