UTM

UNIVERSITI TEKNOLOGI MALAYSIA

SECB3203-PROGRAMMING FOR BIOINFORMATICS

SEMESTER 1 2024/25

PROJECT REPORT

Title: Machine Learning-Driven Biomarker Discovery for Early Detection of Lung Cancer

GROUP 3

Name	Matric No
BEH SHU YIE	A22EC0142
ONG SHUN SHENG	A22EC0259
ZIKRY DABIAL BIN MAJUNING	A22EC0298
MUHAMMAD FAWWAZ REDHA BIN MOHD FADZLI	A22EC0210

Table of Contents

1.0 Introduction

According to the WHO statistics, lung cancer is defined as the second most common oncological disease among both women and men in the world. In the year of 2020, 2.26 million cases of lung cancer and 1.8 million deaths were recorded worldwide. If lung cancer patients can be detected earlier, they will have more chances to survive. It is clear that the early detection of lung cancer is extremely significant in order to enable the lung cancer patients to receive operation as early as possible. In this project, we aim to develop a machine learning-driven model that uses metabolomic biomarkers derived from plasma samples to detect NSCLC at an earlier, more treatable stage.

1.1 Problem Background

As we know, lung cancer is one of the leading causes of cancer-related mortality worldwide, and early diagnosis is critical for improving survival rates. The current clinical method for detecting lung cancer like LDCT, is the standard screening technique for high-risk individuals. However, due to the high cost and significant false discovery rate, the uptake and fulfilment of LDCT screening are unsatisfactory and the current diagnostic methods usually detect the disease at later stages when treatment options are limited.

There is a growing interest in blood-based screening methods for cancer diagnosis, such as detecting metabolic biomarkers in plasma. Previous studies have explored miRNAs as biomarkers for various cancers, including breast and prostate cancer, but these approaches have limitations, such as high cost, technical challenges, and limited sensitivity and accuracy.

In response to these challenges, this study focuses on identifying a new set of metabolic biomarkers in Chinese patients' plasma that could serve as early-stage lung cancer diagnostic tools. By combining advanced metabolomics technology and machine learning methods, the researchers have identified a combination of six metabolites that significantly improve diagnostic accuracy, offering high sensitivity (98.1%) and perfect specificity (100%) for early lung tumor detection. These findings suggest the potential for developing a more sensitive, specific, and minimally invasive screening tool for early lung cancer diagnosis.

1.2 Problem Statement

1. 2.26 million cases of lung cancer and 1.8 million deaths were recorded worldwide in 2020.
2. Over 70% patients are diagnosed when their tumor are developed to the advanced stages, and most of them are not suitable for receiving operation
3. CT screening programs are currently underutilized due to their significant cost and a high false-discovery rate, limiting their effectiveness in clinical settings.

1.3 Aim and Objectives

Aim

To develop a machine learning model that identifies metabolic biomarkers for early-stage lung cancer detection and supports personalized treatment strategies based on individual protein profiles.

Objectives

1. To analyze and identify metabolic biomarkers for early detection of lung cancer.
2. To design and create a machine learning model for lung cancer detection.
3. To evaluate the effectiveness and accuracy of the proposed framework.

1.4 Scopes

This project has a clear and targeted focus, aiming to detect lung cancer in order to improve diagnostics through metabolomic profiling but it also comes with some limitations.

This project will utilize pre-treatment plasma samples from NSCLC patients and healthy people in this project. The type of data considered is metabolomic data, which focus on metabolic changes in amino acids, acylcarnitines, and tryptophan metabolism, that serve as potential biomarkers. The genomic and transcriptomic data will not be considered in this project as we are focusing on the metabolic alterations in cancer patients. The attributes to be used include metabolite concentrations and metabolic ratios derived from the patient's plasma. However, characteristics such as age, gender, unhealthy habits like smoking, and cancer stages are additional factors but are not the primary focus of our analysis. The project is focused solely on NSCLC which represents the most common form of lung cancer that has 80-85% of cases, excluding small cell lung cancer (SCLC). By targeting NSCLC, the biomarkers that signal early-stage lung cancer will be detected.

In this project, the targeted metabolic profiling of patient plasma samples will be used to identify through the application of quantitative HPLC–MS/MS analysis. This technique is essential to identify metabolic disruptions that could indicate the presence of NSCLC. Besides that, supervised machine learning techniques like logistic regression, support vector machines (SVM), random forests and other algorithms are used to build diagnostic models and choose the best diagnostic accuracy model. These models utilize metabolite concentration data to classify patients as either NSCLC or non-cancer individuals.

The project begins by collecting plasma samples from two groups—NSCLC patients and healthy individuals. These samples are subjected to metabolomic profiling using quantitative HPLC-MS/MS analysis, followed by computational modeling to detect significant biomarkers. The measurement tools used are multivariate analyses to preprocess and feature select the metabolomic data. Metabolites showing significant differences between NSCLC and control groups are then used as inputs for machine learning models. The dataset will randomly split into training (70%) and validation (30%) subsets using Stratified K-Fold Cross Validation to ensure robust model evaluation. Performance metrics such as accuracy, classification reports and confusion matrix are used to assess the predictive power of the models.

2.0 Related Works

Many recent studies have focused on the problem of lung cancer, especially non-small cell lung cancer (NSCLC), which represents a major global health challenge. Most cases are diagnosed at late stages, leading to poor survival outcomes. Hence, early detection tools are highly needed to identify lung cancer at an earlier stage, which would allow patients to receive early treatment and improve their survival rates. In this project, a machine-learning model will be developed, utilizing biomarkers and advanced machine-learning techniques to facilitate the early detection of NSCLC.

In January 2019, the National Central Cancer Registry of China (NCCRC) released its latest nationwide tumor statistics of population-based tumor registry data gathered from 368 tumor registries in 2015. According to this report, lung cancer ranks top for its incidence among malignant tumors in China. Moreover, it has been reported that early-stage patients who received proper treatment could have a 5-year survival rate of around 40%. Unfortunately, over 70% of patients are diagnosed when their tumor is developed to the advanced stages like stage 3 or stage 4, and this situation is really hard for them to recover which finally leads to death (Ying Xie et al., 2021). Beside that, one of the previous studies validated that early detection of early-stage lung cancer patients had contributed to pronounced survival benefits during the decade. Patients with stage I lung cancer, accounted for an increasingly considerable proportion, increasing from 15.28% to 40.25%, coinciding with the surgery rate increasing from 38.14% to 54.25%. Overall, period survival analyses found that 42.69% of patients survived 5 years, and stage I patients had a 5-year overall survival rate of 84.20%. Compared with that in 2009–2013, the prognosis of stage I patients in 2014–2018 was dramatically better, with a 5-year overall survival rate increase from 73.26% to 87.68% (Wang CD et al., 2023).Therefore, it is clear that the detection of lung cancer is so vital for the current situation in order to decrease the number of deaths due to lung cancer and help those victims receive early treatment.

In order to detect early lung cancer disease, high risk people are generally recommended annual radiologic screening by low-dose computed tomography (LDCT). However, fulfilment of CT screening is unsatisfactory because of the significant cost and high false-discovery rate. Therefore, the availability of blood-based screening could increase lung cancer patient uptake, including plasma metabolic biomarkers detection. From previous studies of various kinds of cancer, miRNAs screening has become the biomarkers for

distinguishing cancer patients and different tumor subtypes. Nonetheless, miRNA screening has its limitations, such as high cost and technical monopolization. After that, clinical usual tests of blood antigen also meet some problems which are low sensitivity and low accuracy. Therefore, metabolic biomarkers are chosen for the early stage lung cancer diagnostic biomarkers. The main reason we chose metabolic biomarkers is due to the cancer progression being strongly linked to changes in cellular metabolism. Cancer cells can modify their metabolic pathways to support their rapid growth and division. It will request specific enzymes to catalyze biochemical reactions, to meet the growing need of lung cancer development and progression. Here are some examples of key metabolites and enzymes in lung cancer which are proline dehydrogenase, L- kynurenine, aspartic acid, leucine and others.

There have been numerous significant advances in the use of machine learning algorithms to evaluate biological data and find possible biomarkers that might assist in the early identification of lung cancer. Metabolomics data is routinely analyzed using models such as Support Vector Machines (SVM), Random Forest, Naive Bayes, and K-nearest neighbors (KNN) to distinguish between healthy persons and lung cancer patients. These models are especially beneficial for discovering metabolic indicators that can accurately predict the existence of early-stage lung cancer. The use of machine learning and statistical approaches has shown promise in improving classification accuracy for lung cancer detection, particularly when paired with data from plasma metabolites or genetic markers. The use of algorithms to forecast how patients will react to therapies like immunotherapy is another well-known application of machine learning. Researchers have discovered intricate connections between immune cell behaviour and metabolic states by utilizing methods like unsupervised clustering, which has given them insight into how patients might react to various treatment modalities. These advances in machine learning are speeding up the identification of biomarkers, which will eventually lead to more precise, non-invasive diagnostic instruments and more individualized treatment plans for patients with lung cancer.
In conclusion, this research is developing a machine learning model that focuses on effective biomarkers for early NSCLC detection for people from all over the world by offering a potential non-invasive diagnostic approach that could enhance early treatment and survival outcomes.

3.0 Flow Chart Of The Proposed Approach

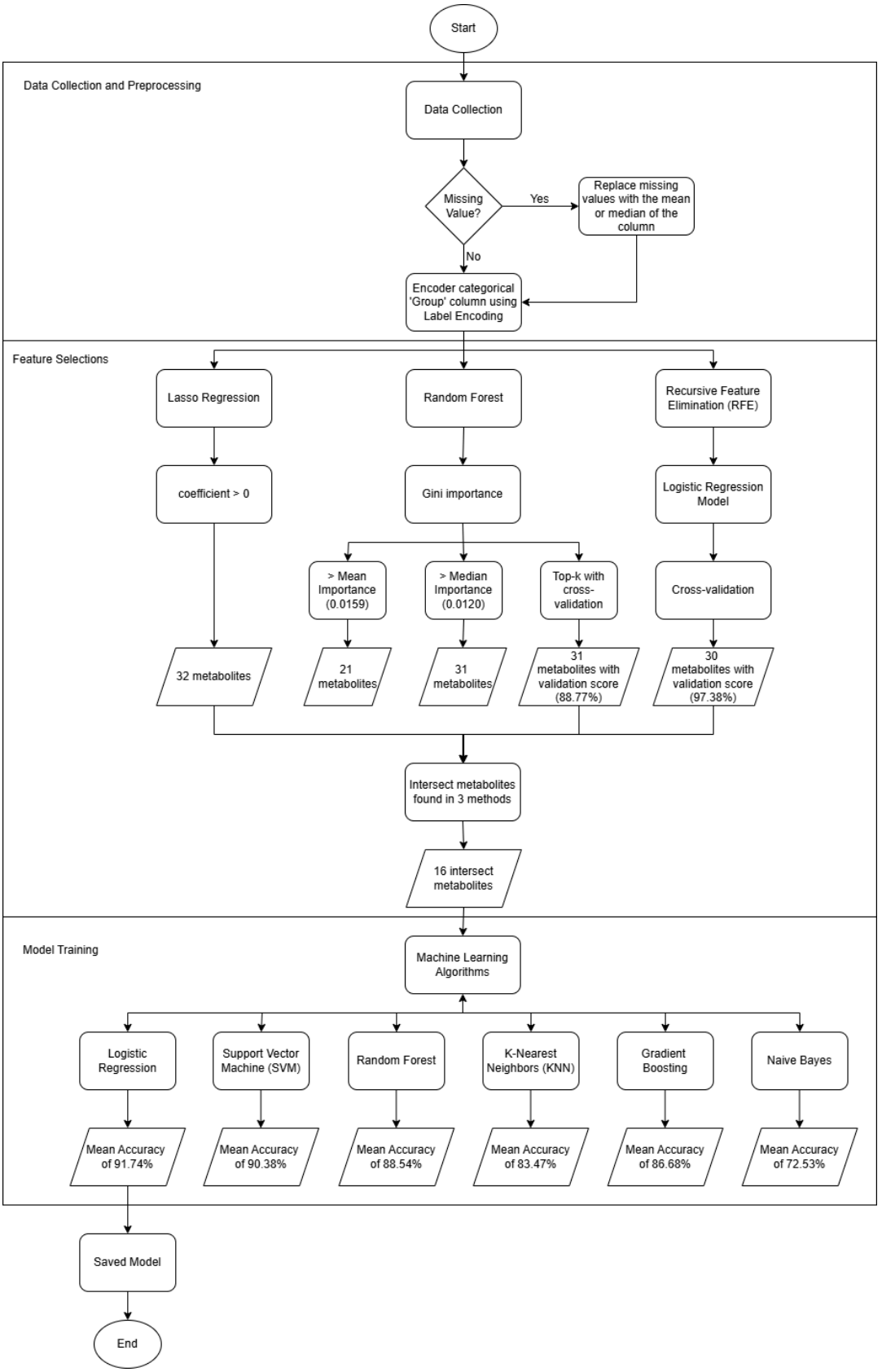


Figure 1: Flow chart of the proposed approach

3.1 Software and Hardware Requirements

Software:

- Windows(for local development)-Alternative for small scale data analysis and development environment.
- Python 3.11.0 until 3.12.9
- StreamLit-Building interactive web applications for data science and machine learning for our projects.
- Visual studio code/Google Colab
- Scikit-learn
- NumPy and Pandas
- Jupyter Notebook
- GitHub

Hardware:

- **Processor**: Intel i7 or AMD Ryzen 7 (8 cores)(i5 and AMD Ryzen 5 also can but the generation of the processor must be above 9 gen and 7000 series.
- **RAM**: 16 GB/32 GB (ideal for handling medium to large datasets)
- **Storage**: 1 TB SSD and above (fast read/write for data loading and model training)
- **GPU**: NVIDIA RTX 3060 or equivalent (for deep learning models)
- **Operating System**: Windows 10 and above
- **Additional**: High-speed internet and external storage for data backup.

