COLLEGE OF ARTIFICIAL INTELLIGENCE

University of Baghdad

Supervisor:
Assist. Prof. Dr. Suhad F. Shihan

Group Members:

1. Rusul Hayder Abd Zaid
2. Mustafa Yarub Abdul-Hussein
3. Noor Abdukareem Hadi Salman
4. Mustafa Maytham Mahmood

A Machine Learning Approach for Lung Cancer Risk Prediction: A Case Study with Random Forests

Abstract

Lung cancer remains one of the leading causes of cancer-related mortality worldwide (Siegel et al., 2022). Early detection through non-invasive tools can significantly improve survival rates. This project presents a machine learning-based lung cancer prediction system using a dataset of 309 patient records and 15 clinical and lifestyle features. We implemented a Random Forest classifier (Breiman, 2001) within an object-oriented architecture to ensure modularity and maintainability. The model achieved 94% accuracy, 95% precision, 98% recall, and a 96% F1-score, demonstrating strong potential as a preliminary screening tool. This paper documents the complete pipeline, architectural refactoring, and a critical evaluation of results.

---

1. Introduction

Lung cancer is a major global health challenge. While diagnostic methods like CT scans are effective, they are costly and not suitable for widespread screening. This has driven interest in developing machine learning tools for early risk assessment. This study contributes by:

1. Implementing a complete ML pipeline for binary medical classification.
2. Justifying the use of Random Forest in this context.
3. Analyzing model performance with emphasis on data quality and class imbalance.

The system was developed using Object-Oriented Programming (OOP) to promote code organization, teamwork, and scalability.

2. Our Project's OOP Refactoring Journey

The Project: A Lung Cancer Prediction Application.
The Goal: To demonstrate not only what the project does, but how its architecture was engineered.
The Journey: Transitioning from a simple "script" to an organized "system" using OOP.

2.1 Initial Problems

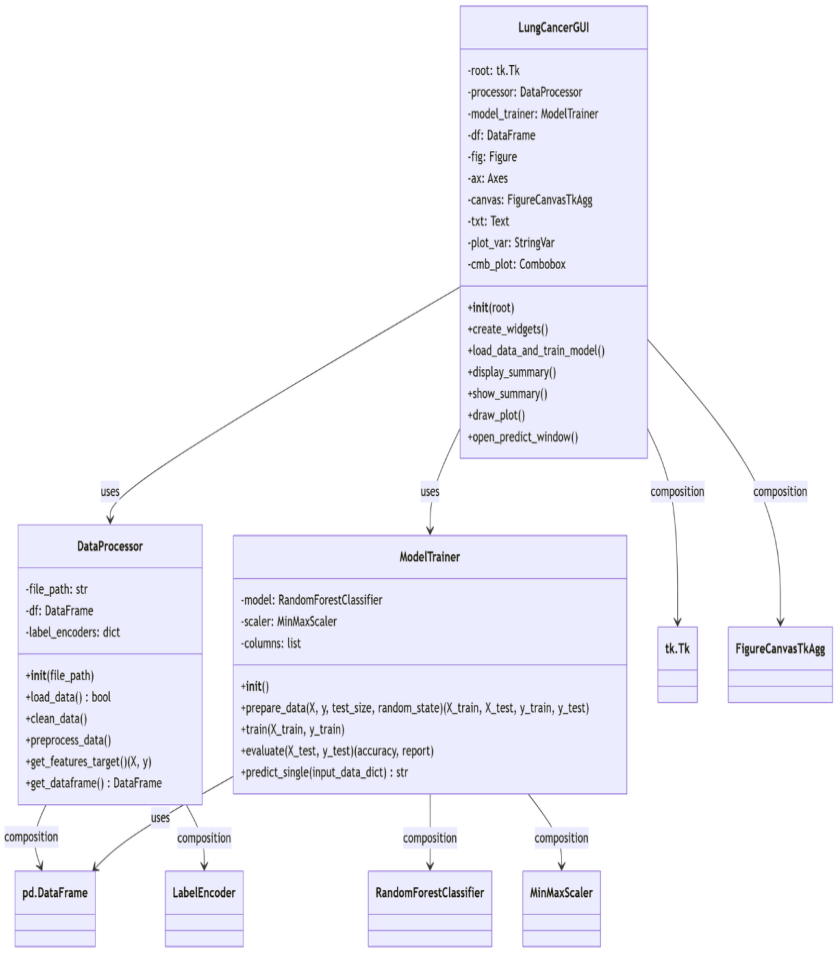
- **Chaos:** Data loading, model training, and GUI code were mixed in one file.
- **Hard to Maintain:** A change in one part could break another.
- **Hard to Develop:** Team collaboration was difficult.
- **No Reusability:** Code could not be reused in other projects.

