Integrating Machine Learning with Plasma Lipidomics to Predict Alzheimer's Disease Progression in Patients with Mild Cognitive Impairment

by Luke Betlow

Abstract

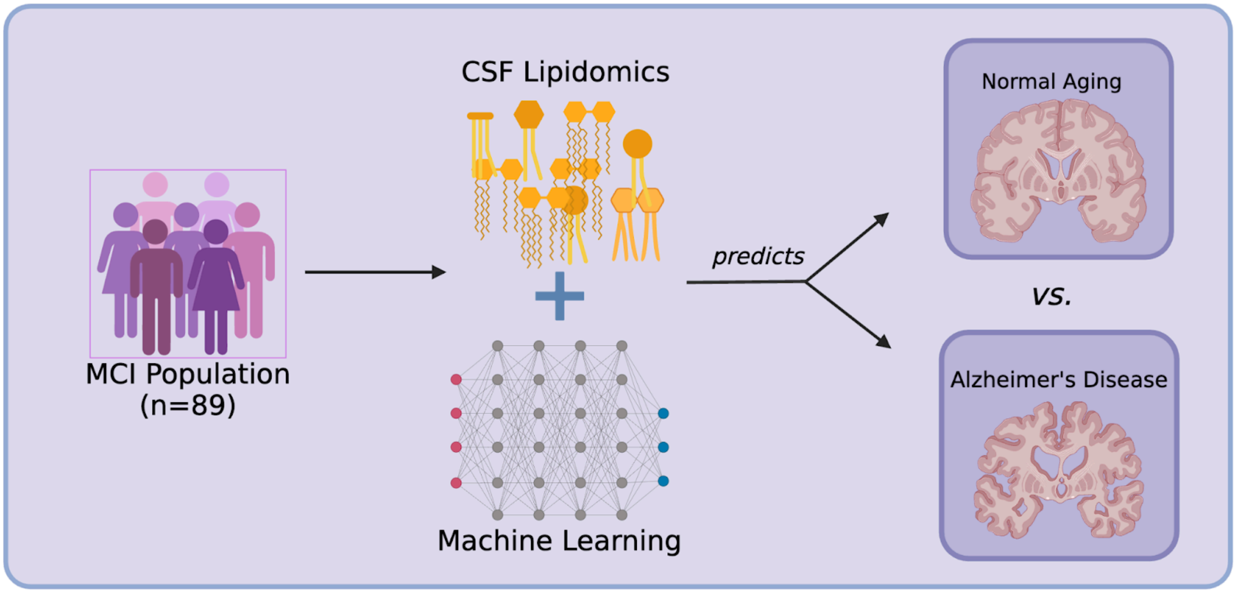

Figure 1: Graphical Abstract of Study

Alzheimer's disease (AD) remains one of the most challenging neurodegenerative disorders, particularly due to the difficulty of early diagnosis and lack of predictability in the progression of the disease in patients who already exhibit mild cognitive impairment (MCI). Recent advances in lipidomics and machine learning offer new avenues for uncovering biological markers that may be predictive of disease development. This study explores whether a machine learning model trained on plasma lipidomic data and select biomarkers can effectively identify MCI patients at higher risk of progressing to AD. We used a publicly available dataset of 212 participants, focusing specifically on a subgroup of 89 MCI patients who progressed to developing AD. Clinical metadata were reduced to retain only lipidomic features and a derived Tau Ratio (CSF p-tau / total tau), and machine learning classifiers were trained to predict binary progression outcomes. Models evaluated included Random Forest, Logistic Regression, Support Vector Machine, Decision Tree, Naive Bayes, and a neural network. The best-performing models (Random Forest and Decision Tree) achieved accuracy scores of 0.7778, with balanced precision and recall scores. Feature importance derived from the Decision Tree model revealed a set of lipidomic variables with high predictive contribution. These findings demonstrate that lipidomic profiles, particularly when enriched with biologically relevant ratios like Tau Ratio, can contribute meaningful signals to classification models. While exploratory in nature, this work supports the utility of machine learning for neurodegenerative disease prediction and offers a reproducible pipeline for future studies aiming to integrate lipidomics into clinical screening tools for AD risk.

