Machine Learning Approaches to Dementia Biomarker Identification

“The Neuromancers”
Casey Dye, Tony Lan, Camaron Mangham

Context

Dementia is a condition with major impact across the globe. Dementia encompasses several conditions, with Alzheimer’s disease being the most prevalent form, that affect an individual's daily functioning with notable impact on memory and thinking. Over 55 million people worldwide live with one of these conditions; up to 24 million have Alzheimer’s disease specifically (Mayeux R, et al 2012). A majority of individuals with dementia live in low or middle income countries.

The impacts of dementia are realized in many ways. It is the seventh leading cause of death and a major cause of disability and dependency. Disproportionately affecting women, dementia causes higher disability-adjusted life years and mortality. Additionally, women provide a majority of care for those with dementia. The overall economic impact is estimated to be $1.3 Trillion (*Dementia* 2023). While there is no known cause of Alzheimer's Disease, the leading hypothesis is that it is caused by a combination of genetic and environmental factors (Mayeux R, et al 2012). As a higher proportion of the population ages, understanding dementia becomes crucial to reduce the impact it has worldwide.

Project Statement

Our aim is to further the understanding of dementia. Our current objective is to identify potential biomarkers using machine learning techniques that can be elaborated upon in future studies to combat dementia. To achieve this, we use a combination of bioinformatic and machine learning methods used in previous dementia studies to analyze genomic and proteomic data from the Aging, Dementia and Traumatic Brain Injury (TBI) Study.

Dataset

The dataset we will be using for this project was developed by the Allen Institute for Brain Science in consortium with the University of Washington and Kaiser Permanente Washington Health Research Institute. These organizations undertook a longitudinal cohort-based study known as the Adult Changes in Thought (ACT) study (*Aging, dementia and Traumatic Brain Injury Study*, *n.d.)*. The data used in our analysis comes from a sample within this broader study. This particular group of participants had either experienced at least one traumatic brain injury with loss of consciousness or were part of the similarity-matched control group (*TECHNICAL WHITE PAPER: OVERVIEW* 2017). For each participant, a post-mortem autopsy was performed that included dissection and banking of frozen brain tissue from fifteen regions. These tissues were used for immunohistochemistry, in situ hybridization, RNA sequencing, targeted proteomic analysis, quantification of free radical injury, gas chromatography-mass spectrometry, and immunoassays (*TECHNICAL WHITE PAPER: QUANTITATIVE DATA GENERATION* 2016).

Documentation on the study and an overview of the data can be found on the Aging, Dementia and Traumatic Brain Injury (TBI) Study website.

Related Work

We began our project by reviewing existing literature to determine methods approved by researchers when working with RNA sequencing (RNA-seq) data. The following papers inspired us as we planned our genetic analysis.

Arzouni et al. (2021) utilized the RNA sequencing data from the Aging, Dementia and Traumatic Brain Injury (TBI) Study for a subset of patients without a prior traumatic brain injury. The researchers utilized the Voom method within the Limma package along with a linear mixed model to identify differentially expressed genes. The top ten differentially expressed genes were used as features in four classification models to classify patients as having had Alzheimer's Disease or not. A random forest classifier was found to generalize best and score highest on the test set with an 83% accuracy.

Moura and de Oliveira (2021) utilized the same dataset to complete a similar analysis. Unlike Arzouni et al., the researchers utilized the entire dataset to compare dementia samples to those without dementia. The DESeq2 algorithm was employed to identify differentially expressed genes. Those genes were then used as features within brain region-specific classification models to identify the presence of dementia. They found that the decision trees performed better than the random forest models, but that different brain regions performed best depending on the evaluation metric one was interested in.

Fiorini et al. (2022) utilized the Aging, Dementia and Traumatic Brain Injury (TBI) Study as well, but completed a sex-stratified analysis focused on outcomes post-TBI comparing participants with prior TBI to those without prior TBI. To do this, they used the limma package to find differentially expressed genes. Researchers then completed a weighted gene co-expression network analysis to describe the ways those genes interact. They used gene modules found in that step to filter to those that correlated with dementia. They also identified significantly correlated gene expression and protein concentration relationships.

Guennewig et al. (2021) also used RNA sequencing data from various brain regions to identify differences between patients with Alzheimer's Disease and control patients. EdgeR was used to identify differentially expressed genes, which were then clustered using hierarchical clustering.

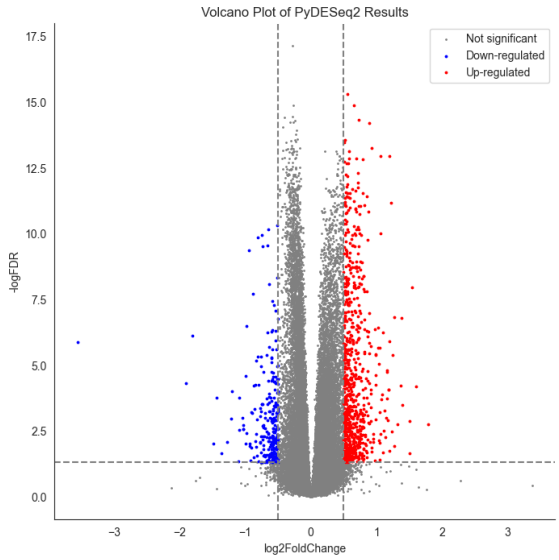
There have also been numerous studies evaluating the proteins associated with dementia. Bloom (2014) reviewed the pivotal studies that established Amyloid-β (Aβ) and Tau in the pathology of Alzheimer's Disease. Recent studies have discovered newer proteins that may be associated with dementia and Alzheimer's such as α-synuclein (Twohig 2019).

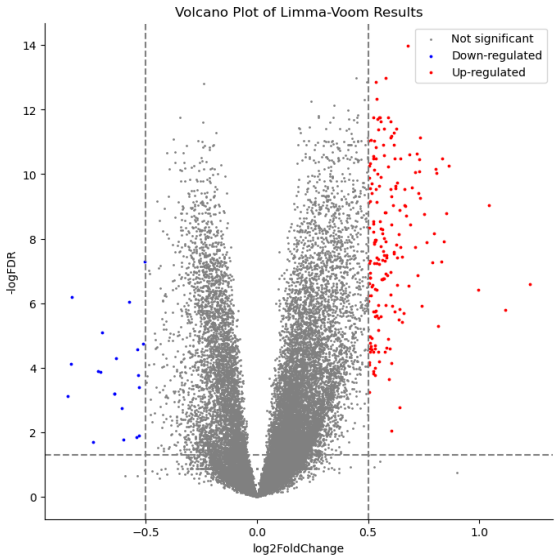
Data Processing
We began by downloading several csv files containing donor information, normalized RNA sequencing data, and protein quantification data from the Allen Brain Atlas Data Portal. The provided API was utilized to create the RNA sequencing raw counts matrix required for the differential gene expression analysis. We prepared this data for analysis by performing cleaning tasks that involved merging dataframes, data type transformation, and removal of some zero-valued rows.

What is Differential Gene Expression Analysis?
As we see in the related literature, determining differentially expressed genes (DEGs) is a common method used in the analysis of gene expression data. The goal is to perform statistical analysis on the read counts obtained from RNA sequencing to determine the quantitative changes in the expression level between two experimental groups. Key statistics provided by the analysis are the log2Fold change, describing the magnitude of overexpression/ underexpression for a given gene, in addition to providing a p-value describing the level of significance. Bcause DEGs can serve as potential biomarkers for various diseases or conditions, we use differential expression methods to compare samples from patients with dementia to those without dementia.

Gene Expression Analysis - Differential Expression
To determine differentially expressed genes, we compared two methods used in previous research studies. DESeq2 and limma-voom are two widely used tools in RNA-seq. Researchers often compare results from different methods to gain a deeper understanding of gene expression changes in their experiments. Because of this, we used both to identify differentially expressed genes from 50,281 genes. We used PyDESeq2, a python implementation of the DESeq2 workflow (Muzellec 2022) and RNAlysis, a python implementation of the limma-voom workflow. The process count matrix was used as the input to the algorithms. These values were normalized to reads per million for the limma-voom pipeline (Teichman, 2021) while DESeq2 performs size-factor based normalization (Love 2014). Genes with an adjusted p-value < .05 and an absolute log fold-change > 0.5 were considered to be differentially expressed to remain consistent with previous work. After filtering, around 1000 and 200 genes remained for DEseq2 and limma-voom, respectively. The top genes were defined by the ten genes with the highest and lowest Wald statistic in the case of DESeq2 and t-statistic in the case of limma voom (Smyth et al., 2002). The top twenty genes from each analysis were combined, resulting in 32 genes for downstream analysis after removal of overlaps.

The volcano plots below provide a concise and visually informative way to identify and prioritize genes that are significantly differentially expressed between two conditions based on both their statistical significance and the magnitude of fold-change. The log2FoldChange is the logarithm (usually base 2) of the ratio of the expression level in one condition to the expression level in another condition. The -logFDR is the negative logarithm of the p-value; more significant values are higher. Below we plot the results of both the pyDESeq2 and Limma-Voom analysis.



