Metabolomics in Alzheimer's Disease- An Investigation into Conflicting Methodologies and Results

By

Victoria Walls

52102718

Project submitted as part of the final examination for the degree of MSc Data Science at the University of Aberdeen, August 2022

Word Count = 3856 words

Abstract

Using a dataset sourced from the Mayo Clinic this paper investigated the metabolites that are statistically different in individuals with Alzheimer's Disease (AD) and Cognitively Normal (CN) controls, and, individuals with Mild Cognitive Impairment (MCI) and CN controls. This was undertaken using a variety of statistical and machine learning methods including principal components analysis, partial least-squares discriminant analysis, and logistic regression. Existing results for metabolites in relation to this disease lack clarity and consensus, so the results from this study were compared against some results from existing literature to see if there was any overlap. This was found, though, only in results from the original paper from which the dataset came from. This misalignment with other existing results suggests that the lack of agreed good practice techniques within the metabolomics field, might be partially responsible for the lack of overlap in existing results. Additionally, the characteristics of the data involved with this field and the frequently low sample sizes indicates there is a potential for there to be a general lack of statistical power in studies to draw robust conclusions. Metabolomics has the capacity to be a greatly influential field however with such an absence of agreement at this point in the case of AD there is much development needed before results can hold any real weight.

Table of Contents

List of Tables

List of Figures

List of Additional Materials

ADDITIONAL DOCUMENT 2 – DATA ANALYSIS CODE

ADDITIONAL DOCUMENT 3 – SUPPLEMENTARY TABLES

ADDITIONAL DOCUMENT 4 - ADMSPOSITIVEIONMODEC18DATA

Introduction

Metabolomics is a young research field - first mentioned in an academic paper in 1998 (Oliver et al., 1998). As with most new scientific areas, development is occurring at rapidly. The result is a field brimming with potential and ideas. Understandably, this fresh approach to biochemistry and data science is not without issue. There is a lack of agreement over "good practice" in data analysis methods, which, along with low statistical power and high dimensional, noisy data, can leave results inconsistent. It is also thought that these inconsistencies could be caused by sample size, lack of diversity (and reproducible diversity) within study populations, inappropriate exclusion of metabolites by multiple testing adjustments and (in the case of Alzheimer's Disease) the heterogeneous Alzheimer's phenotypes used in different studies (Huo et al. 2020).

There have been several studies identifying metabolites relating to Alzheimer's Disease (AD) but there remain conflicting views on which metabolites are directly related. Huo et al., (2020) specifically mentions that the metabolic basis underlying the disease and the metabolic markers predictive of its risk remains to be determined. This work further goes on to express that no conclusive metabolites had been identified. An illustration of this lack of agreement can be found in Supplementary Table 1 which shows that of the 71 named metabolites found in the preparatory literature read for this study, only Hypoxanthine appears in more than one work (Trushina et al., 2013, Wang et al., 2014).

Supplementary Table 2 demonstrates that of these studies 44 different analysis techniques were adopted – of which only eight were shared. These methods are shaded in grey and were influential in the workflow plan for this paper.

Lack of agreed good practice also goes beyond the traditional approaches to analytical research. Many datasets do not cover the entire human metabolome. The dataset used in this study, for example, contains 1909 metabolites. According to the Human Metabolome Database, there are currently 253,245 metabolites in their system, with 3,444 which have been detected and quantified. This clearly limits any current study and the lack of quantified compounds may also be impacted by the mass spectrometry techniques used to extract metabolite data from samples.

Metabolomics has the potential to be a greatly influential field in medicine and medical research, thanks to the minimal and non-invasive methods in which samples can be collected – blood, urine or cerebral spinal fluid (CSF). With AD specifically, metabolic differences associated with the disease are thought to be measurable up to 25 years prior to symptoms showing (Trushina et al., 2013). If a definitive panel of metabolites associated to the disease existed, one's risk and even prognosis could be predicted well in advance from routine testing. Individuals would gain time to prepare and the reassurance that the medical field would have time to further develop treatments prior to their symptoms showing.

Despite a lack of clarity in current results, using metabolomics for diagnostic testing is considered to be equally as if not more accurate than current neuropsychological and imaging techniques as well as being more cost efficient and safer to the patient (Lee and Hu, 2019). This has led to some treatment methods attempted based on metabolomics – e.g., supplementation of carnitines and acylcarnitines to regulate cellular bioenergetics. This approach has also been considered for treatment and exploration of other illnesses affecting cognitive status such as schizophrenia (Cao et al., 2019).

AD is increasing in commonality - in the US alone the number of individuals affected has been projected to grow from 5.3 million in 2017 to nearly 14 million by 2050. Against this backdrop, the current success rate for drug development to treat the disease has been poor. Between 2002 and 2012 the success rate for advancing drug compounds for FDA approval was 0.4% (Cummings 2017). Building a consensus of metabolite results will enhance the accuracy of these clinical trials by allowing diagnosis to be confirmed with measurable biomarkers. Trushina et al., (2013) believes the current FDA-approved drugs do not provide a cure because they are administered too late in the process of the disease. This also provides incentive for building a consensus to facilitate early diagnosis.

This study aims to do a small amount of testing on a dataset containing samples taken from individuals with AD, Mild Cognitive Impairment (MCI), and cognitively normal controls (CN). Results will then be compared with the significant metabolites found by existing literature. The hope was to either provide clarity by aligning with existing results, or to illustrate the lack of agreed “good practice” techniques within the field of metabolomics creates an environment of in a reproducible and inconsistent results.

Methods

All of the data analysis techniques undertaken in this study were written in the Python programming language. A full list of the methods and Python libraries used can be found in the README file (Additional Document 1) and a runnable version of the code used for this analysis can be found in Jupiter Notebook format in Additional Document 2.

The data used in this research was published by an American study (Trushina et al. 2013) and made available through the Metabolomics Workbench Database (plasma dataset ST000046). The original study had four datasets as it covered different sample types however this research was only used the data file named “ADMSPositiveIonModeC18Data”. A copy of this can be found in csv format in Additional Document 4.

A dataset of blood samples was chosen because this could be more easily compared to other literature - metabolites can be significantly influential in different ways across the different sample sources. It was preferred that the data set was based of human samples both for comparison purposes and due to the various problems associated with animal models illustrated by Cummings (2017). These include a poor randomization of samples, external influences on animal behaviour and the lack of most aspects of human AD beyond the changes in the amyloid metabolism. C18 data was selected over HILIC due to the general advantages of Cogent TYPE – C columns which include efficiency and adaptability (MicroSolv Technology Corporation, 2012). It can be difficult to know without testing whether positive or negative ion mode is more suitable for a dataset however for 36% of results tested by Liigand et al., (2018) both modes were found comparable.

Visualising the data showed each metabolite presented as a row, and each sample as a column. The samples were labelled according to group one, two or three, with group numbers referencing the cognitive status of the participants – (AD, MCI and CN) and number (1-15 in each group). Once the dataset was checked and the index set, it was transposed to better suit the analysis methods and inspected for missing values of which there was 2.88%. A visualisation of these missing values can be seen in Figure 1 which supports the results of the tests to measure the number of missing values. A dendrogram illustrating the relationships between the missing values can be seen in Figure 2. None of the samples are connected at a nullity correlation level below 6, which indicates that none of the missing values were significantly influential on each other.

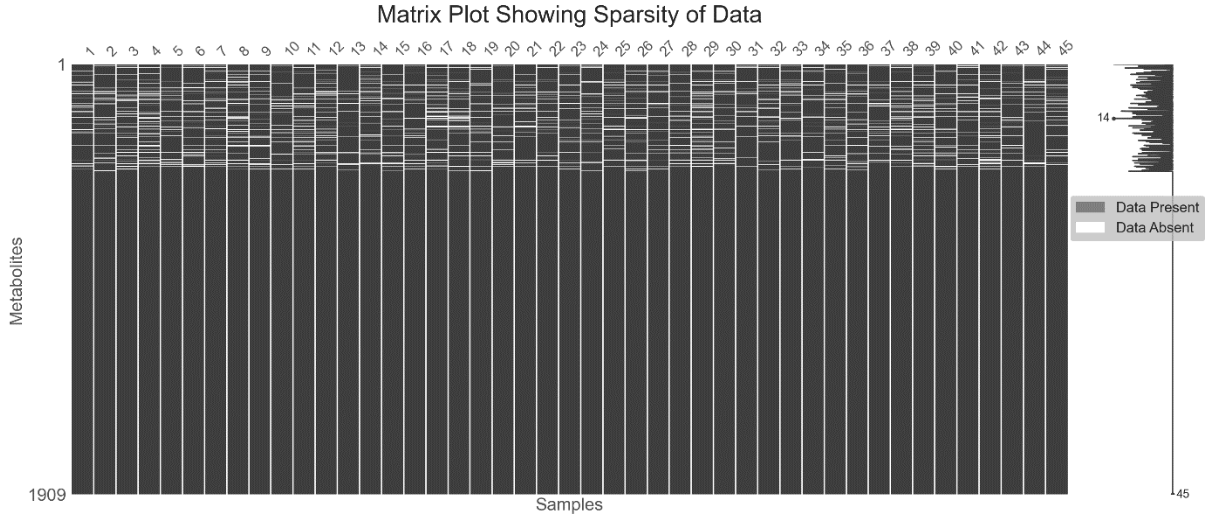


Figure 1 Matrix Plot Showing Sparsity of Data. Missing values ranged from 1.83 – 5.50% in each sample (columns). Sparkline indicates that two metabolites with the greatest number of missing values only had 14 of 45 (10_12_15-octatrienoicacid and D-Glucose.