1. Introduction

Alzheimer’s disease (AD) is the most common form of dementia, currently affecting over 50 million people worldwide, with projections estimating that number will triple by 2050 (Nichols et al., 2022). One of the most critical challenges in addressing AD is the early and accurate identification of patients likely to progress from mild cognitive impairment (MCI) to full clinical AD. Traditional methods, such as cognitive tests and imaging, while valuable, are limited by their accessibility, invasiveness, and cost. Blood-based biomarkers—particularly lipidomics ones—have gained attention as non-invasive, scalable alternatives that may reflect key aspects of AD pathophysiology.

Lipids play essential roles in cellular structure, signaling, and metabolism that are often impacted in health and disease. Dysregulation in lipid pathways has been implicated in neurodegeneration, with recent studies suggesting that changes in plasma lipid profiles may be associated with amyloid and tau pathologies, as well as clinical progression from MCI to AD (Dakterzada et al., 2023). For instance, altered sphingomyelin (e.g., SM(36:0), SM(40:1)) and diglyceride (e.g., DG(44:3)) levels may affect amyloid precursor protein processing and Aβ42 aggregation, while reductions in ether-linked lipids are associated with oxidative stress and impaired tau clearance. These disruptions not only correlate with biomarker features but may also contribute directly to disease progression from MCI to AD (Dakterzada et al., 2023). Furthermore, the development of machine learning (ML) models has made it possible to analyze complex datasets and extract patterns not easily detected through conventional statistical approaches.

This study evaluates whether a lipidomics-based ML model can successfully predict progression to AD in patients with MCI. By training multiple classifiers on lipidomic data and a biologically grounded Tau Ratio (CSF p-tau/total tau), we test the capacity of data-driven models to distinguish between MCI patients at higher and lower risk of disease conversion. Our findings propose supplementation to early Alzheimer’s detection and diagnostic protocol while additionally contributing to the growing literature on the use of machine learning for high-dimensional omics in neurology.

2. Methods

2.1 Data Source and Ethics

The dataset used in this study was obtained from the Dataverse CSUC repository (DOI: 10.34810/data614). It was originally developed by Dakterzada et al. (2023) as part of a study examining lipidomic *changes* related to Alzheimer’s disease (AD) progression. The data includes 213 participants: 104 with a diagnosis of AD, 89 with mild cognitive impairment (MCI), and 20 cognitively healthy controls. Participants were assessed using cognitive testing, cerebrospinal fluid (CSF) biomarkers (Aβ42, total tau, phosphorylated tau), and genotyping for APOE ε4 status. The age distribution of participants is illustrated in the bar plot below.

