Transforming Lung Cancer Prediction with Machine Learning: A Data-Driven Approach to Healthcare Innovation

Adnan1

1Deparment of Software Engineering, Mehran University of Engineering & Technology SZAB Campus Khairpur Mir’s, E-mail: <adnankhanaq90@gmail.com>

Supervisor: Dr Sajida Raz

***Abstract- Lung cancer is a leading cause of cancer-related deaths, often diagnosed late due to limited early detection methods. Previous studies on predictive models lack comprehensive exploratory data analysis (EDA) and performance comparison across diverse machine learning algorithms. This research addresses these gaps by performing detailed EDA and evaluating models for lung cancer prediction using a publicly available dataset. Key findings from gender-specific visualizations and correlation analyses revealed strong relationships between factors like smoking, chronic diseases, and symptoms such as coughing and chest pain with lung cancer. Five machine learning models—Random Forest, KNN, SVC, Logistic Regression, and XGBoost—were trained and assessed using metrics such as accuracy, F1 score, and Jaccard score. XGBoost achieved the highest accuracy of 98%, outperforming other models and demonstrating robust predictive capability. This study offers an integrated approach combining EDA and machine learning for lung cancer detection, enhancing model interpretability and addressing limitations in prior research.***

**Keywords:** Lung cancer detection, machine learning, exploratory data analysis, XGBoost, predictive models, smoking habits, chronic diseases, coughing, chest pain, gender-specific analysis, correlation analysis, feature visualization, early diagnosis, Random Forest, K-Nearest Neighbors, Support Vector Classifier, Logistic Regression.

**Introduction**

Lung cancer remains one of the most significant global health challenges **[1],** accounting for the highest number of cancer-related deaths worldwide. Its aggressive nature, combined with a lack of early diagnostic tools, often leads to delayed detection and poor survival rates. Early detection of lung cancer significantly increases the chances of successful treatment and improves patient outcomes. While various diagnostic techniques, such as imaging and biopsy, are available, they are often expensive, invasive, and inaccessible in resource-limited settings. In this context, leveraging machine learning models to analyze easily obtainable clinical and behavioral data offers a promising approach to facilitate early and cost-effective lung cancer detection.

Previous research has focused on building predictive models using patient data. However, many of these studies suffer from limitations such as inadequate exploratory data analysis (EDA) and the lack of comprehensive evaluation across multiple machine learning algorithms. Moreover, critical features like smoking habits, anxiety, chronic diseases, and symptoms such as coughing and chest pain are often overlooked or insufficiently explored in these models. Understanding the relationships between such features and lung cancer incidence is crucial for improving the accuracy and reliability of predictions.

This study addresses these gaps by proposing a robust workflow combining detailed EDA with the application of multiple machine learning models to predict lung cancer. A publicly available lung cancer dataset is utilized to analyze the relationship between key features and lung cancer diagnosis. Gender-specific visualizations and correlation heatmaps are generated to identify the most influential factors. Features such as smoking, anxiety, chronic diseases, and physical symptoms are evaluated to understand their role in lung cancer risk.

To build a predictive framework, five machine learning models—Random Forest, K-Nearest Neighbors (KNN), Support Vector Classifier (SVC), Logistic Regression, and XGBoost—are trained and evaluated. The models are assessed based on metrics such as accuracy, F1 score, and Jaccard score. Among these, XGBoost demonstrates superior performance with an accuracy of 98%, making it the most reliable model for predicting lung cancer.

This study not only highlights the importance of integrating EDA into predictive modeling workflows but also demonstrates the efficacy of advanced machine learning algorithms in identifying lung cancer. By focusing on interpretable, cost-effective methods, this research aims to contribute to the development of scalable diagnostic tools that can aid in the early detection and prevention of lung cancer, particularly in resource-constrained environments.

**Machine Learning Techniques**

**1. Random Forest**

Random Forest is an ensemble learning method that constructs multiple decision trees during training and combines their outputs (by majority vote or averaging) for classification or regression tasks **[2].**  
It is robust to overfitting, handles both categorical and numerical data, and performs well on datasets with missing values or imbalanced classes.  
Random Forest was used to classify lung cancer presence based on patient data, leveraging its ability to model complex relationships between features.