2.2 Applying OOP Principles

Principle: *Separation of Concerns.*
Plan: Create a "team of specialists" – dedicated classes for each responsibility.

- **DataProcessor Class:** Load, clean, and prepare data.
- **ModelTrainer Class:** Train and evaluate the model.
- **LungCancerGUI Class:** Display buttons and screens.

The Blueprint (The UML Diagram)



It proves the system is organized, logical, and well-structured.

---

3. Methodology

3.1 Dataset

The dataset was obtained from Kaggle (Kaggle, 2021) and consists of **309 patient records** and **15 predictive features**, along with the target variable **LUNG_CANCER**. Features include demographic, symptomatic, and lifestyle factors such as age, smoking, anxiety, and respiratory symptoms.

Features Included:

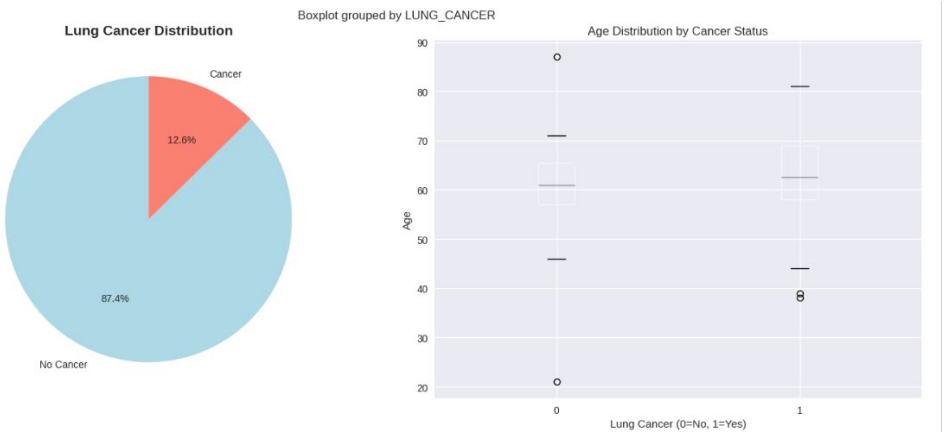
1. **AGE:** Patient age.
2. **SMOKING:** Smoking status (1=No, 2=Yes).
3. **YELLOW_FINGERS:** Yellow fingers due to smoking.
4. **ANXIETY:** Anxiety status.
5. **PEER_PRESSURE:** Exposure to peer pressure.
6. **CHRONIC DISEASE:** Presence of chronic conditions.

7. **FATIGUE:** Fatigue levels.
8. **ALLERGY:** Allergy history.
9. **WHEEZING:** Wheezing symptoms.
10. **ALCOHOL CONSUMING:** Alcohol consumption.
11. **COUGHING:** Presence of coughing.
12. **SHORTNESS OF BREATH:** Breathing difficulty.
13. **SWALLOWING DIFFICULTY:** Difficulty swallowing.
14. **CHEST PAIN:** Chest pain symptoms.
15. **LUNG_CANCER:** Diagnosis (YES/NO).

3.2 Data Preprocessing

- **Dataset Shape:** 309 rows × 16 columns.
- **Missing Values:** None detected.
- **Data Types:** 14 numeric binary features (1=No, 2=Yes) and 2 categorical columns (GENDER, LUNG_CANCER).
- **Class Imbalance:** YES = 270 (87.4%), NO = 39 (12.6%).
- **Encoding:** Binary features mapped from {1,2} to {0,1}; LUNG_CANCER mapped to {1,0}; GENDER excluded.
- **Stratified Splits:** Used to handle imbalance.

