Project 5

Kaining Zhao

October 2024

Abstract

Alzheimer's disease (AD), a principal cause of dementia, lacks definitive early diagnostic markers, complicating effective treatment strategies. This study harnesses the potential of metabolomics to uncover biomarkers that can differentiate between Alzheimer's disease and non-diseased states. Utilizing a dataset comprising metabolic profiles from 127 patients, the study employed t-tests to identify significant metabolites and performed Principal Component Analysis (PCA) to refine data handling and improve analytical accuracy, revealing that the first 20 components captured the majority of the variance within the dataset, with PC1 being particularly influential in the classification models. Machine learning models, including Random Forest and Support Vector Machines (SVM), were then applied to classify the samples, with performance validated through confusion matrices. The results demonstrated that the Random Forest performed better with an accuracy of 78%, and five metabolites have emerged as likely important biomarkers driving the observed group differences. Overall, the study aims to harness advanced metabolomic profiling and machine learning techniques to identify and validate potential biomarkers for Alzheimer's disease.

Introduction

Alzheimer's disease (AD) is a progressive neurodegenerative disorder that primarily affects the elderly population, leading to significant cognitive decline and memory loss. It is the most common cause of dementia, accounting for an estimated 60-70% of cases. Despite extensive research, the pathophysiological mechanisms of Alzheimer's disease remain not fully understood, and early diagnosis continues to be a challenge in clinical settings.

Metabolomics, the comprehensive analysis of small molecules in biological systems, has emerged as a potent tool for uncovering novel insights into the biochemical changes associated with Alzheimer's disease. By profiling the metabolic alterations that occur as the disease progresses, metabolomics offers the potential to identify biomarkers that can signal the onset and advancement of AD before clinical symptoms become apparent.

The search for reliable biomarkers for early detection and monitoring of Alzheimer's disease progression is critical. Biomarkers can provide a measurable indicator of the biological state or condition and are pivotal in diagnosing diseases before the manifestation of clear clinical symptoms. In this context, metabolomics—the comprehensive study of small molecules in biological systems—emerges as a powerful approach. Metabolomics offers a unique opportunity to observe global changes in metabolic profiles that correlate with disease states. These metabolic alterations, when mapped to specific biochemical pathways, can explain the disease mechanisms and potentially guide the development of targeted therapies.

This project aims to utilize metabolomics data to discover potential biomarkers for Alzheimer's disease by employing sophisticated machine learning techniques and statistical analyses to distinguish between normal and disease states. Through this approach, I hope to identify key metabolites that change in concentration in relation to the progression of Alzheimer's disease and validate their efficacy as biomarkers for early diagnosis and progression monitoring. The study initially preprocessed and merged metabolomics data with sample metadata, and conducted t-tests to identify significant metabolites, and then performed Principal Component Analysis (PCA) to reduce dimensionality and improve model efficiency. Following this, the study applied machine learning techniques, including Random Forest and SVM, to classify samples into Alzheimer's disease and control groups.

Methodology

Feature Selection

In this study, feature selection was conducted in two key steps to ensure robust identification of relevant biomarkers for classification. First, a differential analysis was performed using two-sample **t-tests** for each feature (metabolite) to evaluate its significance between groups (e.g., NORMAL vs. MCI/AD). For each feature x_i, the null hypothesis (H_0) assumes that the mean values between the two groups are equal:

$$H_0 : \mu_{\text{Group1}} = \mu_{\text{Group2}}, \quad H_a : \mu_{\text{Group1}} \neq \mu_{\text{Group2}}$$

The test statistic for the i-th feature is calculated as:

$$t_i = \frac{\bar{x}_{\text{Group1}} - \bar{x}_{\text{Group2}}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where $\bar{x}_{\text{Group1}}$ and $\bar{x}_{\text{Group2}}$ are the sample means, s_1^2 and s_2^2 are the variances, and n_1 and n_2 are the sample sizes for Group 1 and Group 2, respectively. Features with p-values below 0.05 after False Discovery Rate (FDR) correction were considered significant.

Subsequently, significant features identified through differential analysis were further reduced using Principal Component Analysis (PCA) to capture major sources of variance while reducing dimensionality. PCA transforms the original feature set into a new set of uncorrelated variables called principal components (PCs). The transformation is defined as:

$$Z_k = \sum_{i=1}^{p} w_{ki} x_i, \quad k = 1, 2, \ldots, K$$

where Z_k is the k-th principal component, w_{ki} are the loadings, and x_i are the original features. The number of retained PCs was determined based on the proportion of explained variance (>90%).