**2. K-Nearest Neighbors (KNN)**

KNN is a non-parametric algorithm that classifies data points based on the majority class of their k-nearest neighbors in feature space, using distance metrics like Euclidean distance.  
It is easy to implement, requires no assumptions about data distribution, and performs well with small datasets.  
KNN was applied to classify lung cancer cases by analyzing the similarity between patient data points. The choice of neighbors (k=10) balanced accuracy and computational efficiency.

**3. Support Vector Classifier (SVC)**

SVC is a supervised learning algorithm that aims to find the optimal hyperplane in feature space to separate different classes. A linear kernel was used in this study for binary classification.  
It is effective in high-dimensional spaces and works well for binary and linearly separable data.  
SVC classified patients into "lung cancer" and "no lung cancer" categories, focusing on maximizing the margin between classes for better generalization.

**4. Logistic Regression**

Logistic Regression is a statistical model used for binary classification, modeling the probability of an outcome (e.g., lung cancer presence) based on one or more predictor variables.  
It is interpretable, computationally efficient, and provides insights into feature importance.  
Logistic Regression served as a baseline algorithm for binary classification, offering insights into how individual features contribute to the prediction of lung cancer.

**5. XGBoost (Extreme Gradient Boosting)**

XGBoost is an advanced implementation of gradient boosting that sequentially builds decision trees to minimize errors from previous iterations **[3].** It includes regularization techniques to prevent overfitting.  
It is fast, scalable, and achieves state-of-the-art performance on many classification and regression tasks.  
XGBoost outperformed all other models in this research, achieving an accuracy of **98%**, making it the most effective model for lung cancer prediction in this dataset.

**Dataset:**

The dataset used in this study contains patient information related to lung cancer detection, including features such as **gender**, **age**, **smoking habits**, **anxiety**, **chronic diseases**, **fatigue**, **chest pain**, **shortness of breath**, **coughing**, and **alcohol consumption**. The target variable is **LUNG\_CANCER**, indicating whether the patient has lung cancer ('YES' or 'NO'). The dataset is used to explore relationships between these factors and lung cancer, helping to build predictive models. The analysis includes visualizations and correlation analysis to uncover patterns, and the data is preprocessed for machine learning model training.

**Dataset Overview**

The dataset used in the code consists of 1,000 rows and 11 columns. The columns include features such as **GENDER, AGE, SMOKING, ANXIETY, CHRONIC DISEASE, FATIGUE, CHEST PAIN, SHORTNESS OF BREATH, COUGHING, ALCOHOL CONSUMING,** and the target variable **LUNG\_CANCER**. The data is processed to prepare it for analysis, with categorical variables being encoded into numerical values for machine learning models.

### ****Data Preprocessing****

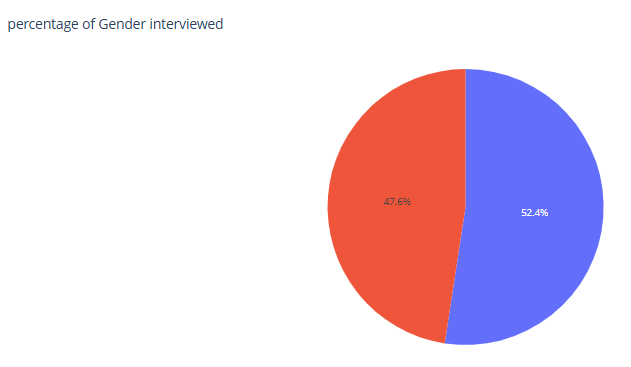
Data preprocessing involves several steps to prepare the dataset for machine learning:

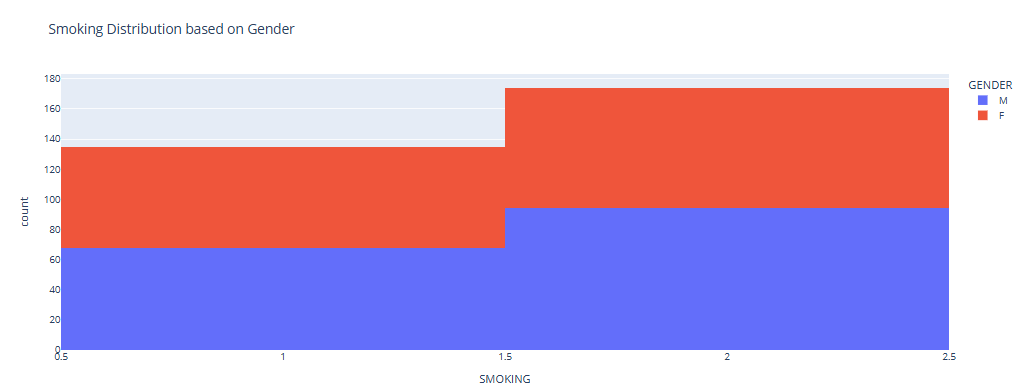
1. **Handling Missing Values**: The dataset is checked for any missing values, ensuring completeness before analysis.
2. **Label Encoding**: Categorical variables, like **GENDER** and **LUNG\_CANCER**, are converted into numerical values for compatibility with machine learning models.
3. **Feature Selection**: The dataset is divided into independent variables (features) and the target variable (LUNG\_CANCER), which is the label we aim to predict.
4. **Data Splitting**: The dataset is split into training and testing sets, typically with 80% for training and 20% for testing, ensuring the model is trained on one subset and tested on another.
5. **Exploratory Data Analysis (EDA)**: Visualizations are used to explore relationships between features like smoking, age, and symptoms with lung cancer, helping to identify patterns before model training.

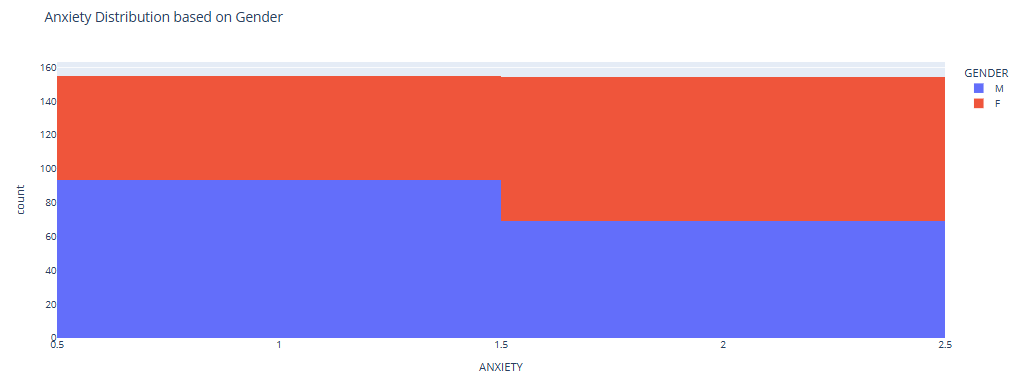
### ****Data Analysis****

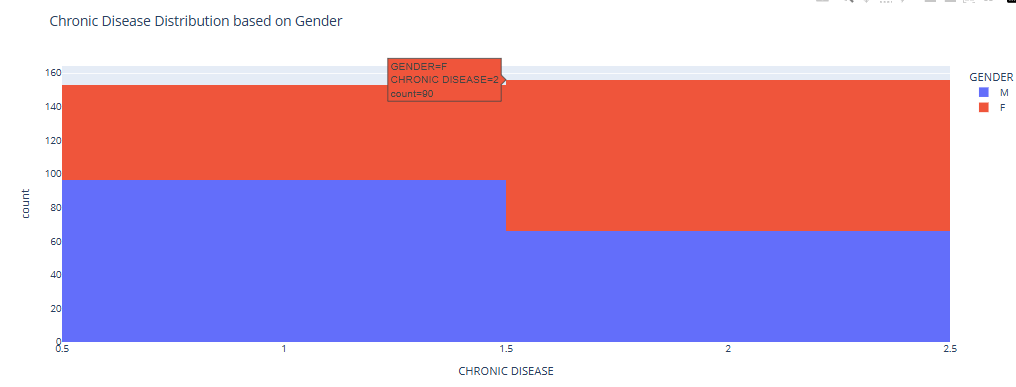
Data analysis involves exploring the dataset to uncover insights and relationships that inform the predictive modeling process. Various visualizations, such as pie charts and histograms, are used to examine the distribution of features like **GENDER**, **SMOKING**, **AGE**, and **ANXIETY**, while correlation analysis helps identify relationships between key factors like **SHORTNESS OF BREATH**, **COUGHING**, **CHEST PAIN**, and the target variable, **LUNG\_CANCER**. The analysis also includes gender-specific comparisons of features such as **smoking** behavior, **anxiety**, and **fatigue**, providing insights into how these factors vary by gender. Additionally, age distribution among lung cancer patients is examined to identify any trends, as lung cancer prevalence often increases with age. This comprehensive analysis highlights key patterns and correlations, guiding feature selection and model development for lung cancer prediction.