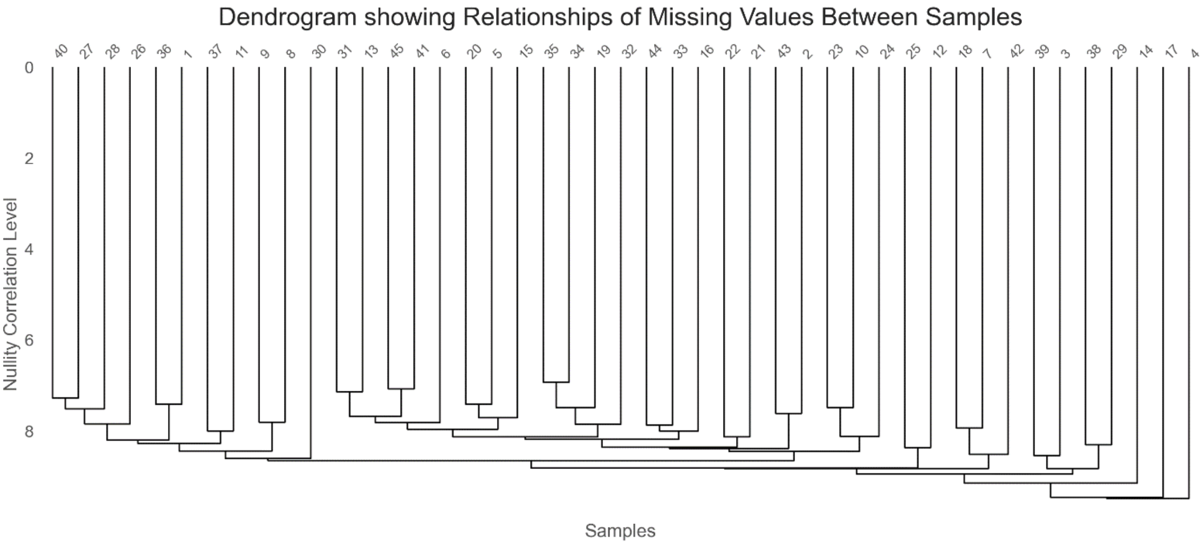


Figure 2 Dendrogram showing Relationships of Missing Values Between Samples

Whilst the field still lacks consensus on best practices, there is a pattern of using k Nearest Neighbours (kNN) as a method of imputation to deal with missing values. Whilst there have been attempts to improve on this method (such as the No Skip k Nearest Neighbour method (Lee and Styczynski 2018)) kNN remains the most common and accessible method, and thus was used. Figure 3 shows a plot grid displaying the distribution of the data as it was scaled and normalised throughout the imputation process.

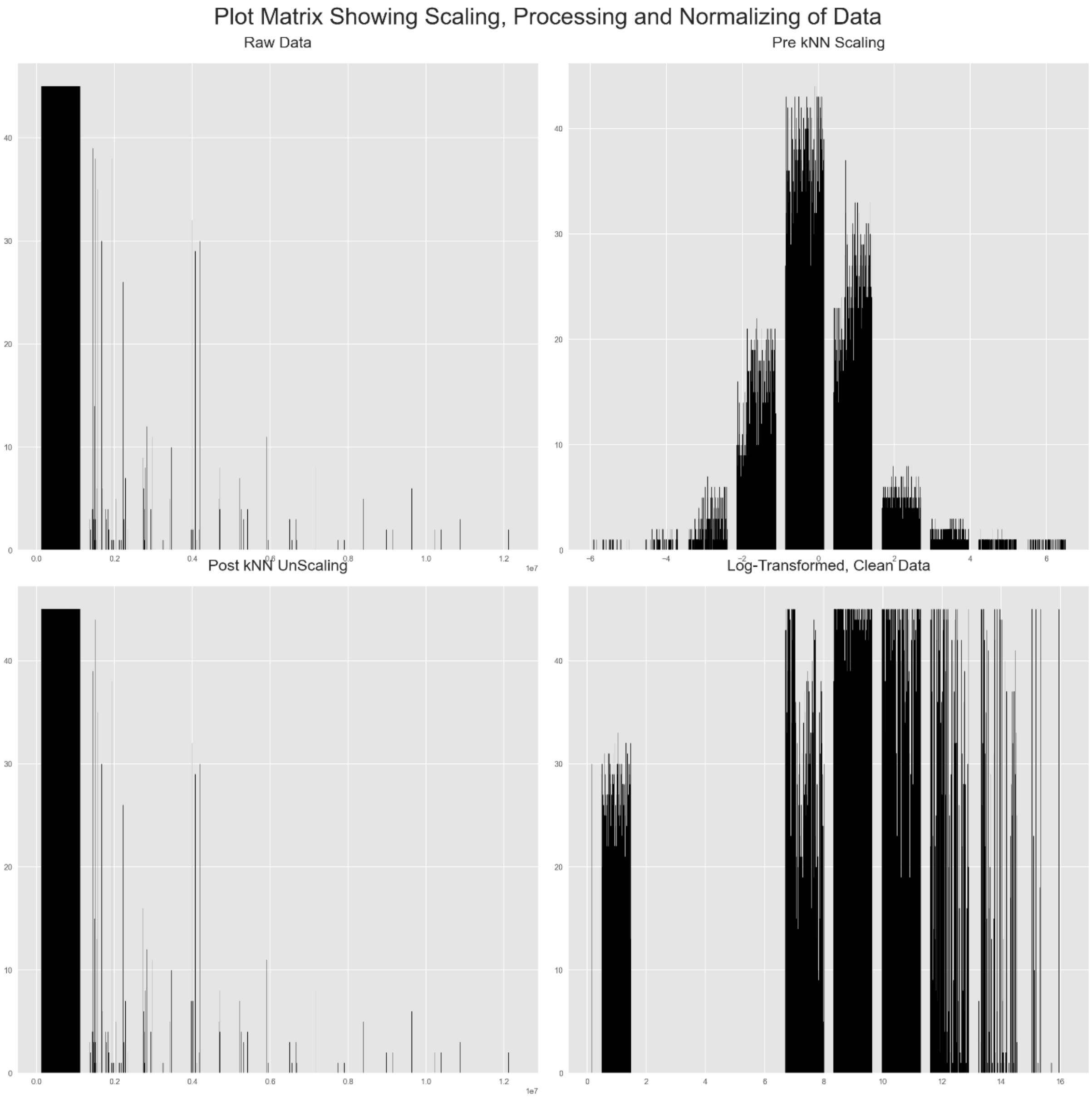


Figure 3 Plot Matrix Showing Scaling, Processing and Normalizing of Data

Before analysis, the data was tested for multicollinearity and initially attempted by visualising the dataset using a cluster map (Figure 4), however with such a large number of metabolites it was clear that it would not be clear to interpret. To combat this an interactive network graph with an adjustable correlation threshold between the values of 0.1 and 1 was built. This allowed direct comparisons of correlated metabolites between different levels of correlation thresholds. Visualisation of direct relationships between correlated metabolites was made accessible using panning, zoom and selection tools. An example of this graph, set to threshold of 0.95, can be seen in (Figure 5), and the adjustable graph can be found in Additional Document 2 (see further details in the README file). Following this, a Variance Inflation Factor (VIF) test was undertaken to produce measurable scores for data multicollinearity (Miles 2005).

