# Lung Cancer Risk Prediction using Machine Learning, Deep Learning, and XAI

Prepared by: M. Teja  
Roll No: 2303A52298  
Batch: 32

## Introduction

This report focuses on building predictive models to classify the risk of lung cancer using a dataset obtained from Kaggle. The study involves applying exploratory data analysis (EDA), machine learning (ML) algorithms, deep learning (DL) models, and explainable artificial intelligence (XAI) methods. The dataset used here was restricted to 1,000 samples, containing features like age, smoking history, alcohol usage, and family background. The primary goal is to assess model performance and apply XAI to improve interpretability of the predictions.

## Dataset Summary

The dataset, after filtering to 1,000 records, showed a small percentage of missing data, mostly in the alcohol\_consumption column. Missing values were handled using SimpleImputer. An imbalance in the target variable (lung\_cancer) was identified and balanced using SMOTE. Numerical variables such as age and pack\_years were distributed across ranges, while categorical features were converted into numeric form for model training.

## Performance of Machine Learning Models

Seven ML algorithms were tested: Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Gradient Boosting, and XGBoost. Gradient Boosting achieved the best results with an accuracy of 72% and F1-score of 0.78. XGBoost closely followed with similar performance. Models like KNN and Decision Tree showed weaker results due to overfitting and sensitivity to dataset size.

## Performance of Deep Learning Models

Three DL models were trained: Multi-Layer Perceptron (MLP), 1D Convolutional Neural Network (CNN), and Long Short-Term Memory (LSTM). The LSTM model provided the best results (accuracy 67.5%, F1-score 0.75). MLP achieved similar performance, while CNN performed slightly worse. Due to the small dataset, DL models did not surpass the top-performing ML models.

## XAI Interpretations

To improve interpretability, XAI techniques were applied:  
- Random Forest feature importance highlighted age and pack\_years as the most significant features.  
- SHAP values explained feature contributions, with smoking exposure having the strongest link to higher risk.  
- LIME provided local explanations for individual predictions, showing the influence of key variables.  
- A Partial Dependence Plot (PDP) confirmed the direct relationship between smoking (pack\_years) and risk levels.

## Comparison of ML and DL Results

ML Models:  
- Logistic Regression: Accuracy 67%, F1 0.74  
- Decision Tree: Accuracy 62.5%, F1 0.70  
- Random Forest: Accuracy 64.5%, F1 0.73  
- SVM: Accuracy 66.5%, F1 0.72  
- KNN: Accuracy 60.5%, F1 0.68  
- Gradient Boosting: Accuracy 72%, F1 0.78  
- XGBoost: Accuracy 71.5%, F1 0.78  
  
DL Models:  
- MLP: Accuracy 67%, F1 0.74  
- 1D CNN: Accuracy 64%, F1 0.67  
- LSTM: Accuracy 67.5%, F1 0.75

## Conclusion

This study demonstrated the successful application of ML and DL models for lung cancer risk classification. Among ML models, Gradient Boosting showed the best results, while LSTM led among DL models. However, ML models were easier to interpret using XAI. SHAP and LIME provided reliable feature-level and instance-level explanations. For medical usage, Gradient Boosting combined with SHAP explanations is recommended due to its balance between accuracy and interpretability. Future improvements can be made by using larger datasets to leverage DL models' potential and enhance XAI techniques.