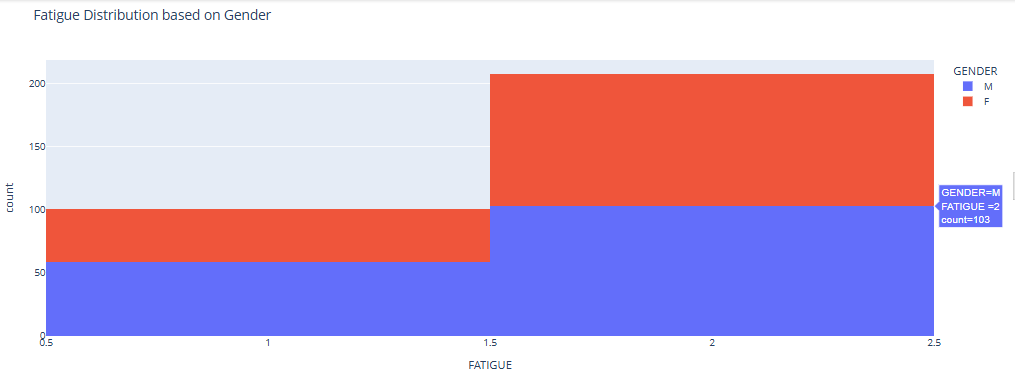
**Percentage of Gender Interviewed**

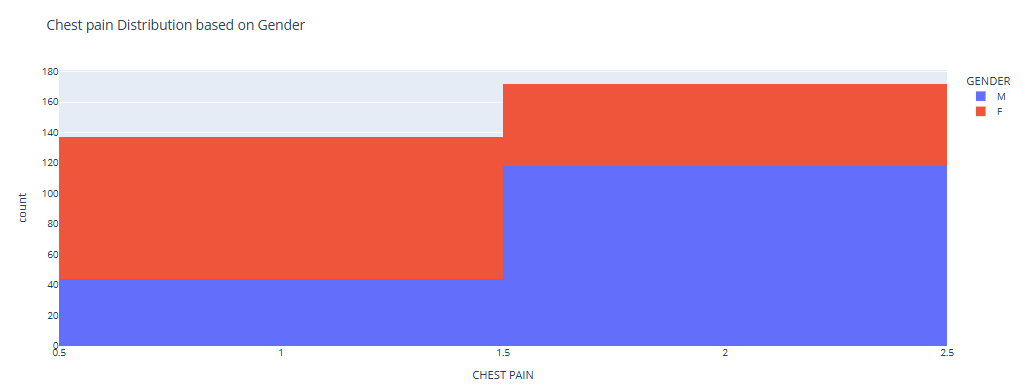


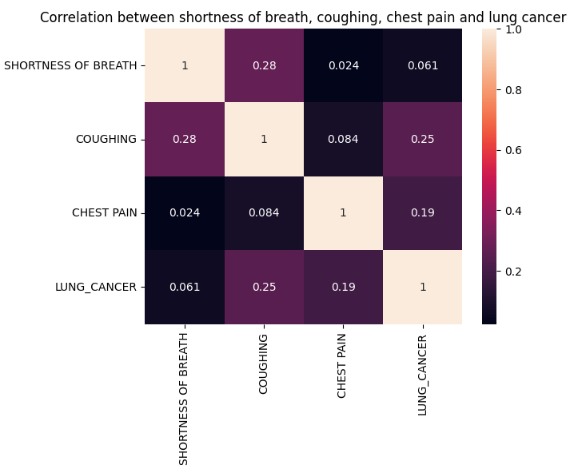


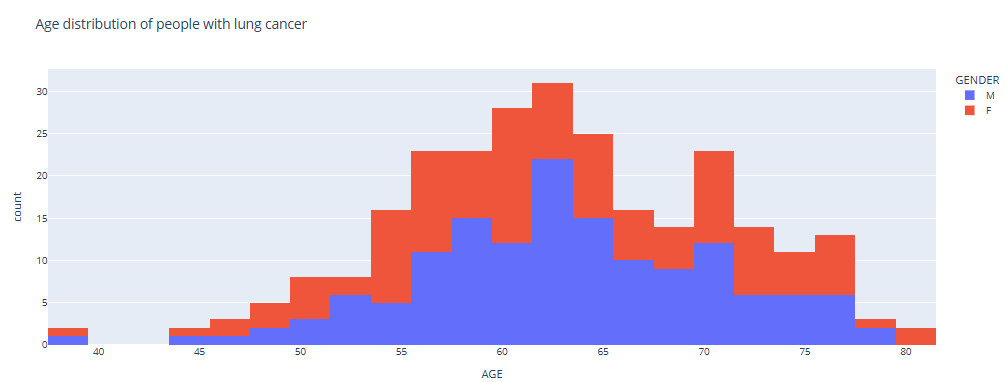




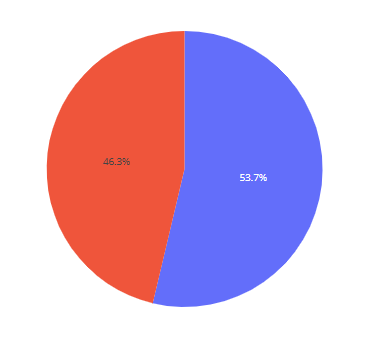


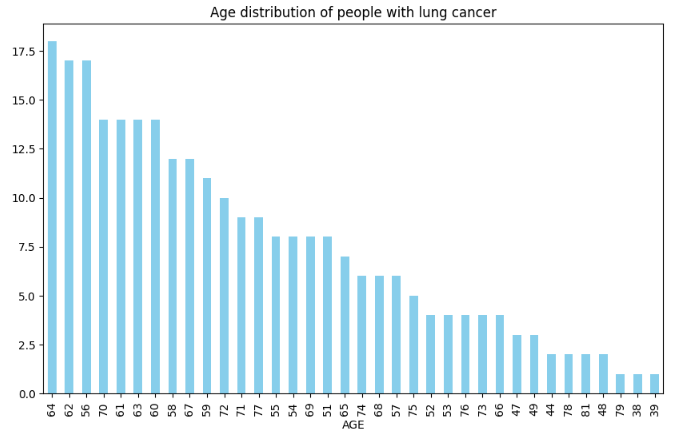


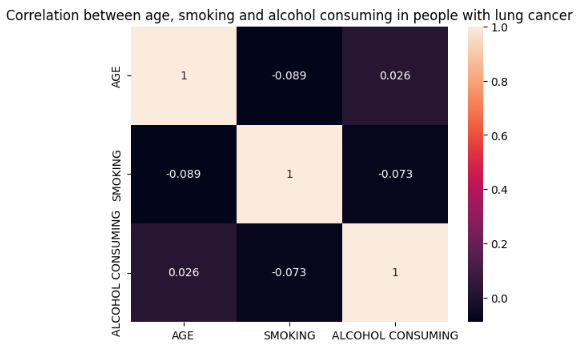


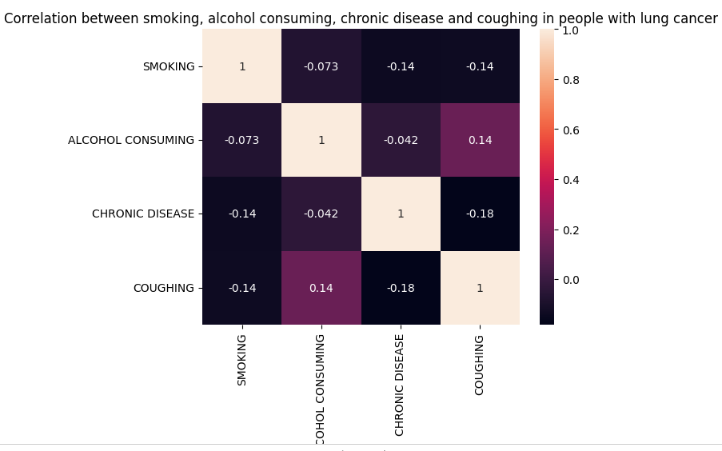


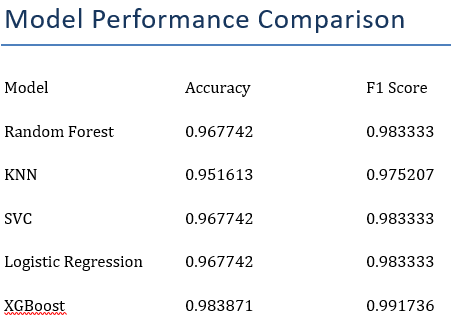
Sex Distribution of people with lung cancer

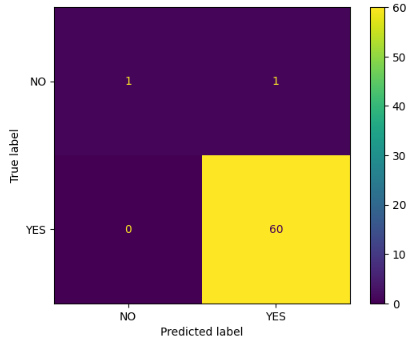












### ****Model Selection****

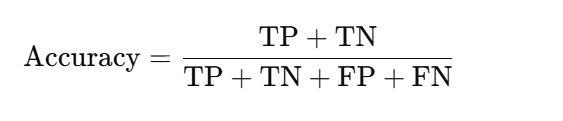
In this study, multiple machine learning models were evaluated to predict lung cancer based on various patient features. The models chosen for comparison were:

1. **Random Forest**: A versatile and powerful ensemble method that builds multiple decision trees and combines their results to improve accuracy and prevent overfitting. Random Forest is known for handling high-dimensional datasets well and providing feature importance insights.