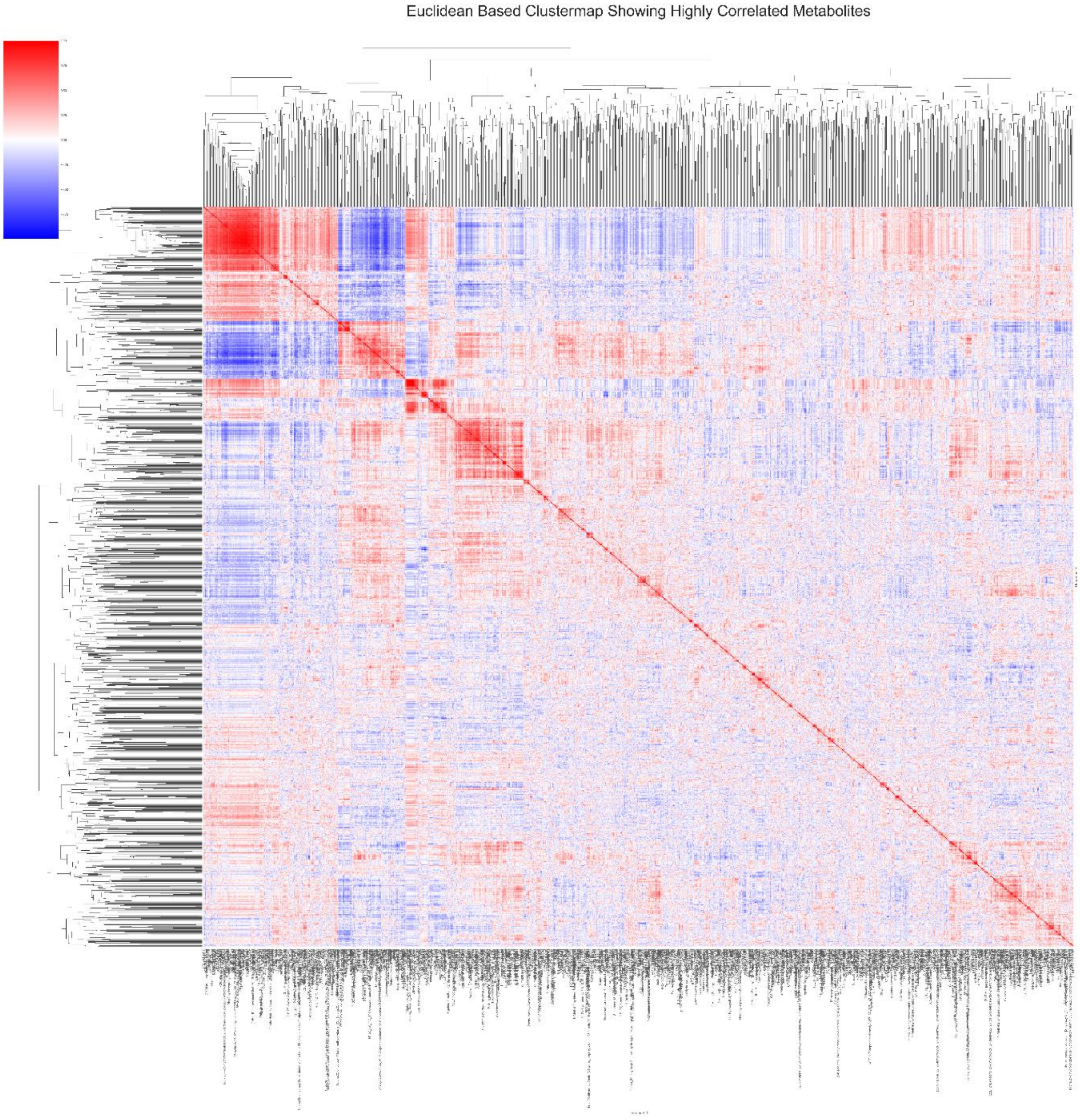


Figure 4 Euclidean Based Clustermap Showing Highly Correlated Metabolites

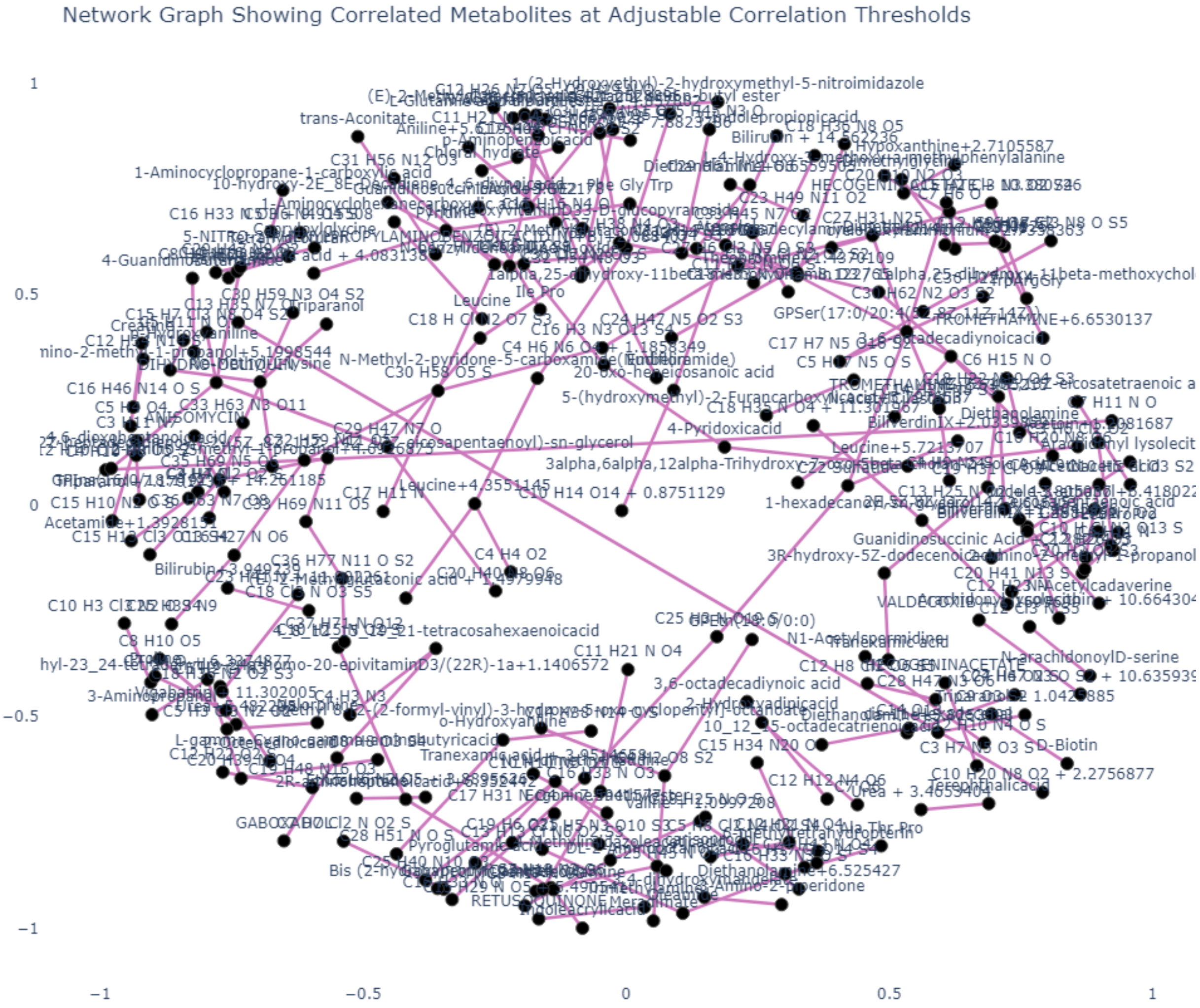


Figure 5 Network Graph Showing Correlated Metabolites at Adjustable Correlation Thresholds

Next a Principal Component Analysis (PCA) was performed and a score plot built to visualise any prevalence of clusters that might appear for the three cognitive groups. Scikit-learn decomposition PCA was used to build one model from which specific numerical values such as the explained variance ratio could be extracted. The model was then rebuilt using Scikit-learn Yellowbrick to produce colour coded scatter plots of the data (by cognitive type) and matplotlib was used to create an elbow plot which compared explained variance against principal components. A new data frame was then produced using the principal components that would explain 95% of the variance in the original data to be used in other tests.

Metabolic differences among the cognitive groups were visualised with Partial least-squares discrimination analysis (PLS-DA) using the new PCA data frame. PLS-DA is a supervised machine learning method used to determine whether known groups are genuinely different and which features best describe these differences. This method is commonly used not only within the field of metabolomics but in the research area of omics as a whole, which was one of the motivating factors behind it being used in this analysis. This is because metabolomics datasets are characterised by having a large volume of missing data, noise and a large number of features (as well as a small number of samples compared to the number of features). With the exception of the proportion of missing data, these characteristics can all be found in the dataset used in this study.

Logistic Regression using the scikit-learn library was used to classify metabolites associated to AD in comparison with CN controls, as well as individuals with MCI in comparison to CN controls. Two data frames were built from the original dataset, and the cognitive status column values were replaced with dummy values - with CN values remaining at zero in both data frames and the respective Alzheimer's severity level having a value of one. The data was then split into the appropriate X and Y objects before being further split into train and test data. These were then used to build an initial version of the logistic regression model. The accuracy of the model was measured using Cross Validation, and the coefficient values for each metabolite was extracted in order to view preliminary results.

In order to further develop the accuracy of the logistic regression models, feature selection using the feature importances property of the scikit-learn extra trees classifier was undertaken (Supplementary Table 3, Supplementary Table 4). Figure 6 shows a chart with the top 50 features according to this method for the AD vs CN data frame, and Figure 7 shows a chart with the top 50 features according to this method for the MCI vs CN data frame. The models were then rerun with these new data frames containing the 50 most important features, and retested using Cross Validation for accuracy. The coefficients for the metabolites were then extracted.