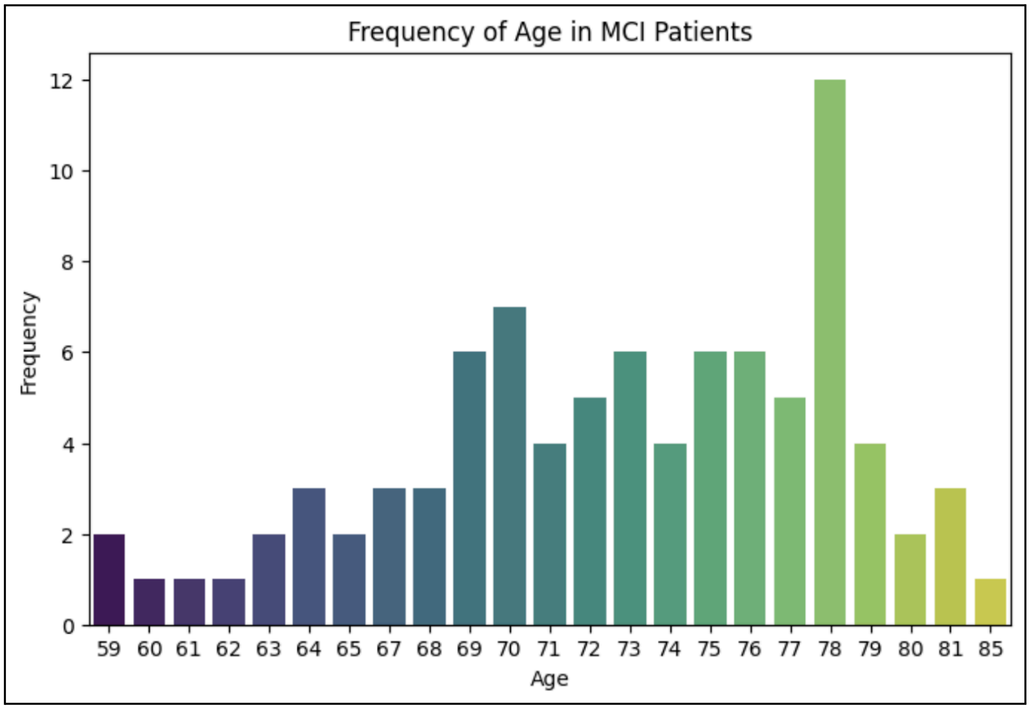

Figure 2: Bar Plot illustrating Age Frequency in Subject Population

Lipidomic profiling for the dataset was performed on plasma samples using liquid chromatography–mass spectrometry (LC-ESI-QTOF-MS/MS). Since the dataset is publicly available and fully de-identified, no additional ethical approval was required for our analysis.

2.3 Cohort Selection and Data Preparation

For the purpose of our analysis, we focused on participants who had a baseline diagnosis of MCI. Within this group, we examined whether each individual eventually progressed to AD. Participants with missing progression labels were excluded, resulting in a refined dataset of 89 participants containing only MCI patients with a known outcome — they either did progress to develop AD or did not.

To better isolate the predictive value of plasma lipidomic data, we removed most conventional clinical and demographic variables, such as age, sex, MMSE scores, APOE status, and raw CSF biomarker concentrations. However, we retained a single engineered feature: the Tau Ratio, calculated by dividing phosphorylated tau by total tau in CSF. This ratio is widely used as a marker of tau pathology specific to AD, rather than reflecting general neurodegeneration. Including the Tau Ratio provided a reference point for evaluating how lipidomic patterns relate to established AD-associated biochemical changes, without reintroducing the full clinical profile.

No scaling or normalization was applied to the features, and missing data were handled by simply removing rows with any missing values. Categorized outcome labels ("Yes"/"No" for AD progression) were retained in their original form.

2.4 Exploratory Correlation Analysis

As part of our early analysis, we explored how lipid features correlated with established AD biomarkers (CSF Aβ42, p-tau, total tau), cognitive score (MMSE), and age. We used Pearson correlation to identify features with consistent associations across multiple clinical variables. Features with an absolute correlation above 0.3 in at least three comparisons were flagged for further attention, serving as a rough filter for analytical relevance. The correlation visualization did not reveal patterns or associations as shown below. The absence of strong linear correlations further justifies the use of nonlinear classifiers that can detect higher-order interactions among features.