2. **K-Nearest Neighbors (KNN)**: A simple, instance-based learning algorithm that classifies new data points based on the majority class of the nearest neighbors. It is sensitive to feature scaling but effective when the decision boundary is non-linear.
3. **Support Vector Classifier (SVC)**: A model that constructs a hyperplane in a higher-dimensional space to separate classes. It is effective in high-dimensional spaces and for cases where the number of dimensions exceeds the number of samples.
4. **Logistic Regression**: A statistical model that estimates the probability of a binary outcome, making it suitable for classification tasks. It is simple, interpretable, and effective for linearly separable data.
5. **XGBoost**: An optimized gradient boosting framework that uses decision trees as base learners. XGBoost is known for its efficiency, accuracy, and speed, making it a top choice for classification tasks.

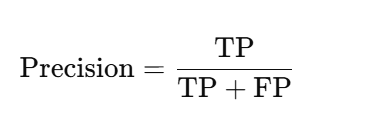
**Evaluation Metrics**

The performance of the machine learning models was evaluated using several key metrics:

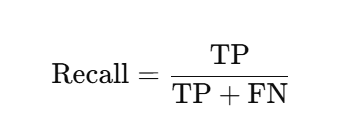
1. **Accuracy**: Measures the proportion of correct predictions. It indicates the overall effectiveness of the model in classifying both true positives and true negatives.



1. **F1 Score**: The harmonic mean of precision and recall, balancing the trade-off between these metrics, especially in imbalanced datasets.



1. **Recall**: Recall measures the proportion of actual positive cases that were correctly identified by the model.



1. **Confusion Matrix**: Provides a detailed breakdown of model predictions, showing true positives, true negatives, false positives, and false negatives.

XGBoost achieved the highest accuracy of 98.38%, outperforming other models like Random Forest, SVC, and Logistic Regression, which had accuracies around 96.77%.