4.0 Data Collection And Preprocessing

The data used in this study comprised metabolomic profiles obtained from participants, encompassing both healthy individuals and those diagnosed with non-small cell lung cancer (NSCLC). The dataset included 63 metabolites relevant to lung cancer diagnosis, ensuring a comprehensive representation of metabolic changes.

Status	Number of data
Lung Cancer	118
Healthy	100

Table 1: Number of data between lung cancer patients and non-cancer patients

To prepare the data for analysis, preprocessing steps were applied to ensure consistency and reliability. Initially, the raw dataset was cleaned by addressing missing values and outliers, which could otherwise compromise the integrity of the analysis. Missing values will be replaced with the mean or median of the column to minimize information loss. Outliers were detected and treated based on domain-specific thresholds to prevent distortion in model performance. The categorical column Group was encoded using LabelEncoder from sklearn.preprocessing. Group lung cancer was encoded as 1 while group healthy was encoded as 0. The new dataset was saved as “Preprocessing_Dataset.xlsx” after preprocessing was done.

5.0 Model Development

In model development, we loaded the preprocessed dataset into a DataFrame called data. The data are separated into features (x) and target (y). Features (X) are created by dropping the Group column from the dataset. This ensures all other columns are used as input variables. The target (y) is defined as the Group column, which likely contains labels or categories for classification. Next, a StandardScaler is applied to standardize the features in X. This scales the data to have a mean of 0 and a standard deviation of 1, which helps improve the performance and stability of many machine learning algorithms. The data is divided into training and testing sets using train_test_split(). 70% of the data (X_train and y_train) is used for training the model, while 30% (X_test and y_test) is reserved for testing. The split is controlled by a random_state of 42 to ensure reproducibility of results.

Next, feature selection was conducted using multiple approaches to assess the importance of individual features. The analysis started with the identification of features using Lasso regression, Random Forest, and Recursive Feature Elimination (RFE). For Lasso regression, it is a regularization method that can help identify important features by adding a penalty for the number of features in the model. It performs both feature selection and regularization, ensuring that less important features are eliminated by driving their coefficients to zero. This method is particularly useful when there are many features and we want to avoid overfitting. Research shows that Lasso is effective in high-dimensional datasets, such as metabolomics, by selecting biomarkers associated with disease outcomes (Tibshirani, 1996). Therefore, in Lasso regression, features with non- zero coefficients were selected.

For Random Forest, it is an ensemble learning method that uses decision trees and is known for its effectiveness in feature selection. It can identify which features contribute most to prediction by measuring the decrease in model accuracy when each feature is randomly permuted. This approach is particularly useful for handling complex, non-linear relationships between features. Research has demonstrated that Random Forest can accurately rank features based on importance in biomedical datasets (Liaw & Wiener, 2002). In Random Forest, the feature importance scores were evaluated, and features exceeding the mean or median importance thresholds were considered significant and selected. Besides that, Validation is performed by experimenting with top-K features (from 1 to 63) to find the optimal number of features.

Lastly, RFE is an iterative method that recursively removes the least important features based on model performance. It is particularly useful when we want to reduce the number of features while maintaining predictive accuracy. RFE has been widely applied in biomarker discovery as it helps identify the most informative features for classifying diseases like cancer (Guyon et al., 2002). The concept of RFE used in our model is using Logistic Regression as the base model, and it will rank features based on the coefficients. Then, the RFE process is looping through feature sets from 1 to 63. The optimal number of features was determined using cross-validation techniques.

After that, intersection and union analyses were conducted on the features identified through the different selection methods. The overlap of the top features from Lasso regression, Random Forest, and RFE ensured robustness in selecting metabolomic markers. The combinations of features were assessed for their predictive performance. Visualizations such as Venn diagrams highlighted feature overlaps among the different selection approaches. The intersect metabolites will be used to build the machine learning models.

6.0 Model Evaluation

Multiple machine learning models were developed and evaluated to assess the predictive power of the selected metabolomic biomarkers. The following algorithms were implemented, which are Logistic Regression, Support Vector Machine (SVM), Random Forest Classifier, K-Nearest Neighbors (KNN), Gradient Boosting Classifier, and Naive Bayes Classifier. Each model's performance was measured using several quality metrics, including accuracy, precision, recall, F1-score, and confusion matrix, ensuring a comprehensive evaluation of diagnostic potential.

Logistic Regression

Logistic Regression was chosen for its simplicity and interpretability. It performed well when features exhibited a linear relationship with the target. Evaluation metrics, such as accuracy and F1-score, confirmed its effectiveness, while its coefficients provided insights into feature importance. However, its performance was constrained in capturing non-linear relationships.

Support Vector Machine (SVM)

SVM, particularly with the radial basis function (RBF) kernel, demonstrated the ability to handle non-linear separations. The model achieved high accuracy and precision, particularly for well-separated classes. However, its computational demands increased with larger datasets, and proper scaling via StandardScaler was critical to its success.

Random Forest

Random Forest excelled in handling high-dimensional datasets with complex interactions. It showed robustness against overfitting due to its ensemble approach and provided feature importance scores that aligned with the selected metabolites. Performance metrics, such as accuracy and recall, validated its capability in biomarker-based classification.

K-Nearest Neighbors (KNN)

KNN, as a non-parametric algorithm, was straightforward to implement and performed effectively for small datasets. Its accuracy and recall metrics were highly dependent on the choice of K and distance metrics. Uniformly scaled data ensured optimal performance.

Gradient Boosting Classifier

Gradient Boosting Classifier combined multiple weak learners to improve performance iteratively. It demonstrated high accuracy and F1-scores, particularly in handling imbalanced datasets. However, it required careful hyperparameter tuning to prevent overfitting.

Naive Bayes

Naive Bayes was computationally efficient and well-suited for high-dimensional data. Despite its simplicity, it achieved reasonable accuracy and provided baseline comparisons for other models.

Each model undergoes K-Fold Cross Validation, where the dataset is divided into multiple folds. Each fold takes a turn as the validation set while the remaining folds are used for training. This approach ensures the model is evaluated across multiple splits, reducing reliance on a single random train-test split. By using Stratified K-Fold, the class distribution in each fold remains similar, which is especially beneficial for imbalanced datasets. Aggregating results across all folds provides a more reliable and stable estimate of model performance, minimizing the impact of specific train-test splits.

6.1 Result Analysis And Discussion

6.1.1 Result Analysis

The present study comprised a total of 218 plasma samples from NSCLC and NC patients (Table 1). Figure 1 represents the flow chart to identify potential metabolites as well as develop the highest accuracy diagnostic model of NSCLC. Metabolic profiling that included the 63 metabolites was feature selected through some multivariate feature selection methods, consisting of lasso regression, random forest and recursive feature elimination (RFE). At the end of the report, the supervised machine learning model used to select a combination of metabolites as potential biomarkers for detecting lung cancer was performed.

Lasso Regression

The 63 metabolites undergo lasso regression which shrinks the coefficients of less important features toward zero. Features with non-zero coefficients are considered important and retained. Figure 2 showed the result of 32 metabolites with non-zero coefficients.

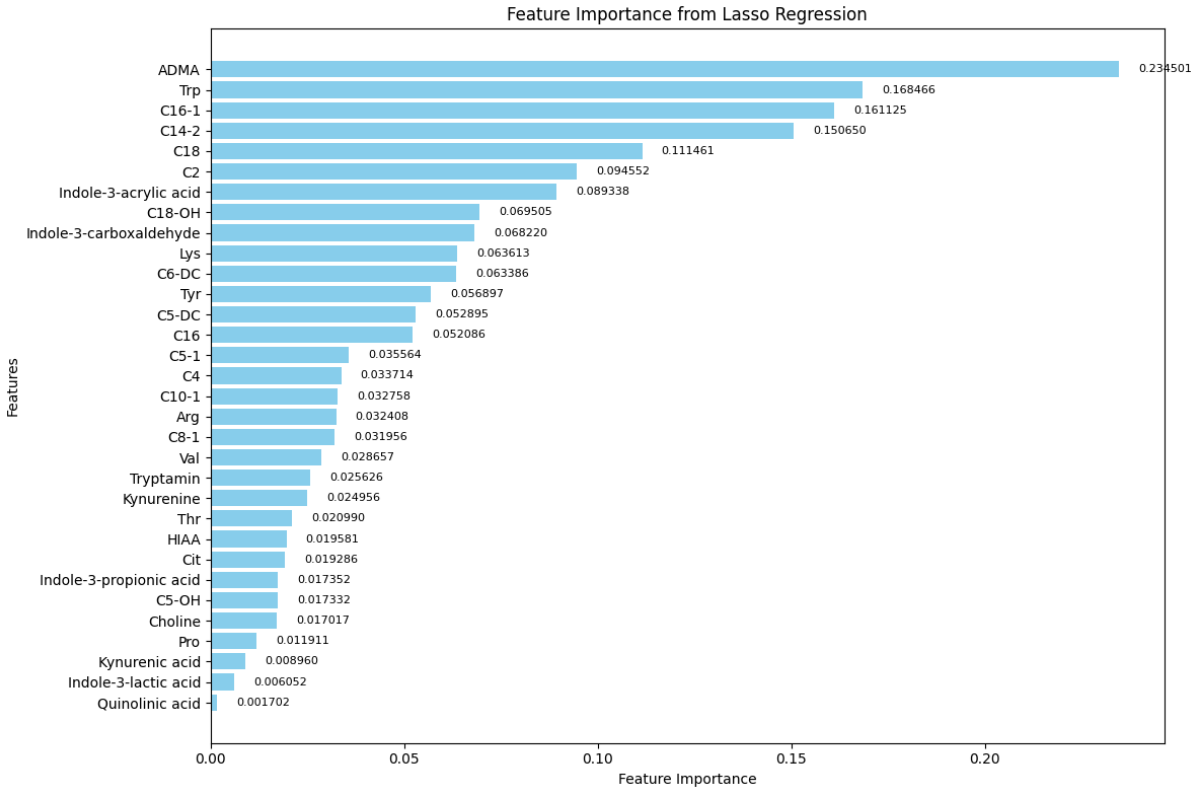


Figure 2: Feature Importance from Lasso Regression

Hence, the 32 metabolites with an importance greater than 0 are ADMA, Trp, C16-1, C14-2, C18, C2, Indole-3-acrylic acid, C18-OH, Indole-3-carboxaldehyde, Lys, C6-DC, Tyr, C5-DC,

C16, C5-1, C4, C10-1, Arg, C8-1, Val, Tryptamin, Kynurenine, Thr, HIAA, Cit, Indole-3-propionic acid, C5-OH, Choline, Pro, Kynurenic acid, Indole-3-lactic acid and Quinolinic acid.