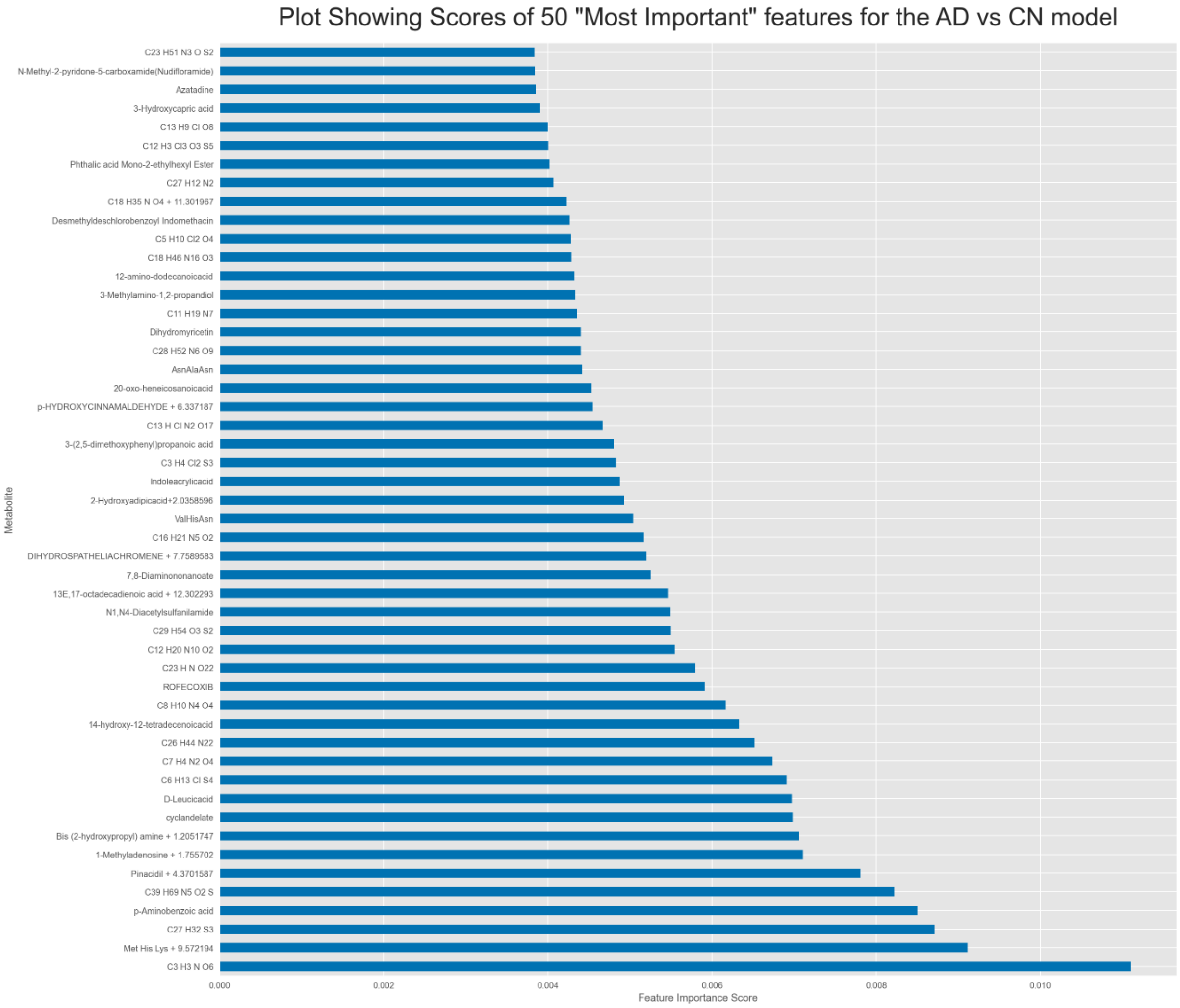


Figure 6 ADvsCN Logistic Regression Feature Selection by Importance

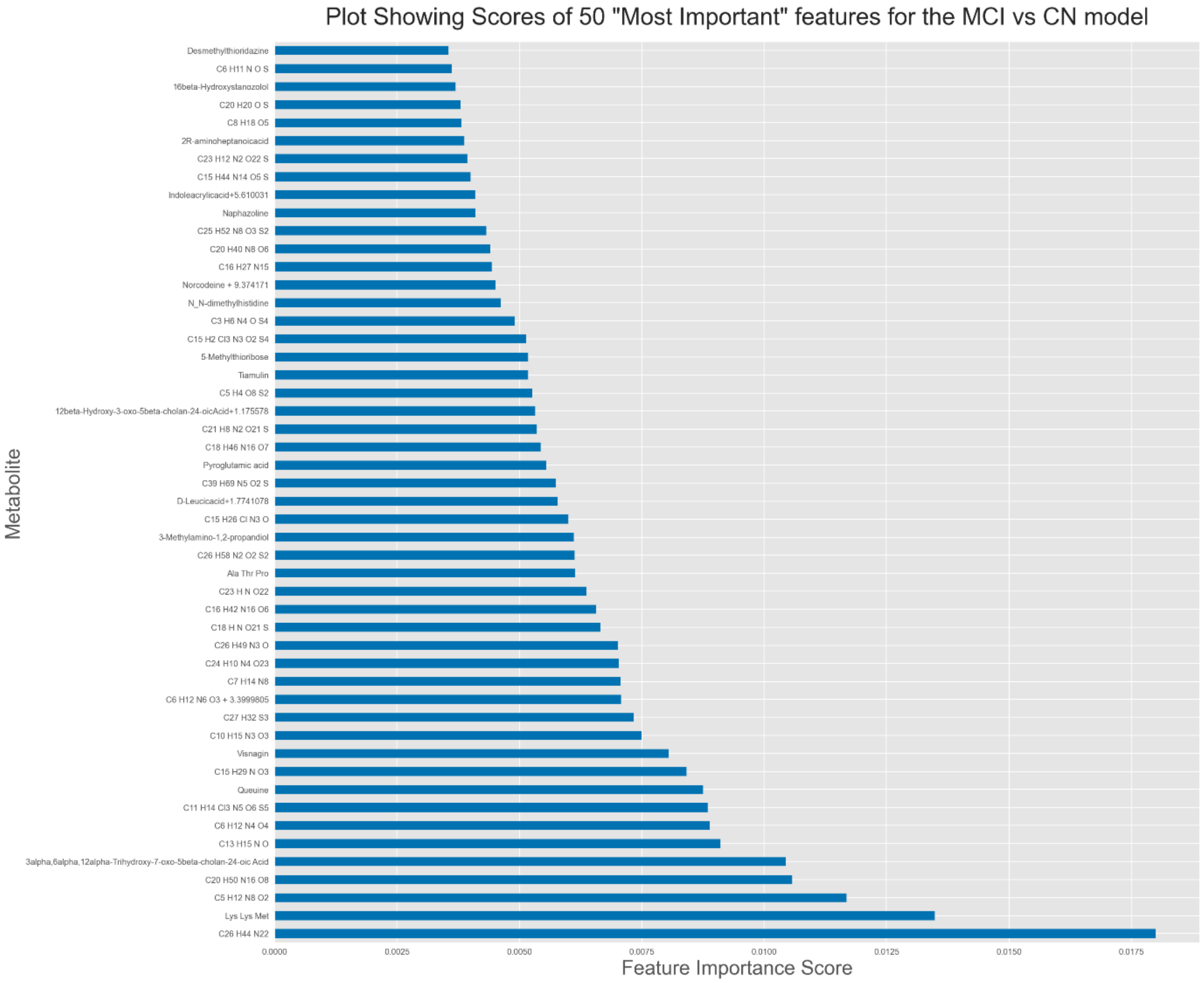
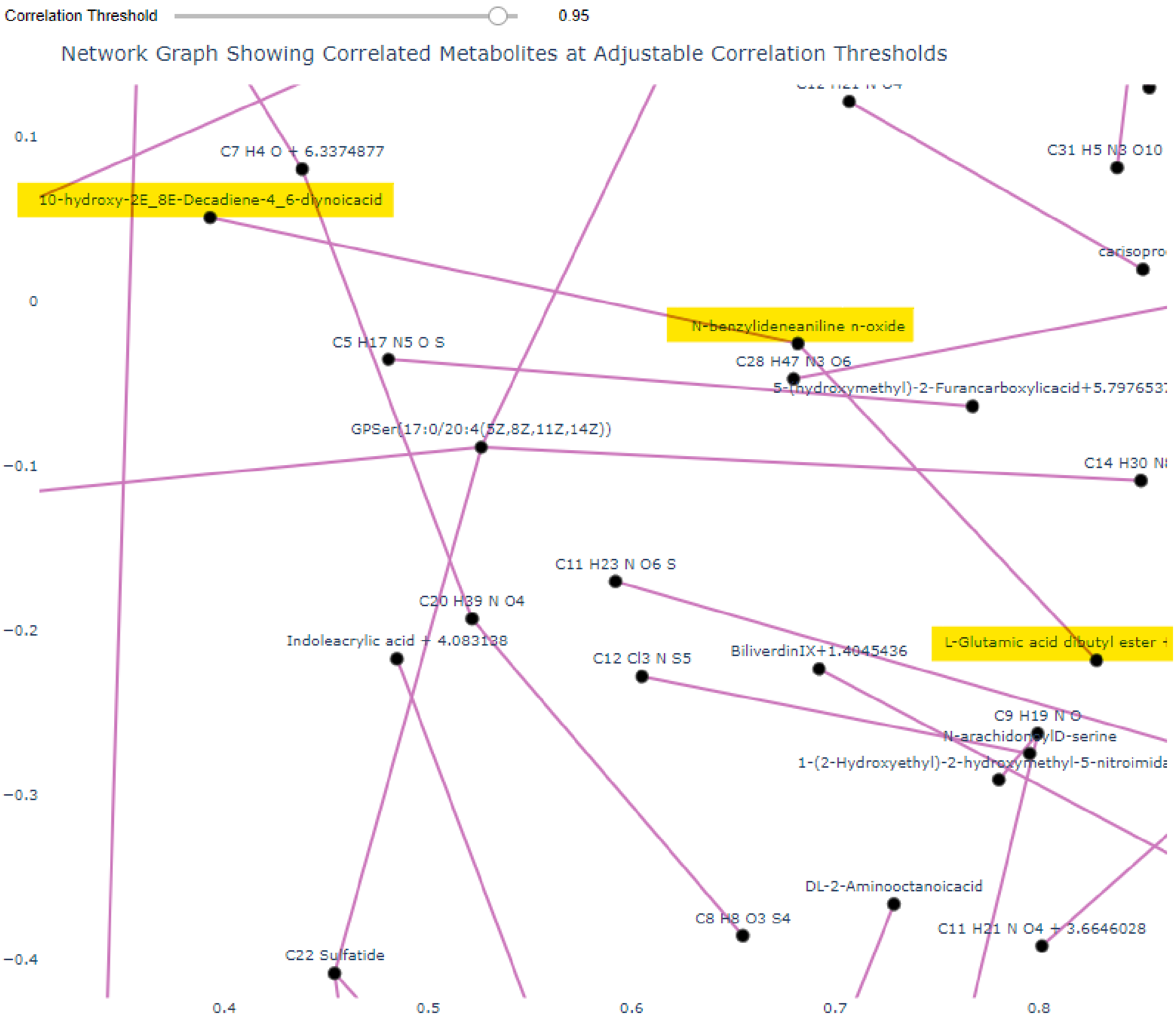


Figure 7 MCIvsCN Logistic Regression Feature Selection by Importance

Results

Due to the sheer volume of metabolites in the dataset, using the cluster map visualisation as a method for testing multicollinearity was unsuccessful as it was impossible to determine exactly which metabolites any visual clusters referred to. Through interactive features such as zoom and panning, the network graph (also available in Supplementary Document 2) allowed direct and visual access to see specific relationships between the correlations of specific metabolites which e.g., a direct relationship between L-Glutamic-acid-dibutyl-ester+4.857882, N-benzylideneaniline-n-oxide, and 10-hydroxy-2E_8E-Decadiene-4_6-diynoicacid (Figure 8). To find a numerical value for the multicollinearity a VIF test was run. This unfortunately resulted with all the variables being given a value of infinity – likely caused by the dimensions of the data set (1909 variables with only 45 observations).



Correlation Threshold 0.95

Figure 8 Network Graph (0.95 Correlation Threshold) Showing Relationship Between Three Named Metabolites

The maximum number of principal components that could be used by the PCA was 45, due to this being the number of observations in the dataset. Upon running, the model found that the first 10 components only explained 40.85% of the variance, in order to explain 95%, the first 40 components needed to be used. Upon assessing the visualisation's built by the model, the elbow plot (Figure 9) confirmed that there was a gradual increase in the variance explanation as the principal components increased - rather than an obvious difference which would identify an optimum number of components. The scatter graph (Figure 10)which was colour-coded by cognitive status showed that there were no clusters in the data.