Classification Model

Two machine learning models, Random Forest (RF) and Support Vector Machine (SVM), were employed for classification of samples using the features identified in the previous step. Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the class with the highest vote across all trees. For a feature vector $\mathbf{x}$, the class prediction $\hat{y}$ is determined as:

$$\hat{y} = \text{mode}\{T_1(\mathbf{x}), T_2(\mathbf{x}), \ldots, T_M(\mathbf{x})\}$$

where $T_1, T_2, \ldots, T_M$ are individual decision trees, and M is the total number of trees. The importance of features was also evaluated based on the mean decrease in Gini impurity.

In addition to Random Forest, a Support Vector Machine was trained to classify the samples by finding the optimal hyperplane that maximizes the margin between classes. For a linear SVM, the decision boundary is defined as:

$$f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b$$

where $\mathbf{w}$ is the weight vector, $\mathbf{x}$ is the feature vector, and b is the bias term. The optimization problem is formulated as:

$$\min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|^2 \quad \text{subject to} \quad y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1, \ \forall i$$

where y_i is the class label for sample i. A radial basis function (RBF) kernel was used to extend the SVM to non-linear decision boundaries.

Performance of the classifiers was evaluated using accuracy, precision, recall, and the area under the receiver operating characteristic curve (AUC). Cross-validation was applied to avoid overfitting and ensure the generalizability of the models.

Performance Assessment

The classification models were evaluated using confusion matrices to determine the accuracy of predictions across different classes. Through SVM analysis, I identified several principal components (PCs) that significantly impact sample differentiation. From these PCs, I further identified top features that most influence each principal component. Building on this, I retrained the SVM model using these selected PCs and compared the performance of linear and non-linear kernels.

Results

Data Visualization

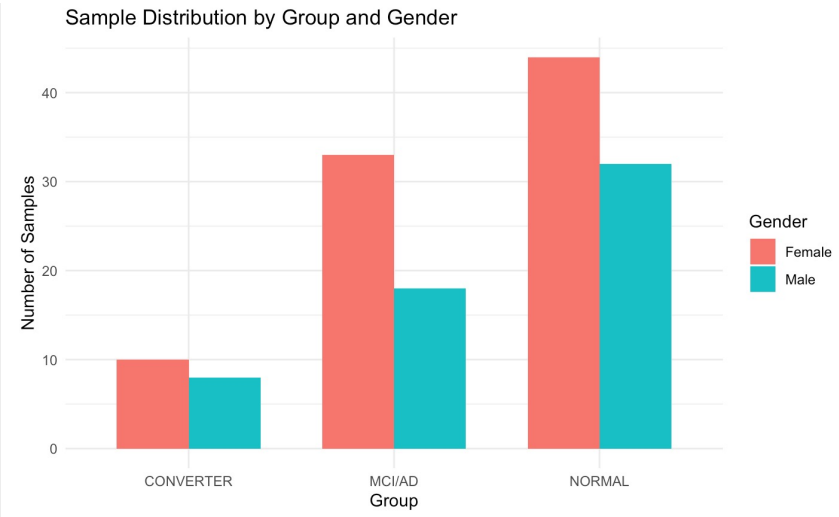

Figure 1: Sample Distribution by Group and Gender

Figure 1 illustrates the distribution of samples across different groups (CONVERTER, MCI/AD, NORMAL) stratified by gender (Male and Female). It is evident that the NORMAL group contains the largest number of samples, with a slightly higher proportion of females compared to males. Similarly, in the MCI/AD group, the number of female samples exceeds that of males. The CONVERTER group has the smallest number of samples overall, but a higher representation of females is still observed. This indicates an overall imbalance in the gender distribution, with a predominance of female samples across all groups.

