Early Lung Cancer Prediction Using Machine Learning and Explainable AI

Dasari Surya Vamsi

SE25MBDS011

https://github.com/CarrierHULK7/Early-Lung-Cancer-Prediction-Using-Machine-Learning-and-Explainable-AI-.git

Abstract

Lung cancer is one of the most fatal cancers worldwide, primarily due to late diagnosis. Early detection plays a crucial role in improving survival rates and treatment outcomes. This project presents an end-to-end machine learning pipeline for early lung cancer prediction

using survey-based patient data. Multiple machine learning models were trained and evaluated using appropriate performance metrics, and the best-performing model was selected. To ensure transparency and trust in predictions, Explainable Artificial Intelligence (XAI) techniques using SHAP were applied to interpret both global and individual predictions. The results demonstrate that machine learning, combined with explainability, can effectively support early lung cancer screening.

---

1. Introduction

Lung cancer accounts for a significant proportion of cancer-related deaths globally. One of the main reasons for its high mortality rate is late-stage diagnosis, where treatment options become limited. Early screening and risk assessment can greatly improve patient outcomes.

With advancements in data-driven approaches, machine learning has emerged as a powerful tool for predictive healthcare analytics. However, medical decision-making requires not only accurate predictions but also interpretable models that clinicians can trust. This project aims to develop an explainable machine learning system for early lung cancer prediction using patient symptoms and lifestyle data.

---

2. Dataset Description

The dataset used in this project is the Survey Lung Cancer dataset obtained from Kaggle. Each record in the dataset represents an individual patient and contains information related to demographic factors, lifestyle habits, and symptoms associated with lung cancer.

Dataset Characteristics

- Patient symptoms such as coughing, wheezing, fatigue, chest pain, and shortness of breath
- Lifestyle factors such as smoking and alcohol consumption
- Demographic attribute: gender
- Target variable indicating the presence or absence of lung cancer

The target variable LUNG_CANCER is a binary label:

- YES indicates the presence of lung cancer
- NO indicates the absence of lung cancer

This dataset is suitable for binary classification tasks in healthcare analytics.

---

3. Data Preprocessing

Raw healthcare data often contains categorical values and inconsistent encodings that are unsuitable for machine learning algorithms. Therefore, several preprocessing steps were applied:

1. Target Encoding

   The target variable LUNG_CANCER was converted into numerical form:

   - YES → 1
   - NO → 0
2. Gender Encoding

   Gender was encoded as:

   - Male → 1
   - Female → 0
3. Binary Feature Conversion

   Many symptom features were originally encoded as 1 and 2. These were converted into binary values:

   - 1 → 0 (absence of symptom)
   - 2 → 1 (presence of symptom)
4. Handling Missing Values

   Any rows containing invalid or missing values were removed to ensure data consistency.

These preprocessing steps ensured that all features were numerical and suitable for machine learning models.

---

4. Feature and Target Separation

After preprocessing, the dataset was divided into:

- Feature matrix (X): all patient symptoms and lifestyle variables
- Target vector (y): lung cancer label

This separation allows the model to learn relationships between patient features and the likelihood of lung cancer.

---

5. Train–Test Split

To evaluate the model's performance on unseen data, the dataset was split into:

- 80% training data
- 20% testing data

Stratified sampling was used to preserve the proportion of lung cancer and non-cancer cases in both sets. This is particularly important in medical datasets to avoid biased evaluation.

---

6. Feature Scaling

Feature scaling was performed using StandardScaler, which standardizes features to have zero mean and unit variance. Although tree-based models such as XGBoost do not require scaling, this step was necessary for fair training of distance-based models like KNN, SVM, and neural networks. Scaling was applied only after the train–test split to prevent data leakage.

