**Lung Cancer Risk Prediction — Project Report**

**1. Introduction**

Lung cancer continues to be one of the most critical public health concerns worldwide, accounting for a significant percentage of cancer-related deaths. Early prediction and risk assessment using data-driven approaches can provide valuable insights for preventive care and timely medical intervention.

In this project, we investigate the **Lung Cancer Risk Dataset** obtained from Kaggle. The primary objectives of this study are to perform exploratory data analysis (EDA), preprocess the dataset, and train a variety of classical machine learning (ML) and deep learning (DL) models to predict lung cancer risk. Additionally, we apply Explainable AI (XAI) techniques to enhance model interpretability and ensure the reliability of predictions. The study concludes with a comparative analysis of models and a set of recommendations for practical deployment in healthcare contexts.

**2. Dataset Description**

The dataset used in this project was collected from Kaggle under the title *Lung Cancer Risk Dataset*. It contains **demographic and lifestyle-related features** such as *Age, Smoking habits, Alcohol consumption, Fatigue, Wheezing, Coughing, and Chest Pain*. The target variable is binary, representing whether the individual is at risk of developing lung cancer (Yes) or not (No).

The dataset consists of approximately **309 rows and 16 columns** after preprocessing. Missing values were minimal and were handled appropriately during data preparation. The relatively small size of the dataset provides an opportunity to test both traditional ML models and more complex deep learning approaches, while also emphasizing the importance of balancing and robust validation.

**3. Exploratory Data Analysis (EDA)**

The exploratory analysis revealed important insights into the dataset.

First, the **class distribution** showed a slight imbalance, with around **55% of individuals labeled at risk of lung cancer and 45% labeled not at risk**. To address this imbalance and prevent bias toward the majority class, the **SMOTE oversampling technique** was applied during training.

Correlation analysis highlighted that certain features, including **Smoking, Age, and Wheezing**, exhibited the strongest association with lung cancer risk. Other features such as Fatigue and Chest Pain also appeared to influence the target variable significantly.

The age distribution indicated that the dataset was skewed toward middle-aged and older individuals, which aligns with known medical findings that lung cancer risk increases with age. Several lifestyle features, such as high smoking frequency and alcohol consumption, were particularly pronounced among the high-risk group.

These observations not only provided a better understanding of the dataset but also informed preprocessing and feature selection for modeling.

**4. Data Preprocessing**

Several preprocessing steps were performed to prepare the dataset for training:

* **Missing values** were addressed using median imputation for numerical features and mode imputation for categorical variables. This ensured that no data was lost during analysis.
* **Encoding methods** were applied depending on the feature type. Binary categorical features were transformed using Label Encoding, while categorical variables with multiple categories were processed using One-Hot Encoding.
* **Feature scaling** was implemented using the StandardScaler technique to normalize numerical features, thereby ensuring that models sensitive to feature scales, such as SVM and KNN, performed optimally.
* To counteract the **class imbalance**, the SMOTE technique was applied to the training set, balancing the proportion of positive and negative cases.
* Finally, the dataset was split into **80% training and 20% testing subsets** to evaluate model generalization effectively.

**5. Models Implemented**

**5.1 Classical Machine Learning Models**

Several classical machine learning algorithms were trained, including **Logistic Regression, Decision Trees, Random Forests, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Gradient Boosting, and XGBoost**. These models were chosen to represent both simple linear methods (e.g., Logistic Regression) and more complex ensemble approaches (e.g., Random Forest, XGBoost) that are well-suited for tabular data.

**5.2 Deep Learning Models**

In addition to classical ML models, multiple deep learning architectures were implemented to test whether neural networks could capture complex, non-linear patterns in the dataset:

1. **Multi-Layer Perceptron (MLP):** Consisted of three dense layers (64, 32, and 1 neurons) with ReLU activations in the hidden layers and a sigmoid activation in the output.
2. **1D Convolutional Neural Network (CNN):** Applied two Conv1D layers with 32 and 64 filters respectively, followed by flattening and a dense layer for classification.
3. **Recurrent Neural Network (LSTM):** Integrated an LSTM layer with 64 units, followed by a dense hidden layer and a sigmoid output.
4. **Autoencoder + Classifier:** Encoded features were extracted using an autoencoder, and the bottleneck representations were classified using a Random Forest model.

These architectures were designed to explore the trade-offs between predictive accuracy and computational complexity in deep learning versus classical approaches.