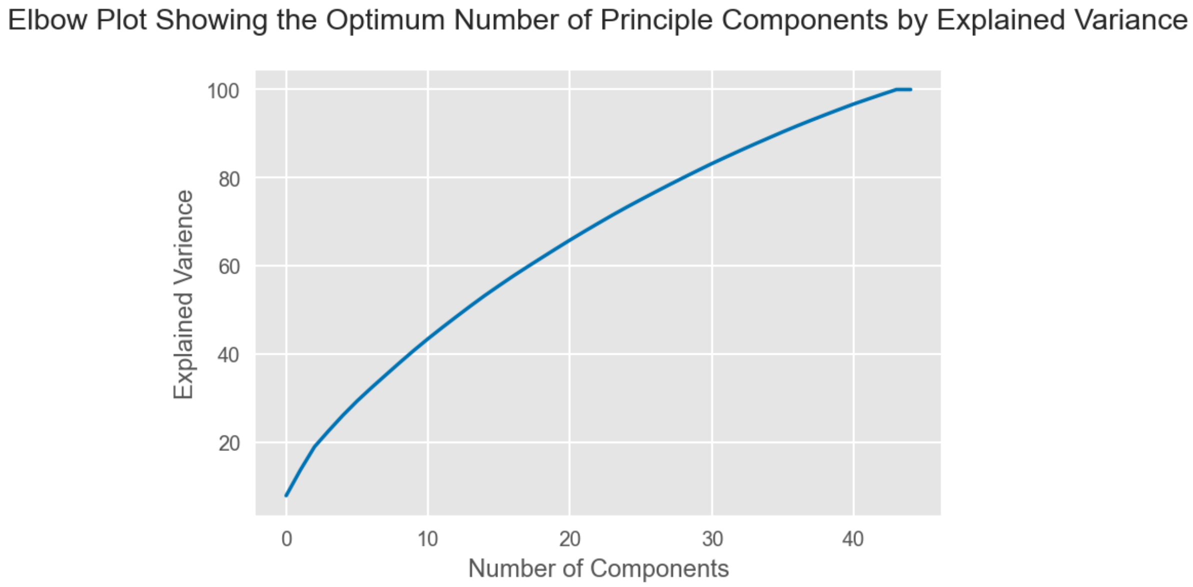


Figure 9 Principal Components Analysis Elbow Plot of Explained Variance

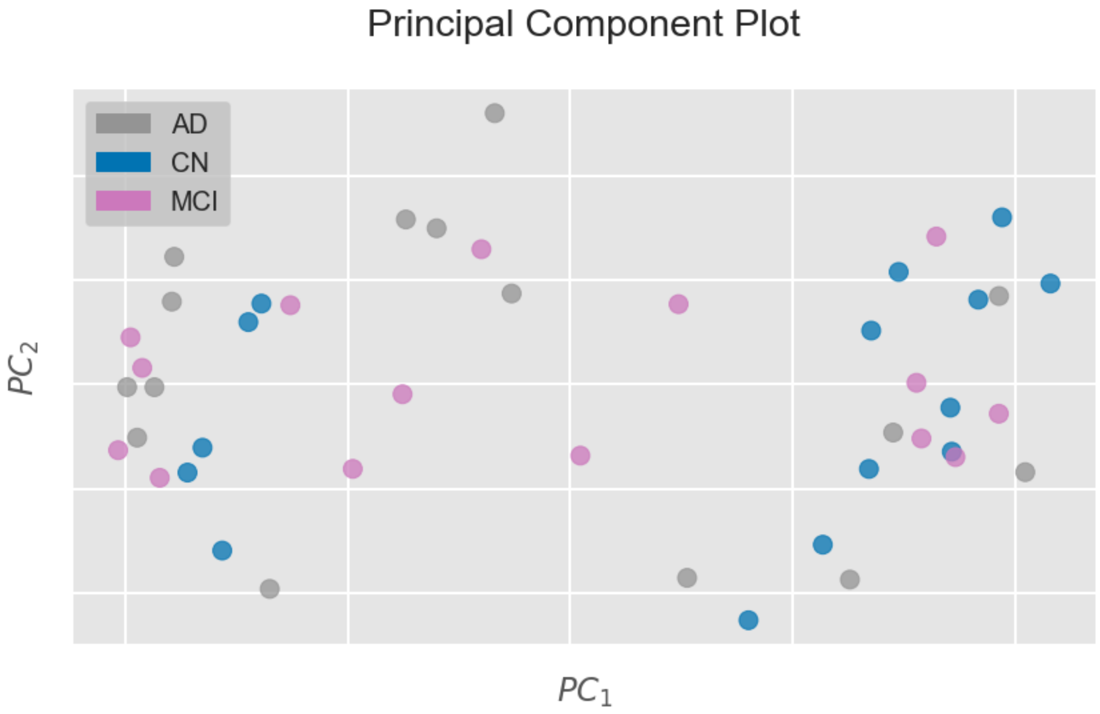


Figure 10 Scatter Plot of Principle Components by Cognitive Type

The initial visual comparison (Figure 11) of components prior to the PLS-DA of AD versus CN data suggested that there were not any obvious outliers between the two groups of clear points on the right-hand side of the graph where CN appears to spike higher than AD. Upon plotting the PLS regression scores (Figure 12), there is visible separation between AD and CN values on the latent variable (LV) 1 axis.

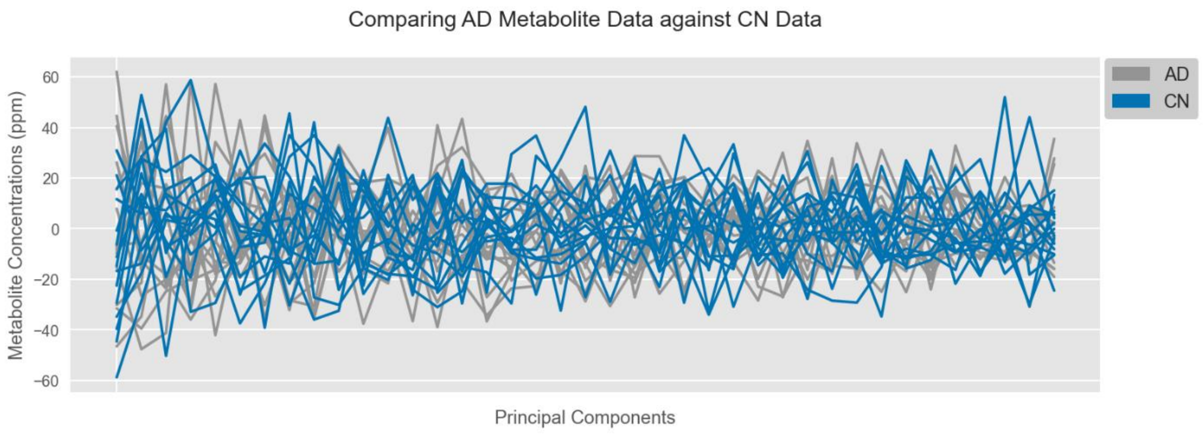


Figure 11 PLS-DA Comparing AD Metabolite Data against CN Data

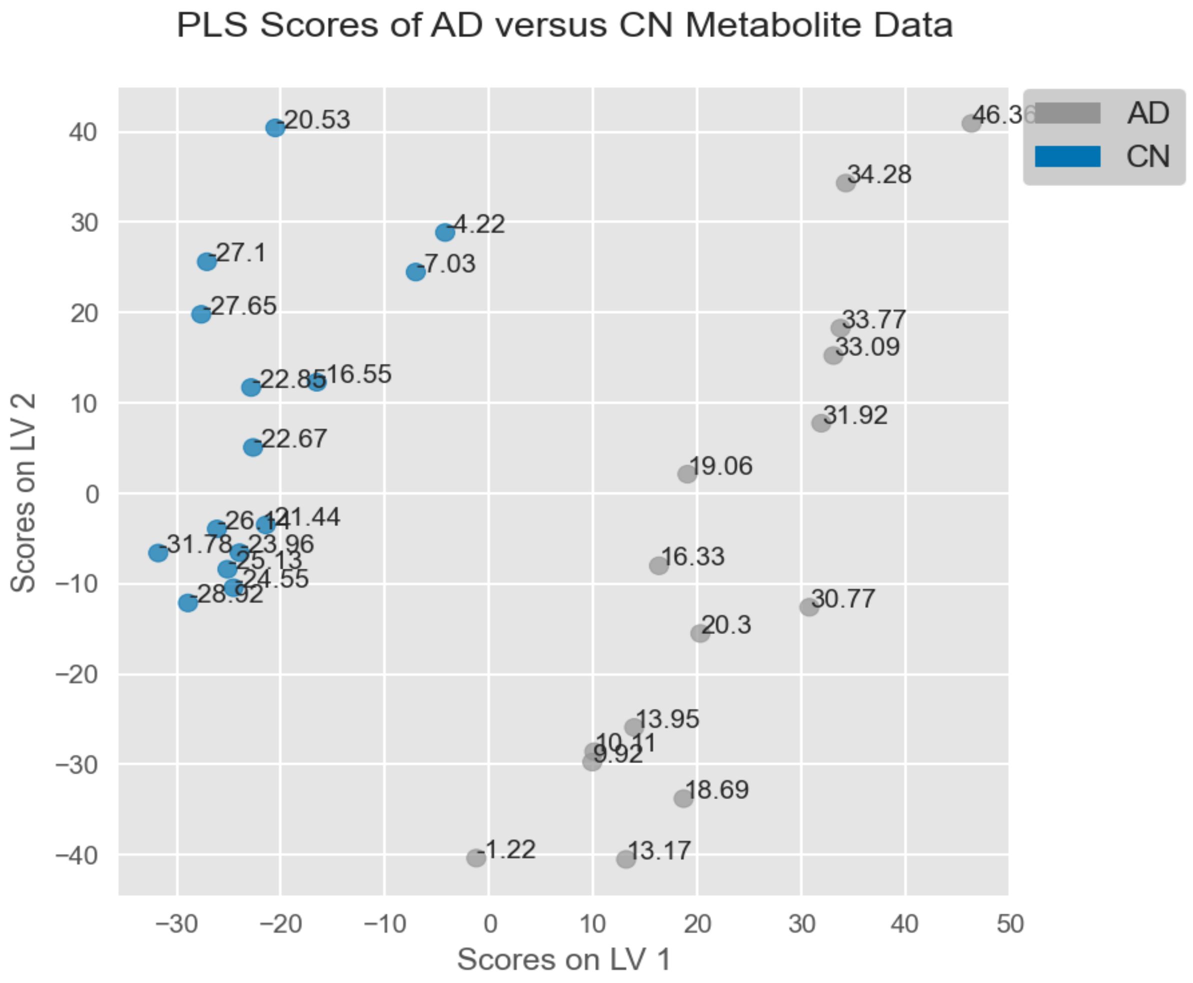


Figure 12 PLS Scores of AD versus CN Metabolite Data

Figure 13 shows the initial visual comparison of MCI metabolite data against CN data prior to the PLS-DA model, and whilst there are not any specifically obvious outliers there are at least three potential points where CN appears to spike at significantly different levels to MCI. These potential outliers however are not visible once the PLS regression scores have been plotted (Figure 14), and there is again clear visual separation between the two cognitive types on the LV1 axis.