Random Forest

Furthermore, the 63 metabolites undergo random forest. First, we use the threshold based on mean and median feature importance to select features. The calculated mean was 0.0159 while the calculated median was 0.0120. From Figure 3, the results showed that 21 metabolites above the mean importance (which is the red dotted-line) while 31 metabolites above the median importance (which is the blue dotted-line).

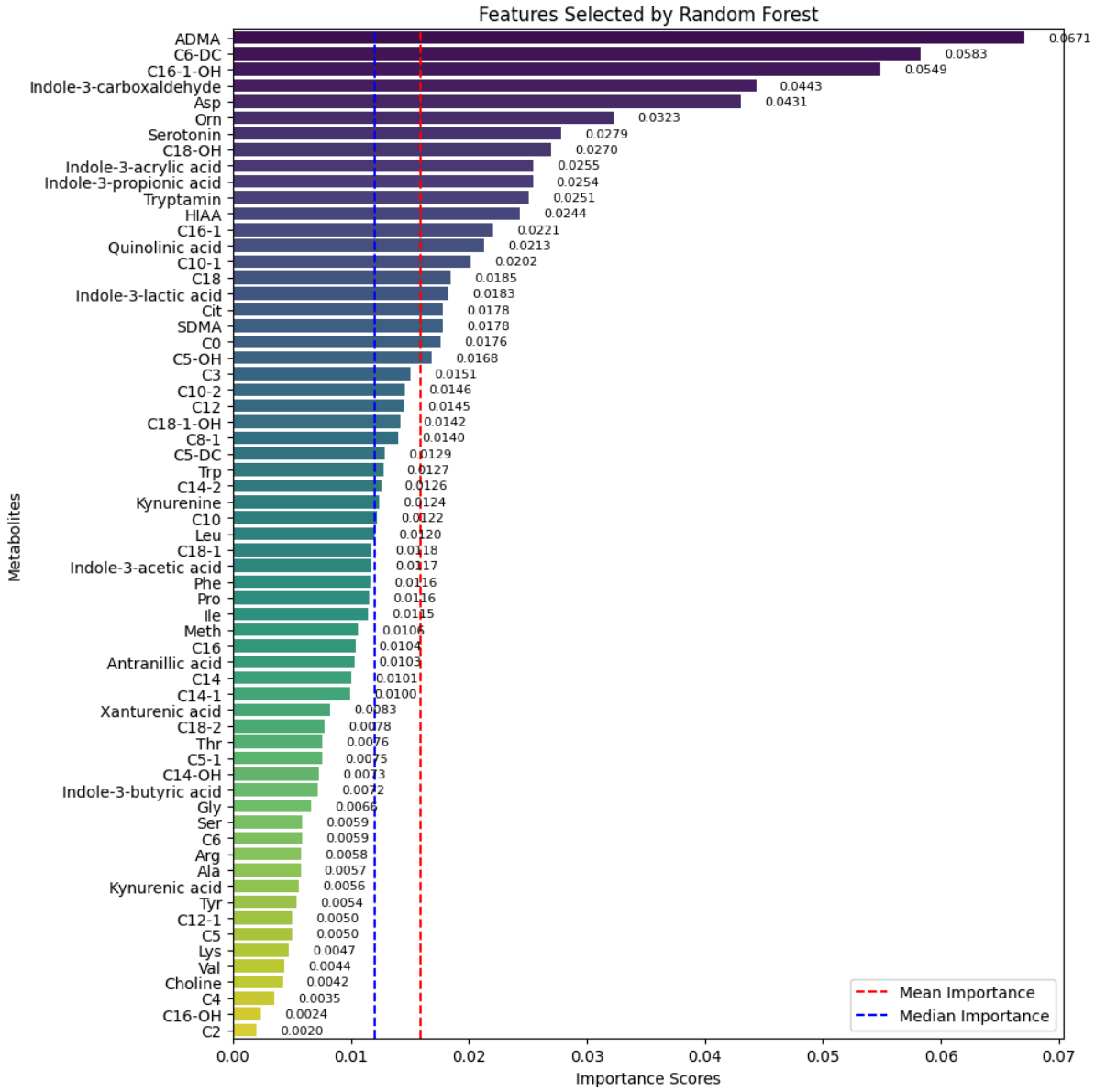


Figure 3: Features Selected by Random Forest

Next, we experimented with top-K features (1-63) to validate model performance using cross-validation. From Figure 4, we found that the highest cross-validation score with 31 features was 88.77%. The top 31 metabolites based on validation score showed in Figure 5.

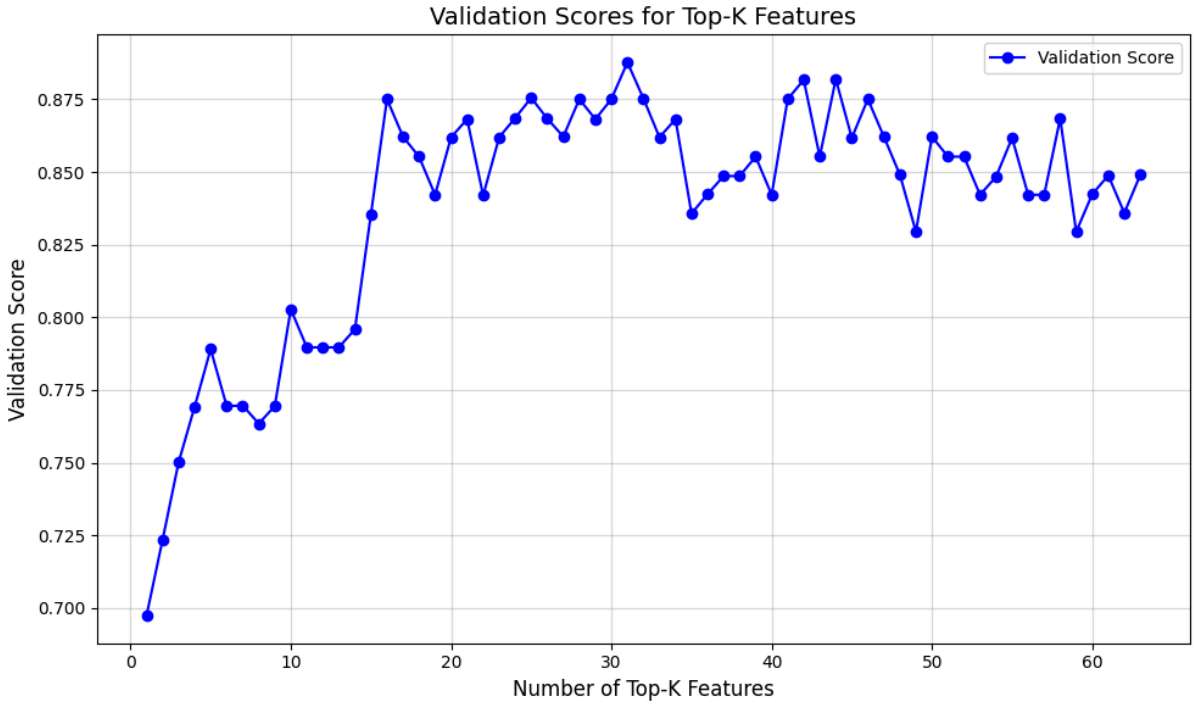


Figure 4: Validation Score for Top-K Features

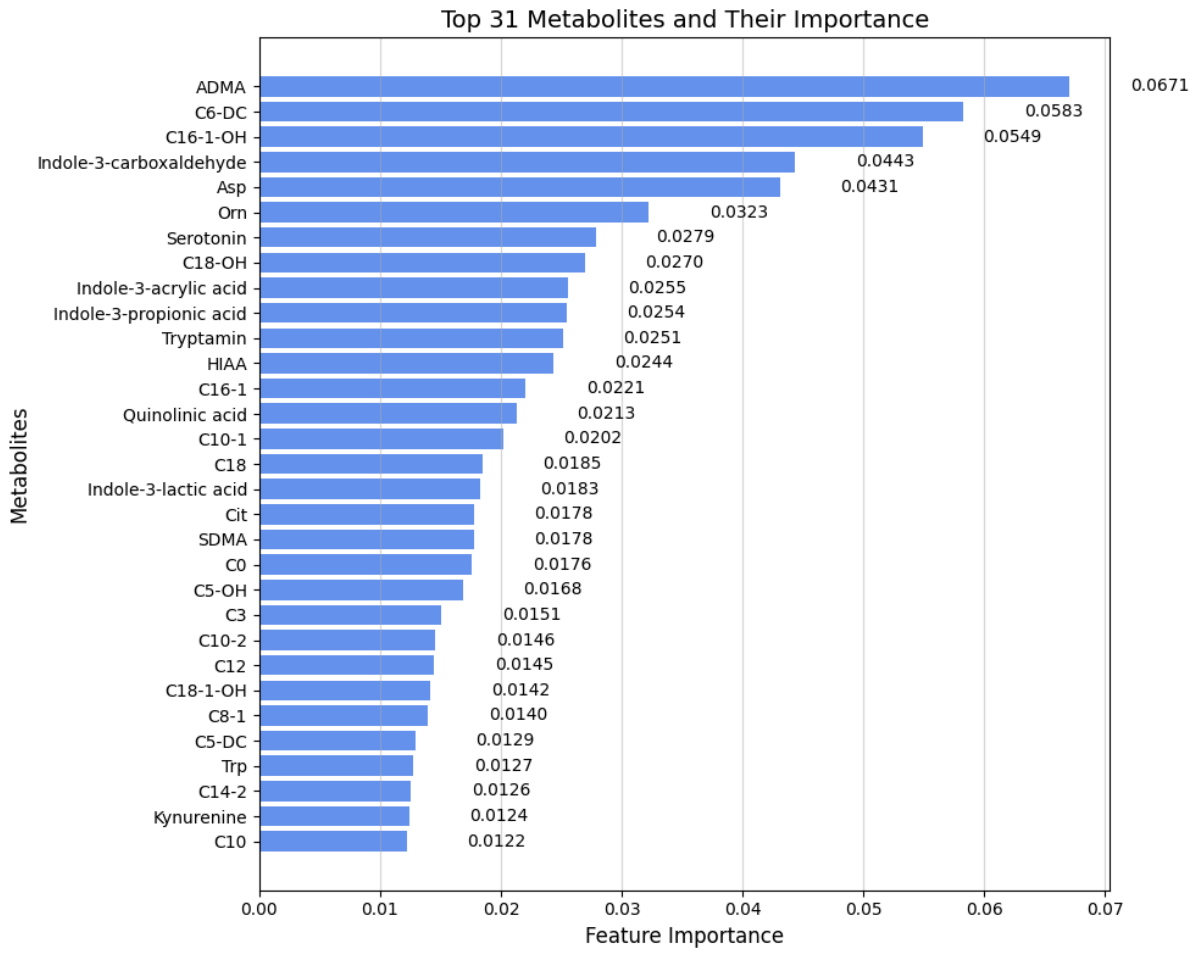


Figure 5: Top 31 selected metabolites and Their Importance

The decision to select the 31 metabolites based on the top-k with the highest validation score of 88.77% is driven by the performance of the model. While both the 21 metabolites above mean importance and the 31 metabolites above median importance provide relevant features, the 31 metabolites with the highest validation score (88.77%) offer the best model performance. By prioritizing validation score over feature importance, we ensure that the selected features contribute most effectively to the model's predictive accuracy. This approach enhances the robustness of the model by focusing on the subset of features that yield optimal results through cross-validation, ultimately improving model performance. Hence, the top 31 metabolites with highest validation score are ADMA, C6-DC, C16-1-OH, Indole-3-carboxaldehyde, Asp, Orn, Serotonin, C18-OH, Indole-3-acrylic acid, Indole-3-propionic acid, Tryptamin, HIAA, C16-1, Quinolinic acid, C10-1, C18, Indole-3-lactic acid, Cit, SDMA, C0, C5-OH, C3, C10-2, C12, C18-1-OH, C8-1, C5-DC, Trp, C14-2, Kynurenine and C10.

Recursive Feature Elimination (RFE)

In the RFE (Recursive Feature Elimination) method, we use Logistic Regression to evaluate the importance of features based on their coefficients, then loop through feature sets ranging from 1 to 63 features. Figure 6 showed the result of the highest validation score between 1 to 63 features is 97.38% validation score with 30 features.