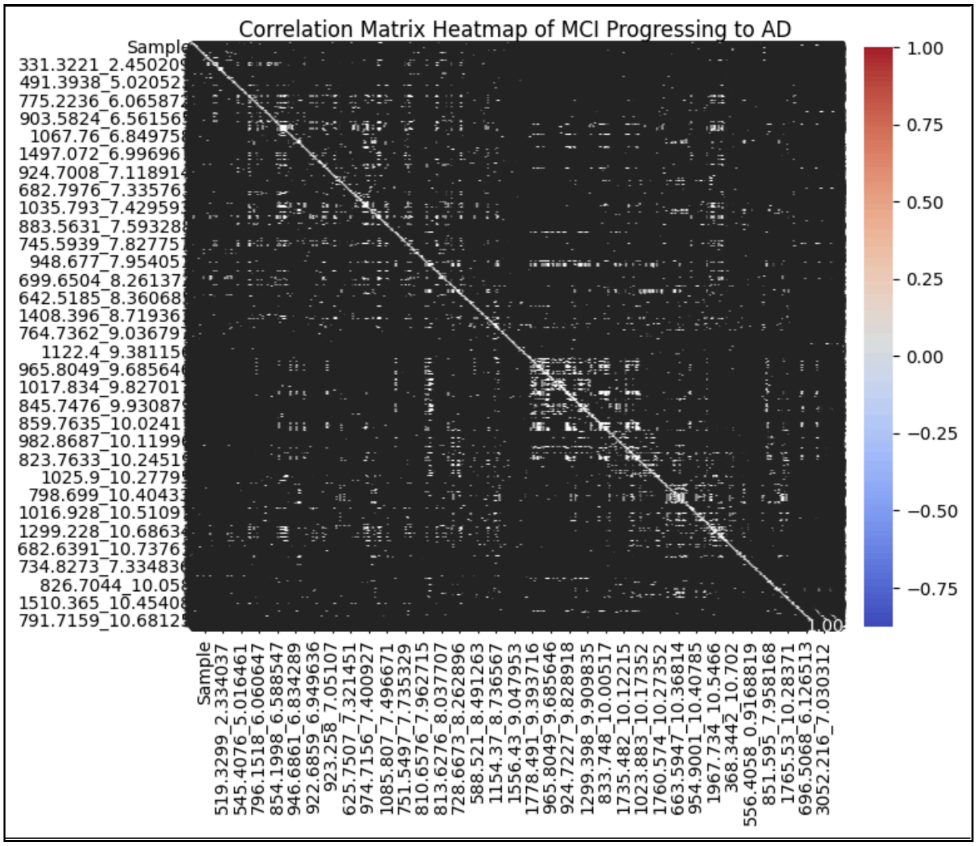

Figure 3: Correlation Matrix Heatmap

2.5 Machine Learning Models

To evaluate the predictive value of lipidomic data in determining which MCI patients progressed to Alzheimer's disease, we trained six common classification models: Logistic Regression, Random Forest, Support Vector Machine (SVM), Decision Tree, Naive Bayes, and a feedforward Neural Network. The dataset was split 80/20 into training and testing sets. Model performance was assessed using accuracy, precision, recall, and F1-score, with separate scores reported for each class (progressors vs. non-progressors).

Logistic Regression is a linear classification algorithm that models the probability of an outcome using a logistic function. It is widely used for its interpretability and low computational cost, particularly when data relationships are approximately linear (Hosmer et al., 2013). However, it can underperform on complex, nonlinear patterns, which are common in biological data.

Random Forest is an ensemble model that builds multiple decision trees and averages their outputs to reduce overfitting and improve generalizability (Breiman, 2001). It handles high-dimensional data well, tolerates multicollinearity, and provides interpretable feature importance rankings. Its main drawback is reduced transparency compared to single-tree models, as the ensemble structure obscures specific decision pathways.

Support Vector Machines (SVMs) classify data by finding the optimal separating hyperplane with maximum margin. SVMs perform well in high-dimensional settings and are effective even with limited data, but they require careful parameter tuning and are less interpretable than tree-based models (Cortes & Vapnik, 1995). Their performance can also suffer when class distributions are imbalanced.

Decision Trees create a sequence of binary splits based on input features to predict outcomes. They are easy to visualize and interpret, making them useful for uncovering which features contribute most to classification. However, they are prone to overfitting, especially when used without pruning or regularization (Quinlan, 1986).

Naive Bayes classifiers apply Bayes' theorem with the assumption that all input features are conditionally independent. This assumption rarely holds in real-world data, especially in biological systems where many variables are correlated, but the method remains popular for its simplicity and speed (Rish, 2001). In our case, its performance was significantly lower than other models, likely due to the interdependence among lipidomic features.

Neural Networks, implemented here via a Multi-layer Perceptron (MLP), consist of layers of interconnected nodes that allow the model to learn nonlinear and complex relationships. While powerful and flexible, they require larger datasets, are sensitive to parameter selection, and are generally considered less interpretable than other methods used here (Goodfellow et al., 2016).

2.6 Feature Importance

To better understand which features most heavily influenced the model's predictions, we trained a Decision Tree model and used its built-in feature importances method (see Fig. 3). This approach scores each feature based on how much it reduces impurity in the data when used to split nodes in the tree. Decision Trees were chosen for this analysis because of their interpretability, simplicity, and ability to handle unscaled, high dimensional data like lipidomics.

3. Results

We evaluated six different classification models for predicting progression to Alzheimer's disease in MCI patients, using lipidomic features and the Tau Ratio as predictors. The models included Random Forest, Logistic Regression, SVM, Decision Tree, Naive Bayes, and a MLP neural network.