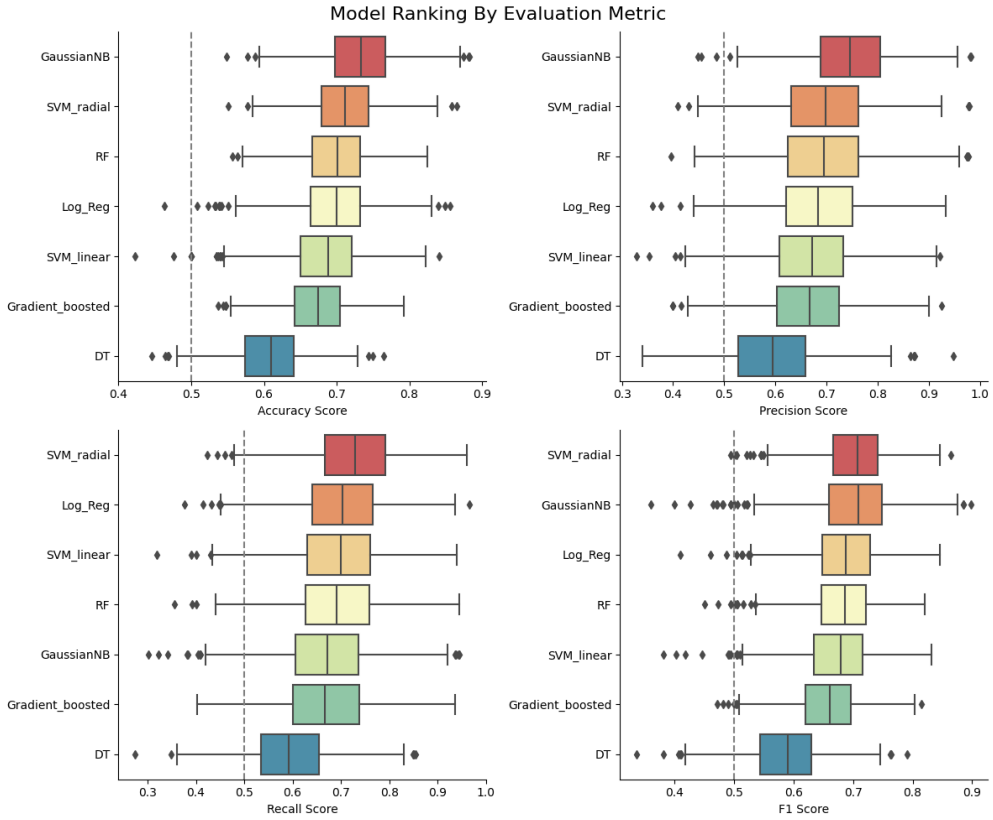


Gene Expression Analysis - Model and Gene Feature Evaluation

We used the 32 top genes as features in supervised classification models to identify the genes that may serve as most impactful biomarkers of the disease. We used several models that were used in similar studies (specifically, Support vector machines, random forests, and decision trees) in addition to using several popular classification models from the well known scikit-learn python package.

We begin by preparing the data appropriately for machine learning models. There are several samples per patient/donor, so we use a custom splitting function to separate the samples across training and testing sets by donor to prevent potential data leakage.

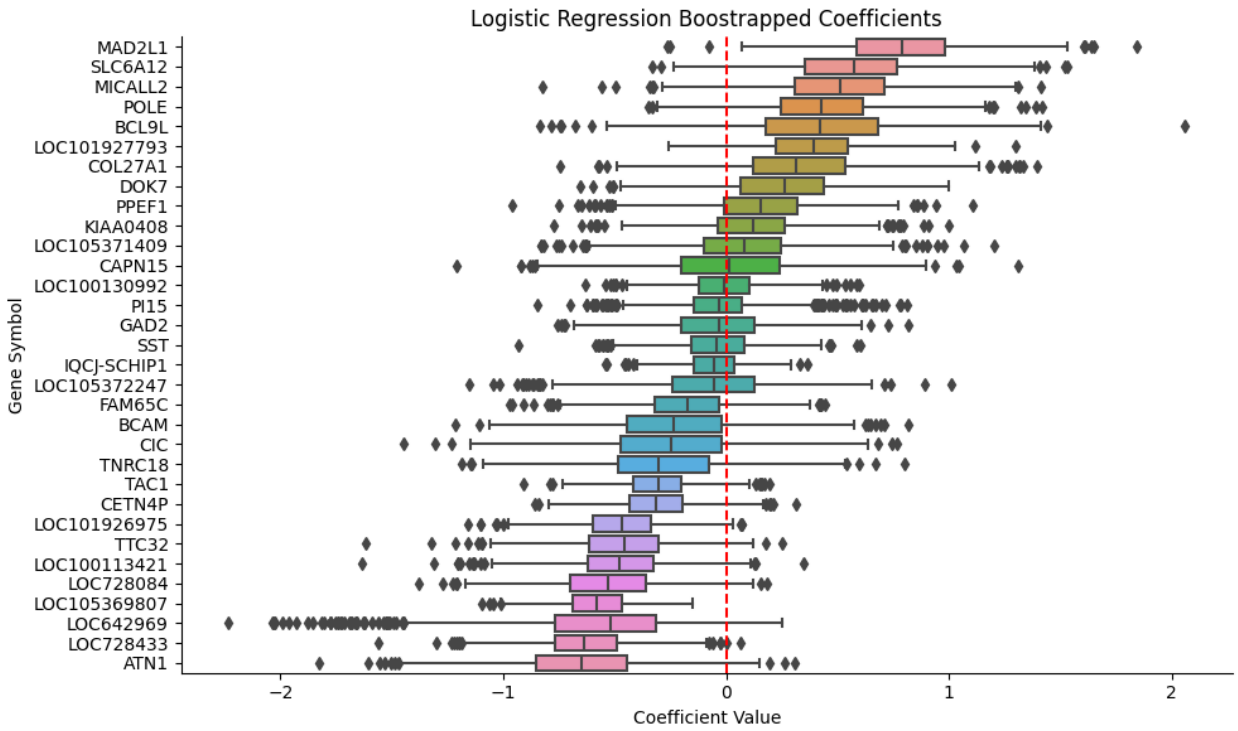
Principal component analysis of randomly sampled test train splits revealed high variability in the feature distribution of the test set. To evaluate the overall performance of each model, we train the base models on 1000 iterations of random data splits and evaluate their performance by accuracy, precision, recall, and f1 score.



Overall, the gaussian naive bayes model showed the best accuracy and precision scores, while the support vector machine with a radial kernel showed the best recall and F1 score.

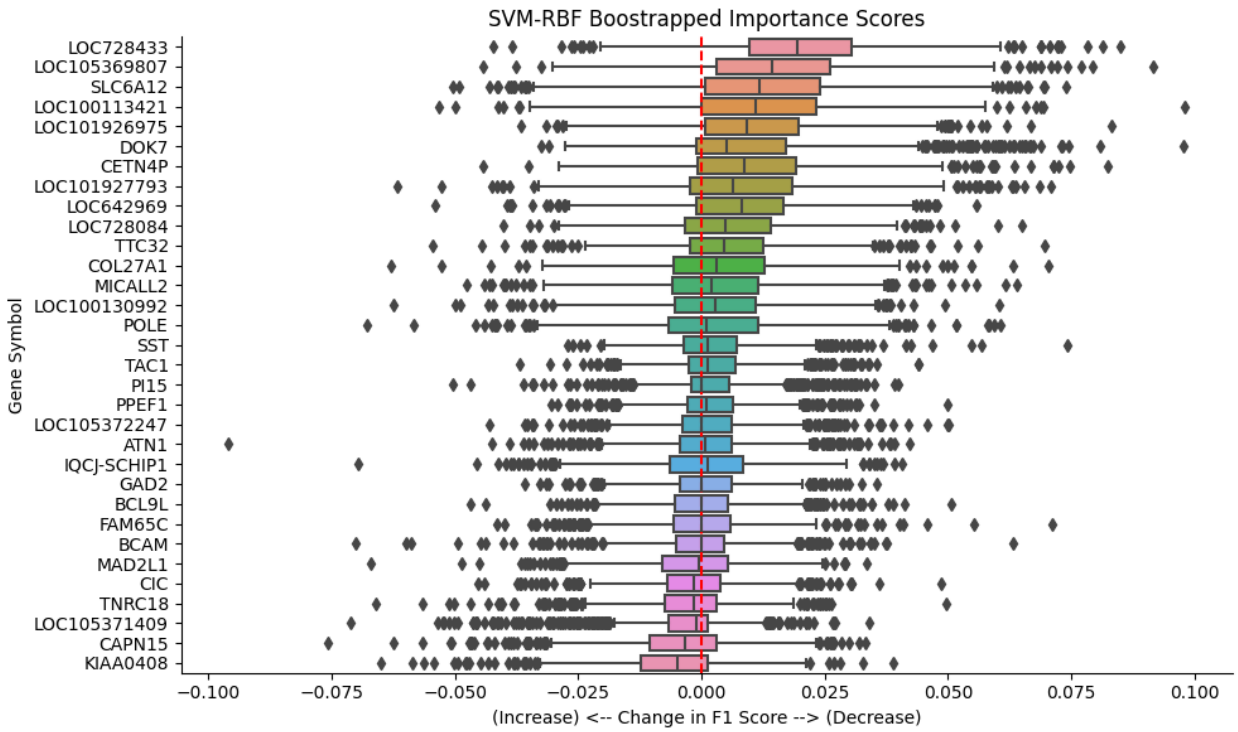
The logistic regression model is the most common classification model and is easily interpretable. So as a baseline, we take a look at the bootstrapped regression coefficients from this model on the training data. After training the model over many iterations (train F1 = 0.81, test F1 = 0.69, means), we visualize a mean ranked distribution of the bootstrapped coefficients. The highest positive coefficient is for the MAD2L1 gene (MAD2 mitotic arrest deficient-like 1). This logistic regression model suggests that higher read counts for this gene the model is more likely to label a sample as dementia. The lowest coefficient is LOC728433 (fibroblast growth factor 7 pseudogene). The model suggests that with higher read counts for this gene the model is less likely to label the sample as dementia.