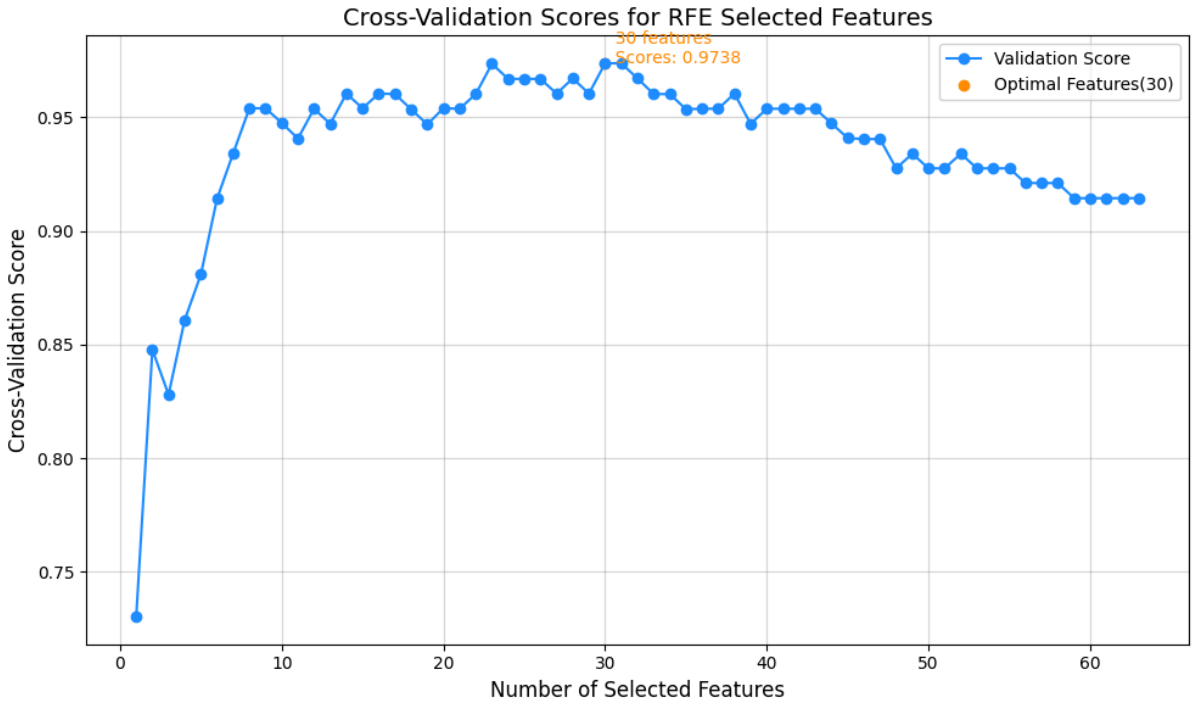


Figure 6: Cross-Validation Scores for RFE Selected Features

Therefore, the 30 metabolites with highest validation score are Quinolinic acid, Indole-3-carboxaldehyde, Indole-3-acrylic acid, Indole-3-propionic acid, Indole-3-butyric acid, Pro, Val, Asp, Arg, Cit, Thr, Lys, Trp, Tyr, C2, C4, C6, C5-OH, C5-DC, C8-1, C6-DC, C10-1, C12, C14-2, C16-1, C16, C18-1, C18, C18-OH and ADMA.

In order to identify the potential metabolites as biomarkers, we intersect the selected metabolites from the three methods. Figure 8 is a Venn Diagram showing the 46 union metabolites (all the colors included) and 16 intersect metabolites (grey color).

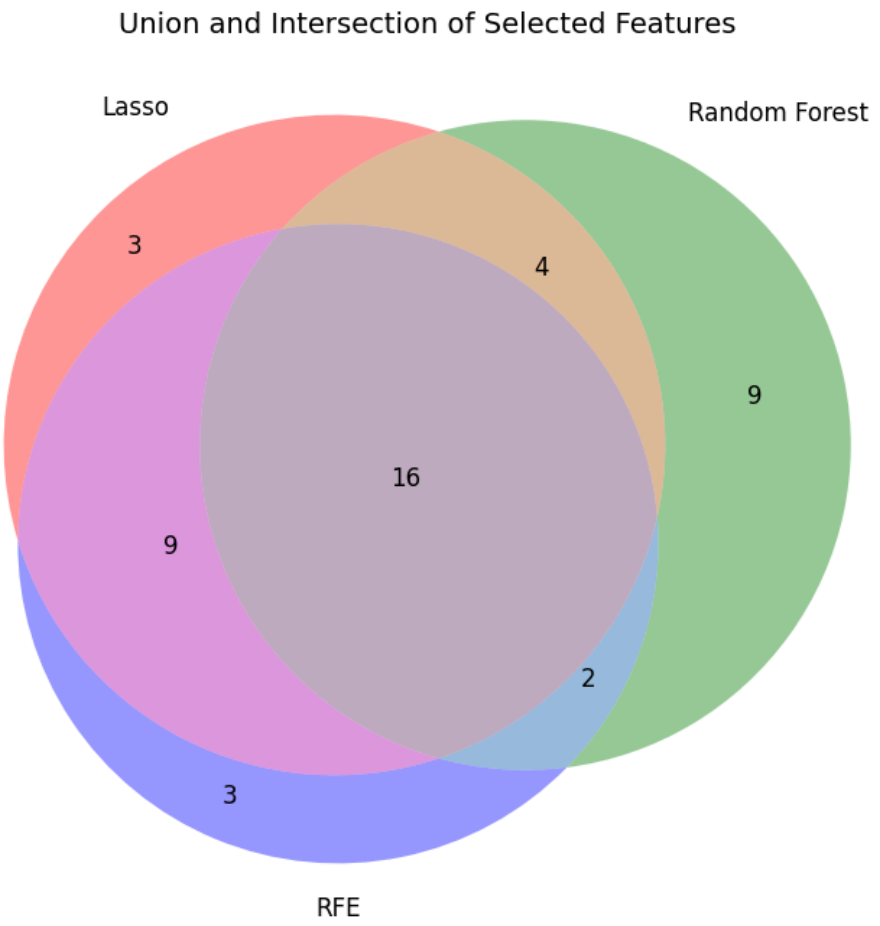


Figure 7: Union and intersection of Selected Metabolites

The 40 union metabolites identified across all three methods are Indole-3-lactic acid, C16, C0, C5-1, C8-1, HIAA, Arg, C10, Indole-3-butyric acid, ADMA, C16-1, Tryptamin, Serotonin, C6-DC, Trp, C10-2, C4, Tyr, SDMA, C10-1, Quinolinic acid, C6, C5-OH, C12, C18, C16-1-OH, Thr, Kynurenic acid, C18-1-OH, Cit, Orn, Kynurenine, C14-2, Indole-3-propionic acid, C2, C18-1, Pro, C5-DC, Val, C3, Indole-3-carboxaldehyde, Lys, Choline, Indole-3-acrylic acid, Asp and C18-OH. The 16 intersecting metabolites identified across all three methods are C6-DC, Indole-3-carboxaldehyde, Trp, C8-1, C18-OH, Quinolinic acid, C10-1, C14-2, Indole-3-propionic acid, ADMA, Indole-3-acrylic acid, C18, C5-DC, Cit, C16-1 and C5-OH. The 16 intersecting metabolites are considered potential biomarkers for detecting lung cancer due to their consistent selection.

We used 16 potential metabolic biomarker to build model using 6 machine learning algorithms which are Logistic Regression, Support Vector Machine (SVM), Random Forest

Classifier, K-Nearest Neighbors (KNN), Gradient Boosting Classifier and Naive Bayes Classifier and compare their accuracy in order to choose the best machine learning model.

Logistic Regression

The Logistic Regression model shows a mean accuracy of 0.9174.

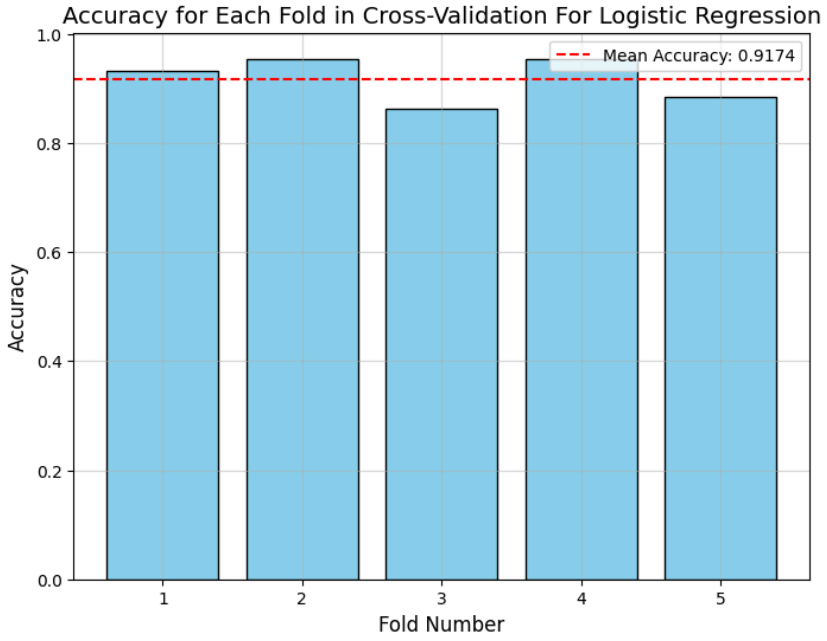


Figure 8: Accuracy for Each Fold in Cross-Validation for Logistic Regression

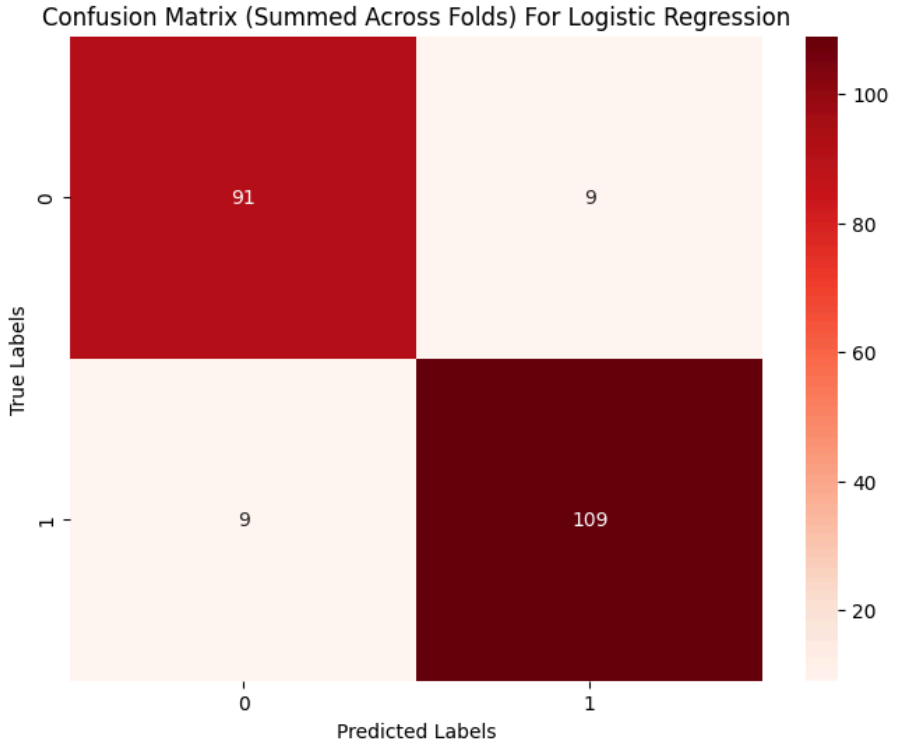


Figure 9: Confusion Matrix (Summed Across Folds) for Logistic Regression

Support Vector Machine (SVM)

The SVM model shows a mean accuracy of 0.9038.