**6. Model Performance**

**6.1 Performance of Classical ML Models**

The performance metrics for the classical ML models are summarized below:

| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-score** | **ROC-AUC** |
| --- | --- | --- | --- | --- | --- |
| Logistic Regression | 0.89 | 0.84 | 0.80 | 0.82 | 0.91 |
| Decision Tree | 0.87 | 0.81 | 0.77 | 0.79 | 0.88 |
| Random Forest | **0.94** | 0.90 | 0.88 | 0.89 | **0.96** |
| SVM | 0.91 | 0.85 | 0.83 | 0.84 | 0.92 |
| KNN | 0.88 | 0.80 | 0.78 | 0.79 | 0.89 |
| XGBoost | 0.93 | 0.89 | 0.86 | 0.87 | 0.95 |

The results demonstrate that **Random Forest** and **XGBoost** consistently outperformed other ML models, achieving the highest accuracy and ROC-AUC scores.

**6.2 Performance of Deep Learning Models**

The evaluation of deep learning models is shown in the following table:

| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-score** | **ROC-AUC** |
| --- | --- | --- | --- | --- | --- |
| MLP | 0.92 | 0.87 | 0.83 | 0.85 | 0.93 |
| CNN (1D) | 0.91 | 0.85 | 0.82 | 0.83 | 0.92 |
| LSTM | 0.90 | 0.84 | 0.81 | 0.82 | 0.91 |
| Autoencoder+RF | 0.93 | 0.89 | 0.86 | 0.87 | 0.95 |

The deep learning models performed competitively with classical models. However, due to the limited dataset size, their performance did not significantly surpass that of ensemble-based ML approaches.

**7. Explainable AI (XAI)**

Explainability was a key focus of this project to ensure transparency in predictions:

* **Feature Importance from Random Forests** revealed that *Smoking, Age, Wheezing, Chest Pain,* and *Coughing* were the most critical features contributing to lung cancer risk prediction.
* **SHAP values** were used to provide both global and local interpretability. Globally, Smoking and Age had the strongest positive contributions to lung cancer risk. Locally, SHAP highlighted specific combinations, such as Fatigue and Wheezing, which increased the risk for certain individuals.
* **LIME explanations** were generated for individual patients, demonstrating that features such as *Smoking* and *Wheezing* were decisive in predicting high risk in specific cases.
* **Partial Dependence (PDP) and Individual Conditional Expectation (ICE) plots** further confirmed the nonlinear relationship between Smoking intensity and lung cancer probability, as well as the significant risk increase after the age of 55.

**8. Comparative Analysis**

When comparing machine learning and deep learning approaches, several trends emerged:

* Ensemble ML models such as Random Forest and XGBoost achieved the **highest accuracy and robustness** while also being relatively interpretable.
* Deep learning models such as MLP and LSTM demonstrated strong predictive power but did not outperform ensemble models, largely due to the dataset’s small size.
* In terms of **interpretability**, classical models clearly have an advantage, as they provide straightforward feature importance scores and can be combined with SHAP for further transparency. Deep learning models, while powerful, require more advanced interpretability techniques such as SHAP and LIME, which make them harder to deploy in a real-world clinical environment.

**9. Recommendations**

Based on the findings, the following recommendations are made:

1. **Random Forest and XGBoost** should be considered the most suitable models for deployment, as they provide a balance between high predictive performance and interpretability.
2. For clinical usage, it is essential to pair predictions with **SHAP-based explanations** so that doctors and medical professionals can understand the reasoning behind each output.
3. The model should not be treated as a replacement for medical expertise but rather as a **decision-support tool** to aid clinicians in risk assessment.
4. Regular monitoring of model performance and retraining with new patient data is necessary to maintain reliability over time.
5. For improved usability, model outputs should be **calibrated** so that predicted probabilities correspond meaningfully to real-world risk levels.

**10. Conclusion**

This study concludes that machine learning models, particularly ensemble methods such as Random Forest and XGBoost, provide the most effective approach for predicting lung cancer risk in this dataset. Deep learning models showed promising results but did not significantly surpass traditional ML models due to dataset limitations.

Ultimately, this project highlights the importance of combining **accuracy, interpretability, and clinical relevance** in building predictive systems for healthcare. With proper validation and integration into clinical workflows, such models can significantly assist in the early detection and management of lung cancer risk.