More details about these genes can be found at the gene database resource of the National Center for Biotechnology Information (NCBI): MAD2L1; LOC728433



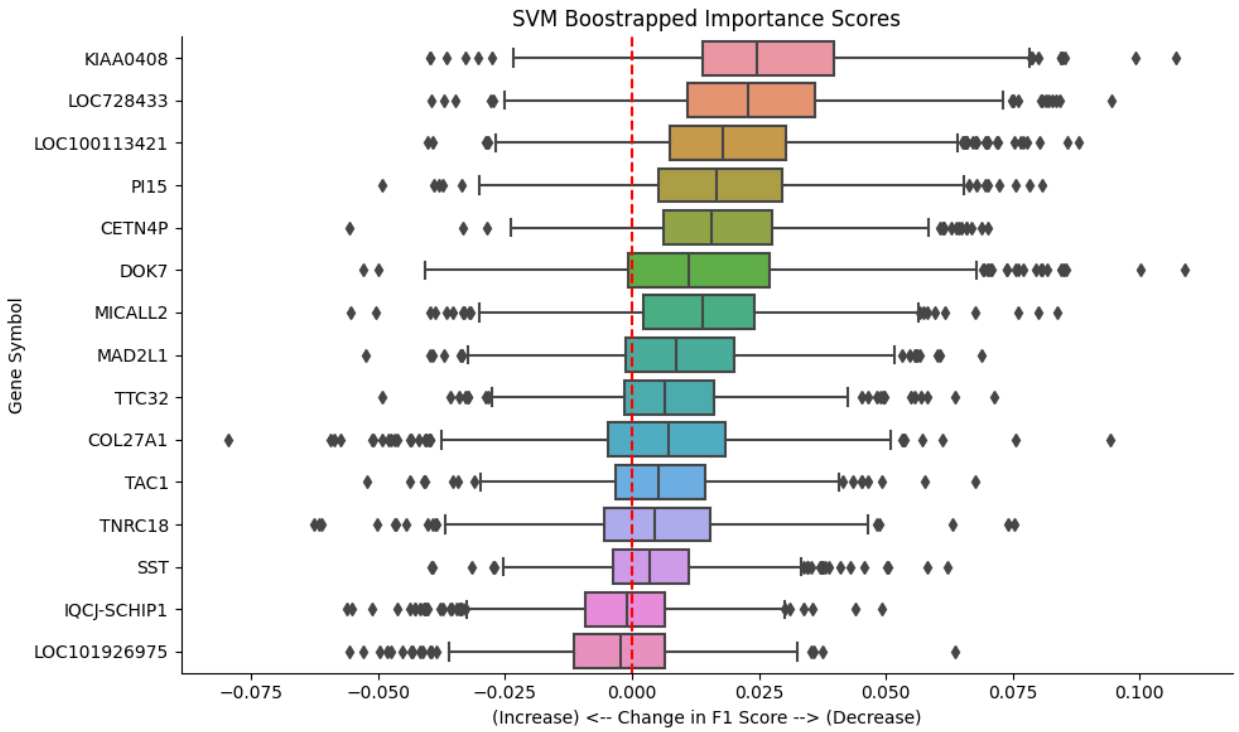
Overall, the model suggests that these genes and their closest ranked neighbors are important features in dementia classification on this dataset.

Next we perform a similar analysis on the two top performing models, Gaussian Naive Bayes (train F1 = 0.73, test F1 = 0.7, mean), and the Support Vector Machine radial kernel (train F1 = 0.85, test F1 = 0.7, mean). To evaluate the performance of these models, we utilize the sklearn “permutation feature importance” method to assess the impact each feature has on the overall score. Though the effect is small, both models show that the LOC728433 gene has an impact on the F1 score. In addition, the MAD2L1 gene, also suggested to be important by the logistic regression model, has little effect on the model overall all iterations. Overall all features appear to have minimal impact on the models.



We consider that collinearity could be a reason for the low feature importance found in the permutation analysis. A multicollinearity analysis reveals the relationship between groups of genes. Within the 32 genes there are 2 distinct groups with several subgroups. After thresholding and sampling a single gene from each cluster, the GaussianNB and SVM analysis was performed once more. Both analyses still show LOC728433 as an important gene, but KIAA0408, and LOC100113421 are also top contenders. Notably the scores of both models' performance improved slightly with the reduced set of genes (train F1 = 0.77, test F1 = 0.74, Gaussian NB) (train F1 = 0.85, test F1 = 0.72, SVM). See the appendix for additional figures.

More information regarding these genes: LOC728433, KIAA0408, LOC100113421



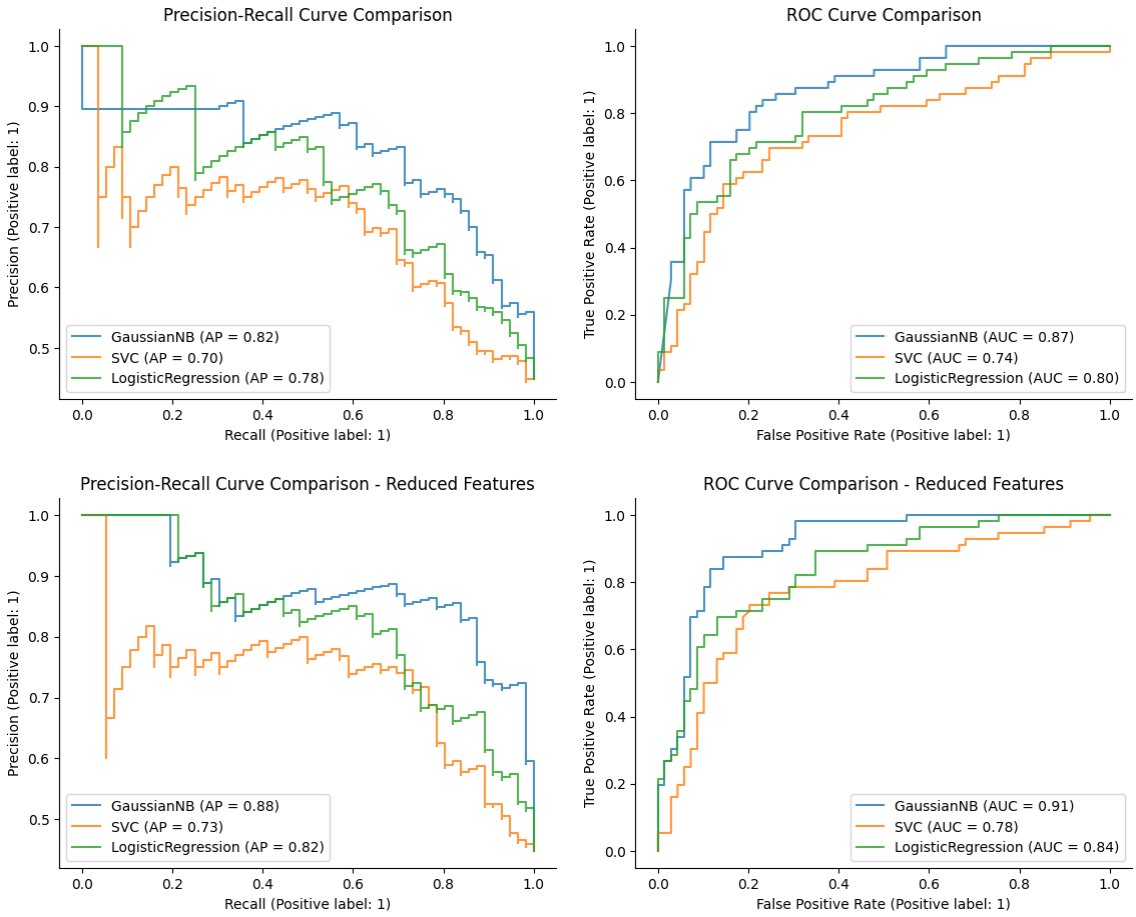
Gene Expression Analysis - Predictive Modeling for Dementia Detection

We also aimed to tune models for dementia detection and compare their performance with all 32 genes vs the reduced gene set found in the previous analysis. We summarize their performance with both precision-recall curves and ROC curves. After performing hyperparameter tuning for a logistic regression and SMV model on a static data split, the best models only showed minor improvement. Surprisingly, when reducing the feature set, the performance of each model improved a modest amount. The Gaussian Naive Bayes Model performed best of all models on these metrics.

We also attempted to tune other models such as the gradient boosted decision tree, but also resulted in worse performance on the holdout set. The boosted decision tree generally underfitt the validation data when iterating across learning rate and max depth.

It is important to note that this analysis was performed on a single static data split, and there is opportunity to examine the robustness of the models' performance by training and evaluating on many random samples.

