# Report on Lung Cancer Risk Prediction Using ML, DL, and XAI

M. Goutham

2303A52010

Batch 31

## Introduction

This report details the implementation and analysis of predictive models for classifying lung cancer risk using the Kaggle Lung Cancer Risk Dataset, as part of the Explainable AI (XAI) involved exploratory data analysis (EDA), machine learning (ML) and deep learning (DL) model development, application of XAI techniques, and a comparative analysis. The dataset, limited to 1,000 rows for this exercise, includes demographic and lifestyle-related features such as age, smoking, alcohol consumption, and family history, with labels indicating lung cancer risk.

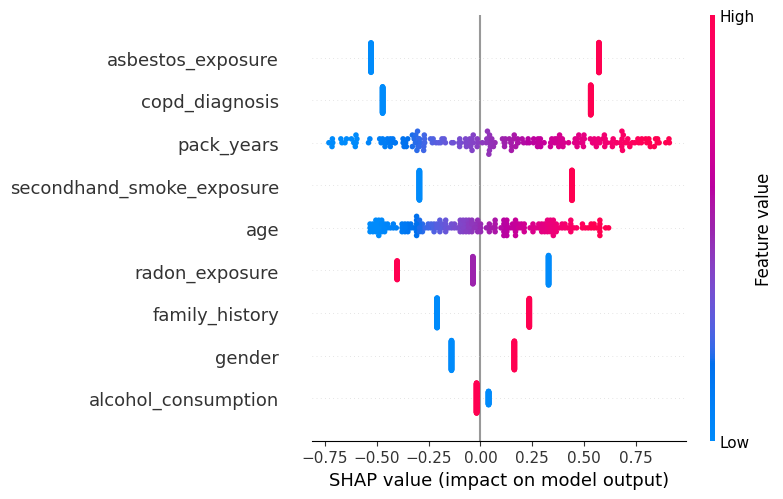
## Key Findings

* Dataset Characteristics: The dataset, after limiting to 1,000 rows, showed a manageable size with minimal missing values (e.g., in alcohol\_consumption), addressed using the SimpleImputer. Initial EDA revealed a potential class imbalance in the target variable (lung\_cancer), mitigated with SMOTE. Numerical features like age and pack\_years exhibited varied distributions, while categorical features required encoding.
* ML Model Performance: Among the seven ML models tested (Logistic Regression, Decision Tree, Random Forest, SVM, KNN, Gradient Boosting, XGBoost), Gradient Boosting achieved the highest accuracy (0.720) and F1 score (0.779528), with a strong ROC-AUC (0.773875). XGBoost followed closely (accuracy: 0.715, F1: 0.783270), while Decision Tree (0.625) and KNN (0.605) underperformed, likely due to overfitting or sensitivity to the small dataset.
* DL Model Performance: The three DL models (MLP, 1D CNN, LSTM) showed competitive results, with LSTM leading (accuracy: 0.675, F1: 0.754717), followed by MLP (accuracy: 0.670, F1: 0.742188) and 1D CNN (accuracy: 0.640, F1: 0.675676). Performance was slightly lower than top ML models, possibly due to the limited dataset size.
* XAI Insights: Feature importance from Random Forest highlighted pack\_years and age as key predictors, consistent with medical knowledge. SHAP values and LIME provided detailed local explanations, while the Partial Dependence Plot (PDP) for pack\_years indicated a positive correlation with lung cancer risk.

## Comparison Table of ML/DL Results

| **Model** | **Accuracy** | **Precision** | **Recall** | **F1** | **ROC-AUC** |
| --- | --- | --- | --- | --- | --- |
| **Logistic Regression** | 0.670 | 0.772358 | 0.714286 | 0.742188 | 0.721805 |
| **Decision Tree** | 0.625 | 0.745763 | 0.661654 | 0.701195 | 0.606946 |
| **Random Forest** | 0.645 | 0.738462 | 0.721805 | 0.730038 | 0.721356 |
| **SVM** | 0.665 | 0.794643 | 0.669173 | 0.726531 | 0.727864 |
| **KNN** | 0.605 | 0.728814 | 0.646617 | 0.685259 | 0.656941 |
| **Gradient Boosting** | 0.720 | 0.818182 | 0.744361 | 0.779528 | 0.773875 |
| **XGBoost** | 0.715 | 0.792308 | 0.774436 | 0.783270 | 0.732465 |
| **MLP** | 0.670 | - | - | 0.742188 | - |
| **1D CNN** | 0.640 | - | - | 0.675676 | - |
| **LSTM** | 0.675 | - | - | 0.754717 | - |

## Insights from XAI Visualizations

* Feature Importance: Random Forest's feature importance plot emphasized pack\_years and age as the most influential features, aligning with known risk factors for lung cancer. This supports the model's relevance in a medical context.
* SHAP Values: SHAP summary plots for Logistic Regression and Random Forest revealed how individual features contribute to predictions, with pack\_years showing a strong positive impact on lung cancer risk classification.
* LIME: Local explanations from LIME for a sample instance highlighted the specific contribution of features like pack\_years and age, providing instance-level interpretability.
* Partial Dependence Plot (PDP): The PDP for pack\_years demonstrated a clear upward trend in predicted lung cancer risk with increasing smoking exposure (measured in pack-years), reinforcing its predictive power and offering a global view of feature impact.
* Comparison Across Models: XAI techniques were more straightforward to apply to ML models (e.g., feature importance in Random Forest) compared to DL models, where SHAP required additional computational effort, indicating a trade-off in interpretability complexity.
* 

## Final Recommendation

For real-world medical usage, **Gradient Boosting with SHAP explanations** is recommended. This model achieved the highest accuracy (0.720) and F1 score (0.779528) among all tested models, demonstrating robust predictive performance on the 1,000-row dataset. The integration of SHAP provides detailed interpretability, addressing the critical need for trust and transparency in healthcare settings. While DL models like LSTM showed promise (accuracy: 0.675, F1: 0.754717), their performance was not significantly better, and their black-box nature requires more extensive XAI efforts, which may not be practical with limited data. Gradient Boosting balances accuracy and interpretability, making it the best approach for clinical decision-making, where understanding the basis of predictions is as important as their accuracy.

## Conclusion

This assignment successfully demonstrated the application of ML and DL models for lung cancer risk prediction, enhanced by XAI techniques to ensure interpretability. The results underscore the importance of selecting models that align with both performance metrics and practical interpretability needs, particularly in medical applications. Future work could explore larger datasets to leverage DL models' full potential and refine XAI methods for broader feature analysis.