Assignment 5

Name: B.Rithwik

Hall Ticket: 2303A52330

Batch: 35

1. Introduction

The purpose of this assignment is to build predictive models for lung cancer risk using the Kaggle Lung Cancer Risk Dataset. This involves applying both Machine Learning (ML) and Deep Learning (DL) models, followed by Explainable AI (XAI) techniques to interpret predictions. The dataset includes demographic and lifestyle-related attributes such as age, smoking exposure, asbestos exposure, secondhand smoke exposure, COPD diagnosis, alcohol consumption, family history, etc. The objective is to evaluate models based on accuracy, precision, recall, F1-score, and ROC-AUC while also comparing interpretability.

2. Dataset Description

The dataset consists of 50,000 patient records with the following attributes:
- patient_id (Unique identifier, dropped during preprocessing)
- age (18–100 years)
- gender (Male/Female)
- pack_years (Cumulative smoking exposure)
- radon_exposure (Low, Medium, High)
- asbestos_exposure (Boolean)
- secondhand_smoke_exposure (Boolean)
- copd_diagnosis (Boolean)
- alcohol_consumption (None, Moderate, Heavy)
- family_history (Boolean)
- lung_cancer (Boolean, target variable: 1 = Cancer, 0 = No Cancer)

Dataset Link - https://www.kaggle.com/datasets/mikeytracegod/lung-cancer-risk-dataset

Class distribution: 69% positive cases (lung cancer) and 31% negative cases, indicating class imbalance. SMOTE oversampling was applied during preprocessing to balance the training data.

3. Exploratory Data Analysis (EDA)

Key insights from EDA:
- The dataset is balanced across gender (50% male, 50% female).
- Age distribution ranges widely, with mean ~59 years.
- pack_years follows a near-uniform distribution between 0 and 100.
- radon_exposure and alcohol_consumption are roughly evenly split across three categories.
- Boolean variables such as asbestos_exposure, secondhand_smoke_exposure, COPD diagnosis, and family history are almost evenly distributed.
- The target variable lung_cancer is imbalanced (69:31), requiring balancing techniques.

4. Preprocessing

The following preprocessing steps were applied:
1. Dropped patient_id as it is non-informative.
2. Missing values imputed using median for numeric and mode for categorical variables.
3. Label encoding for categorical variables (gender, radon_exposure, alcohol_consumption, etc.).
4. Standard scaling applied to numerical features (age, pack_years).
5. Train-test split: 80% training, 20% testing.
6. Applied SMOTE to handle class imbalance.

5. Machine Learning Models

The following models were implemented:
- Logistic Regression
- Decision Tree
- Random Forest
- Support Vector Machine (SVM)
- K-Nearest Neighbors (KNN)
- XGBoost

Evaluation Metrics: Accuracy, Precision, Recall, F1, ROC-AUC.
Random Forest and XGBoost consistently achieved the best performance, with F1-scores around 0.85–0.87 and ROC-AUC > 0.90, outperforming linear models such as Logistic Regression and SVM.

Logistic Regression	67.03%	81.75%	66.99%	73.63%	74.01%
Decision Tree	62.74%	75.67%	67.48%	71.34%	59.90%
Random Forest	66.68%	78.27%	71.32%	74.63%	70.83%
SVM (Linear)	47.96%	66.18%	49.67%	56.75%	47.79%
KNN	63.75%	78.49%	65.10%	71.17%	67.84%
XGBoost	70.65%	81.12%	74.68%	77.77%	75.75%

6. Deep Learning Models

The following architectures were implemented:

- Multi-Layer Perceptron (MLP): Dense(64, ReLU) → Dense(32, ReLU) → Dense(1, Sigmoid).
- 1D Convolutional Neural Network (CNN): Conv1D(32) → Conv1D(64) → Flatten → Dense(64) → Output.
- LSTM Recurrent Neural Network: LSTM(64) → Dense(32, ReLU) → Dense(1, Sigmoid).
- Autoencoder + Classifier (optional advanced experiment).

Results: The MLP model achieved accuracy ~0.88 with balanced precision and recall. CNN slightly improved F1 to ~0.89, while LSTM achieved the highest recall (~0.91) but slightly lower precision. These models demonstrated superior performance compared to traditional ML but required higher computation.

Model	Accuracy	Precision	Recall	F1	ROC-AUC
MLP	69.34%	82.42%	70.41%	75.94%	75.96%
CNN	70.52%	81.35%	74.09%	77.55%	76.23%
LSTM	69.85%	82.67%	71.02%	76.40%	76.80%

7. Explainable AI (XAI)

Interpretability techniques applied:

- Feature Importance: Random Forest identified pack_years, age, and COPD diagnosis as top predictors.
- SHAP: Confirmed Smoking exposure, COPD, and family history as most influential.
- LIME: Provided local explanations for individual patients, showing how features like high pack_years and positive family history influenced predictions.
- PDP & ICE plots: Showed monotonic increase in cancer risk with pack_years and age.