Model Performance on Single Test-Train-Val Split



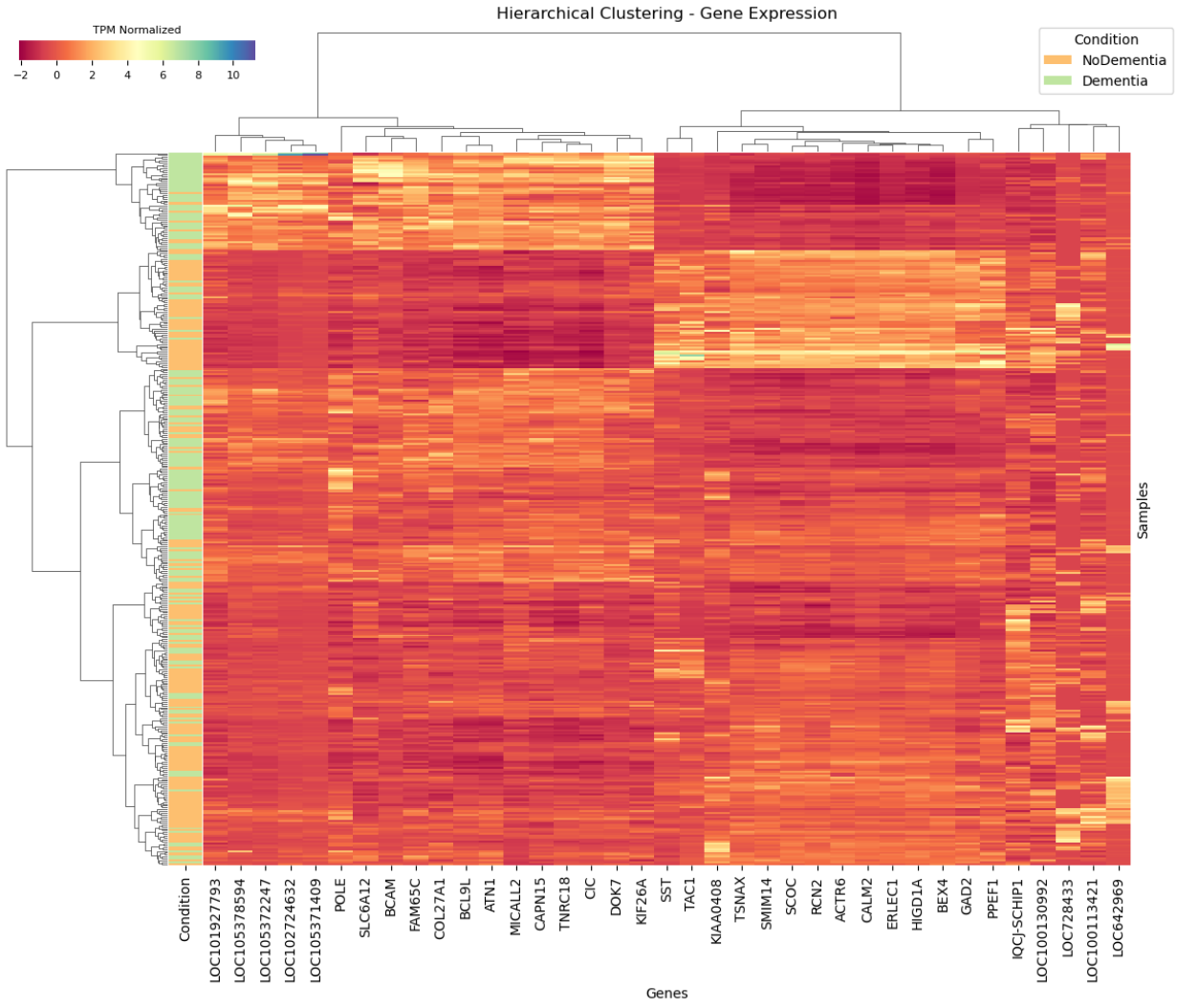
Gene Expression Analysis - Unsupervised Learning

We were interested in seeing whether the gene expression profiles for the differentially expressed genes would reveal additional patterns in the data. Like Guennewig et al. (2021), we performed hierarchical clustering, but, to build upon the work, we also attempted clustering using k-means, HDBSCAN, and k-medoids. We began with the full gene expression profiles for the 377 samples. We included the genes that were identified as most differentially expressed (both up-regulated and down-regulated). We chose the normalization method of TPM. TPM offers the benefit of normalizing for sequencing depth and gene length, but also allows you to compare between samples (*How to choose normalization methods (TPM/RPKM/FPKM) for mrna expression* 2023).

Gene Expression Analysis - Hierarchical Clustering

Hierarchical clustering was performed with z-scores calculated for each column on both the samples and genes and visualized in the clustermap below. When the dendrogram is cut to 4 clusters of samples, one group of samples that are mainly in the dementia group form their own branch. The remaining three clusters do seem to primarily contain samples within

the same status of dementia, but are branched together. When the dendrogram is cut to 4 clusters of genes, we can see groupings that exhibit similar expression patterns. This allows us to identify groups of samples and gene expression that are interesting. We can see that, for the clusters that primarily contain dementia samples, there is higher normalized TPM of 'LOC101927793', 'LOC105378594', 'LOC105372247', 'LOC102724632', 'LOC105371409', "POLE', 'SLC6A12', 'BCAM', 'FAM65C', 'COL27A1', 'BCL9L', 'ATN1', 'MICALL2', 'CAPN15', 'TNRC18', 'CIC', 'DOK7', and 'KIF26A' and lower normalized TPM of 'SST', 'TAC1', 'KIAA0408', 'TSNAX', 'SMIM14', 'SCOC', 'RCN2', 'ACTR6', 'CALM2', 'ERLEC1', 'HIGD1A', 'BEX4', 'GAD2', 'PPEF1'. The pattern is not as clear for the remaining genes, so more work will need to be done to identify why that is.

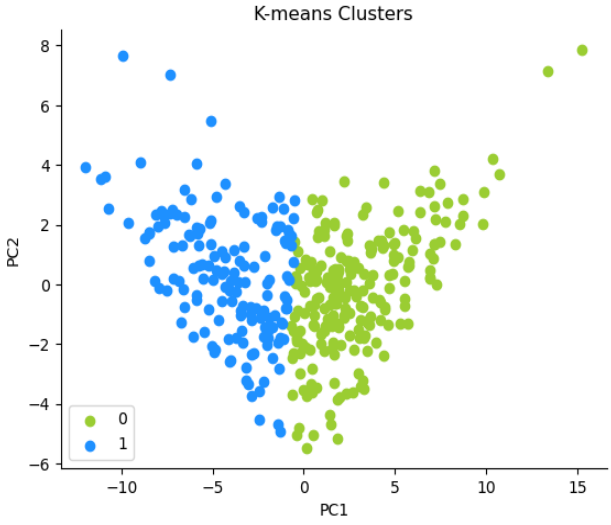


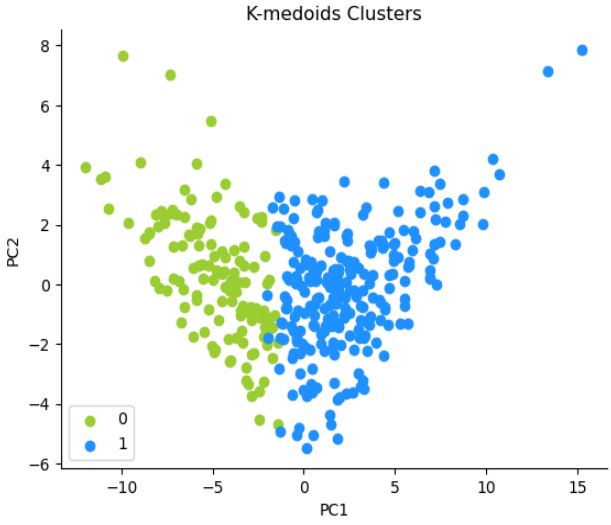
Gene Expression Analysis - Unsupervised Clustering of Samples

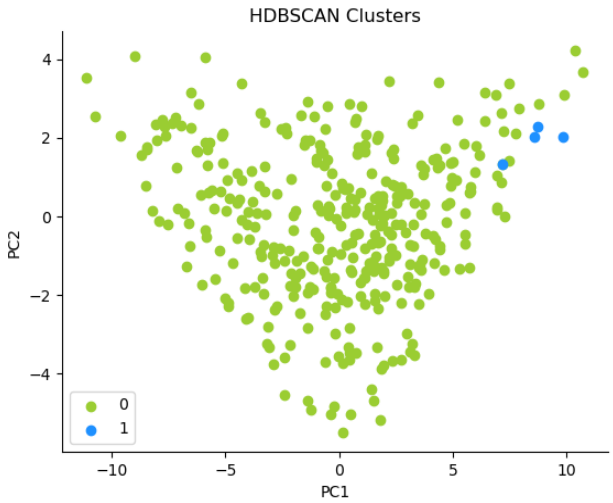
We also chose to perform clustering on samples using K-means, K-medoids and Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) to see whether any interesting patterns emerged. Prior to performing clustering, we used StandardScaler to standardize the data to a common scale. PCA was performed to visualize the clusters using the first two components.

We first needed to determine the appropriate number of clusters for each method. HDBSCAN does not require a specified number of clusters as the algorithm finds the optimal number of clusters by integrating DBSCAN over multiple epsilon values (scikit-learn developers, *n.d.*). Two clusters were returned using this methodology. For K-means and K-medoids, we looked at the Davies-Bouldin score, Calinski-Harabasz score, silhouette score, and the elbow method using inertia for cluster sizes of two to nineteen. Two clusters performed best for the three scores, so this was chosen as the optimal number of clusters.