---

7. Machine Learning Models Used

To identify the most suitable algorithm, multiple machine learning models were trained and compared:

- Logistic Regression
- K-Nearest Neighbors (KNN)
- Naive Bayes
- Decision Tree
- Random Forest
- Support Vector Machine (SVM)
- Multi-layer Perceptron (Neural Network)
- XGBoost (Extreme Gradient Boosting)

This comparative approach ensures that the final model selection is data-driven rather than assumption-based.

---

8. Model Training and Evaluation

Each model was trained using the training dataset and evaluated on the test dataset using the following metrics:

- Accuracy: Overall correctness of predictions
- Precision: Proportion of predicted cancer cases that are correct
- Recall (Sensitivity): Ability to correctly identify cancer cases
- F1-score: Balance between precision and recall
- ROC-AUC: Ability to distinguish between cancer and non-cancer cases

In medical applications, recall and ROC-AUC are particularly important to minimize false negatives.

---

9. Model Evaluation Results

9.1 ROC Curve Analysis

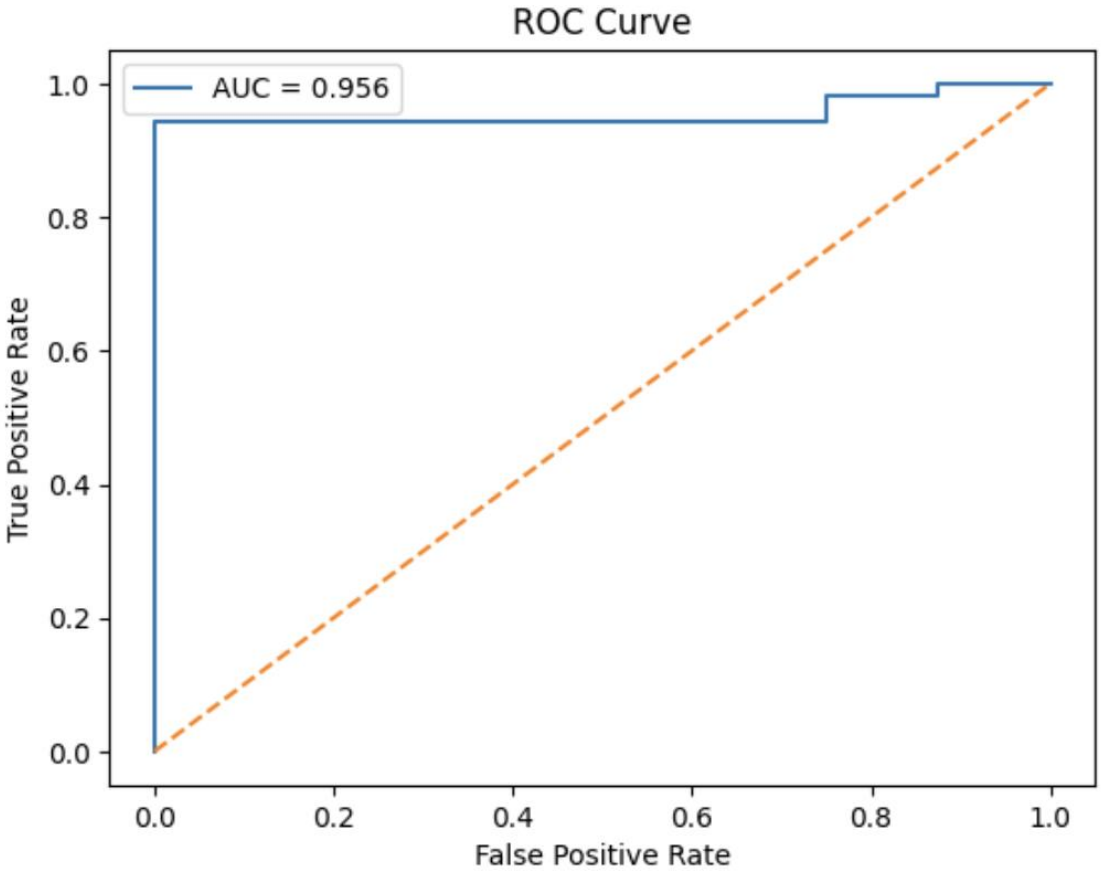

Figure 1 shows the ROC curve for the best-performing model. The curve demonstrates a strong trade-off between sensitivity and specificity, with a high area under the curve (AUC), indicating effective discrimination between lung cancer and non-cancer cases.