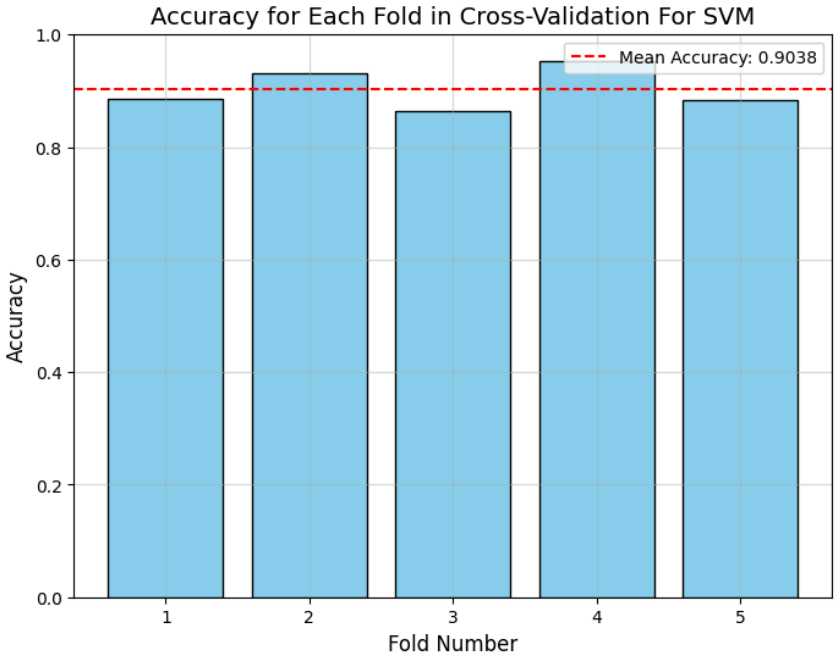


Figure 10: Accuracy for Each Fold in Cross-Validation for SVM

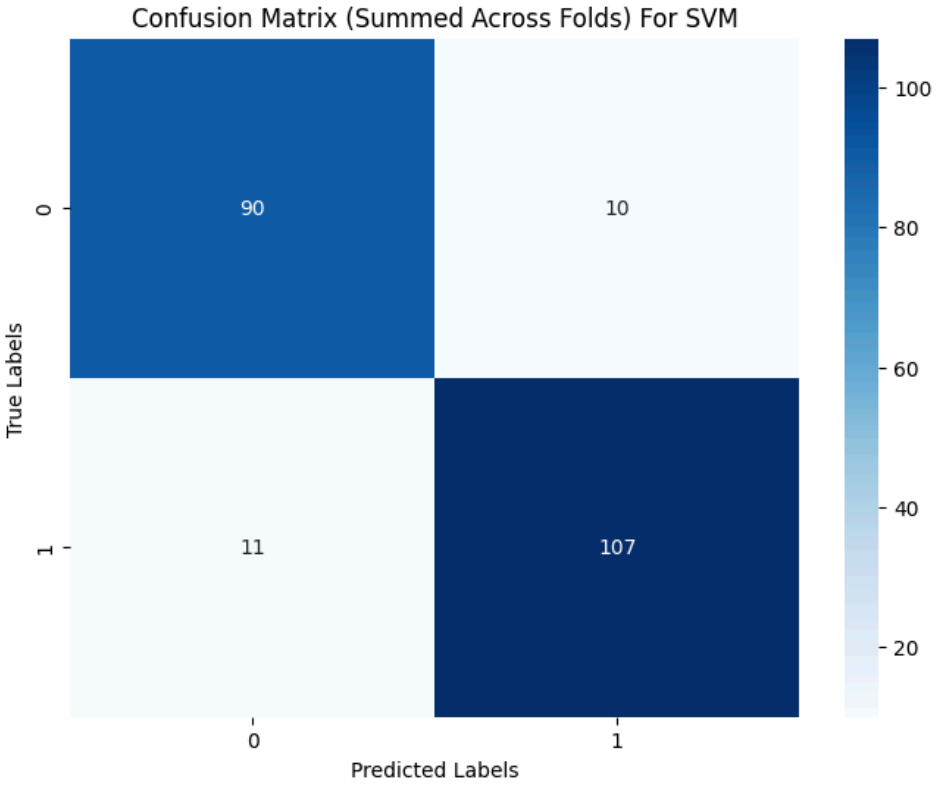


Figure 11: Confusion Matrix (Summed Across Folds) for SVM

Random Forest Classifier

The Random Forest model shows a mean accuracy of 0.8808 .

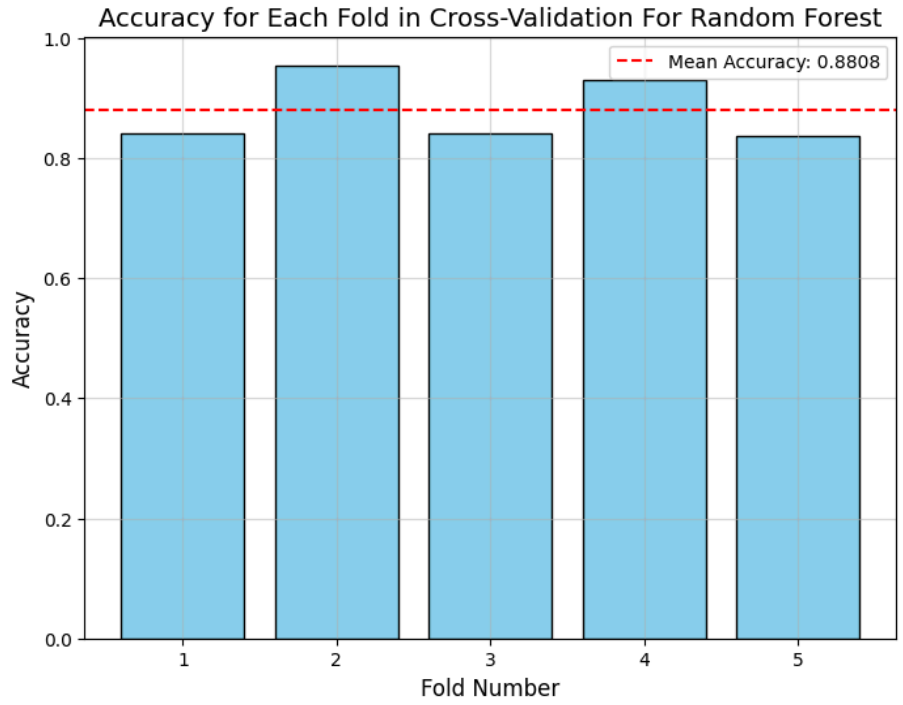


Figure 12: Accuracy for Each Fold in Cross-Validation for Random Forest

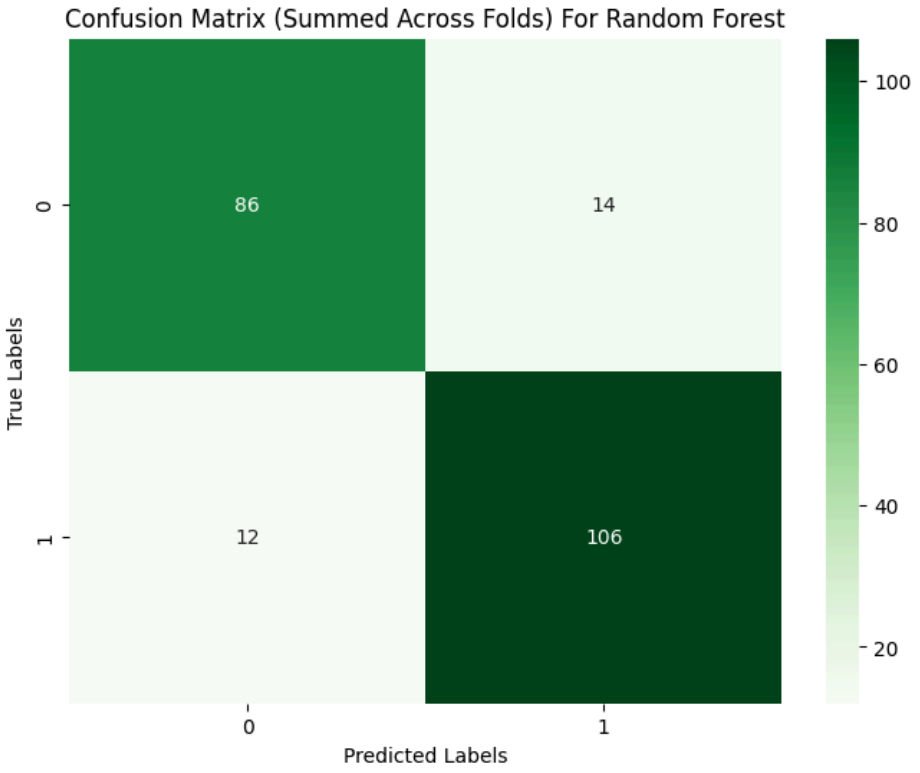


Figure 13: Confusion Matrix (Summed Across Folds) for Random Forest

K-Nearest Neighbors (KNN)

The KNN model also shows a mean accuracy of 0.8347.

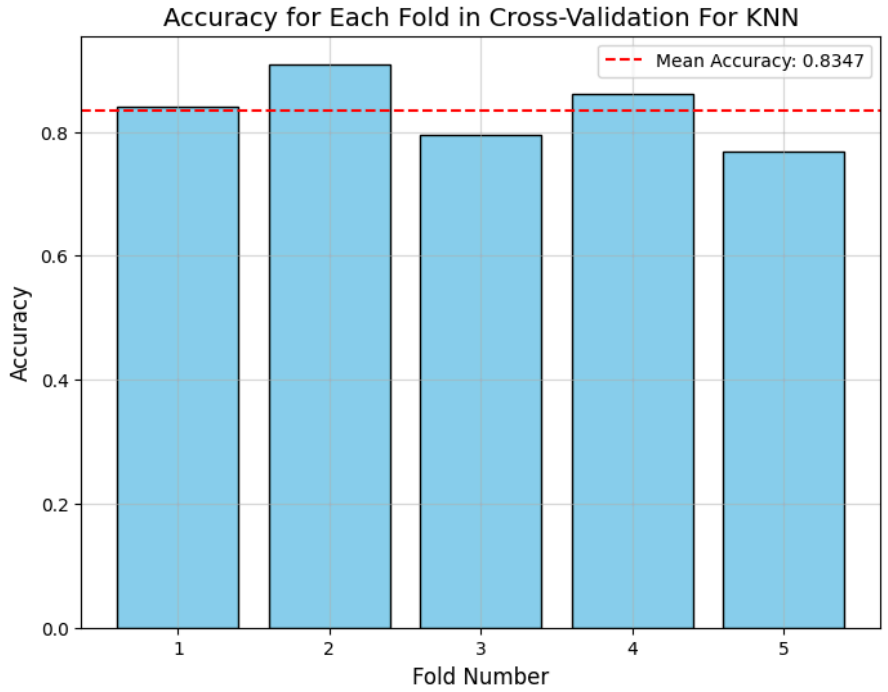


Figure 14: Accuracy for Each Fold in Cross-Validation for KNN

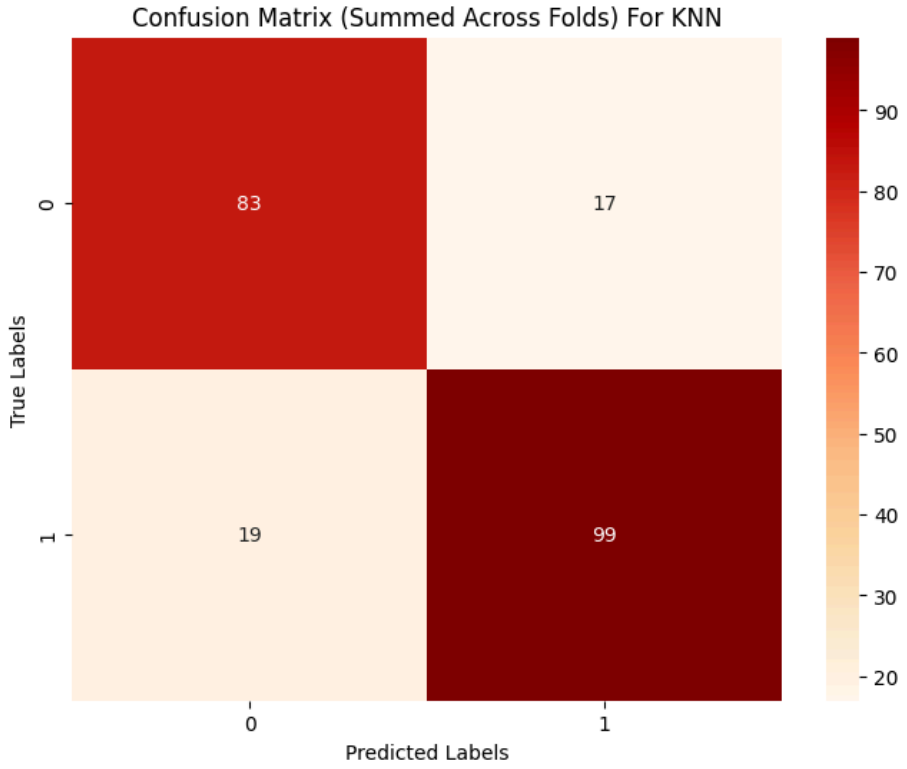


Figure 15: Confusion Matrix (Summed Across Folds) for KNN

Gradient Boosting Classifier

The Gradient Boosting model shows a mean accuracy of 0.8667.