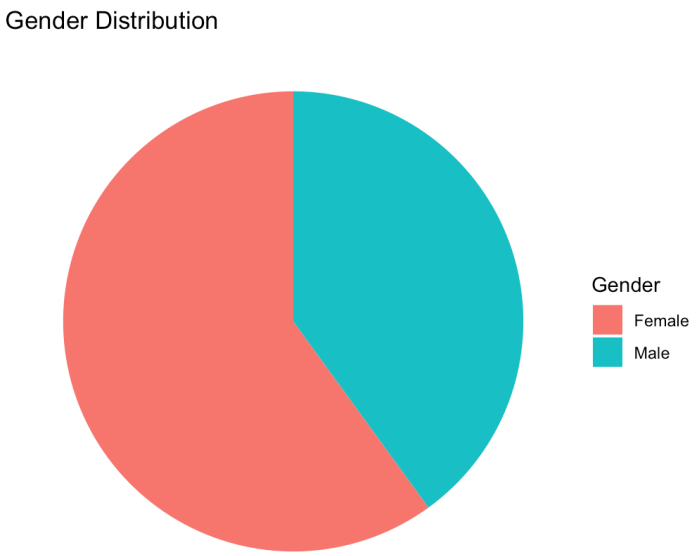

Figure 2: Gender Distribution

Figure 2 presents the overall gender distribution among all samples. The pie chart shows that female samples constitute approximately two-thirds of the total dataset, significantly outnumbering male samples. This disparity in gender distribution should be taken into consideration during the analysis, as it may introduce potential biases. Adjustments for gender as a confounding variable may be necessary in subsequent analyses to ensure robust results.

Differential Expression Analysis Results

The volcano plot illustrates the differential expression of metabolites between two groups based on their statistical significance and magnitude of change. The x-axis represents the Log_2 Fold Change, indicating the magnitude of upregulation (positive values) or downregulation (negative values), while the y-axis shows the $-\log_{10}$(p-value), reflecting the statistical significance of each metabolite.

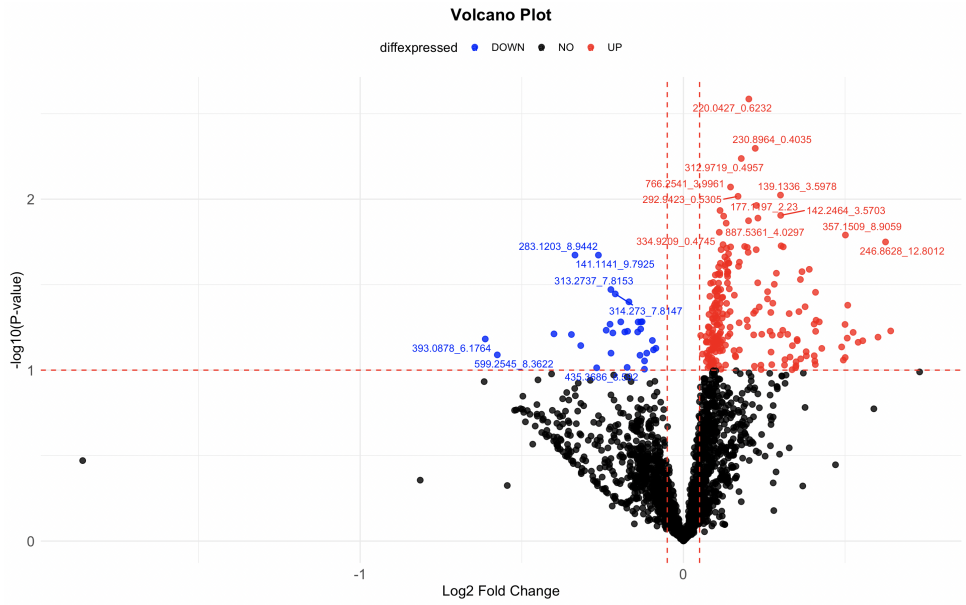

Figure 3: Volcano Plot

Metabolites with significant upregulation (p-value < 0.1 and Log_2 Fold Change > 0.05) are marked in red, whereas significantly downregulated metabolites are displayed in blue. Black points indicate metabolites that do not meet the significance thresholds. The red dashed lines highlight the thresholds for statistical significance (p-value < 0.05) and fold change, providing clear separation between significantly altered and unaltered metabolites. In total, 251 metabolites were identified as significantly different between the groups.

PCA Results

Principal Component Analysis (PCA) was performed to reduce the dimensionality of the dataset and identify major sources of variance across the metabolites. Figure X illustrates the cumulative variance explained by the principal components (PCs). The first few principal components capture the majority of the variability in the dataset. Specifically, the first 10 PCs account for approximately 80 of the total variance, while the first 20 PCs capture over 90. Beyond 20 PCs, additional components contribute minimally to the explained variance, indicating diminishing returns for including further components in the analysis.