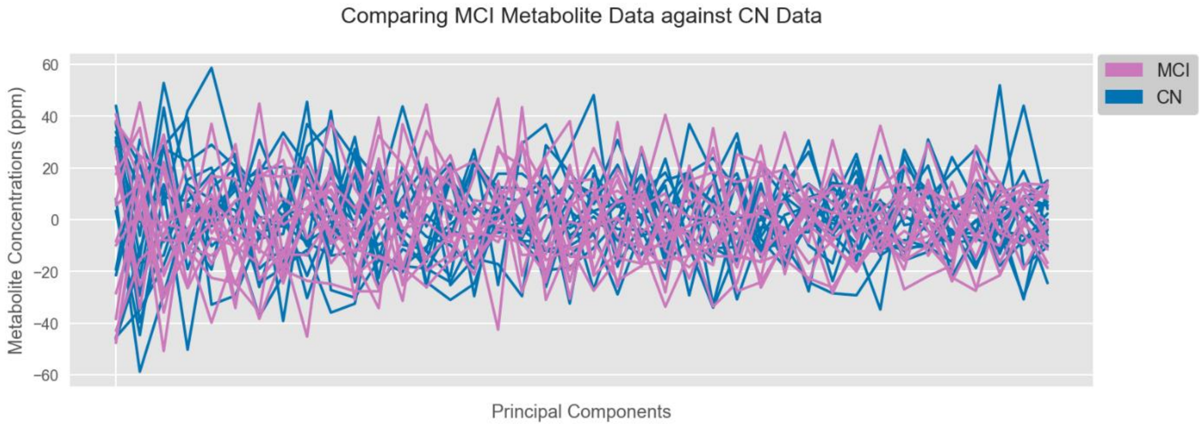


Figure 13 PLS-DA Comparing MCI Metabolite Data against CN Data

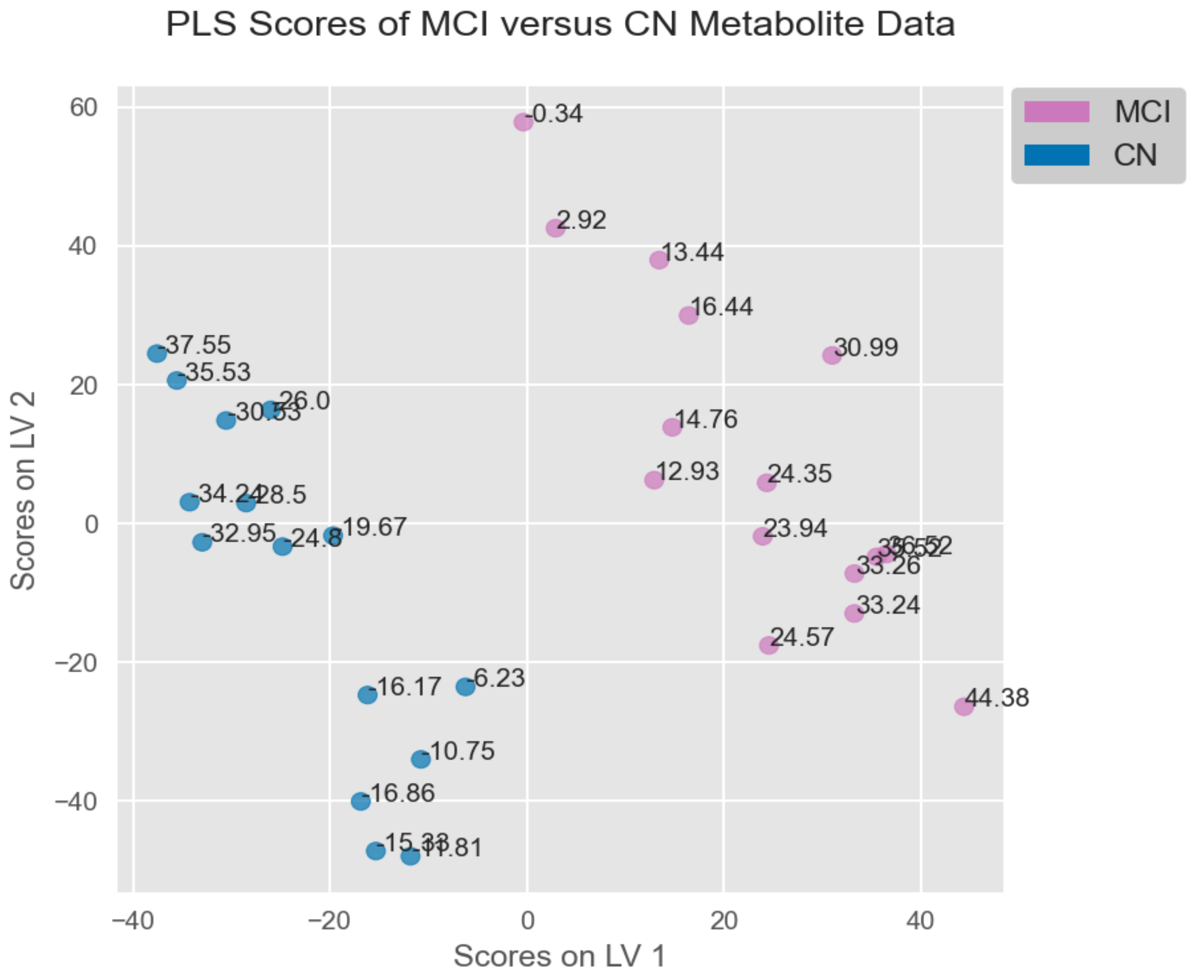


Figure 14 PLS Scores of MCI versus CN Metabolite Data

Finally, Figure 15 shows the initial visualisation between MCI and AD, and indicates that there are some potential outliers in the AD data which can specifically be seen on the left-hand side of the graph where there are at least four spikes that are higher than any other on the plot. Once again however, these outliers are not visible when the PLS regression scores are plotted against each other (Figure 16). This plot shows that there is clear and significant visible separation between the two cognitive types on the LV1 axis.

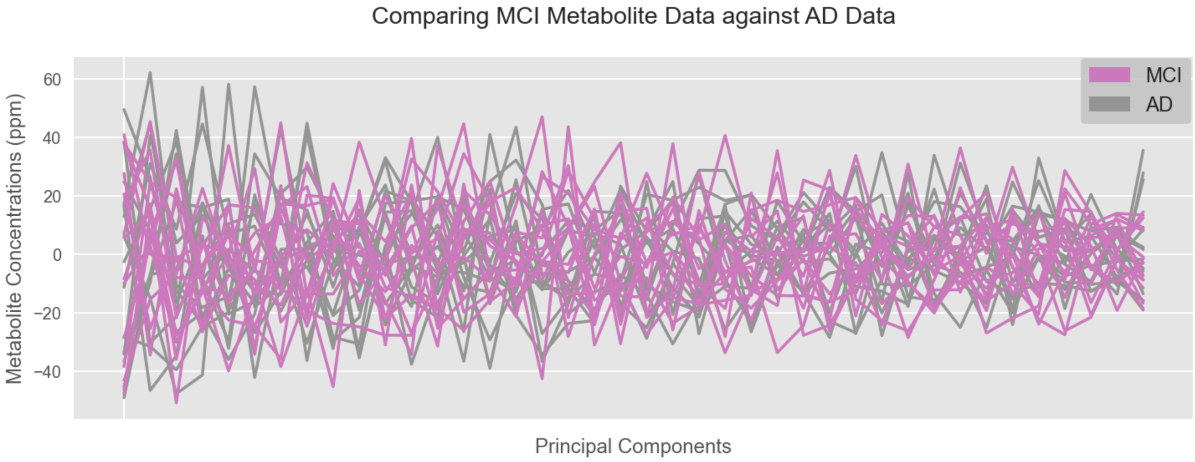


Figure 15 PLS-DA Comparing MCI Metabolite Data against AD Data

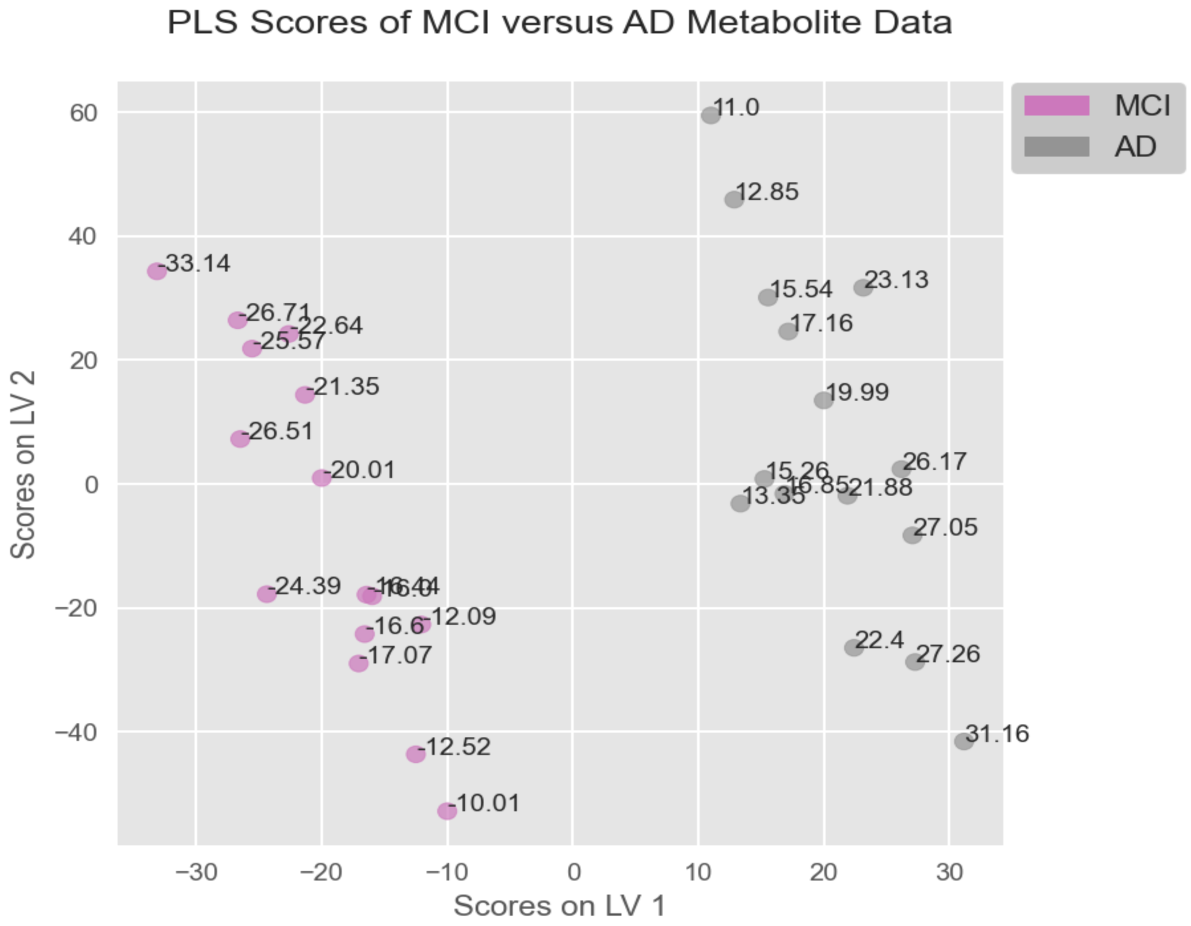


Figure 16 PLS Scores of MCI versus AD Metabolite Data

The original logistic regression model for AD vs CN showed a test accuracy using the classifier scores of 0.62. After Cross Validation testing it was found that the range of accuracy was between 0.34 and 1.0 with a mean of 0.57. This initial model found p-Aminobenzoic acid, GPSer(16:0/20:0), and Lys Lys Met to be the most significant metabolites with a positive coefficient, and C6 H12 N6 O3 + 3.3999805, C4 H9 Cl N2 O and C6 H12 N6 O3 to be the most significant metabolites with a negative coefficient (further details can be seen on Tables 1 and 2).