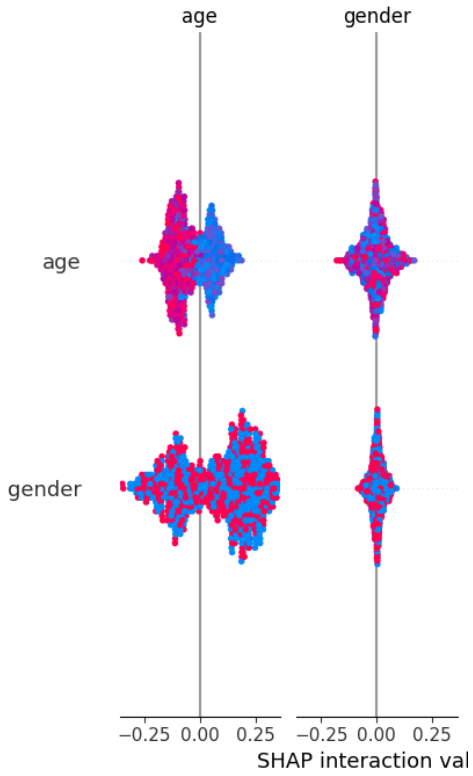
These techniques provided critical transparency into both ML and DL predictions, essential for clinical use.

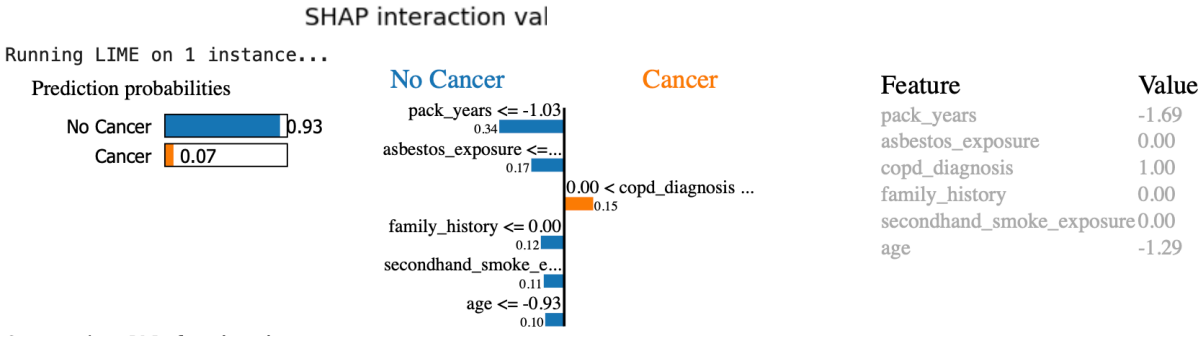
8. Comparative Analysis

Comparison across ML and DL methods:
- ML models (Random Forest, XGBoost) performed strongly and offered good interpretability through feature importance.
- DL models (CNN, LSTM) achieved slightly better predictive performance, especially in recall, but were more complex and less interpretable.
- XAI tools such as SHAP and LIME helped bridge interpretability gaps for DL.

Trade-off: While DL provides marginally higher accuracy, ML models are easier to interpret and may be preferable in real-world medical applications where explainability is critical.

	accuracy	precision	recall	f1	roc_auc	Accuracy	Precision	Recall	F1	ROC-AUC
Logistic Regression	67.03%	81.75%	66.99%	73.63%	74.01%	NaN	NaN	NaN	NaN	NaN
Decision Tree	62.74%	75.67%	67.48%	71.34%	59.90%	NaN	NaN	NaN	NaN	NaN
Random Forest	66.68%	78.27%	71.32%	74.63%	70.83%	NaN	NaN	NaN	NaN	NaN
SVM (Linear)	47.96%	66.18%	49.67%	56.75%	47.79%	NaN	NaN	NaN	NaN	NaN
KNN	63.75%	78.49%	65.10%	71.17%	67.84%	NaN	NaN	NaN	NaN	NaN
XGBoost	70.65%	81.12%	74.68%	77.77%	75.75%	NaN	NaN	NaN	NaN	NaN
MLP	NaN	NaN	NaN	NaN	NaN	69.34%	82.42%	70.41%	75.94%	75.96%
CNN	NaN	NaN	NaN	NaN	NaN	70.52%	81.35%	74.09%	77.55%	76.23%
LSTM	NaN	NaN	NaN	NaN	NaN	69.85%	82.67%	71.02%	76.40%	76.80%





9. Conclusion

This study successfully implemented both ML and DL models for lung cancer risk prediction. Random Forest and XGBoost emerged as the best-performing ML models, while CNN and LSTM showed superior DL performance. Explainable AI techniques (SHAP, LIME, PDP, ICE) provided transparency into model predictions. For medical practice, Random Forest combined with SHAP/LIME offers an optimal balance of accuracy and interpretability. However, DL methods could be deployed in high-stakes cases with strong interpretability support.