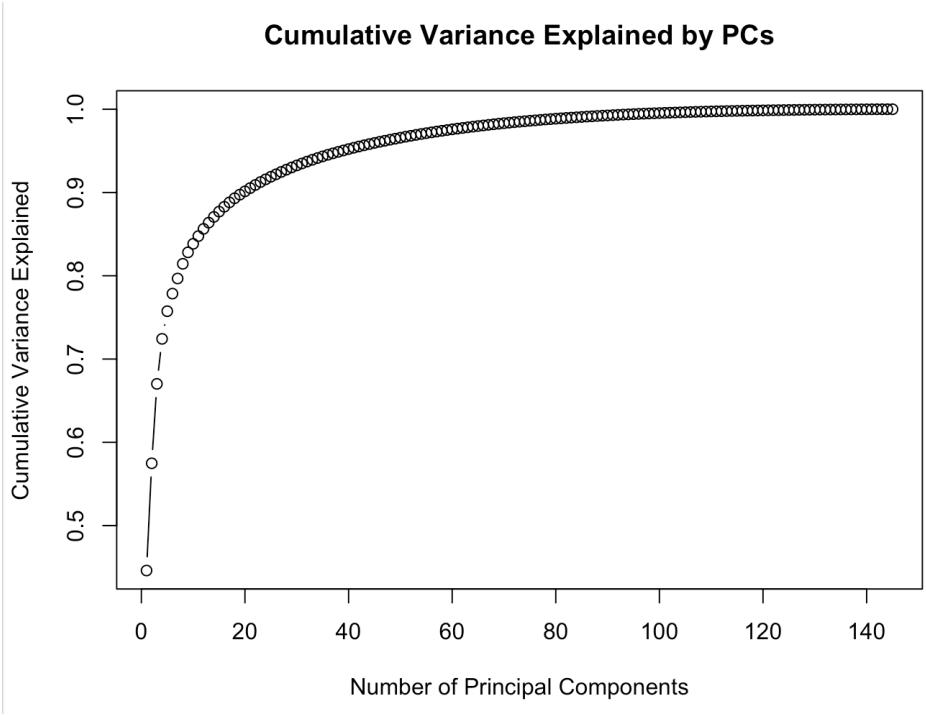

Figure 4: Cumulative Variance Explained by PCs

Based on this plot, a cutoff of 20 PCs was selected for subsequent analyses, as it balances dimensionality reduction with the retention of significant biological information. These selected PCs are likely to contain the most biologically relevant patterns in the data, minimizing noise and redundancy. This step ensures that the downstream modeling and interpretation are both efficient and meaningful.

Classification Model Performance

Table 1: Performance Metrics of Classification Models

Metric	**Random Forest (RF)**	**Support Vector Machine (SVM)**
Accuracy (%)	78.57 (95% CI: 59.05–91.70)	64.29 (95% CI: 44.07–81.36)
Balanced Accuracy (%)	74.44	54.44
Sensitivity (%)	88.89	88.89
Specificity (%)	60.00	20.00
Positive Predictive Value (%)	80.00	66.67
Negative Predictive Value (%)	75.00	50.00
Kappa Statistic	0.5116	0.1026
McNemar's Test p-value	0.6831	0.1138

The classification performance of the Random Forest (RF) and Support Vector Machine (SVM) models is summarized in Table 1.

The Random Forest model achieved an overall accuracy of 78.57 (95% CI: 59.05%–91.70%), with a balanced accuracy of 74.44. The sensitivity (true positive rate) was 88.89, and the specificity (true

negative rate) was 60.00, indicating a strong ability to classify both positive and negative classes. The Kappa statistic of 0.5116 suggests moderate agreement between predicted and true labels.

In contrast, the SVM model achieved an accuracy of 64.29 (95% CI: 44.07%–81.36%) and a balanced accuracy of 54.44. While maintaining a sensitivity of 88.89, its specificity was much lower at 20.00, highlighting difficulties in correctly identifying the negative class. The Kappa statistic of 0.1026suggests minimal agreement between predictions and true labels.

Overall, the Random Forest model outperformed the SVM model in terms of accuracy, specificity, balanced accuracy, and agreement, making it the more reliable classifier for this dataset.