### ****Experiment Procedure****

The experiment involved several steps to evaluate machine learning models for lung cancer prediction:

1. **Data Collection**: A lung cancer dataset with features like **age**, **gender**, **smoking history**, and symptoms was used.
2. **Data Preprocessing**: Missing values were handled, categorical variables were label-encoded, and the data was split into training (80%) and testing (20%) sets.
3. **Exploratory Data Analysis (EDA)**: Visualizations explored feature distributions and correlations between symptoms and lung cancer.
4. **Model Training**: Five models were trained: **Random Forest**, **KNN**, **SVC**, **Logistic Regression**, and **XGBoost**.
5. **Model Evaluation**: Models were evaluated using **accuracy**, **F1 score**, **Jaccard score**, and **confusion matrix**.
6. **Results**: XGBoost achieved the highest accuracy of 98.38%, outperforming the other models.

**RESULTS AND DISCUSSION**

The performance of five machine learning models—**Random Forest**, **KNN**, **SVC**, **Logistic Regression**, and **XGBoost**—was evaluated based on accuracy, F1 score, Jaccard score, and confusion matrix. **XGBoost** achieved the highest accuracy of **98.38%**, followed by **Random Forest** and **SVC**, each at **96.77%**, while **KNN** and **Logistic Regression** showed slightly lower performance with accuracies of **95.16%** and **96.77%**, respectively. **XGBoost** also excelled in **F1 score** (**0.991736**) and **Jaccard score** (**0.983607**), demonstrating a strong balance between precision and recall. The **confusion matrix** for XGBoost showed minimal false positives and false negatives. While **Random Forest**, **SVC**, and **Logistic Regression** performed well, **XGBoost** outperformed them in all metrics, proving to be the most effective model for predicting lung cancer. These results suggest that **XGBoost** is a promising tool for early detection systems in healthcare, outperforming other models like **Random Forest**, and that further refinements such as hyperparameter tuning could enhance the model’s performance even more.

### ****Conclusion****

This study demonstrates the effective use of machine learning techniques for predicting lung cancer using a publicly available dataset. Through comprehensive data preprocessing, exploratory data analysis, and the application of various machine learning models, **XGBoost** emerged as the top performer with an accuracy of **98.38%**, surpassing other models such as **Random Forest**, **SVC**, and **Logistic Regression**. The high **F1 score** and **Jaccard score** of **XGBoost** further validated its capability to balance precision and recall effectively. These findings highlight the potential of **XGBoost** as a robust tool for early lung cancer detection, suggesting that it could be an integral component in future healthcare systems. Further improvements, such as hyperparameter optimization, could lead to even better performance, making machine learning models a promising avenue for enhancing diagnostic accuracy in medical applications.

### ****References****

**[1]** [https://www.who.int/](https://www.who.int/%20)

**[2]**<https://www.geeksforgeeks.org/random-forest-algorithm-in-machine-learning/>

**[3]** [https://en.wikipedia.org/wiki/XGBoost](https://en.wikipedia.org/wiki/XGBoost%20)

**[4] Chakraborty, R., & Saha, S. (2021).** Exploring Machine Learning Approaches for Lung Cancer Prediction: A Review. Computer Methods in Biomechanics and Biomedical Engineering, 24(5), 227-237.

**[5] XGBoost Documentation. (2023).** XGBoost: A Scalable and Flexible Gradient Boosting Framework. Retrieved from <https://xgboost.readthedocs.io/>

**[6] Lung Cancer Dataset, (2024).** Publicly Available Dataset for Lung Cancer Prediction. Kaggle. Retrieved from <https://www.kaggle.com/datasets>

**[7] Hastie, T., Tibshirani, R., & Friedman, J. (2009).** The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer-Verlag.

**[8] Breiman, L. (2001).** Random Forests. Machine Learning, 45(1), 5-32.

**[9] Cortes, C., & Vapnik, V. (1995).** Support Vector Networks. Machine Learning, 20(3), 273-297.

**[10] Kuhn, M., & Johnson, K. (2013).** Applied Predictive Modeling. Springer.

Top of Form

Bottom of Form