**CSC 586 U16 - Data Mining Methods**

**Project Title: Visual Analytics for Lung Cancer Risk Factors**

**Lalitha Priya Bijja [1]**Computer Science  
University of South Dakota

**Uday Kumar Dasari [2]**Computer Science  
University of South Dakota

**Maharshi Devarasetty [3]**Computer Science  
University of South Dakota

**I. Introduction**

Lung cancer is one of the leading causes of cancer-related deaths worldwide. Early detection and understanding the risk factors are critical in improving patient outcomes and reducing mortality rates [13]. With advancements in medical data collection and analysis, there is a growing opportunity to explore large datasets related to lung cancer patients. These datasets can reveal valuable insights into common symptoms, risk factors (like smoking or chronic diseases), and demographic information (like age or gender) that contribute to lung cancer diagnosis.

Data analytics and machine learning have emerged as powerful tools in the healthcare industry. Through exploratory data analysis (EDA), visualization, and predictive modelling, it is possible to better understand how certain risk factors contribute to lung cancer development. Additionally, machine learning models can enhance early detection by identifying patterns that may be overlooked by traditional diagnostic methods.

**Problem Statement:** The project addresses the critical need for efficient lung cancer diagnosis by combining exploratory data analysis, data visualization, and the development of predictive models such as logistic regression, SVM, and XGBoost to support decision-making in healthcare.

**II. Objectives**

1. It understands the key factors contributing to lung cancer development by analyzing a dataset of lung cancer patients and exploring correlations between various symptoms, habits (like smoking and alcohol consumption), and demographic features.

2. It improves early detection and diagnosis of lung cancer by using machine learning techniques to predict the likelihood of lung cancer based on patient data. This will aid in building models that help predict lung cancer cases based on various risk factors, thereby enhancing early intervention efforts.

**III. Literature review**

Lung cancer remains one of the deadliest cancers globally, accounting for over 2 million new cases and approximately 1.8 million deaths annually, as reported by the World Health Organization (WHO). Despite advances in treatment, early detection of lung cancer is critical to improving patient survival rates. Traditional diagnostic methods, such as chest X-rays and CT scans, are highly effective but often expensive and not readily available to all patients. Thus, there is increasing interest in exploring how data-driven methods and machine learning algorithms can improve the detection of lung cancer and its associated risk factors using readily available patient data.

1. Epidemiological Studies on Lung Cancer Risk Factors

Numerous epidemiological studies have identified the primary risk factors associated with lung cancer. Smoking is by far the most significant contributor, accounting for about 85% of cases globally. *Ezzati et al. (2003)* conducted an extensive study that demonstrated the correlation between smoking and lung cancer mortality. Other studies, such as *Mukherjee et al. (2020),* have highlighted additional factors like chronic respiratory diseases, family history, and environmental exposures, including pollutants and occupational hazards. These studies underline the importance of assessing lifestyle factors and health history when evaluating lung cancer risk.

2. Machine Learning in Lung Cancer Detection

Machine learning has become an increasingly crucial tool in the early diagnosis of lung cancer. *Sujatha et al. (2019)* evaluated various machine learning algorithms for lung cancer prediction, including logistic regression, decision trees, random forest, and SVM (Support Vector Machines). They found that, while traditional models like logistic regression were highly interpretable, more complex models like XGBoost often yielded better predictive performance due to their ability to handle non-linear interactions between variables. Studies by *Liu et al. (2020)* further confirmed that ensemble learning methods, such as XGBoost and Random Forest, provide superior accuracy when compared to other machine learning techniques in healthcare.

3. Handling Imbalanced Datasets in Medical Diagnosis

One of the challenges in developing machine learning models for medical diagnosis is dealing with imbalanced datasets. In lung cancer datasets, the number of non-cancerous cases often far outweighs the number of cancer cases, leading to biased models that favour the majority class. *Chawla et al. (2002)* introduced SMOTE (Synthetic Minority Over-sampling Technique) as a solution to this problem. SMOTE is a widely used method for generating synthetic examples of the minority class to balance the dataset, and its effectiveness in improving model performance has been demonstrated in several studies, including those focused on lung cancer diagnosis (*Tang et al., 2021).*

4. Data Visualization for Healthcare Analysis

Data visualization plays a critical role in understanding patterns in medical data. *Wang et al. (2018)* demonstrated the importance of visualizing risk factors using techniques such as histograms, bar charts, and heatmaps. Visual tools help both clinicians and data scientists identify trends in age, gender, and lifestyle factors that contribute to lung cancer risk. For example, *Belkin et al. (2019)* used visualizations to reveal that younger patients, especially non-smokers, were more likely to be diagnosed with non-small-cell lung cancer, a less aggressive form of the disease. This emphasizes the role of visualization in revealing hidden patterns in complex datasets.

5. Predictive Models and Correlation Analysis

Understanding the relationships between various risk factors is crucial in building accurate predictive models. *Zhang et al. (2017)* conducted correlation analyses on lung cancer datasets and found strong links between smoking habits, chronic diseases, and lung cancer diagnosis. Their work highlighted the need for feature selection and engineering to enhance predictive accuracy. Similarly, *Hastie et al. (2009)* emphasized the role of correlation heatmaps in identifying multicollinearity and reducing noise in predictive models. Research in this area suggests that focusing on a select number of features significantly improves model performance.

6. Gender Differences in Lung Cancer

Research has also explored the role of gender in lung cancer incidence and outcomes. *Wang et al. (2020)* conducted a study on gender-based differences in lung cancer and found that while males are more likely to develop lung cancer due to higher rates of smoking, females with lung cancer often present different symptoms and risk profiles. This difference makes it necessary to model gender as a factor in lung cancer predictions. *Oberije et al. (2015)* used machine learning to predict survival rates in lung cancer patients and found that gender was a significant predictor, influencing not only incidence rates but also patient outcomes.