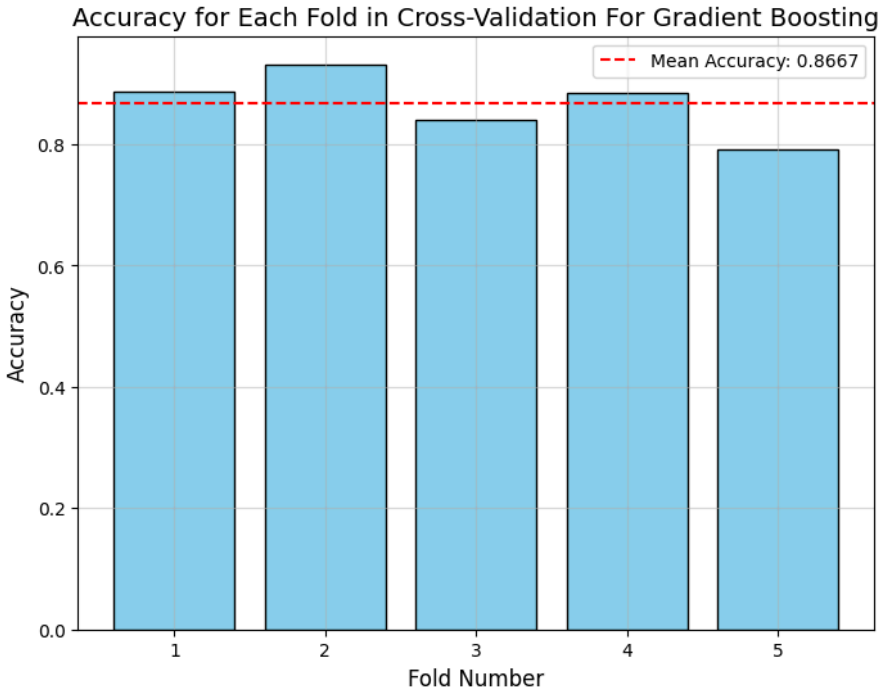


Figure 16: Accuracy for Each Fold in Cross-Validation for Gradient Boosting

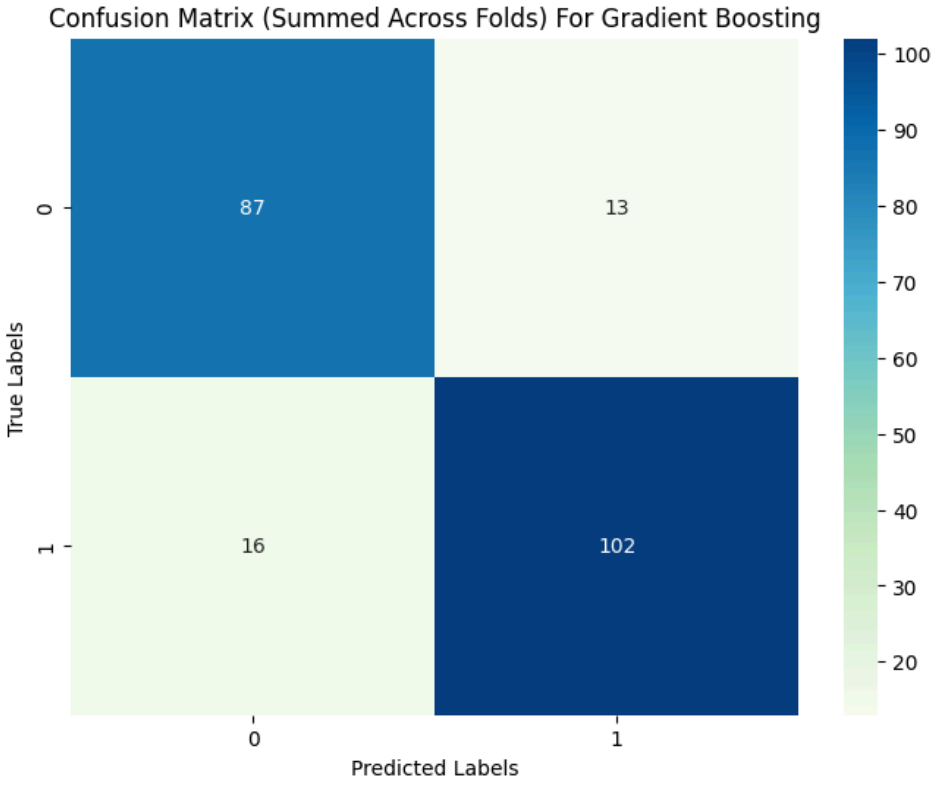


Figure 17: Confusion Matrix (Summed Across Folds) for Gradient Boosting

Naive Bayes Classifier

The Naive Bayes model shows a mean accuracy of 0.7253.

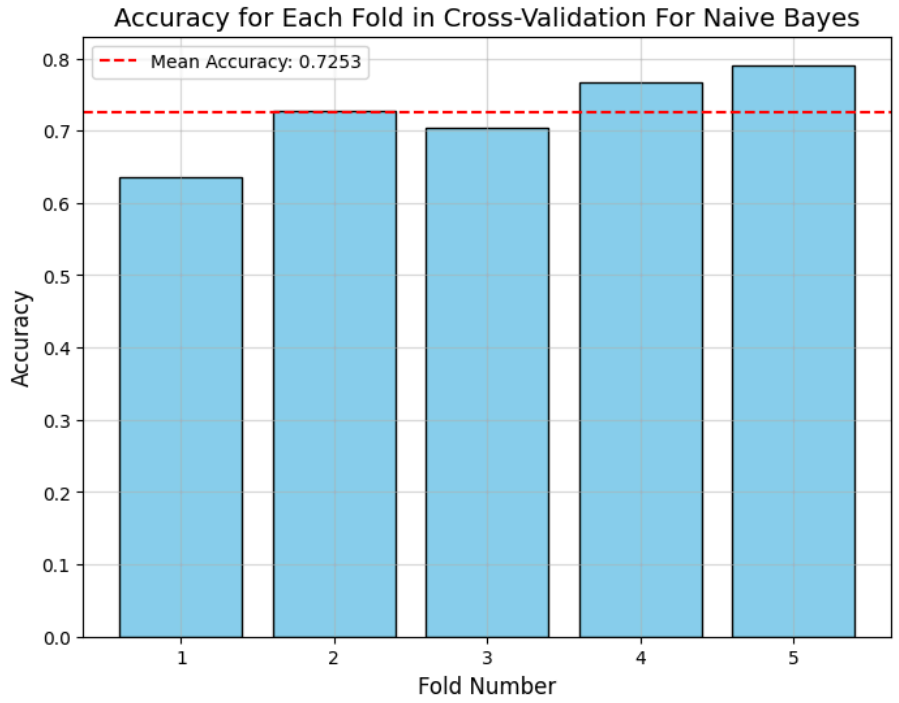


Figure 18: Accuracy for Each Fold in Cross-Validation for Naive Bayes

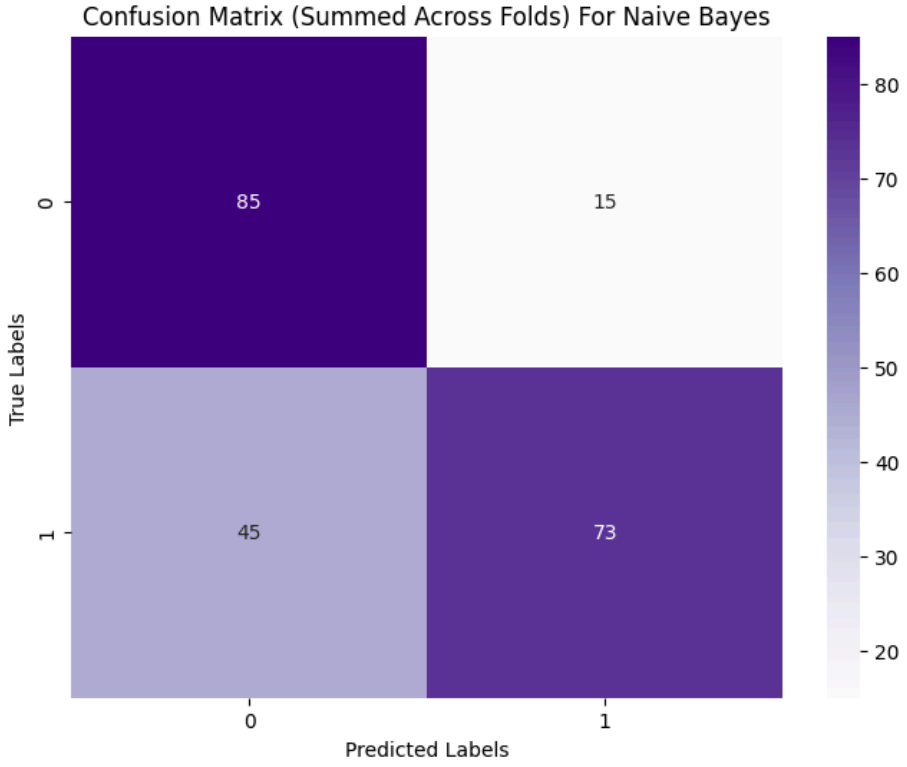


Figure 19: Confusion Matrix (Summed Across Folds) for Naive Bayes

In summary, the Logistic Regression model showed the highest mean accuracy. Table 2 shows the comparison of mean accuracy and confusion matrix between 6 machine learning models.

Algorithm	Mean Accuracy	Confusion Matrix (Summed Across Folds) (tp fp fn fn)	
Logistic Regression	0.9174	91	9
		9	109
Support Vector Machine	0.9038	90	10
		11	107
Random Forest	0.8808	86	14
		12	106
K-Nearest Neighbors	0.8347	83	17
		19	99
Gradient Boosting	0.8667	87	13
		16	102
Naive Bayes	0.7253	85	15
		45	73

Table 2: Comparison of mean accuracy and confusion matrix between between 6 models

6.1.2 Discussion

Importance of Metabolomics in Lung Cancer Research

Lung cancer remains a leading cause of cancer-related deaths worldwide, with its high mortality rates often attributed to late-stage diagnosis. The complexity and uncertainty surrounding its etiology and mechanisms compel researchers to adopt innovative approaches for deeper understanding. While traditional methods like histopathology and imaging play a vital role in diagnosis, they are often limited to detecting cancer in its later stages. This delay significantly impacts survival rates and treatment options.

Metabolomics, as a subset of OMICs technologies, offers a novel avenue for investigating the biochemical underpinnings of diseases, including lung cancer. By quantitatively profiling small molecules, metabolomics provides real-time snapshots of the biochemical processes occurring within the body. Unlike genomics or transcriptomics, which focus on static DNA or RNA information, metabolomics captures dynamic changes influenced by genetic, environmental, and lifestyle factors. This makes it particularly effective in understanding diseases characterized by metabolic dysregulation, such as cancer.

In this study, plasma metabolites were chosen as the primary focus due to their accessibility and ability to reflect systemic metabolic alterations. Plasma serves as a medium that integrates metabolic signals from different tissues, making it a rich source for identifying potential biomarkers. Tumor-associated changes in metabolite concentrations can reveal critical insights into cancer biology, including tumor growth, progression, and metastasis. These features highlight the potential of metabolomics to complement traditional diagnostic methods, providing a more comprehensive understanding of lung cancer's molecular landscape.