9.2 Confusion Matrix Analysis

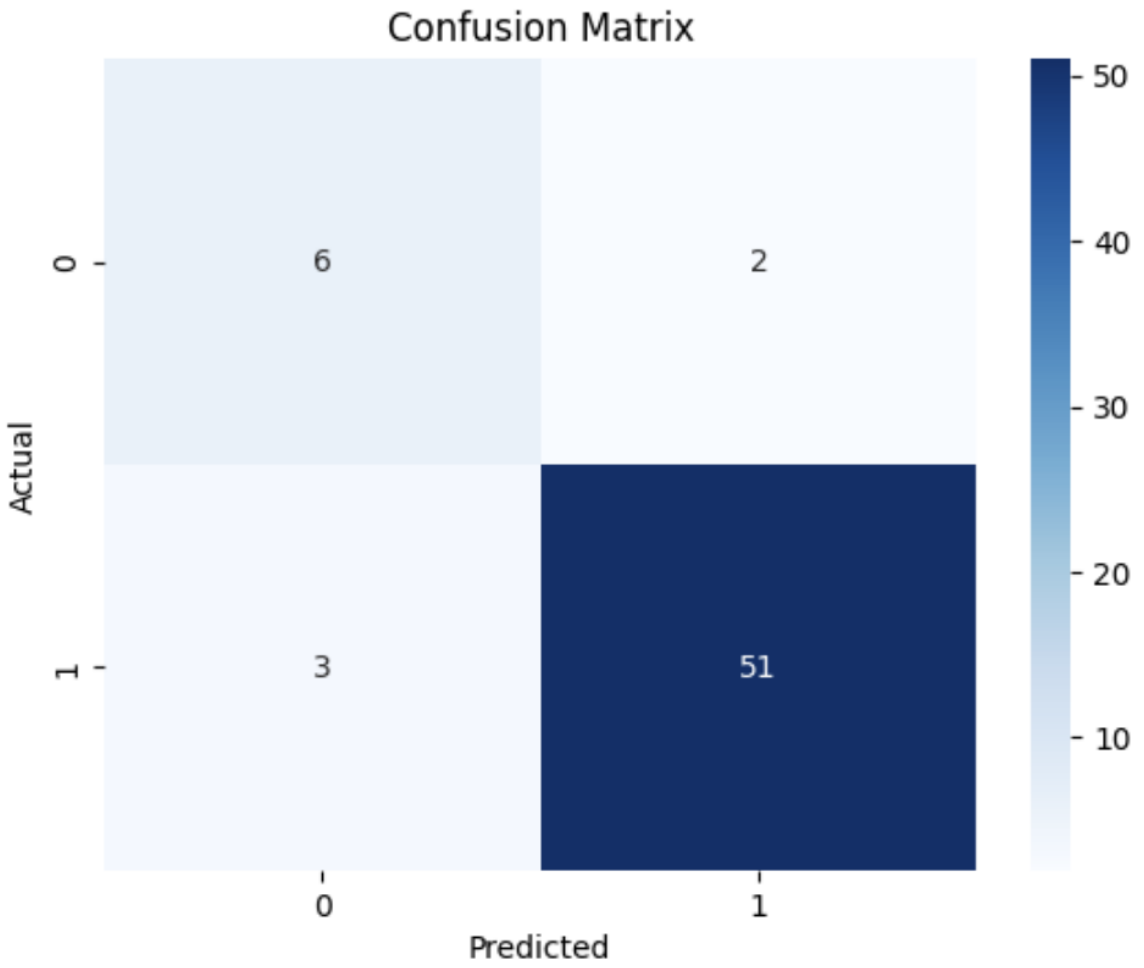

Figure 2 presents the confusion matrix of the final model. It provides insight into true positives, true negatives, false positives, and false negatives. In healthcare applications, minimizing false negatives is critical, as undetected cancer cases can have severe consequences.