Model performance (as shown in the table below) varied, with the highest classification accuracy (0.7778) achieved by both the Random Forest and Decision Tree models. Logistic Regression and the MLP both achieved 0.7222 accuracy, followed by SVM (0.6667) and Naive Bayes (0.5000). The Random Forest model demonstrated strong class balance, with perfect precision for non-progressors (1.00) and high recall for progressors (1.00), yielding F1-scores of 0.78 for both classes. The Decision Tree model also showed strong recall for non-progressors (0.91) but had slightly lower precision for progressors (0.80).

Table 1: Derived Accuracy and Efficacy of Different ML Models

ML Model	*Accuracy*	*Precision (no)*	*Precision (yes)*	*Recall (no)*	*Recall (yes)*	*F1-score (no)*	*F1-score (yes)*
Random Forest	0.7778	1.00	0.64	0.64	1.00	0.78	0.78
Logistic Regression	0.7222	0.88	0.60	0.64	0.86	0.74	0.71
SVM	0.6667	0.78	0.56	0.64	0.71	0.70	0.62
Neural Network	0.7222	0.88	0.60	0.64	0.86	0.74	0.71
Decision Trees	0.7778	0.77	0.80	0.91	0.57	0.83	0.67
Naive Bayes	0.5000	0.62	0.40	0.45	0.57	0.53	0.47

We evaluated feature importance using the Decision Tree model. Features with the highest Gini importance scores were lipidomic markers labeled by their mass-to-charge ratio and retention time. The top ten most influential features, supporting the interpretation that lipidomic variation independent of traditional clinical markers, contains signals relevant to AD risk (Figure 4). As shown in Figure 4, seven specific lipid features account for nearly all the predictive importance in the Decision Tree ML model.

