Lung Cancer Risk Prediction – Report

1. Dataset Overview

The dataset used for lung cancer risk prediction consists of **1,000 samples** with **10 features**, including both numerical and categorical data. The features are:

- age, gender, pack_years, radon_exposure, asbestos_exposure, secondhand_smoke_exposure, copd_diagnosis, alcohol_consumption, family_history, and the target lung_cancer.

There were no missing values except for the alcohol_consumption feature, which had **334 missing entries**. The class distribution showed **652 'Yes'** and **348 'No'** labels for lung cancer, indicating some imbalance. After applying **SMOTE**, both classes were balanced with **519 samples each** in the training set.

2. Summary Statistics

- **Age** ranged from 18 to 100 years, with a mean of **56.99 years**.
- **Pack years** ranged from approximately **0.4 to 99.9**, with a mean of **49.09**.

This shows that the dataset spans a wide age range and includes diverse smoking histories.

3. Performance Overview

Among the Machine Learning (ML) models, **Gradient Boosting** achieved the highest accuracy (**0.720**), F1 score (**0.779528**), and ROC-AUC (**0.773763**), indicating strong overall performance in classifying lung cancer risk. **XGBoost** followed closely with an accuracy of **0.715** and an F1 score of **0.783270**, suggesting robust predictive capability. **Logistic Regression** and **SVM** also performed reasonably well with accuracies of **0.670** and **0.665**, respectively, along with balanced F1 scores. In contrast, **Decision Tree (0.615 accuracy)** and **KNN (0.605 accuracy)** underperformed, likely due to overfitting or sensitivity to the dataset's structure, given the limited size of 1,000 rows. For Deep Learning (DL) models, **MLP** led with an accuracy of **0.695** and an F1 score of **0.768061**, slightly outperforming **LSTM (0.675 accuracy, 0.750958 F1)** and **1D CNN (0.660 accuracy, 0.719008 F1)**, though all DL models showed comparable or slightly lower performance than the top ML models.

4. Interpretability Trade-offs

ML models like Logistic Regression and Random Forest offer inherent interpretability through feature importance and coefficients, which is crucial for medical applications where understanding model decisions is essential. Although **Gradient Boosting** and **XGBoost** provided the best predictive accuracy, they are less transparent without the use of additional XAI techniques such as SHAP or LIME. On the other hand, DL models like **MLP, 1D CNN, and LSTM** are treated as black-boxes and require advanced explainability tools to interpret their predictions, potentially reducing trust in clinical decision-making despite their promising results.

5. XAI Insights

Feature Importance

The Random Forest Feature Importances highlighted that:

1. **pack_years** is the most influential feature.
2. **age** is the second most significant predictor.
3. Other features like **radon_exposure**, **asbestos_exposure**, and **copd_diagnosis** also contributed meaningfully to the model's predictions.

This aligns with known medical risk factors, where prolonged smoking exposure (pack_years) and increasing age are major contributors to lung cancer.

Partial Dependence Plot

The PDP for pack_years confirms that as smoking exposure increases, the risk of lung cancer rises, which is consistent with clinical knowledge. The plot shows a strong positive correlation between pack_years and the predicted risk.

These insights are crucial for fostering trust in AI-assisted healthcare solutions by explaining how individual risk factors contribute to the predictions.

6. Comparative Analysis

ML models, particularly **Gradient Boosting** and **XGBoost**, outperformed DL models on this small dataset, likely because deep learning requires larger datasets to fully realize its potential. While DL models such as **MLP** and **LSTM** showed promising results, they lagged behind the top-performing ML algorithms. The results highlight a common trade-off between accuracy and interpretability, suggesting that ML models are better suited for medical classification tasks when data is limited and explainability is critical.

7. Recommendation

For real-world clinical applications, **Random Forest** with SHAP explanations is recommended. Despite its slightly lower accuracy (**0.635**) compared to Gradient Boosting and XGBoost, it offers superior interpretability by clearly highlighting key features like **‘pack_years’** and **‘age’**, which clinicians can understand and trust. This balance between performance and transparency makes it the most suitable model for healthcare settings where explainability and patient safety are paramount.