Role of Early Detection Biomarkers

Early detection biomarkers play an instrumental role in identifying individuals at risk or detecting cancer during its initial stages—often before clinical symptoms appear. Early-stage detection is crucial for improving survival rates, as treatment is more effective and less invasive when administered during these stages. This project specifically emphasizes the identification of biomarkers that can detect lung cancer at an early stage, aligning with the broader goal of reducing mortality through timely intervention.

The clinical significance of early detection is underscored by the limitations of current diagnostic modalities. Techniques such as low-dose computed tomography (LDCT) are effective but may lead to false positives, overdiagnosis, and increased healthcare costs. In contrast, metabolomics-based biomarkers offer a non-invasive, cost-effective, and potentially more accurate alternative for early diagnosis. By focusing on metabolites associated with early-stage cancer, this study aims to provide critical diagnostic information that can guide clinical decisions before advanced imaging or symptomatic manifestation. The proposed method leverages metabolomics data to compare the biochemical profiles of healthy individuals and patients with lung cancer.

Feature Selection

Metabolomics data often involves high-dimensional datasets with complex interactions between metabolites. For robust biomarker identification, multivariate approaches were employed over univariate methods. Unlike univariate techniques, which evaluate each feature independently, multivariate methods capture interactions and interdependencies, reflecting the inherent complexity of biological systems.

Feature selection was conducted using three different methods which are Lasso Regression, Random Forest Feature Importance, and Recursive Feature Elimination (RFE). Each method evaluates feature significance from a unique perspective.

Lasso Regression

Lasso Regression identified 32 metabolites with importance scores above a defined threshold. By penalizing the absolute size of regression coefficients, Lasso automatically shrinks less significant coefficients to zero, effectively performing feature selection. This method is particularly well-suited for high-dimensional data, such as metabolomics, where the number of features often exceeds the number of samples.

In addition to reducing model complexity, Lasso mitigates the risk of overfitting, ensuring that the selected features are both statistically significant and biologically meaningful. This is crucial for metabolomics datasets, which are characterized by high levels of noise and interdependence among features. The ability of Lasso to handle correlated variables further underscores its utility in this context, as many metabolites are part of interconnected biochemical pathways.

Random Forest Feature Importance

Random Forest identified 31 key metabolites, leveraging its ability to model non-linear relationships and interactions between features. Unlike Lasso, which is a linear method, Random Forest captures the complex, non-linear dependencies that often exist in biological data. By averaging the results of multiple decision trees, Random Forest provides stable and interpretable rankings of feature importance.

Feature importance scores were calculated using Gini Importance, which measures each feature's contribution to reducing uncertainty (impurity) in the model's predictions. Two thresholds were applied: the mean and median importance scores. Features with importance scores above the mean (0.0159) have 21 metabolites while features with importance scores above the median (0.0120) have 31 metabolites. However, no validation has been applied yet to confirm if this is the optimal subset. Therefore, we tried another method where cross-validation experiments which demonstrated that the top 31 features yielded the highest predictive accuracy (88.77%).

Recursive Feature Elimination (RFE)

RFE starts with all features and iteratively removes the least important ones based on the Logistic Regression model's coefficients. Then, cross-validation is used to evaluate the performance of the model with the selected features and identifies the optimal number of features based on the cross-validation scores. The results indicate that using the top 30 features produced the best performance (accuracy of 97.38%).

The iterative approach of RFE provides a thorough evaluation of feature subsets, allowing the identification of the most relevant features while avoiding overfitting. RFE's strength lies in its ability to refine the selection process systematically, resulting in a more accurate and generalizable model. It is particularly useful when the goal is to identify a minimal set of features that provide the highest predictive power.

Intersect Selected Metabolites as Potential Biomarkers

The identification of a consensus set of 16 metabolites through the intersection of Lasso Regression, Random Forest, and Recursive Feature Elimination (RFE) highlights a robust approach to biomarker discovery for lung cancer detection. These metabolites, validated through diverse analytical techniques, provide both statistical and biological significance,

ensuring their relevance in clinical applications. By combining complementary strengths of the feature selection methods, the study increases confidence in the selected features and emphasizes their role in metabolic, immune, and signaling dysregulation in lung cancer.

While research findings have provided detailed insights into some of these metabolites, not all 16 have been extensively studied. For the selected metabolites with existing research, such as **Tryptophan (Try)**, **Citrulline (Cit)**, and several selected metabolites, their roles in lung cancer progression and metabolic reprogramming have been highlighted. However, further exploration is required to fully understand the biological and clinical significance of the remaining metabolites. This represents a promising avenue for future studies to uncover additional information and validate the roles of all selected biomarkers.

Among the investigated metabolites, **Tryptophan (Try)** emerges as a critical player in lung cancer progression. As an essential amino acid, Tryptophan and its metabolic pathways regulate the tumor environment, immune suppression, and drug resistance, fostering cancer progression. Notably, therapeutic advances targeting the kynurenine pathway—an offshoot of Tryptophan metabolism—show promise in inhibiting lung cancer growth, with ongoing clinical trials exploring their potential. This highlights the metabolite's dual diagnostic and therapeutic importance.

Similarly, **Citrulline (Cit)** plays a significant role in enhancing lung cancer cell viability, proliferation, migration, and invasion. It facilitates tumor progression by stimulating glycolysis and activating the IL-6/STAT3 signaling pathway via binding to the RAB3C protein. Mouse model experiments further validate Citrulline’s role in tumor growth acceleration, underscoring its utility as a potential biomarker and therapeutic target.

Lipids and fatty acid metabolites also feature prominently in the identified biomarkers. For instance, **Palmitoleic Acid (C16:1)** and **Stearic Acid (C18)** are directly linked to cancer aggressiveness and metabolic reprogramming. Elevated levels of these fatty acids, often driven by overexpression of enzymes like Stearoyl-CoA desaturase 1 (SCD1), signify altered lipid metabolism in lung adenocarcinoma. These metabolic changes are integral to supporting rapid cancer cell growth and survival, making them critical indicators of cancer progression.

Several metabolites, including **Octenoylcarnitine (C8-1)**, **Hydroxyhexanoylcarnitine (C5-OH)**, and **Tetradecadienoylcarnitine (C14-2)**, emphasize the importance of β-oxidation in lung cancer. These metabolites reflect the heightened energy demands of cancer cells, with

alterations in their levels indicating mitochondrial dysfunction and disrupted fatty acid metabolism. Such metabolic reprogramming is a hallmark of cancer progression and serves as a foundation for potential therapeutic interventions.

Another critical biomarker, **Asymmetric Dimethylarginine (ADMA)**, highlights the intersection of metabolic and vascular dysregulation in non-small cell lung cancer (NSCLC). Elevated ADMA levels are associated with endothelial dysfunction and immune suppression, key features of the tumor microenvironment. This metabolite serves as a marker of systemic dysregulation, further cementing its role in cancer pathophysiology.

While the findings from this study provide significant insights into the roles of several key metabolites, the biological significance of some of the identified biomarkers remains underexplored. Future research will focus on characterizing these lesser-studied metabolites, further validating their roles, and uncovering their mechanisms in lung cancer progression. By building on the groundwork laid by this study, the aim is to create a comprehensive understanding of all selected metabolites, ultimately paving the way for precision medicine in lung cancer care.

Model Development

The 16 selected metabolites were used to construct six machine learning models using six different algorithms for classifying NSCLC and healthy samples. Logistic regression was chosen for its outstanding performance which scored a mean accuracy of 91.74%, demonstrating its effectiveness in distinguishing between lung cancer patients and healthy individuals. As a linear model, logistic regression offers interpretability, allowing for direct insights into the relationship between predictors and outcomes. This makes it particularly suitable for clinical applications, where transparency and simplicity are often prioritized. Additionally, the high accuracy of the model underscores the diagnostic value of the selected metabolites, providing a practical tool for early-stage lung cancer detection.

Implications for Early Detection and Personalized Medicine

The identification of these 16 potential biomarkers has significant implications for both clinical practice and research. Their integration into diagnostic pipelines could revolutionize early detection strategies, reducing the reliance on invasive procedures and imaging

techniques. Furthermore, these biomarkers provide a foundation for personalized medicine, where metabolic profiles could guide tailored treatment plans.

By incorporating metabolomics into lung cancer research, this study highlights the shift toward precision medicine. The use of robust feature selection techniques establishes a reproducible framework for biomarker discovery, paving the way for similar approaches in other diseases.

7.0 Conclusions, Limitations, Future Works

7.1 Conclusion

In conclusion, this study underscores the transformative potential of metabolomics in advancing lung cancer diagnostics, particularly in its early stages. The identification of 16 robust metabolites as biomarkers for non-small cell lung cancer (NSCLC) highlights the critical role metabolomics plays in bridging the gap between molecular insights and clinical applications. These metabolites provide a comprehensive biochemical fingerprint of lung cancer, offering a non-invasive, cost-effective, and accurate diagnostic alternative. By integrating data-driven feature selection techniques—Lasso Regression, Random Forest Feature Importance, and Recursive Feature Elimination (RFE)—this research ensures the reliability and robustness of the identified biomarkers. Each method contributed unique strengths, culminating in a consensus that strengthened the credibility of the results. Logistic regression emerged as the best-performing model, achieving an impressive mean accuracy of 91.74% in distinguishing between lung cancer patients and healthy individuals. The success of this model lies in its simplicity, interpretability, and ability to effectively utilize the selected metabolites for early-stage detection.

The findings align with the growing demand for precision medicine, where early intervention and personalized treatment strategies are becoming the standard in healthcare. Beyond diagnostic improvements, the study paves the way for developing targeted therapies, as these metabolites may reveal pathways critical to cancer progression. However, the study's implications extend beyond lung cancer; it sets a methodological precedent for biomarker discovery in other diseases characterized by metabolic dysregulation. While significant progress has been made, the study acknowledges its limitations and areas requiring further investigation. Expanding the sample size and diversity, integrating multi-omics data, and translating findings into practical diagnostic tools are essential steps toward clinical implementation. Ultimately, this research not only contributes to the scientific understanding of lung cancer but also highlights the broader promise of metabolomics in transforming healthcare, offering hope for improved outcomes through earlier and more accurate diagnosis.

7.2 Limitations

Despite its promising findings, this study is subject to several limitations that warrant consideration. The limited sample size is a primary constraint, as smaller datasets can introduce biases and reduce the generalizability of results. While the identified 16 metabolites demonstrated strong diagnostic potential within the analyzed cohort, their performance across diverse populations remains uncertain. Variability in lifestyle, diet, and genetic backgrounds may influence metabolite levels, potentially affecting the reproducibility of the findings in broader settings. Another limitation lies in the reliance on logistic regression as the primary classification model. Although it achieved high accuracy and interpretability, it may not capture complex non-linear relationships inherent in metabolomics data. Advanced machine learning techniques such as ensemble methods or neural networks could potentially enhance predictive performance by uncovering deeper patterns. Additionally, this study focused exclusively on plasma metabolites, which, while accessible and systemic, may not fully capture the localized metabolic changes within lung tissues.

Furthermore, the study does not explore the mechanistic roles of the identified metabolites in lung cancer pathophysiology. Understanding these mechanisms could provide deeper insights into disease progression and inform therapeutic strategies. Finally, while the feature selection methods employed were robust, combining Lasso, Random Forest, and RFE may introduce biases inherent to each technique. Independent validation of these biomarkers using external datasets is crucial to confirm their clinical utility. Addressing these limitations through expanded datasets, advanced modeling approaches, and mechanistic studies will strengthen the findings and facilitate their translation into clinical applications.

7.3 Future Works

Building on the promising results of this study, several avenues for future research are proposed to enhance its clinical relevance and impact. First, independent validation using larger and more diverse cohorts is critical to establish the generalizability and robustness of the identified biomarkers. Such validation should encompass individuals from various geographic regions and demographic backgrounds, capturing the influence of genetic, environmental, and lifestyle factors on metabolite profiles. This step will not only confirm the reliability of the biomarkers but also refine their predictive accuracy in real-world settings.