Index	Metabolite	Coefficient
621	p-Aminobenzoic acid	0.031606
1900	GPSer(16:0/20:0)	0.027982
1422	Lys Lys Met	0.026026

Table 1 AD vs CN Logistic Regression 1 - top 3 positive metabolites

Index	Metabolite	Coefficient
846	C6 H12 N6 O3 + 3.3999805	-0.028289
620	C4 H9 Cl N2 O	-0.027782
847	C6 H12 N6 O3	-0.025803

Table 2 AD vs CN Logistic Regression 1 - top 3 negative metabolites

The updated logistic regression model post feature selection had an initial accuracy level of 1.0, and after Cross Validation testing it was found that the range was between 0.5 and 1.0 with a mean of 0.94. This model found C29 H54 O3 S2, p-Aminobenzoic acid and Dihydrospatheliachromene + 7.7589583 to be the most significant metabolites with a positive coefficient (Table 3). The most significant metabolites with a negative coefficient were Desmethyldeschlorobenzoyl Indomethacin, 3-HYDROXYCAPRIC acid and Bis (2-hydroxypropyl) amine + 1.2051747 (Table 4).

Index	Metabolite	Coefficient
18	C29 H54 O3 S2	0.306401
3	p-Aminobenzoic acid	0.236046
22	DIHYDROSPATHELIACHROMENE + 7.7589583	0.196452

Table 3 AD vs CN Logistic Regression 2 - top 3 positive metabolites

Index	Metabolite	Coefficient
40	Desmethyldeschlorobenzoyl Indomethacin	-0.221207
46	3-Hydroxycapric acid	-0.218719
7	Bis (2-hydroxypropyl) amine + 1.2051747	-0.157478

Table 4 AD vs CN Logistic Regression 2 - top 3 negative metabolites

The initial logistic regression model for MCI vs CN had an accuracy using the classifier scores of 0.50. After Cross Validation testing it was found that the range of accuracy was between 0.34 and 1.00 with a mean of 0.54. It found Lys Lys Met, C26 H44 N22 and 2-Amino-3-methyl-1-butanol + 1.0206223 to be the most significant metabolites with a positive coefficient (Table 5). The most significant metabolites of the negative coefficient were C6 H12 N6 O3 + 3.3999805, C20 H32 N2 O20 S3 and VALDECOXIB + 1.1699699 (Table 6).

Index	Metabolite	Coefficient
1422	Lys Lys Met	0.040013
1841	C26 H44 N22	0.029207
542	2-Amino-3-methyl-1-butanol + 1.0206223	0.028181

Table 5 MCI vs CN Logistic Regression 1 - top 3 positive metabolites

Index	Metabolite	Coefficient
846	C6 H12 N6 O3 + 3.3999805	-0.043538
1881	C20 H32 N2 O20 S3	-0.030086
1157	VALDECOXIB + 1.1699699	-0.024109

Table 6 MCI vs CN Logistic Regression 1 - top 3 negative metabolites

The updated logistic regression model after feature selection had an initial accuracy level of 1.0, and after Cross Validation testing it was found that the range was between 0.5 and 1.0 with a mean of 0.9. This model found C13 H15 N O, C26 H49 N3 O and Visnagin to be the most significant metabolites with a positive coefficient (Table 7). The most significant metabolites of the negative coefficient were C6 H12 N6 O3 + 3.3999805, C5 H4 O8 S2 and C10 H15 N3 O3 (Table 8).

Index	Metabolite	Coefficient
5	C13 H15 N O	0.235669
16	C26 H49 N3 O	0.219942
10	Visnagin	0.172322

Table 7 MCI vs CN Logistic Regression 2 - top 3 positive metabolites

Index	Metabolite	Coefficient
13	C6 H12 N6 O3 + 3.3999805	-0.340717
30	C5 H4 O8 S2	-0.184574
11	C10 H15 N3 O3	-0.166148

Table 8 MCI vs CN Logistic Regression 2 - top 3 negative metabolites

Discussion

None of the initial 12 metabolites found by the logistic regression classifier for AD versus CN overlap with the named metabolites found in the literature. Of the top 50 most important metabolites found by the feature importance testing, only Indoleacrylic Acid overlapped with any of the metabolites in Supplementary Table 1. This same metabolite was found and named as significant by Trushina et al., (2013). There also did not appear to be any overlap from the initial 12 metabolites for MCI versus CN, however $C_{10}H_{15}N_3O_3$ – also known as Pentanedioic Acid Imidazolyl Ethanamide may overlap with the Pentadecanoic Acid found by Wang et al., (2014). Of the 50 most important metabolites found by the feature importance testing, both Indoleacrylic Acid and Norcodine (also Trushina et al., 2013) overlapped with the metabolites in Supplementary Table 1.

One particularly interesting outcome was finding a large number of vitamin D and vitamin D3 derivatives amongst the metabolites with the highest values in the correlation matrix built in the initial testing phase. This stands out because both vitamin D derivatives and vitamin D3 derivatives were named by Trushina et al., (2013) as having a significant influence on predicting individuals with AD.

Whilst there are some indications of consistency with existing literature, only metabolites found by the paper from which this dataset originated from definitively shared any overlap. This suggests and supports the idea that there is no scientific consensus on which metabolites are directly relevant to AD.

More clear presence of consistency appears in a 2014 PCA on blood samples performed by Wang et al., whose score plot did not show separate clusters for the three cognitive groups, indicating a potential expectation that PCA performed on the dataset in this study would also lack of clusters. This was proved to be the case and was clearly visualised using the colour coded scatterplot.

Challenges

By choosing to use an existing dataset, this study faced several constraints, which included managing potential issues relating to quality, dimensions and overall size. In this instance the dimensions of the data caused issues, with so few individual participant samples in comparison to the number of metabolites some statistical tests such as VIF test and logistic regression using the stats models library were affected, which meant looking for other means to learn about the data.

The dataset only contained 45 of these samples (15 individuals per cognitive type) which, given that treatment studies are recommended to have at least 100 individuals just in the placebo group (Cummings 2017), appears to be very low. This obviously impacts the robustness of the results - a significant issue in a field where results across different studies are already lacking in consensus regarding significant metabolites. This lack of samples is also the cause of the dimensional issues mentioned above. It would be beneficial for future studies to use these analysis methods with datasets which have significantly larger sample groups.

The selection of metabolites was also limited. One could argue that until all 253,245 metabolites can be detected and quantified any metabolomics research results are potentially unreliable. In saying this however, balance must be maintained in the dimensions, increasing the number of metabolites without significantly increasing the number of observations (samples) will lead to statistical error.

This study was also limited by the usage of a pre-existing dataset. Aside from small sample sizes the creators of this dataset draw attention to the lack of diversity in the biological sex of their study participants (predominantly male). Whilst it is claimed that the selection of the individuals was designed to mimic experimental groups on “demographic factors”, there is no clarification on what this means, and no mention of age or race of the participants. It is now common knowledge in the medical field that individuals with different demographic characteristics (such as sex, age and race) present and experience various medical conditions differently (Song et al., 2019, Hu et al., 2021, Skurla et al., 2022). In some cases (Chester 2019) lack of inclusion of these characteristics has caused lasting harm. It is therefore our responsibility as producers of research to include these characteristics and acknowledge the limitations of not doing so.

With the field of metabolomics being so new, there is a clear lack of agreement on good practice when undertaking studies such as this one. This can lead to potential inconsistencies with results and also means that the results in this study may not be comparable to the results of other studies performed with different methods. One attempt designed to manage this issue was the use of the NS-kNN method by Lee and Styczynski (2018), and this study attempted to use such a method. The goal was to produce a more accurate imputation method than kNN which assumes the majority of missing values in a dataset are missing completely at random. The NS-kNN method addresses datasets with large proportions of missing data being not at random – a potentially common occurrence in a field that uses mass spectrometry in metabolite measurements.

Unfortunately, the NS-kNN code did not produce the expected results so the analyses for this project were conducted on data using the standard kNN method. It is recommended that future studies attempt to translate the Lee and Stycznski code from Matlab into Python as there are still clear theoretical benefits to using this technique.

Additional Notes

A number of papers read in the preparatory literature for this project indicated the potential for there to be links between AD and schizophrenia (White and Cummings 1996, Ribe et al., 2015, Rubin 2016, Cao et al., 2019, Huo et al., 2020). At the time of writing there was no publicly available datasets covering metabolites in individuals with schizophrenia therefore these links could not be tested. It is recommended for future research should such a dataset become available that this link be tested.

It is recommended that further investigation on the PLS – DA technique within Metabolomics as a whole be undertaken as it can be a risky technique to use by itself. As mentioned by Ruiz-Perez et al. (2020), a large enough dataset can be separatable but most of the separating hyper planes can be noise and thus would make the technique inappropriate. This seems particularly relevant given the typical characteristics of datasets in this field.

Concluding Remarks

Overall, the above discussion shows that whilst metabolomics for AD clearly shows promise, there remains a lack of consistency in methods and results. This is evidenced in the results in this paper, which echo the conflicting results found in existing literature. There is much opportunity for continued refinement and development of research methods in order to enable this field to add significant weight to our understanding of AD.

References

Boccard, J. and Rutledge, D.N. (2013). A consensus orthogonal partial least squares discriminant analysis (OPLS-DA) strategy for multiblock Omics data fusion. *Analytica Chimica Acta*, [online] 769, pp.30–39. doi:10.1016/j.aca.2013.01.022.

Brereton, R.G. and Lloyd, G.R. (2014). Partial least squares discriminant analysis: taking the magic away. *Journal of Chemometrics*, [online] 28(4), pp.213–225. doi:10.1002/cem.2609.

Cao, B., Wang, D., Pan, Z., Brietzke, E., McIntyre, R.S., Musial, N., Mansur, R.B., Subramanieapillai, M., Zeng, J., Huang, N. and Wang, J. (2019). Characterizing acyl-carnitine biosignatures for schizophrenia: a longitudinal pre- and post-treatment study. *Translational Psychiatry*, [online] 9(1). doi:10.1038/s41398-018-0353-x.

Chester, V. (2019). Autistic women and girls: under-recognised, under-researched and under-served. *Advances in Autism*, [online] 5(1), pp.1–1. doi:10.1108/AIA-01-2019-049.

Clish, C.B. (2015). Metabolomics: an emerging but powerful tool for precision medicine. *Cold Spring Harbor Molecular Case Studies*, [online] 1(1). doi:10.1101/mcs.a000588.

Cummings, J. (2017). Lessons Learned from Alzheimer Disease: Clinical Trials with Negative Outcomes. *Clinical and Translational Science*, [online] 11(2), pp.147–152. doi:10.1111/cts.12491.