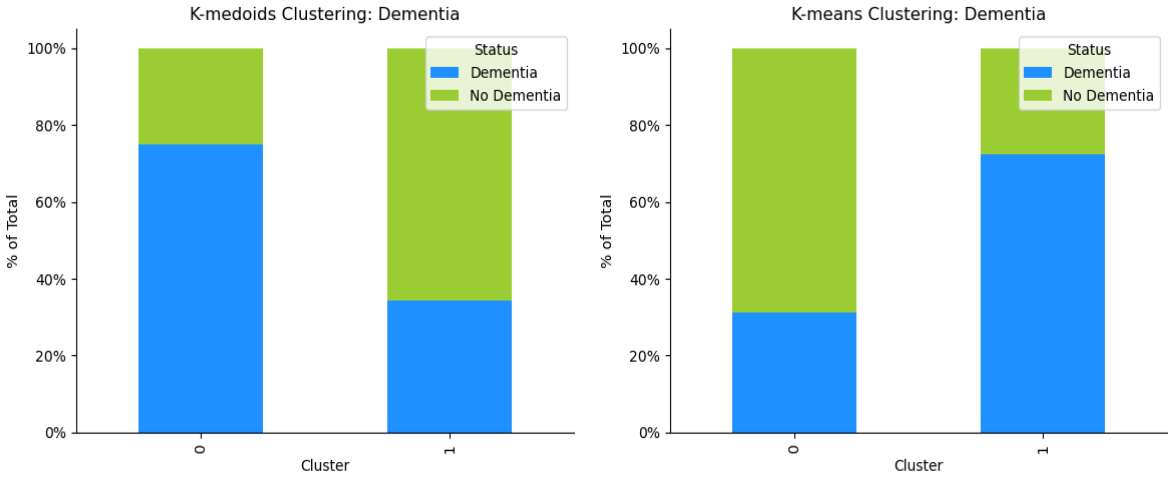
In comparing the three methodologies, we first visually inspected the clusters. HDBSCAN did not seem to find meaningful clusters, as most samples were clustered together with a very small second cluster returned. The clusters defined by K-means and K-medoids were very similar - the biggest difference seemed to be some points near the boundary. This was also validated by which scores each algorithm performed better on. K-medoids had a better silhouette score and Davies-Bouldin score whereas K-means had a better Calinski-Harabasz score.



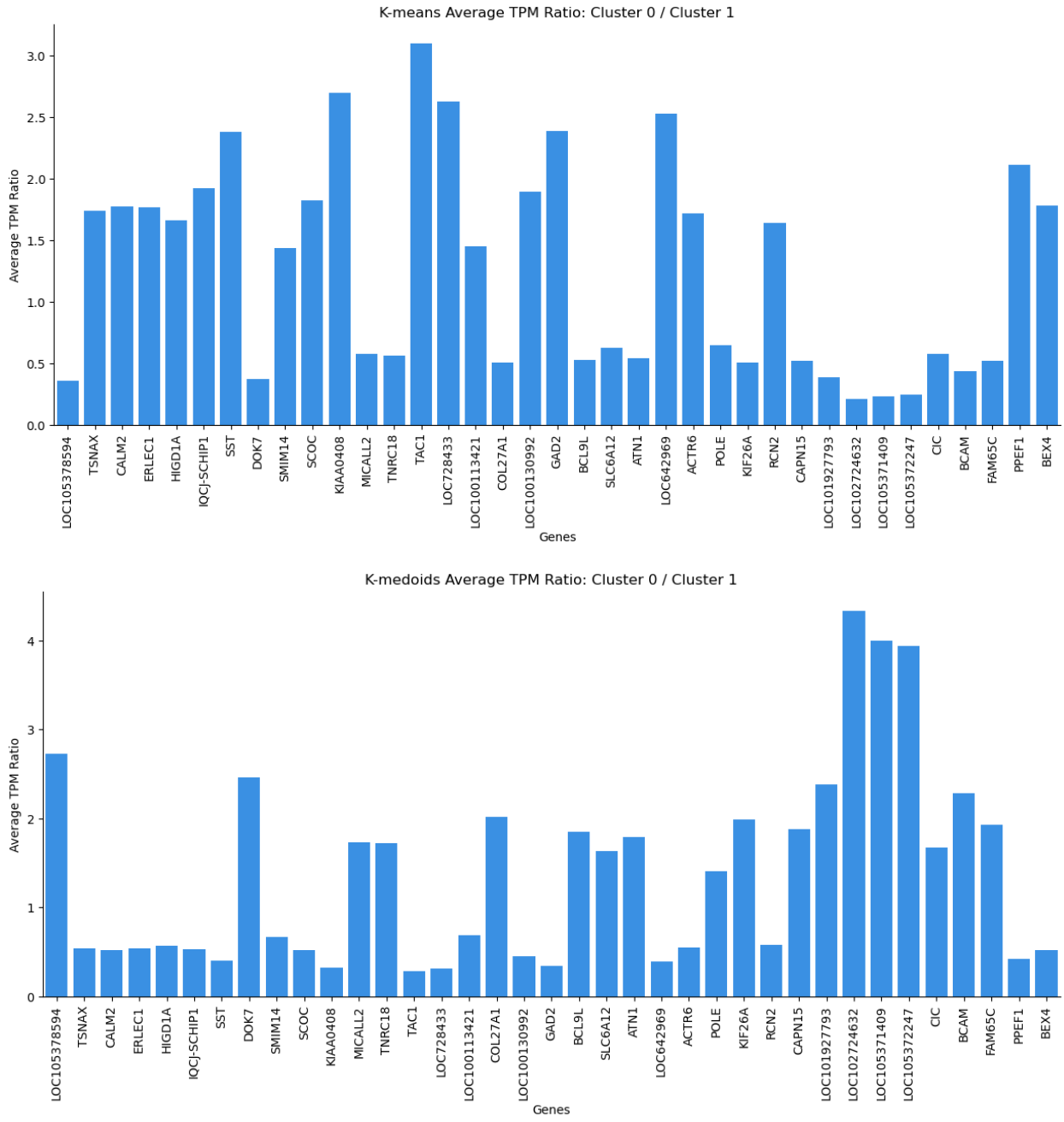




We compared the samples within each cluster returned by K-means and K-medoids to gain a better understanding of the composition of each cluster. We started by looking at various demographic differences. One particular feature that stood out was ‘act_demented’-this was our variable used to segment samples for the differential expression analysis. Because we used the genes determined to be differentially expressed for dementia, it would make sense that the clusters identify samples with dementia and those without. We can see this does occur somewhat, and both algorithms perform similarly. K-means cluster 1 had primarily dementia samples whereas K-medoids cluster 0 had primarily dementia samples. There are still some samples that end up in an unexpected cluster.



We also analyzed the ratio of each gene’s average expression in one cluster to the other by each algorithm. This seemed to be directionally the same although in an inverse fashion. We can see that the most drastic relative differences between the two clusters were seen in the 'LOC102724632', 'LOC105371409', 'LOC105372247', 'LOC105378594', 'DOK7', 'GAD2', 'LOC642969', 'LOC728433', 'KIAA0408', and 'TAC1' genes.



The results seem to confirm that the genes identified with the differential expression analysis can help to partition samples into clusters that contain a majority of samples with dementia or without dementia. More analysis could be done to identify the samples that end up in unexpected clusters to determine the gene expression that seems to be related to their cluster label.

Protein Data Analysis

Protein Analysis Objective

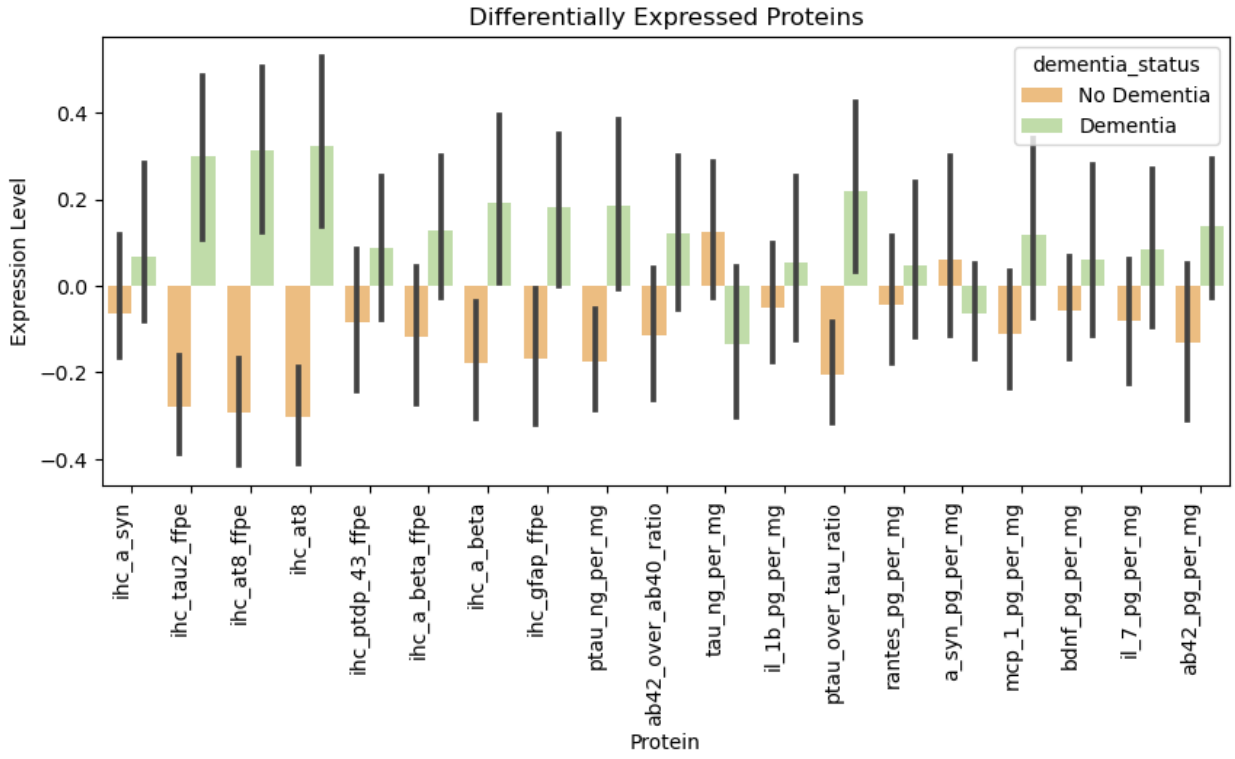
The primary objective of the protein analysis was to understand the importance of various proteins in dementia. A secondary objective was to examine any correlations between gene expression and protein expression for key proteins..