---

10. Best Model Selection

Based on the evaluation metrics, XGBoost achieved the highest ROC-AUC score among all tested models. Therefore, XGBoost was selected as the final model for lung cancer prediction. Its ensemble-based approach allows it to capture complex patterns in the data effectively.

---

11. Explainable AI using SHAP

11.1 Overview of SHAP

SHAP (SHapley Additive exPlanations) is an explainable AI technique based on game theory. It assigns each feature a contribution value toward a model's prediction, ensuring fair and consistent explanations.

Explainability is especially important in healthcare, where clinicians need to understand why a model makes a particular prediction.

11.2 SHAP Summary Plot

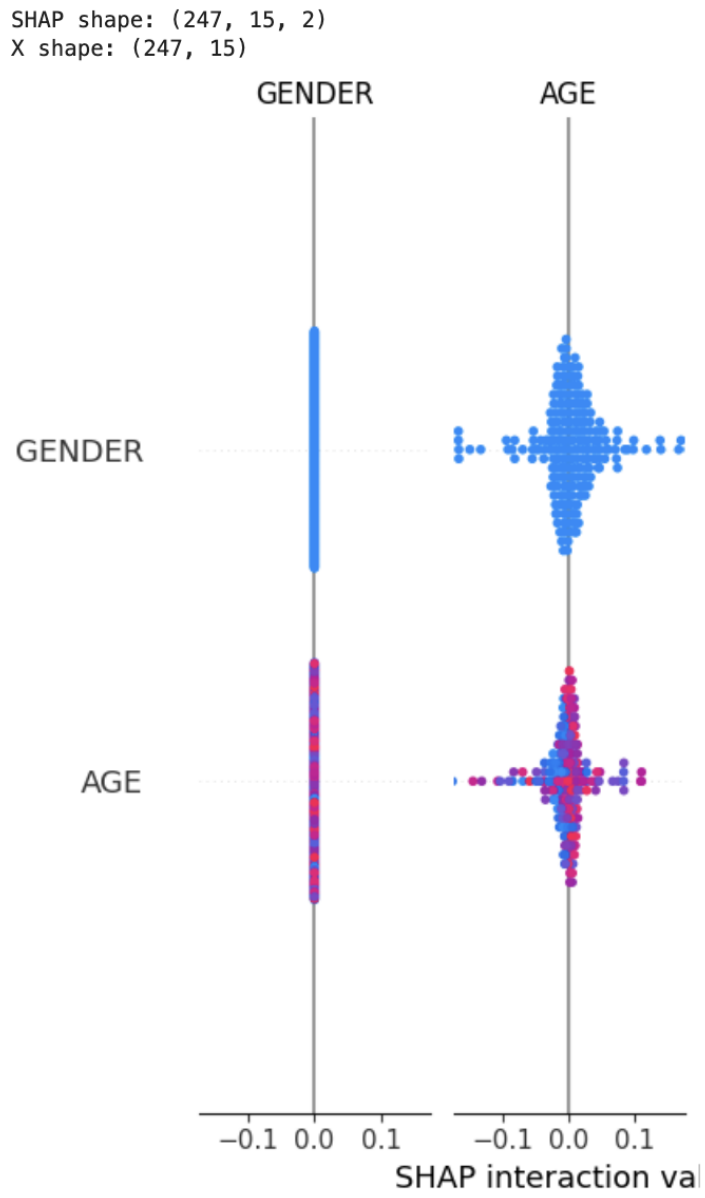

Figure 3 shows the SHAP summary plot, which provides a global explanation of feature importance across all patients. Features related to smoking and respiratory symptoms were found to have the highest influence on lung cancer prediction, aligning with established medical knowledge.