**IV. Methodology**

This project employs a combination of data analysis, data preprocessing, machine learning algorithms, and data visualization to predict lung cancer risk based on a dataset containing patient information and symptoms. The approach is divided into multiple stages, utilizing the following techniques and tools:

**1. Data preprocessing** – Data Cleaning, Label Encoding, Balancing the dataset.

**2. Exploratory Data Analysis (EDA) and Visualization**

**a. Exploratory Data Analysis:**  
- Gender distribution in lung cancer cases  
- Age-wise breakdown of positive lung cancer cases.  
- Lifestyle risk factors such as smoking etc.  
- symptoms like coughing, chest pain, etc.

**b. Visualization techniques:**  
- Histograms and Bar charts  
- Stacked bar plots  
- Correlation Heatmaps

**3. Machine Learning Models** – Logistic Regression, Support Vector Machine (SVM), XGBoost (Extreme Gradient Boosting)

**4. Model Evaluation and Cross-Validation** – Cross-Validation, Performance metrics such as Accuracy, Precision, Recall, F1-score, AUC-ROC

**5. Tools and libraries used:**  
- Python programming language  
- pandas  
- Matplotlib and Seaborn  
- Scikit-Learn  
- XGBoost  
- SMOTE technique  
- Label Encoder and One-Hot Encoder

**V. Dataset**

The Lung Cancer Dataset used in this project is sourced from Kaggle and consists of various demographic, lifestyle, and health-related attributes that can help predict the presence of lung cancer.

**Dataset:** <https://www.kaggle.com/code/hasibalmuzdadid/lung-cancer-analysis-accuracy-96-4/input>  
**source** – Kaggle (survey lung cancer)  
**Number of Instances (Rows):** ~300 instances  
**Number of Attributes (Columns):** 16 columns (including the target variable)  
**Target variable** – LUNG\_CANCER

Future Work- Future work may include applying deep learning models or refining feature selection using advanced methods.

**VI. Expected Challenges**

**Imbalanced Dataset:** The dataset may have fewer positive cases (those with lung cancer) compared to negative cases. This creates an imbalance problem that will be addressed using techniques like SMOTE to oversample the minority class. *He, H., & Garcia, E. A. (2009)*

**Categorical Data:** The dataset contains many categorical features (e.g., SMOKING, GENDER, ANXIETY), which need to be converted to numerical format for most machine learning algorithms. Improper encoding can lead to unintended biases or model underperformance. *I. H., Frank, E., & Hall, M. A. (2011)*

With a relatively small dataset, the risk of **o**verfitting is high, especially when using complex models like XGBoost. The model may perform well on the training data but fail to generalize to unseen data. Techniques like cross-validation, regularization (e.g., L2 regularization), and early stopping can help mitigate overfitting.

**VII. Timeline**

Week 1 – Project plan and initial EDA report.

Week 2 – Data Pre-processing

Week 3 – Feature Engineered dataset and correlation analysis report

Week 4 – Handling class imbalance

Week 5 – Building Models Logistic Regression & SVM

Week 6 – Building Models XGBoost & Ensemble Techniques with performance comparison report.

Week 7 – Final model performance report and interpretability analysis.

Week 8 – Final report, visualizations, and conclusions.

**References**

[1] Ezzati, M., Henley, S. J., Thun, M. J., & Lopez, A. D. (2003). Role of Smoking in Global and Regional Cancer Mortality. *Annals of Oncology*, 14(6), 953–956.

[2] Mukherjee, S., et al. (2020). Lung cancer risk from occupational exposures and smoking: The synergy between personal and environmental factors. *Environmental Research*, 191.

[3] Sujatha, R., Chatterjee, J. M., et al. (2019). An overview of machine learning models in lung cancer prediction. *Journal of Computational Science*, 36, 101–108.

[4] Liu, C., Yin, G., et al. (2020). Boosting-based algorithms for predicting lung cancer survival. *Journal of Biomedical Informatics*, 107, 103455.

[5] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16, 321–357.

[6] Tang, Z., Shen, H., et al. (2021). SMOTE-based analysis in predicting lung cancer classification. *IEEE Access*, 9, 102413-102425.

[7] Wang, L., et al. (2018). Data visualization tools for risk factor analysis in lung cancer: Exploring demographic and clinical data. *Journal of Medical Imaging and Health Informatics*, 8(5), 925–931.

[8] Belkin, A., et al. (2019). Visualizing risk factors and symptoms of lung cancer in younger patients: A comparative study. *Lung Cancer Journal*, 134, 58–62

[9] Zhang, H., et al. (2017). Correlation-based feature selection and prediction of lung cancer in high-risk populations. *Scientific Reports*, 7, 564.

[10] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). *Springer*.

[11] Wang, Z., Wang, Y., et al. (2020). Gender differences in lung cancer risk, symptoms, and outcomes: A systematic review. *Journal of Cancer Research and Clinical Oncology*, 146(5), 1159–1168.

[12] Oberije, C., et al. (2015). Gender and risk prediction models for lung cancer survival using machine learning. *Annals of Oncology*, 26(2), 321–326.

[13] <https://www.cancer.org/cancer/lung-cancer/causes-risks-prevention/risk-factors.html>

[14] He, H., & Garcia, E. A. (2009). Learning from imbalanced data. https://doi.org/10.1109/TSMCB.2008.2002909

[15] Witten, I. H., Frank, E., & Hall, M. A. (2011). Data Mining: Practical Machine Learning Tools and Techniques. https://doi.org/10.1016/C2009-0-19715-5