Protein Analysis Data Preparation

The protein data frame contained 377 samples from 107 donors. Initial examination of the data showed that there was missing data. One protein in particular, isoprostane, was missing data in 239 samples, and therefore dropped from the analysis. The remainder of the missing data was dropped resulting in a final dataset that was 74% of the original data. Dropping samples with missing data was chosen over imputing values since biological systems are extremely complex and variability between individuals could be significant. Imputation of missing data may not have accounted for these factors, potentially leading to inaccurate results. The final dataset had 279 samples and 28 protein features. It is important to distinguish between *proteins* and *protein features* in this dataset. There are 14 unique proteins (or family of proteins) represented in the data. However, a protein could be assessed in multiple ways.

Protein Features Differential Expression

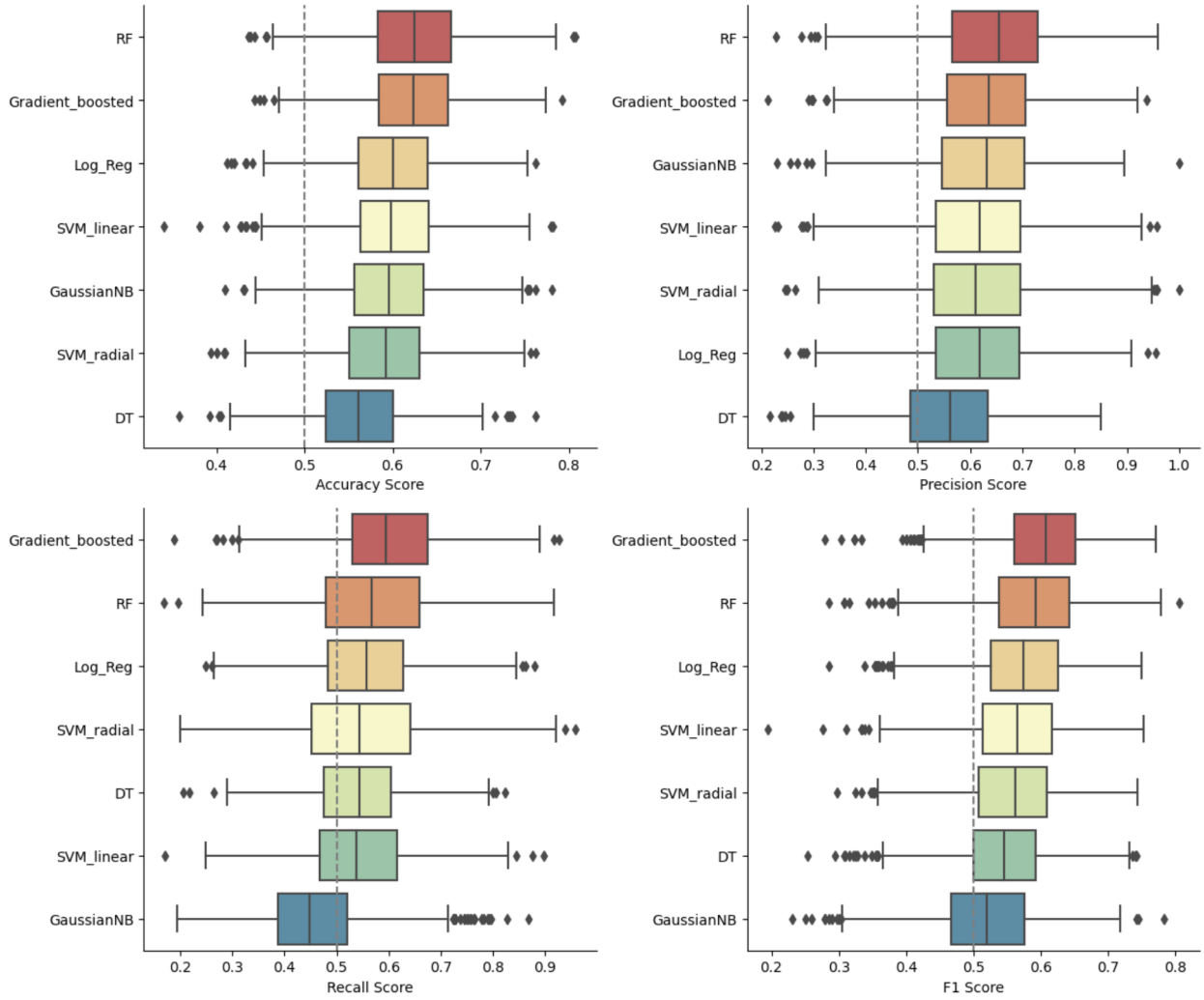
First we explored the data to see which protein features had significant differential expression between Dementia and No Dementia samples.



Of the 28 features in the protein dataset, 19 had significant differential expression ($P < 0.05$). All of the tau features (6/6) and the majority of amyloid features (3/5) are differentially expressed. Visually, the 3 features that stand out the most are IHC staining of Tau2 from FFPE (ihc_tau2_ffpe), IHC staining of anti-Tau antibody from FFPE (ihc_at8_ffpe), and IHC staining of anti-Tau antibody from fresh frozen tissue (ihc_at8). This is not surprising since Tau proteins are a well established biomarker for Alzheimer's Disease.

Protein Analysis Model Evaluation

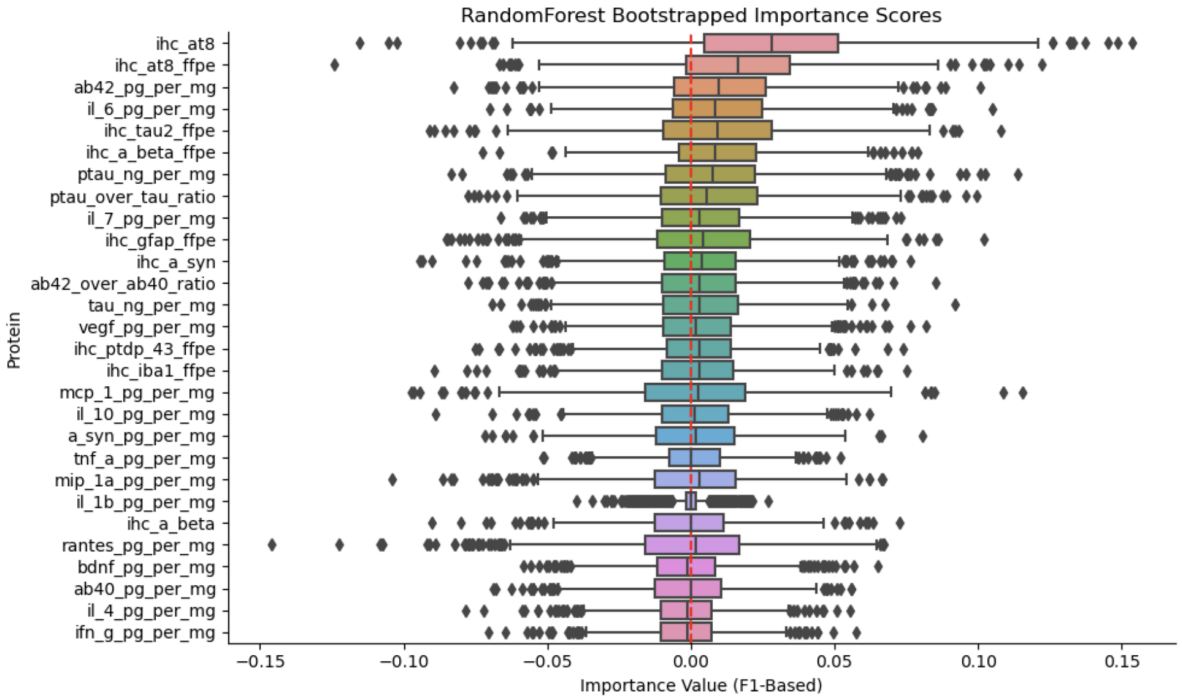
The same algorithms evaluated for gene analysis were used for protein analysis. Results are shown below.