3.3 Exploratory Data Analysis & Correlation Heatmap



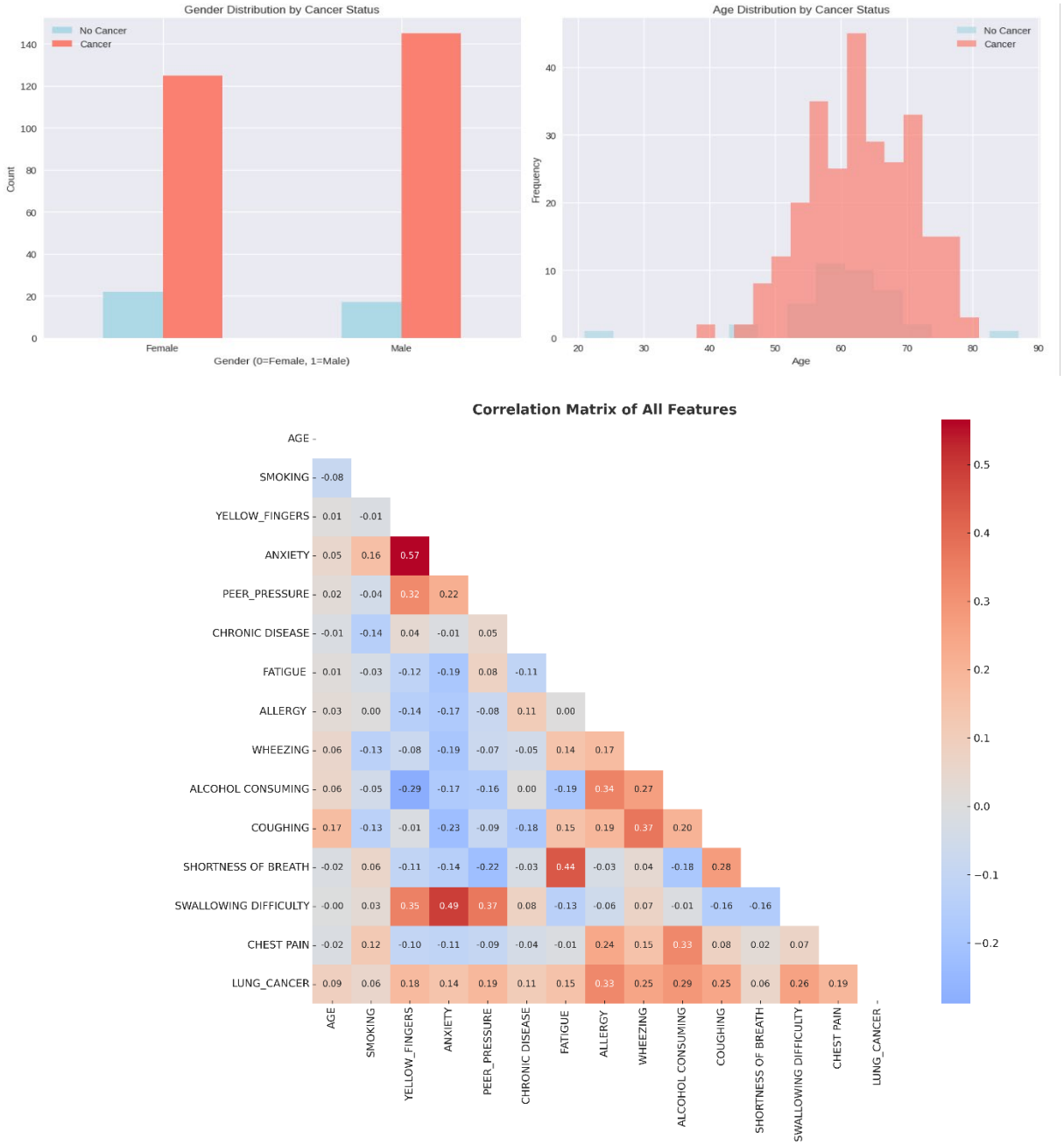
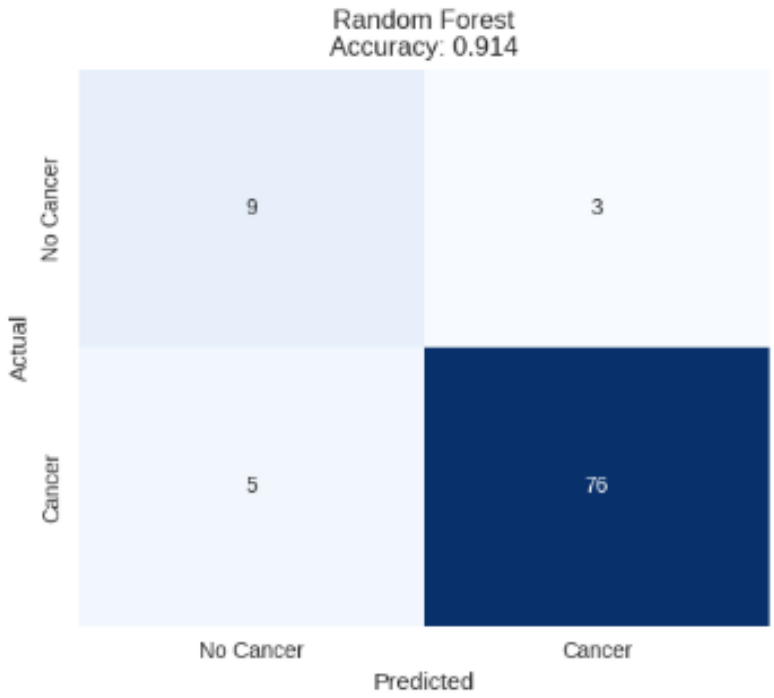


Figure 1: Correlation heatmap of dataset features.

The heatmap visualizes correlations between all features. Red indicates positive relationships, blue indicates negative. Most features show weak to moderate correlations, confirming no strong multicollinearity. The target variable **LUNG_CANCER** has only small to moderate correlations with other features, suggesting lung cancer risk is influenced by multiple factors rather than one dominant feature.