Douaud, G., Groves, A.R., Tamnes, C.K., Westlye, L.T., Duff, E.P., Engvig, A., Walhovd, K.B., James, A., Gass, A., Monsch, A.U., Matthews, P.M., Fjell, A.M., Smith, S.M. and Johansen-Berg, H. (2014). A common brain network links development, aging, and vulnerability to disease. *Proceedings of the National Academy of Sciences*, 111(49), pp.17648–17653. doi:10.1073/pnas.1410378111.

He, Y., Yu, Z., Giegling, I., Xie, L., Hartmann, A.M., Prehn, C., Adamski, J., Kahn, R., Li, Y., Illig, T., Wang-Sattler, R. and Rujescu, D. (2012). Schizophrenia shows a unique metabolomics signature in plasma. *Translational Psychiatry*, [online] 2(8), pp.e149–e149. doi:10.1038/tp.2012.76.

Hu, X., Chehal, P.K., Kaplan, C., Krukowski, R.A., Lan, R.H., Stepanski, E., Schwartzberg, L., Vidal, G. and Graetz, I. (2021). Characterization of Clinical Symptoms by Race Among Women With Early-Stage, Hormone Receptor–Positive Breast Cancer Before Starting Chemotherapy. *JAMA Network Open*, [online] 4(6), pp.e2112076–e2112076. doi:10.1001/jamanetworkopen.2021.12076.

Human Metabolome Database. (n.d.). *Database Statistics: Metabolite Statistics*. [online] Available at: https://hmdb.ca/statistics [Accessed 3 Aug. 2022]. HMD Release 5.0 - January 2022.

Human Metabolome Database. (2021). *Human Metabolome Database: Showing metabocard for Hypoxanthine (HMDB0000157)*. [online] Available at: https://hmdb.ca/metabolites/HMDB0000157 [Accessed 28 Jun. 2022].

Huo, Z., Yu, L., Yang, J., Zhu, Y., Bennett, D.A. and Zhao, J. (2020). Brain and blood metabolome for Alzheimer's dementia: findings from a targeted metabolomics analysis. *Neurobiology of Aging*, [online] 86, pp.123–133. doi:10.1016/j.neurobiolaging.2019.10.014.

Illig, T., Gieger, C., Zhai, G., Römisch-Margl, W., Wang-Sattler, R., Prehn, C., Altmaier, E., Kastenmüller, G., Kato, B.S., Mewes, H.-W., Meitinger, T., de Angelis, M.H., Kronenberg, F., Soranzo, N., Wichmann, H.-E., Spector, T.D., Adamski, J. and Suhre, K. (2010). A genome-wide perspective of genetic variation in human metabolism. *Nature Genetics*, [online] 42(2), pp.137–141. doi:10.1038/ng.507 163 metabolites in plasma.

Kaddurah-Daouk, R. (2006). Metabolic Profiling of Patients with Schizophrenia. *PLoS Medicine*, [online] 3(8), p.e363. doi:10.1371/journal.pmed.0030363.

Lee, J.Y. and Styczynski, M.P. (2018). NS-kNN: a modified k-nearest neighbors approach for imputing metabolomics data. *Metabolomics*, [online] 14(12), p.153. doi:10.1007/s11306-018-1451-8.

Lee, M. and Hu, T. (2019). Computational Methods for the Discovery of Metabolic Markers of Complex Traits. *Metabolites*, [online] 9(4), p.66. doi:10.3390/metabo9040066.

Liigand, P., Kaupmees, K., Haav, K., Liigand, J., Leito, I., Girod, M., Antoine, R. and Kruve, A. (2017). Think Negative: Finding the Best Electrospray Ionization/MS Mode for Your Analyte. *Analytical Chemistry*, [online] 89(11), pp.5665–5668. doi:10.1021/acs.analchem.7b00096.

Manchester, M. and Anand, A. (2017). Chapter Two - Metabolomics: Strategies to Define the Role of Metabolism in Virus Infection and Pathogenesis. *Advances in Virus Research*, [online] 98, pp.57–81. Available at: https://www.sciencedirect.com/science/article/abs/pii/S0065352717300015 [Accessed 1 Jun. 2022].

MicroSolv Technology Corporation. (2012). *Understanding some of the differences between HILIC columns and TYPE-C columns – Tips & Suggestions – MicroSolv Technology Corporation*. [online] Available at: https://kb.mtc-usa.com/article/aa-00206/46/ [Accessed 12 Aug. 2022].

Miles, J. (2005). Tolerance and Variance Inflation Factor. *Wiley StatsRef: Statistics Reference Online*. [online] doi:10.1002/9781118445112.stat06593.

Mittelstrass, K., Ried, J.S., Yu, Z., Krumsiek, J., Gieger, C., Prehn, C., Roemisch-Margl, W., Polonikov, A., Peters, A., Theis, F.J., Meitinger, T., Kronenberg, F., Weidinger, S., Wichmann, H.E., Suhre, K., Wang-Sattler, R., Adamski, J. and Illig, T. (2011). Discovery of Sexual Dimorphisms in Metabolic and Genetic Biomarkers. *PLoS Genetics*, [online] 7(8), p.e1002215. doi:10.1371/journal.pgen.1002215 Metabolite inclusion criteria for 'Schizophrenia shows a unique metabolomics signature in plasma'.

Oliver, S.G., Winson, M.K., Kell, D.B. and Baganz, F. (1998). Systematic functional analysis of the yeast genome. *Trends in Biotechnology*, [online] 16(9), pp.373–378. doi:10.1016/s0167-7799(98)01214-1.

Petersen, R.C. (2004). Mild cognitive impairment as a diagnostic entity. *Journal of Internal Medicine*, [online] 256(3), pp.183–94. doi:10.1111/j.1365-2796.2004.01388.x.

Proitsi, P., Kim, M., Whiley, L., Simmons, A., Sattlecker, M., Velayudhan, L., Lupton, M.K., Soininen, H., Kloszewska, I., Mecocci, P., Tsolaki, M., Vellas, B., Lovestone, S., Powell, J.F., Dobson, R.J.B. and Legido-Quigley, C. (2016). Association of blood lipids with Alzheimer's Disease: A comprehensive lipidomics analysis. *Alzheimer's & Dementia*, [online] 13(2), pp.140–151. doi:http://dx.doi.org/10.1016/j.jalz.2016.08.003.

Ribe, A.R., Laursen, T.M., Charles, M., Katon, W., Fenger-Grøn, M., Davydow, D.S., Chwastiak, L., Cerimele, J.M. and Vestergaard, M. (2015). Long-term Risk of Dementia in Persons With Schizophrenia: A Danish Population-Based Cohort Study. *JAMA Psychiatry*, [online] 72(11), pp.1–7. doi:10.1001/jamapsychiatry.2015.1546.

Rubin, E. (2016). *The Relationship Between Schizophrenia and Dementia | Psychology Today*. [online] www.psychologytoday.com. Available at: https://www.psychologytoday.com/us/blog/demystifying-psychiatry/201603/the-relationship-between-schizophrenia-and-dementia#:~:text=Many%20had%20clinical%20features%20seen%20in%20persons%20with [Accessed 24 May 2022].

Ruiz-Perez, D., Guan, H., Madhivanan, P., Mathee, K. and Narasimhan, G. (2020). So you think you can PLS-DA? *BMC Bioinformatics*, [online] 21(S1). doi:10.1186/s12859-019-3310-7.

Schneider, L.S. and Sano, M. (2009). Current Alzheimer's Disease clinical trials: Methods and placebo outcomes. *Alzheimer's & Dementia*, [online] 5(5), pp.388–397. doi:10.1016/j.jalz.2009.07.038.

Skurla, S.E., Friedman, E.R., Park, E.R., Cannon, S., Kilbourne, G.A., Pirl, W.F. and Traeger, L. (2022). Perceptions of somatic and affective symptoms and psychosocial care utilization in younger and older survivors of lung cancer. *Supportive Care in Cancer*, [online] 30(6), pp.5311–5318. doi:10.1007/s00520-022-06926-6.

Song, T.-J., Cho, S.-J., Kim, W.-J., Yang, K.I., Yun, C.-H. and Chu, M.K. (2019). Sex Differences in Prevalence, Symptoms, Impact, and Psychiatric Comorbidities in Migraine and Probable Migraine: A Population-Based Study. *Headache: The Journal of Head and Face Pain*, [online] 59(2), pp.215–223. doi:10.1111/head.13470.

Szymańska, E., Saccenti, E., Smilde, A.K. and Westerhuis, J.A. (2011). Double-check: validation of diagnostic statistics for PLS-DA models in metabolomics studies. *Metabolomics*, [online] 8(S1), pp.3–16. doi:10.1007/s11306-011-0330-3.

Trushina, E., Dutta, T., Persson, X.-M.T., Mielke, M.M. and Petersen, R.C. (2013). Identification of Altered Metabolic Pathways in Plasma and CSF in Mild Cognitive Impairment and Alzheimer's Disease Using Metabolomics. *PLoS ONE*, [online] 8(5), p.e63644. doi:10.1371/journal.pone.0063644.

Wang, G., Zhou, Y., Huang, F.-J., Tang, H.-D., Xu, X.-H., Liu, J.-J., Wang, Y., Deng, Y.-L., Ren, R.-J., Xu, W., Ma, J.-F., Zhang, Y.-N., Zhao, A.-H., Chen, S.-D. and Jia, W. (2014). Plasma Metabolite Profiles of Alzheimer's Disease and Mild Cognitive Impairment. *Journal of Proteome Research*, [online] 13(5), pp.2649–2658. doi:10.1021/pr5000895.

White, K.E. and Cummings, J.L. (1996). Schizophrenia and Alzheimer's Disease: Clinical and pathophysiologic analogies. *Comprehensive Psychiatry*, [online] 37(3), pp.188–195. doi:10.1016/s0010-440x(96)90035-8.

Yu, Z., Kastenmüller, G., He, Y., Belcredi, P., Möller, G., Prehn, C., Mendes, J., Wahl, S., Roemisch-Margl, W., Ceglarek, U., Polonikov, A., Dahmen, N., Prokisch, H., Xie, L., Li, Y., Wichmann, H.-E., Peters, A., Kronenberg, F., Suhre, K. and Adamski, J. (2011). Differences between Human Plasma and Serum Metabolite Profiles. *PLoS ONE*, [online] 6(7), p.e21230. doi:10.1371/journal.pone.0021230 163 Metabolites in Plasma (NB: Some authors the same as Schizophrenia Data source. Potentially questionable credibility given authors referenced themselves).