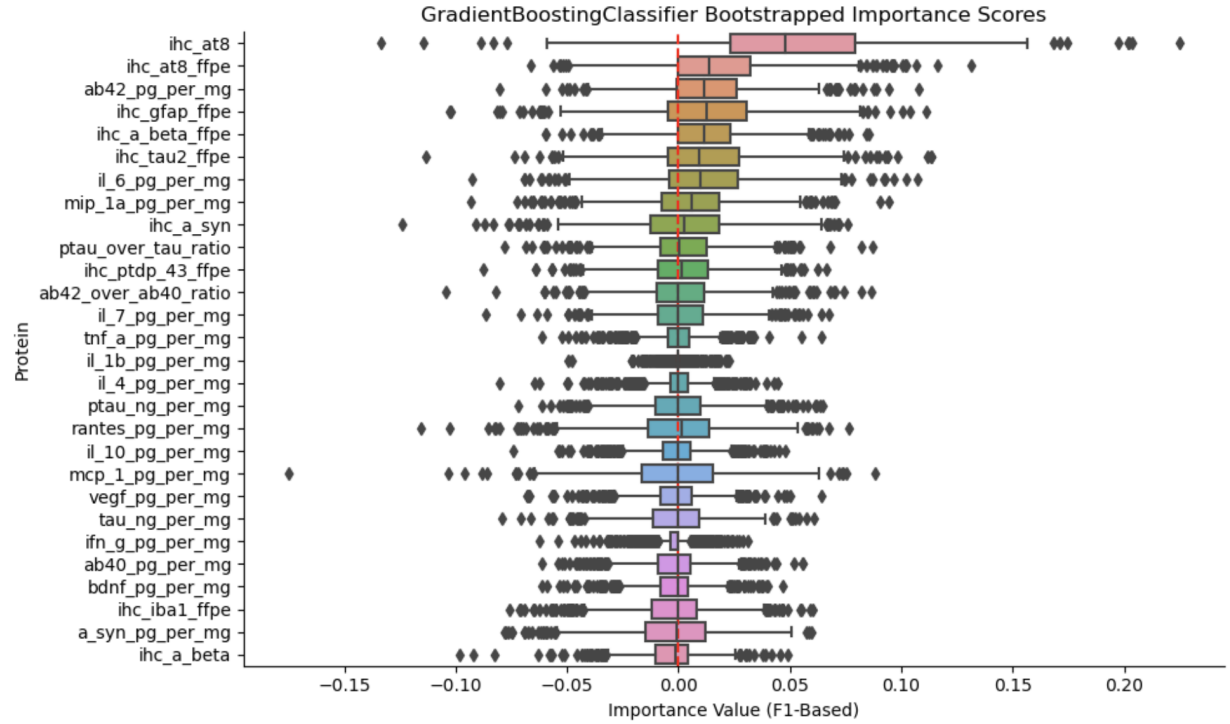


Random Forest and Gradient Boosting Classifiers achieved the highest scores across evaluation metrics and were used for further analysis.

Protein Analysis Feature Importance

To better understand each feature's importance in the classification algorithm, feature importance analysis was performed. The permutation importance method was used due to the high cardinality of the features.

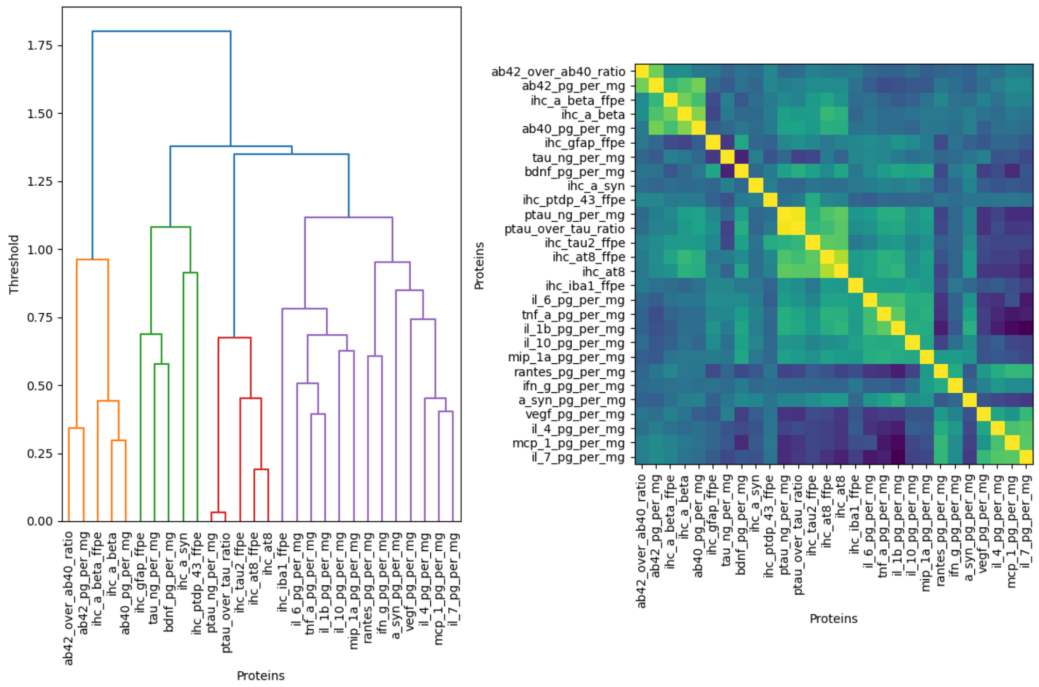




Results were consistent between the two models. With tau or amyloid related features representing 6 of the top 10 important features in each model.

Protein Analysis Multicollinearity

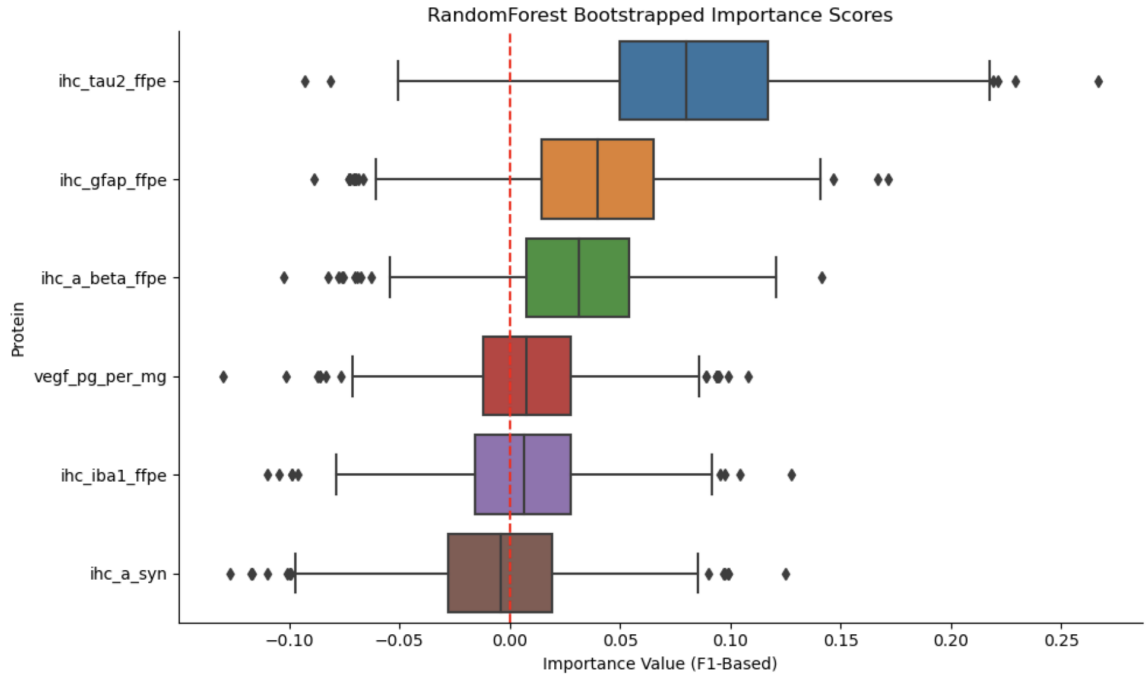
Since there were multiple features associated with the same protein. Multicollinearity was evaluated in the dataset.