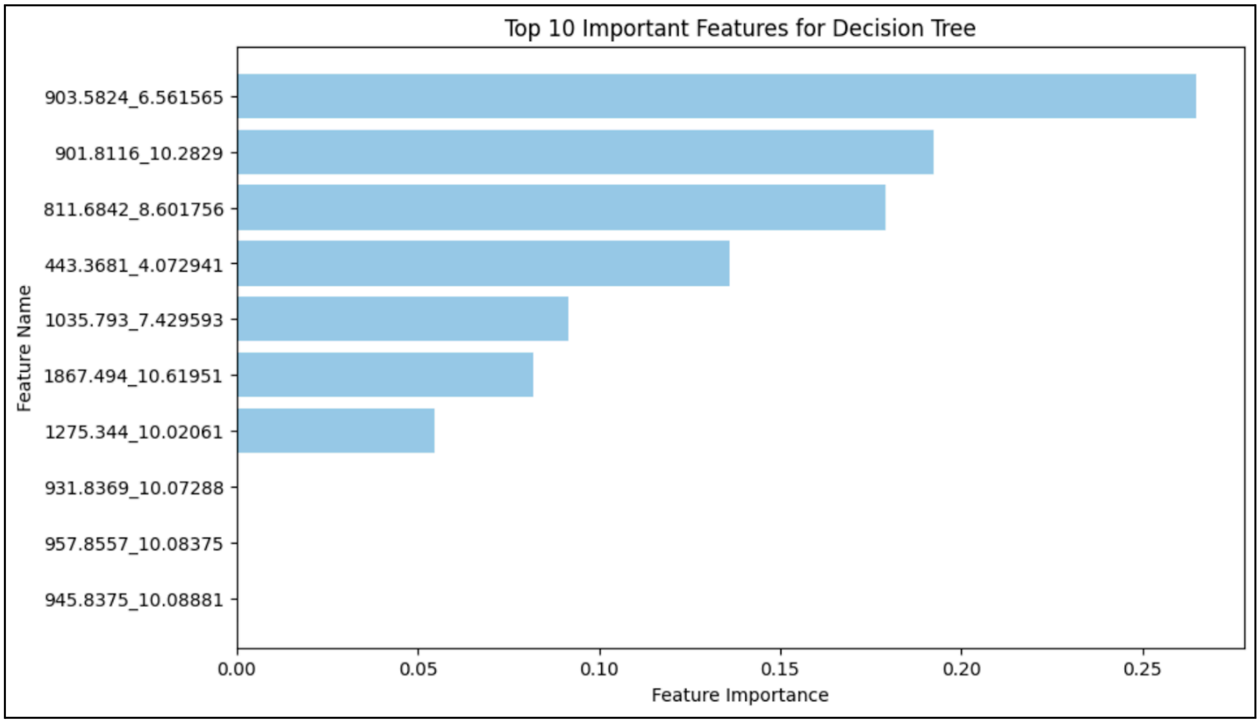

Figure 4: Relative Feature Importance of Lipids in Decision Tree ML Model

4. Discussion and Conclusion

This study supports the feasibility of using plasma lipidomic data to predict Alzheimer's disease progression in patients with MCI using standard machine learning models. While no single model excelled across all metrics, Random Forest and Decision Tree classifiers consistently demonstrated strong classification performance, especially in balancing recall between both progression classes.

Notably, this study excludes traditional clinical and cognitive predictors, instead emphasizing the predictive potential of molecular signatures alone. Our findings mostly align with those of Dakterzada et al. (2023), who identified several neutral and ether-linked lipids (including plasmalogens and triglycerides) as correlated to both AD biomarkers and disease progression. While their work used classical regression models to establish statistical associations, our approach extends their findings by demonstrating that lipidomic patterns alone can effectively train machine learning models that reach high predictive accuracy.

These findings have significant implications for the future of AD screening. Current diagnostic modalities such as neuroimaging (e.g., PET scans), cognitive testing (e.g., MMSE), and CSF analysis are effective but have important drawbacks. They are invasive, time-consuming, costly, or difficult to scale for population-level screening (Jack et al., 2018). In contrast, lipidomic profiling offers a more scalable and cost-effective screening technology. While it is unlikely that lipidomics alone will entirely replace existing clinical tools, our results suggest that it could play a vital supplementary role, particularly in stratifying MCI patients early in the diagnostic process and prioritizing them for further evaluation. This would help address a current bottleneck in early diagnosis, enabling more proactive and personalized treatment interventions (Toledo et al., 2017).

However, several limitations merit discussion. First, this analysis did not incorporate genetic carrier status, such as APOE ε4, despite its well-documented influence on lipid metabolism, amyloid accumulation, and AD risk (Huynh et al., 2017). Excluding APOE status helped us isolate the predictive power of plasma-derived features alone, but future work should test how layering in genetic data may improve model accuracy and help define distinct molecular subtypes of MCI patients. Additionally, the study's modest sample size, lack of external validation cohort, and default hyperparameter settings limit generalizability. Follow-up studies with larger and more diverse populations—ideally across multiple clinical sites—are needed to validate these findings and optimize model performance.

Another relevant clinical question concerns the role of amyloid deposition in the absence of Alzheimer's disease. Amyloid plaques have long been considered a central hallmark of AD, yet numerous studies have shown that some cognitively normal older adults present with amyloid accumulation without ever developing clinical dementia (Aizenstein et al., 2008; Jack et al., 2018). This dissociation complicates biomarker interpretation, especially in the early stages of the disease. Because our approach does not directly depend on amyloid status but instead relies on plasma lipid profiles, it may offer additional value in detecting clinically relevant neurodegeneration, even in individuals with ambiguous amyloid findings. As new research continues to highlight tau as a more direct correlate of symptom progression (Ossenkoppele et al., 2018), lipidomics may serve as an informative adjunct to both amyloid and tau markers in a broader diagnostic framework.

In conclusion, this work demonstrates the potential efficacy of machine learning applied to lipidomic data as a scalable, biologically grounded approach for predicting AD development risk in MCI patients. When integrated with existing clinical diagnostic tools, this method could support earlier intervention and more personalized care strategies while remaining minimally invasive in the preliminary stages of diagnosis. As biomarker research continues to evolve, the fusion of omics-based modeling with established clinical tools may shift Alzheimer's diagnostics toward a multi-modal, precision medicine paradigm — supplemented by advances in machine learning as explored in this study.

5. Figure Legend

Figure 1. **Graphical Abstract of Study**

This graphic illustrates the methodology and conception of the study, showing that the combination of machine learning and lipidomic profiling can predict Alzheimer's disease progression.

