Course Title: Data Mining
Course Number and Section: CS 619 (CRN: 21329)
Term: Spring 2025
Project: Data Mining Exploration and Application
Name: Drona Chaitanya Reddy Gangarapu

Predictive Modeling of Lung Cancer Diagnosis Using Clinical and Demographic Data

Abstract

This report presents a comprehensive study on predicting lung nodule malignancy using routine clinical and demographic data. We explore data preprocessing, exploratory analysis, and classification modeling (Logistic Regression, Decision Tree, SVM) on a dataset of 482 patients with 15 features each. Model evaluation reveals performance at chance levels (≈50% accuracy), highlighting data limitations and motivating richer feature sets and advanced methods.

Introduction

Lung cancer accounts for the highest number of cancer-related deaths worldwide, estimated at 1.8 million annually by the World Health Organization. Timely detection is crucial, yet early-stage disease often eludes standard screening methods. Machine learning offers a promising avenue for non-invasive triage tools that leverage routinely collected data. This study investigates whether simple clinical and demographic features can reliably predict lung nodule malignancy, potentially aiding physicians in prioritizing patients for advanced diagnostic imaging.

Background & Significance

Radiomics—the extraction of high-dimensional data from medical images—has shown promise in distinguishing malignant from benign nodules. However, such approaches require specialized software and expertise. In contrast, our work examines the predictive power of readily available variables, such as age, gender, smoking status, nodule measurements, and blood biomarkers. If successful, this approach could scale more broadly across healthcare settings with limited imaging resources.

Data Description

The dataset comprises 482 complete patient records (after dropping missing values) from an initial 500 entries. Each record contains 15 features, including:

- Demographics: Age, Gender
- Lifestyle: Smoking history (pack-years)
- Imaging measurements: Nodule diameter, texture score, margin sharpness
- Blood biomarkers: C-reactive protein, white blood cell count
- Clinical lab values: Platelet count, hemoglobin level
- Diagnosis label: Benign (0) or Malignant (1)

Descriptive statistics for each feature were computed to assess central tendency, dispersion, and distributional shape. No substantial outliers were detected beyond expected clinical ranges.

Methodology

1. Data Preprocessing

- Missing-value handling: Dropped 18 records containing nulls, resulting in 482 samples.
- Categorical encoding: Transformed gender (M/F) and smoking status (Yes/No) into numeric codes.
- Feature scaling: Applied z-score normalization (mean=0, std=1) to continuous variables.
- Target encoding: Confirmed binary diagnosis labels; no further binarization needed.

2. Exploratory Data Analysis

- Summary statistics: Computed mean, median, standard deviation for each feature.
- Histograms & boxplots: Visualized distributions for key variables (e.g., nodule diameter) to assess symmetry and detect skew.
- Correlation matrix: Generated heatmap; all feature pairs had $|r| < 0.3$, indicating low multicollinearity.

- Feature-target relationships: Scatterplots and violin plots showed considerable overlap between benign and malignant groups.

3. Feature Engineering

Initial attempts at polynomial expansion (degree=2) and interaction terms (e.g., age × nodule diameter) were evaluated but did not yield meaningful performance improvements on the validation set. Hence, primary modeling focused on the original feature set.

4. Predictive Modeling

- Train/Test split: Stratified 80/20 to maintain class balance.
- Algorithms:
  - Logistic Regression (max_iter=1000)
  - Decision Tree Classifier (max_depth tuned via grid search)
  - Support Vector Machine with RBF kernel (C, gamma tuned via CV)
- Hyperparameter tuning: 5-fold cross-validation grid search conducted for tree depth and SVM parameters.
- Evaluation metrics: Accuracy, Precision, Recall, F1 Score, ROC AUC calculated on test set.

Findings

Table 1 presents test-set performance for each model:

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
Logistic Regression	0.499	0.499	0.499	0.499	0.507
Decision Tree	0.514	0.514	0.519	0.516	0.514
SVM (RBF)	0.512	0.512	0.518	0.515	0.503

Key insights:

- All classifiers performed at near-random levels, with ROC AUC ≈0.50.
- The lack of discriminative power suggests feature insufficiency.
- Visual ROC curves and confusion matrices confirm high error rates.

Discussion

The near-random performance across models highlights several critical considerations:

1. Data Quality & Feature Relevance: Basic clinical measures may lack the granularity to capture malignant signatures. Advanced radiomic features or genomic markers could enhance predictiveness.

2. Model vs Data Complexity: Complex models like SVM may overfit on small datasets (n=482). Simpler models also failed, indicating data limitations over model capacity.

3. Challenge of Overlap: Significant distributional overlap in features between classes makes boundary learning difficult.

4. Future directions: Implement ensemble approaches (Random Forest, XGBoost), perform feature selection (LASSO), and gather additional modalities.

5. Ethical & Clinical Implications: A model performing at chance level cannot be deployed clinically. Rigorous validation and regulatory review are essential when integrating ML into healthcare.

Conclusion & Future Work

This study developed a reproducible pipeline for lung nodule malignancy prediction using accessible clinical data. Performance across three classification algorithms remained at chance levels, underscoring the need for richer datasets and refined modeling strategies. Future work will focus on integrating advanced imaging features, leveraging ensemble methods, and conducting robust cross-validation to improve model reliability and clinical applicability.

References

- World Health Organization. (2020). Global Cancer Observatory.

- Aerts, H. J. W. L., et al. (2014). Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nature Communications.

- Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research.