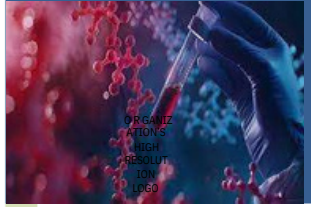


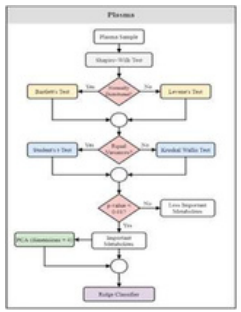
Computational Identification and Validation of Novel Biomarkers for Lung Cancer

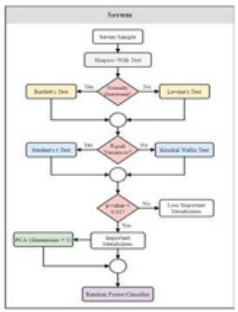
Abstract

Metabolomic biomarkers are essential for the early detection of lung cancer, enhancing diagnostic speed and accuracy. This study presents a systematic approach to identify key metabolites in plasma and serum samples. We utilized statistical tests—including Shapiro–Wilk, Bartlett's, Levene's, Student's t-Test, and Kruskal–Wallis—to uncover potential biomarkers. In the second phase, Recursive Feature Elimination with Random Forest identified the most significant metabolites, achieving prediction accuracies of 100% for plasma and 90.91% for serum samples, surpassing existing methods.

Introduction

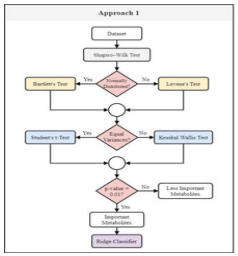
Lung cancer is a leading cause of cancer-related deaths, with over 235,000 new cases and 131,000 deaths estimated in the U.S. in 2021. Early diagnosis is crucial for improving survival rates, as life expectancy drops significantly from 6.16 years at Stage I to just 1.42 years at Stage IV. Traditional diagnostic methods often detect lung cancer at advanced stages, limiting treatment options. This study leverages advancements in metabolomics—the analysis of metabolites in biological samples—to identify unique biomarkers in blood plasma and serum that can effectively differentiate lung cancer patients from healthy individuals.

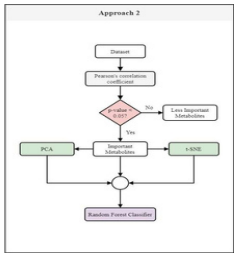




Methods and Materials

The dataset (ST000392) used in this study was generated through GC-TOF-MS analysis of plasma and serum samples from 82 subjects, including 41 lung cancer patients and 41 controls (20 males, 60 females). A total of 158 metabolites were examined, with samples collected in EDTA tubes and stored at −80°C. Data processing was conducted using ChromaTOF software (version 2.32) and validated via the Metabolomics BinBasedatabase, all under ethical guidelines with IRB approval. Feature selection involved statistical tests like Shapiro-Wilk and Student's t-Test to identify significant metabolites, followed by Recursive Feature Elimination (RFE) to refine the selection.

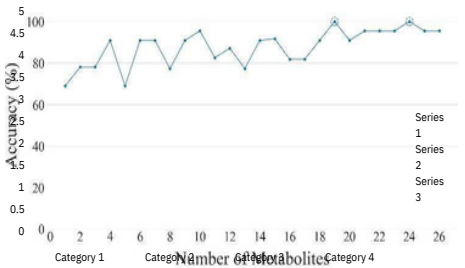


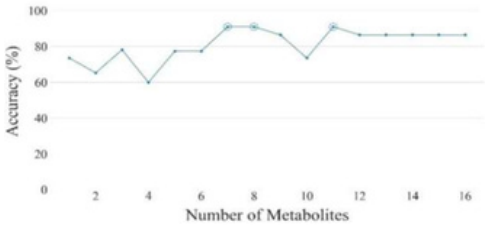


Results

The initial screening using statistical tests like Shapiro-Wilk, Bartlett's, Levene's, Student's t-Test, and Kruskal-Wallis identified 26 important metabolites in plasma and 16 in serum samples. Recursive Feature Elimination (RFE) further refined the selection, revealing that 19 plasma metabolites and 7 serum metabolites provided the best accuracy. These optimal biomarkers were identified by iteratively eliminating features and evaluating the impact on classification performance.

The selected biomarkers achieved prediction accuracies of 100% for plasma and 90.91% for serum samples, surpassing the performance of leading existing methods while using fewer metabolites compared to previous studies. This streamlined approach to biomarker selection enhances the efficiency and reliability of lung cancer detection, paving the way for the development of non-invasive diagnostic tools that could significantly improve early intervention and patient outcomes.





Discussion

After conducting Bartlett's and Levene'stests on the plasma samples, we identified 138 features with consistent variance (parametric), which were subsequently analyzed using Student's t-Test. The remaining 20 features exhibited varying variance (non-parametric) and were evaluated using the Kruskal-Wallis Test.

Conclusions

Cancer remains a leading cause of death worldwide, making early and cost- effective detection crucial for improving survival rates. In this study, we advanced the identification of metabolomic biomarkers for lung cancer using machine learning techniques. By applying the Ridge Classifier, we achieved 100% accuracy in plasma samples, while the XGBoostClassifier delivered 90.91% accuracy for serum samples. Our method not only enhances predictive accuracy but also reduces the number of biomarkers needed, identifying 19 key metabolites in plasma and 7 in serum. Compared to previous methods, our approach is more efficient and effective, with significant implications for early cancer detection that could revolutionize bioinformatics and healthcare.

Thank you!

Team Freemasons: Arnav Sonavane (leader), AarianThakur, HarshalMahale, Tania Vithayathil
GitHub Link: https://github.com/W2SG-smokiee/stanford-biohacks