Figure 2. **Bar Plot Illustrating Age Frequency in Subject Population**

This bar plot displays the age distribution of participants within the MCI subgroup. It was generated using Python's matplotlib and seaborn libraries to visualize the demographic profile of the sample used for model training.

Figure 3. **Correlation Matrix Heatmap of MCI Progressing to AD**

This heatmap shows pairwise Pearson correlations among lipidomic features in MCI patients who progressed to AD. The matrix was created in Python using seaborn.heatmap and helped illustrate a lack of collinearity, thus suggesting the need for more advanced models.

Figure 4. **Relative Feature Importance of Lipids in Decision Tree ML Model**

This bar graph ranks the top 10 most influential lipidomic features used by the Decision Tree model, based on their Gini impurity reduction scores. It was produced using matplotlib and illustrates how specific plasma lipids contributed most strongly to prediction of AD progression.

References

[1] Aizenstein, H. J., Nebes, R. D., Saxton, J. A., Price, J. C., Mathis, C. A., Tsopelas, N. D., ... & Klunk, W. E. (2008). Frequent amyloid deposition without significant cognitive impairment among the elderly. *Archives of Neurology, 65*(11), 1509–1517. https://doi.org/10.1001/archneur.65.11.1509

[2] Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

[3] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297. https://doi.org/10.1007/BF00994018

[4] Dakterzada, F., Jové, M., Huerto, R., Carnes, A., Sol, J., Pamplona, R., & Piñol-Ripoll, G. (2023). Changes in plasma neutral and ether-linked lipids are associated with the pathology and progression of Alzheimer's disease. *Aging and Disease, 14*(5), 1728–1738. https://doi.org/10.14336/AD.2023.0221

[5] Dean, J. M., & Lodhi, I. J. (2018). Structural and functional roles of ether lipids. *Protein & Cell, 9*(2), 196–206. https://doi.org/10.1007/s13238-017-0423-5

[6] Farmer, B. C., Walsh, A. E., Kluemper, J. C., & Johnson, L. A. (2020). Lipid droplets in neurodegenerative disorders. *Frontiers in Neuroscience, 14*, 742. https://doi.org/10.3389/fnins.2020.00742

[7] Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

[8] Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). Wiley.

[9] Huynh, T. P. V., Davis, A. A., Ulrich, J. D., & Holtzman, D. M. (2017). Apolipoprotein E and Alzheimer’s disease: The influence of apolipoprotein E on amyloid-β and other amyloidogenic proteins. *Journal of Lipid Research, 58*(5), 824–836. https://doi.org/10.1194/jlr.R075481

[10] Jack, C. R., Bennett, D. A., Blennow, K., Carrillo, M. C., Dunn, B., Haeberlein, S. B., ... & Silverberg, N. (2018). NIA-AA research framework: Toward a biological definition of Alzheimer’s disease. *Alzheimer’s & Dementia, 14*(4), 535–562. https://doi.org/10.1016/j.jalz.2018.02.018

[11] Nichols, E., Steinmetz, J. D., Vollset, S. E., Fukutaki, K., Chalek, J., Abd-Allah, F., ... & Murray, C. J. L. (2022). Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: An analysis for the Global Burden of Disease Study 2019. *The Lancet Public Health, 7*(2), e105–e125. https://doi.org/10.1016/S2468-2667(21)00249-8

[12] Ossenkoppele, R., Schonhaut, D. R., Schöll, M., Lockhart, S. N., Ayakta, N., Baker, S. L., ... & Rabinovici, G. D. (2018). Tau PET patterns mirror clinical and neuroanatomical variability in Alzheimer's disease. *Brain, 141*(5), 1551–1567. https://doi.org/10.1093/brain/aww027

[13] Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning, 1*(1), 81–106. https://doi.org/10.1007/BF00116251

[14] Rish, I. (2001, August). An empirical study of the naive Bayes classifier. In *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence* (Vol. 3, pp. 41–46). IBM Research.

[15] Toledo, J. B., Arnold, M., Kastenmüller, G., Chang, R., Baillie, R. A., Han, X., ... & Saykin, A. J. (2017). Metabolic network failures in Alzheimer’s disease: A biochemical roadmap. *Alzheimer’s & Dementia, 13*(9), 965–984. https://doi.org/10.1016/j.jalz.2017.01.020