Biomarker candidate

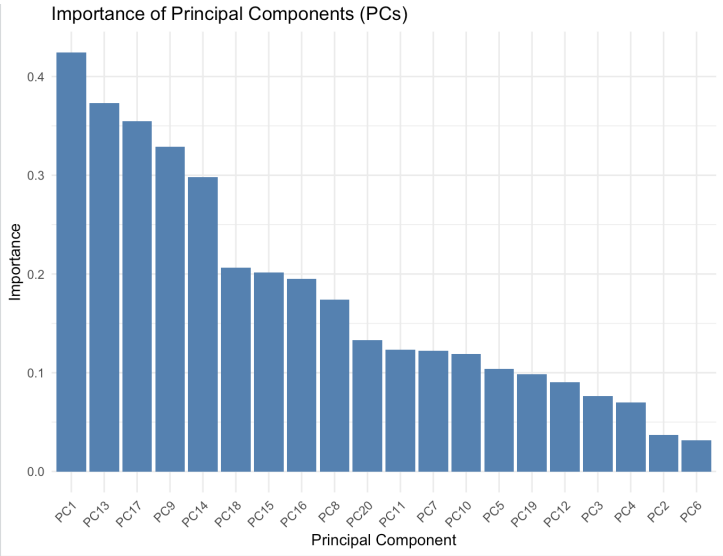

Figure 5: Importance of PCs

The feature importance analysis derived from the classification model highlights the contribution of individual principal components to the model's predictive performance. As shown in Figure 5, PC1 emerged as the most important component, contributing over 40% to the overall model accuracy. This indicates that PC1 captures key variations in the data that are critical for distinguishing between the two groups.

Table 2: Metabolites Contributing to PC1

Metabolite Index	**Metabolite Feature**
1	197.0378_12.5188
2	243.9366_0.5318
3	312.9719_0.4957
4	301.9402_12.6409
5	132.0661_2.7703

The metabolites contributing to PC1 are presented in Table 2. These metabolites likely represent important biomarkers driving the observed group differences. Their prominence in PC1 suggests their potential role in underlying metabolic pathways associated with the classification task. The distribution of feature importance among the remaining PCs diminishes progressively, with PCs such as PC13, PC17, and PC9 also contributing substantially but to a lesser extent.

Discussion and Conclusions

This study has successfully leveraged advanced metabolomic profiling techniques to identify significant metabolites that distinguish Alzheimer's disease from control states. By leveraging Principal Component Analysis (PCA) and machine learning models, particularly Random Forest and Support Vector Machine (SVM), I have identified significant biomarkers and principal components that encapsulate the most variance and biological relevance within the dataset.

The principal components analysis revealed that the first 20 components captured the majority of the variance within the dataset, with PC1 being particularly influential in the classification models. This component was further explored to identify key metabolites that could serve as potential biomarkers for Alzheimer's disease. Notably, five metabolites have emerged as likely important biomarkers driving the observed group differences, suggesting specific targets for further biological and clinical investigation. The Randome Foresr demonstrated superior performance over SVM in terms of accuracy, specificity, and the Kappa statistic, underscoring its suitability for this type of data.

The significant metabolites identified hold potential as biomarkers for Alzheimer's disease. Future work should focus on validating these biomarkers in independent cohorts to confirm their diagnostic and prognostic utility. Validation in a clinical setting can further establish their effectiveness in early diagnosis or in monitoring disease progression. To deepen the understanding of the biological processes underlying the identified biomarkers, further enrichment analysis using databases such as KEGG is imperative, allowing us to map the significant metabolites to specific biochemical pathways, providing insights into the metabolic alterations associated with Alzheimer's disease. The integration of I findings with genetic, transcriptomic, and proteomic data could offer a multi-omic approach to understanding Alzheimer's disease comprehensively.

References

Pan, Yiming, et al. “Metabolites as Frailty Biomarkers in Older Adults.” Proceedings of the National Academy of Sciences of the United States of America, vol. 118, no. 1, 2021, pp. 1–2. JSTOR, https://www.jstor.org/stable/27006404. Accessed 14 Dec. 2024.

Pomfret, Sarah M., et al. “Metabolomics for Biomonitoring: An Evaluation of the Metabolome as an Indicator of Aquatic Ecosystem Health.” Environmental Reviews, vol. 28, no. 1, 2020, pp. 89–98. JSTOR, https://www.jstor.org/stable/26998604. Accessed 14 Dec. 2024.

Pannkuk, Evan L., et al. “Global Metabolomic Identification of Long-Term Dose-Dependent Urinary Biomarkers in Nonhuman Primates Exposed to Ionizing Radiation.” Radiation Research, vol. 184, no. 2, 2015, pp. 121–33. JSTOR, http://www.jstor.org/stable/24546006. Accessed 14 Dec. 2024.