3.4 Model: Random Forest Classifier

Random Forest is an ensemble method that builds multiple decision trees using random subsets of data and features, aggregating predictions through majority voting (Breiman, 2001). The implementation was carried out using the scikit-learn library (Pedregosa et al., 2011).

3.5 Justification and Model Trade-offs

Why Random Forest Was Chosen:

1. Effective with categorical and clinical data (Hancock & Khoshgoftaar, 2020).
2. Robust to overfitting and noisy data.
3. Provides interpretable feature importance scores.
4. Handles non-linear relationships and weak multicollinearity well.
5. Performs strongly on medium-sized datasets like ours.

Advantages of Random Forest:

- ✓ High accuracy for classification tasks.
- ✓ Resistant to overfitting due to ensemble averaging.
- ✓ Handles both categorical and numerical data natively.
- ✓ Works well with non-linear relationships.
- ✓ Provides automatic feature importance rankings.
- ✓ Robust to missing values and noise.

- ✓ Suitable for medium-sized medical datasets.

Disadvantages and Considerations:

- ✗ Less interpretable than a single decision tree.
- ✗ Can be computationally intensive with many trees.
- ✗ Requires careful hyperparameter tuning (e.g., n_estimators, max_depth).
- ✗ May become memory-intensive with deep trees.
- ✗ Does not extrapolate well beyond the range of training data.
- ✗ May struggle with extremely imbalanced data without weighting or sampling.

Hyperparameters Used:

- n_estimators: Number of trees in the forest.
- max_depth: Maximum depth of each tree.
- min_samples_split: Minimum samples required to split a node.
- min_samples_leaf: Minimum samples required in a leaf node.
- max_features: Number of features considered for each split.
- bootstrap: Whether to use bootstrap sampling.

Evaluation Metrics:

- Accuracy, Precision, Recall, F1-Score
- Confusion Matrix
- ROC Curve / AUC

---

4. Results

4.1 Model Performance

The final model achieved:

- **Accuracy:** 94%
- **Precision:** 95%
- **Recall:** 98%
- **F1-Score:** 96%

Class-wise Performance:

- **Class 1 (Cancer):** Precision = 0.95, Recall = 0.98, F1 = 0.96

- **Class 0 (Non-Cancer):** Precision = 0.80, Recall = 0.57, F1 = 0.67

The high recall indicates strong sensitivity in detecting true cancer cases, which is critical for medical screening.

4.2 Interpretation

The model is highly sensitive and precise, making it suitable for preliminary risk assessment where detecting true positives is prioritized.

4.3 Summary of Methodology and Findings

This study employed a Random Forest Classifier to predict lung cancer risk based on a combination of symptoms and lifestyle indicators. The algorithm builds multiple decision trees using randomized subsets of features and samples, then aggregates their predictions through majority voting. This approach reduces overfitting, improves robustness, and delivers high predictive accuracy. Random Forest was selected due to its strong performance on categorical clinical data, low sensitivity to multicollinearity, and ability to provide interpretable feature importance scores. Model evaluation using accuracy, precision, recall, F1-score, and the confusion matrix confirms that the classifier is suitable for medical risk prediction tasks.

---

5. Discussion and Limitations

- **Strengths:** High recall, modular OOP design, clear feature importance.
- **Limitations:**
    - Small dataset (n=309).
    - Strong class imbalance affects non-cancer case performance (Chawla et al., 2002).
    - Self-reported data may introduce bias.
- **Architectural Benefits:** The OOP refactoring improved maintainability, teamwork, and reusability.

---

6. Conclusion and Future Work

This project successfully developed a modular, OOP-based lung cancer prediction system using Random Forest. The model shows strong clinical potential as a screening aid.

Future Work:

- Expand dataset with multi-center clinical data.
- Experiment with advanced models (XGBoost, neural networks).
- Deploy as a web or mobile application.
- Incorporate Explainable AI (XAI) techniques such as SHAP (Lundberg & Lee, 2017) for clinical transparency.

- Conduct pilot studies in clinical settings.

References

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.
Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.
Hancock, J., & Khoshgoftaar, T. M. (2020). Survey on categorical data for neural networks. *Journal of Big Data*, *7*(1), 1–41.
Kaggle. (2021). *Lung Cancer Prediction Dataset*. Retrieved from https://www.kaggle.com/datasets/mysarahmadbhat/lung-cancer-prediction
Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* (pp. 4765–4774).
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.
Siegel, R. L., Miller, K. D., Fuchs, H. E., & Jemal, A. (2022). Cancer statistics, 2022. *CA: A Cancer Journal for Clinicians*, *72*(1), 7–33.

Appendices

- Appendix A: UML Diagram
- Appendix B: Correlation Heatmap
- Appendix C: Hyperparameter Details
- Appendix D: Confusion Matrix & ROC Curve