High-Accuracy Pneumonia Diagnostic Model Using DNA Methylation Biomarkers and Elastic Net Regression

Aahan Singh

Abstract:
Pneumonia continues to lack precise and early diagnostic tools, contributing to delayed treatment and poor outcomes. This study presents a high-accuracy machine learning classifier that leverages DNA methylation (CpG β-values) to distinguish pneumonia patients (n = 126) from healthy controls

Methodology:
Genome-wide methylation data were preprocessed using **StandardScaler** for normalization and **SMOTE** to address class imbalance. Over **200 CpG features** were initially selected and expanded by ~165 additional sites, followed by dimensionality reduction using **Recursive Feature Elimination (RFE)** and model explainability via **SHAP analysis**. The resulting features were functionally annotated using **UCSC Genome Browser** and **GREAT**, identifying key immune- and lung-related genes such as **C1QB, SMAD9, LILRB4, and EPHB2.**
Three models—**Logistic Regression, Random Forest,** and **Elastic Net**—were trained and evaluated for diagnostic performance.

Key Results:
The **Elastic Net model** achieved the highest performance, with:

- **Accuracy:** 98%
- **Precision:** 98%
- **Recall:** 98%
- **F1-score:** 98%

Conclusion:
This methylation-based classifier demonstrates robust potential as a non-invasive diagnostic tool for pneumonia. Future directions include external validation in independent cohorts, incorporation of regional and temporal metadata, and development of a clinical prototype for early detection and risk stratification.

Keywords: DNA methylation, pneumonia diagnosis, Elastic Net, CpG biomarkers, machine learning, epigenetics, SMOTE, SHAP, RFE

Skills Used: Python (pandas, scikit-learn), bioinformatics (CpG-to-gene mapping, pathway analysis), machine learning optimization