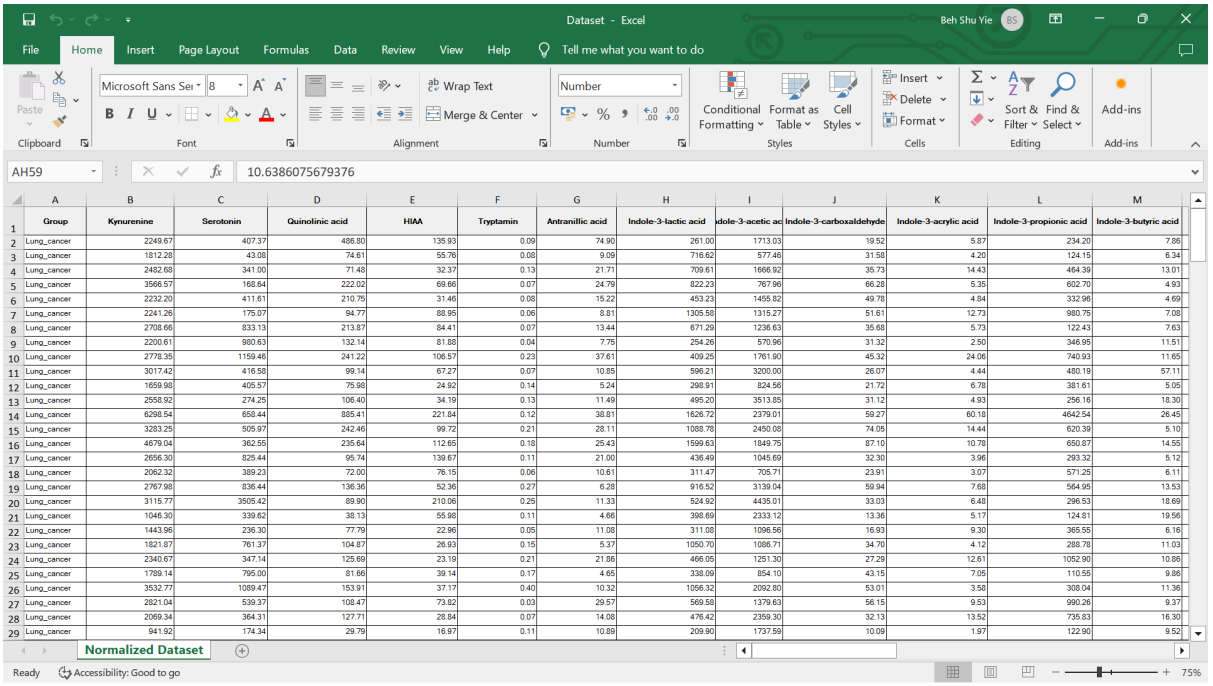
Second, integrating multi-omics data—combining metabolomics with genomics, proteomics, or transcriptomics—could provide a holistic understanding of lung cancer's molecular underpinnings. This integrative approach may uncover additional biomarkers and pathways critical to tumor development, offering new targets for therapeutic intervention. Employing advanced analytical techniques, such as deep learning, could further enhance the interpretation of these high-dimensional datasets, uncovering complex patterns that traditional methods may overlook.

Third, translating the findings into clinical practice requires the development of user-friendly diagnostic tools. Portable metabolomics-based platforms, such as point-of-care devices, could revolutionize lung cancer screening by providing rapid, non-invasive, and cost-effective diagnostics. Collaborations with clinicians, engineers, and industry stakeholders will be essential to design, test, and deploy these tools. Finally, exploring the functional roles of the identified metabolites in lung cancer progression could offer novel insights into tumor biology. Mechanistic studies investigating these metabolites as potential therapeutic targets could complement their diagnostic utility, contributing to both early detection and improved treatment strategies. By addressing these areas, future research can bridge the gap between discovery and application, maximizing the potential of metabolomics to transform lung cancer care and improve patient outcomes.

8.0 Appendix

Dr Chan Weng Howe is our supervisor in this project.

Dataset (Screenshot)



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Group	Kynurenine	Serotonin	Quinolinic acid	HIAA	Tryptamin	Antranillic acid	Indole-3-lactic acid	dole-3-acetic ac	Indole-3-carboxaldehyde	Indole-3-acrylic acid	Indole-3-propionic acid	Indole-3-butyric acid
2	Lung_cancer	2249.67	407.37	486.80	135.93	0.09	74.90	261.00	1713.03	19.52	5.87	234.20	7.86
3	Lung_cancer	1812.28	43.08	74.61	55.76	0.08	9.09	716.62	577.46	31.58	4.20	124.15	6.34
4	Lung_cancer	2482.68	341.00	71.48	32.37	0.13	21.71	709.61	1666.92	35.73	14.43	464.39	13.01
5	Lung_cancer	3566.57	168.64	222.02	69.66	0.07	24.79	822.23	767.96	66.28	5.35	602.70	4.93
6	Lung_cancer	2232.20	411.61	210.75	31.46	0.08	15.22	453.23	1455.82	49.78	4.84	332.96	4.69
7	Lung_cancer	2241.26	175.07	94.77	88.95	0.06	8.81	1305.58	1315.27	51.61	12.73	980.75	7.08
8	Lung_cancer	2708.66	833.13	213.87	84.41	0.07	13.44	671.29	1236.63	35.68	5.73	122.43	7.63
9	Lung_cancer	2200.61	980.63	132.14	81.88	0.04	7.75	254.26	570.96	31.32	2.50	346.95	11.51
10	Lung_cancer	2778.35	1159.46	241.22	106.57	0.23	37.61	409.25	1761.90	45.32	24.06	740.93	11.65
11	Lung_cancer	3017.42	416.58	99.14	67.27	0.07	10.85	596.21	3200.00	26.07	4.44	480.19	57.11
12	Lung_cancer	1659.98	405.57	75.98	24.92	0.14	5.24	298.91	824.56	21.72	6.78	381.61	5.05
13	Lung_cancer	2558.92	274.25	106.40	34.19	0.13	11.49	495.20	3513.85	31.12	4.93	256.16	18.30
14	Lung_cancer	6298.54	658.44	885.41	221.84	0.12	38.81	1626.72	2379.01	59.27	60.18	4642.54	26.45
15	Lung_cancer	3283.25	505.97	242.46	99.72	0.21	28.11	1088.78	2450.08	74.05	14.44	620.39	5.10
16	Lung_cancer	4679.04	362.55	235.64	112.65	0.18	25.43	1599.63	1849.75	87.10	10.78	650.87	14.55
17	Lung_cancer	2656.30	825.44	95.74	139.67	0.11	21.00	436.49	1045.69	32.30	3.96	293.32	5.12
18	Lung_cancer	2062.32	389.23	72.00	76.15	0.06	10.61	311.47	705.71	23.91	3.07	571.25	6.11
19	Lung_cancer	2767.98	836.44	136.36	52.36	0.27	6.28	916.52	3139.04	59.94	7.68	564.95	13.53
20	Lung_cancer	3115.77	3505.42	89.90	210.06	0.25	11.33	524.92	4435.01	33.03	6.48	296.53	18.69
21	Lung_cancer	1046.30	339.62	38.13	55.98	0.11	4.66	398.69	2333.12	13.36	5.17	124.81	19.56
22	Lung_cancer	1443.96	236.30	77.79	22.96	0.05	11.08	311.08	1096.56	16.93	9.30	365.55	6.16
23	Lung_cancer	1821.87	761.37	104.87	26.93	0.15	5.37	1050.70	1086.71	34.70	4.12	288.78	11.03
24	Lung_cancer	2340.67	347.14	125.69	23.19	0.21	21.86	466.05	1251.30	27.29	12.61	1052.90	10.86
25	Lung_cancer	1789.14	795.00	81.66	39.14	0.17	4.65	338.09	854.10	43.15	7.05	110.55	9.86
26	Lung_cancer	3532.77	1089.47	153.91	37.17	0.40	10.32	1056.32	2092.80	53.01	3.58	308.04	11.36
27	Lung_cancer	2821.04	539.37	108.47	73.82	0.03	29.57	569.58	1379.63	56.15	9.53	990.26	9.37
28	Lung_cancer	2069.34	364.31	127.71	28.84	0.07	14.08	476.42	2359.30	32.13	13.52	735.83	16.30
29	Lung_cancer	941.92	174.34	29.79	16.97	0.11	10.89	209.90	1737.59	10.09	1.97	122.90	9.52

8.1 References

Ksenia M. Shestakova, Natalia E. Moskaleva, Andrey A. Boldin, Pavel M. Rezvanov, Alexandr V. Shestopalov, Sergey A. Rumyantsev, Elena Yu. Zlatnik, Inna A. Novikova, Alexander B. Sagakyants, Sofya V. Timofeeva, Yuriy Simonov, Sabina N. Baskhanova, Elena Tobolkina, Serge Rudaz & Svetlana A. Appolonova. (2023, 8 July). Targeted metabolomic profiling as a tool for diagnostics of patients with non-small-cell lung cancer. https://www.nature.com/articles/s41598-023-38140-7#Sec14

Shi-ang Qi, Qian Wu, Zhenpu Chen, Wei Zhang, Yongchun Zhou, Kaining Mao, Jia Li, Yuanyuan Li, Jie Chen, Youguang Huang & Yunchao Huang. (2021, 3 June). High-resolution metabolomic biomarkers for lung cancer diagnosis and prognosis. https://www.nature.com/articles/s41598-021-91276-2

Inman, S. (2024, September 9). Deep learning models expedite biomarker discovery, detection in lung cancer. Cancer Network. https://www.cancernetwork.com/view/deep-learning-models-expedite-biomarker-discovery-detection-in-lung-cancer

Ying Xie, Wei-Yu Meng, Run-Ze Li, Yu-Wei Wang, Xin Qian, Chang Chan, Zhi-Fang Yu, Xing-Xing Fan, Hu-Dan Pan, Chun Xie, Qi-Biao Wu, Pei-Yu Yan, Liang Liu, Yi-Jun Tang, Xiao-Jun Yao, Mei-Fang Wang, Elaine Lai-Han Leung. (January 2021). Early Lung Cancer Diagnostic Biomarker Discovery By Machine Learning Methods. https://www.sciencedirect.com/science/article/pii/S1936523320303995

Pathmanathan Rajadurai, Soon Hin How, Chong Kin Liam, Anand Sachithanandan, Sing Yang Soon, Lye Mun Tho. (March 2020). Lung Cancer in Malaysia. https://www.jto.org/article/S1556-0864(19)33639-1/fulltext

Fangwei Wang, Qisheng Su & Chaoqian Li. (2022, 6 October). Identification of novel biomarkers in non-small cell lung cancer using machine learning. https://www.nature.com/articles/s41598-022-21050-5#data-availability

Wang CD, Shao J, Song LJ, Ren PW, Liu D, Li WM. (2023, June 30). Persistent increase and improved survival of stage I lung cancer based on a large-scale real-world sample of 26,226 cases. https://pmc.ncbi.nlm.nih.gov/articles/PMC10431578/

Robert Tibshirani. (1996). Regression Shrinkage and Selection via the Lasso. https://www.jstor.org/stable/2346178

Andy Liaw, Matthew Wiener. (Nov 2001). Classification and Regression by RandomForest. https://www.researchgate.net/publication/228451484_Classification_and_Regression_by_RandomForest

ChenweiLi, Hui Zhao. (2021, Aug 6). Tryptophan and Its Metabolites in Lung Cancer: Basic Functions and Clinical Significance. https://www.frontiersin.org/journals/oncology/articles/10.3389/fonc.2021.707277/full

Qingjun Meng, Yanguang Li, Zhen Sun, Junfeng Liu. (2024, May 21) Citrulline facilitates the glycolysis, proliferation, and metastasis of lung cancer cells by regulating RAB3C. https://onlinelibrary.wiley.com/doi/abs/10.1002/tox.24326

T. Sajic, S. Arni, R. Aebersold, I. Schmitt-Opitz, S. Hillinger. (Nov 2023). Detection of Possible Targets in Aggressive Lung Cancer Using Molecular Abundances and Catalytic Action of Serine Hydrolases https://www.jto.org/article/S1556-0864(23)01618-0/fulltext

Kristof Fellegi. (2018, Dec 12). Feature selection for biomarker discovery. https://studenttheses.uu.nl/bitstream/handle/20.500.12932/31758/Master_Thesis_Kristof_Fellegi_Reviewed.pdf?sequence=2&utm_source=chatgpt.com

Analytic Labs. (2023, 25 Aug). What are Lasso and Ridge Techniques? https://www.analytixlabs.co.in/blog/lasso-and-ridge-regression/