11.3 SHAP Waterfall Plot

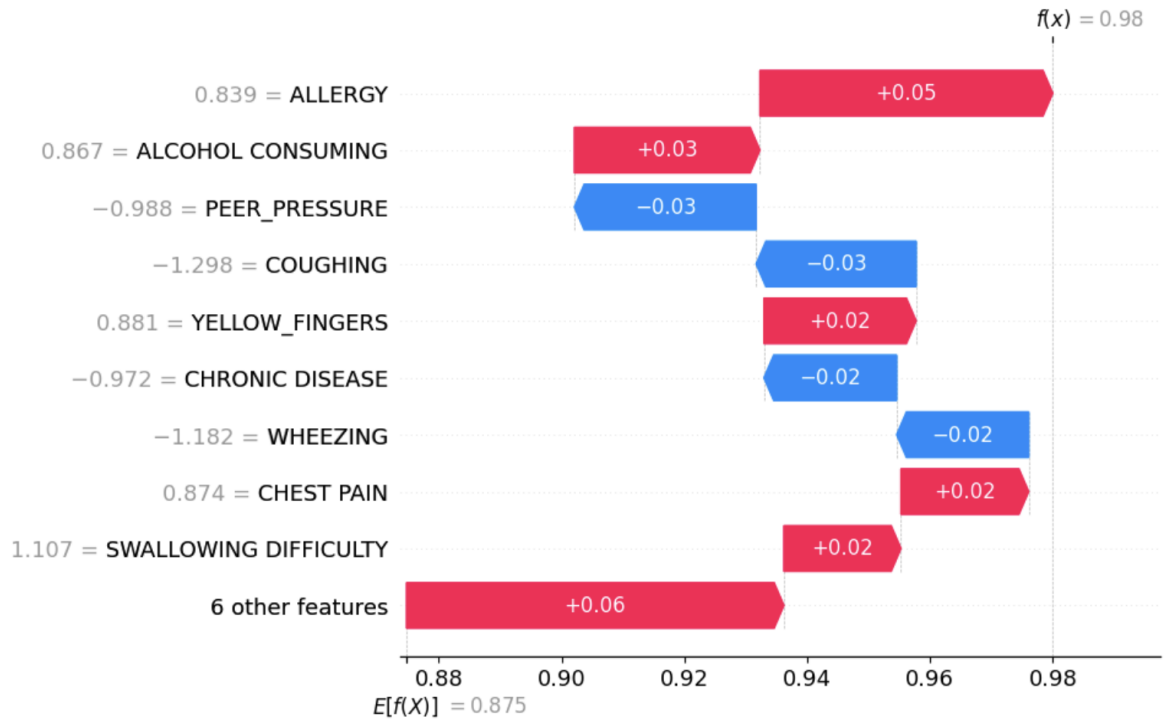

Figure 4 illustrates the SHAP waterfall plot for an individual patient. This plot explains how each feature contributes positively or negatively to the final prediction, offering patient-specific interpretability.

---

12. Model Saving and Reusability

The trained XGBoost model and the feature scaler were saved to disk. This enables the model to be reused for future predictions without retraining and facilitates potential deployment in clinical decision-support systems.

---

13. Conclusion

This project demonstrates a complete machine learning pipeline for early lung cancer prediction using survey-based patient data. Among multiple evaluated models, XGBoost provided the best performance based on ROC-AUC. The integration of SHAP ensured transparency and interpretability, making the model suitable for healthcare applications. While the model is not intended to replace clinical diagnosis, it can assist in early screening and risk assessment.

14. Future Scope

- Use larger and clinically validated datasets
- Apply cross-validation and hyperparameter tuning
- Integrate medical imaging data
- Deploy the model as a web-based clinical decision-support system