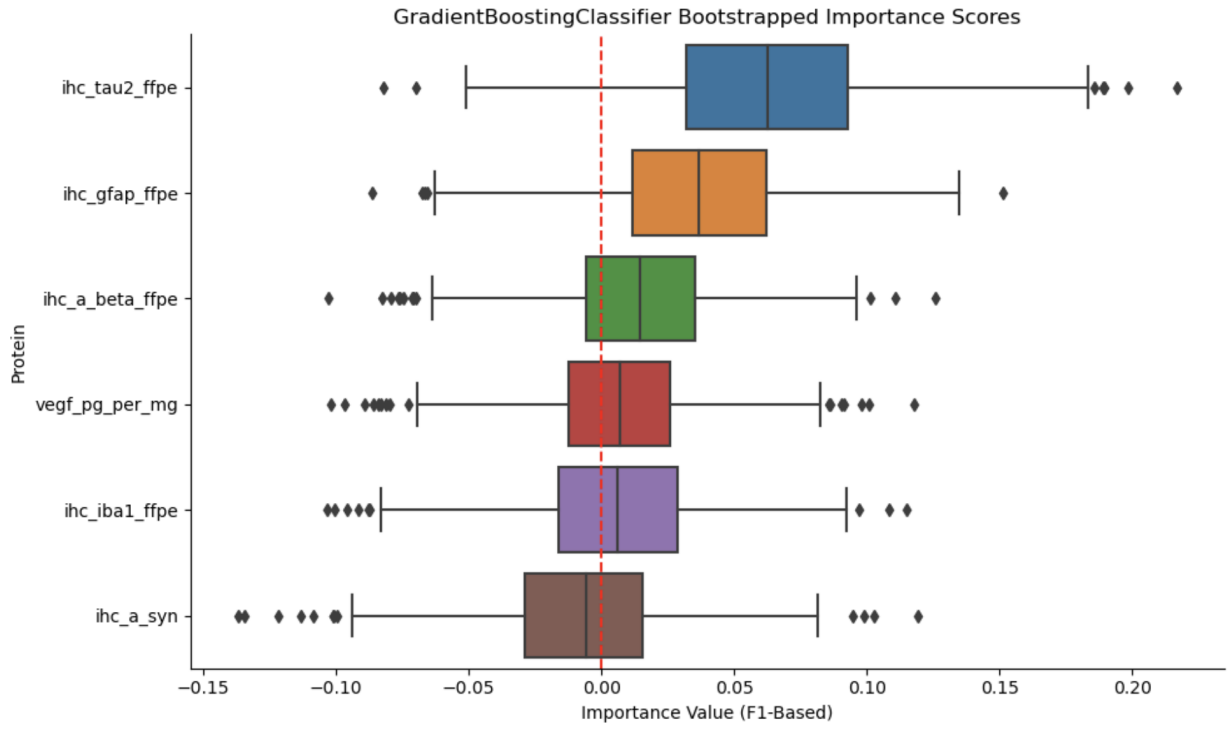


The groupings were generally consistent with expectations, the 5 amyloid beta proteins were grouped together, and 5 of the 6 Tau related proteins were grouped together. Different thresholds were explored with Threshold = 1 yielding the highest model performance (accuracy: 0.663, precision: 0.705, recall: 0.660, f1: 0.681), with the following features selected: 'ihc_a_syn', 'ihc_tau2_ffpe', 'ihc_a_beta_ffpe', 'ihc_iba1_ffpe', 'ihc_gfap_ffpe', 'vegf_pg_per_mg'.

RandomForest and GradientBoosting models were refitted using these 6 selected features, and feature importance was calculated.

Protein Analysis Feature Importance with Selected Features





The resulting feature importance graphs were strongly consistent, with the relative importance of each feature the same across both models. Again, not surprisingly, a Tau

feature was found to be very important. Also of note, GFAP had the second highest importance in both models.

Protein Analysis Model Tuning

Hyperparameter tuning was explored using RandomSearchCV and GridSearchCV. No significant improvements were found through hyperparameter tuning. This could likely be due to the relatively small sample size of the data set.

Protein Analysis Discussion

The differential expression of genes seen in this analysis was consistent with the current scientific literature on biomarkers for dementia and Alzheimer's Disease. It was interesting to see the higher feature importance of Tau features relative to Amyloid Beta features. This is consistent with the currently hypothesized pathological manifestations of these proteins in dementia patients. It is believed that amyloid beta protein accumulation leads to tau protein accumulations. Therefore a patient with high tau proteins is further downstream in the pathophysiology of dementia. Another feature of interest that showed high importance was GFAP (ihc_gfap_ffpe). GFAP has recently been studied as a potential new biomarker for Alzheimer's Disease (Kim 2023). This analysis adds support to the role that GFAP may have in the development of Alzheimer's Disease.

Overall Discussion

Limitations

One of the limitations was the relatively small sample size of our dataset. For both the gene and protein data, there were 377 samples. Because of the limited sample size there was considerable variability in distributions for each run of train test split (as shown in PCA plots). This caused very different results for each run. A mitigation strategy implemented was to use bootstrapping and sampling the dataset 1000 times to arrive at an average result accounting for inter-sample variability.

Broader Impacts

Discovering gene expression or protein expression patterns that could be predictive of dementia could impact society because the conditions could be identified before symptoms progress, allowing researchers to longitudinally study patients and potentially discover treatments to slow or cure dementia. Our analysis focused on samples recovered during autopsy, so this would require adaptation of our analysis to samples that could be taken from living patients. An ethical issue that could arise from using our findings as they exist currently is that they may not be generalizable because all samples came from deceased donors and were chosen because they had a prior TBI or were similarity matched to someone who had a prior TBI. Future work could involve utilizing the provided sample weights to improve generalizability to the broader ACT study population.

Future Work

The Aging, Dementia and Traumatic Brain Injury (TBI) Study dataset is a feature rich dataset that provides the opportunity for much further exploration. In this project we perform a whole-brain analysis of genes, but the data provides the information necessary to hone in on specific brain regions and for brain region specific analysis. In addition, while we used all patients we there is opportunity to hone in on specific subpopulations such as by sex, age, or by excluding individuals with previous history of traumatic brain injury. Further analysis could also be performed with the differentially expressed genes. We performed cursory gene ontology analyses for both differential expression pipelines, but this data could be further explored to gain insight into the gene co-expression network and the ways the differentially expressed genes interact.

Future work could involve exploring the composition of the genes and samples that formed a cluster. There also seems to be debate on quantification for RNA sequencing analysis. While some sources stated that TPM was a better way to compare samples, others, such as Zhao et al. (2021), felt that its normalization of sequencing depth limited the ability to compare samples. More work could be done to explore the difference this choice makes.

Statement of Work

Casey Dye - Limma-Voom differential expression and enrichment analysis, Clustering analysis

Tony Lan - Protein differential expression analysis, Protein multicollinear analysis, Protein machine learning models evaluation, Protein feature importance analysis, Protein gene correlation analysis

Camaron Mangham - DESeq2 differential expression analysis, classification model evaluation, gene feature selection, dementia prediction

All group members contributed equally to literature review, theory, writing and visualization.

References

1. (2016). (tech.). *TECHNICAL WHITE PAPER: QUANTITATIVE DATA GENERATION*. Retrieved December 10, 2023, from https://help.brain-map.org/display/aging/Documentation.
2. (2017). (tech.). *TECHNICAL WHITE PAPER: OVERVIEW*. Retrieved December 10, 2023, from https://help.brain-map.org/display/aging/Documentation.
3. *Aging, dementia and Traumatic Brain Injury Study*. Overview :: Allen Brain Atlas: Aging, Dementia and TBI Study. (n.d.). https://aging.brain-map.org/overview/home
4. Arzouni, N., Matloff, W., Zhao, L., Ning, K., & Toga, A. W. (2020). Identification of Dysregulated Genes for Late-Onset Alzheimer's Disease Using Gene Expression Data in Brain. Journal of Alzheimer's disease & Parkinsonism, 10(6), 498.

5. Barrasa, I. (2020, February 13). Practical RNA-seq analysis - Massachusetts Institute of Technology. http://barc.wi.mit.edu/education/hot_topics/RNAseq_Feb2020/RNASeq_2020_1slidePerPage.pdf
6. Bloom GS. Amyloid-β and Tau: The Trigger and Bullet in Alzheimer Disease Pathogenesis. *JAMA Neurol.* 2014;71(4):505–508. doi:10.1001/jamaneurol.2013.5847
7. Bystrykh L. (2021). Python for gene expression. F1000Research, 10, 870. https://doi.org/10.12688/f1000research.53842.2
8. Chu, C. P., Hokamp, J. A., Cianciolo, R. E., Dabney, A. R., Brinkmeyer-Langford, C., Lees, G. E., & Nabity, M. B. (2017). RNA-seq of serial kidney biopsies obtained during progression of chronic kidney disease from dogs with X-linked hereditary nephropathy. *Scientific Reports*, *7*(1). https://doi.org/10.1038/s41598-017-16603-y
9. Corley, S. M., MacKenzie, K. L., Beverdam, A., Roddam, L. F., & Wilkins, M. R. (2017). Differentially expressed genes from RNA-seq and functional enrichment results are affected by the choice of single-end versus paired-end reads and stranded versus non-stranded protocols. *BMC Genomics*, *18*(1). https://doi.org/10.1186/s12864-017-3797-0
10. *DESeq 2*. NGS Analysis. (2018, January 15). https://learn.gencore.bio.nyu.edu/rna-seq-analysis/deseq-2/
11. Fiorini, M. R., Dilliott, A. A., & Farhan, S. M. K. (2022). Sex-stratified RNA-seq analysis reveals traumatic brain injury-induced transcriptional changes in the female hippocampus conducive to dementia. *Frontiers in neurology*, *13*, 1026448. https://doi.org/10.3389/fneur.2022.1026448
12. How to choose normalization methods (TPM/RPKM/FPKM) for mrna expression. Novogene. (2023, April 12). https://www.novogene.com/us-en/resources/blog/how-to-choose-normalization-methods-tpm-rpkm-fpkm-for-mrna-expression/#:~:text=TPM%20%28transcripts%20per%20kilobase%20million%29%20is%20very%20much,statistic%20when%20calculating%20gene%20expression%20comparisons%20across%20samples.
13. Guennewig, B., Lim, J., Marshall, L., McCorkindale, A. N., Paasila, P. J., Patrick, E., Kril, J. J., Halliday, G. M., Cooper, A. A., & Sutherland, G. T. (2021). Defining early changes in alzheimer's disease from RNA sequencing of brain regions differentially affected by pathology. *Scientific Reports*, *11*(1). https://doi.org/10.1038/s41598-021-83872-z
14. *How to analyze RNA-Seq data? Find differentially expressed genes in your research.* (2016). *YouTube*. Retrieved November 6, 2023, from https://www.youtube.com/watch?v=xh_wpWj0AzM.
15. Kim KY, Shin KY, Chang KA. GFAP as a Potential Biomarker for Alzheimer's Disease: A Systematic Review and Meta-Analysis. Cells. 2023 May 4;12(9):1309. doi: 10.3390/cells12091309. PMID: 37174709; PMCID: PMC10177296.
16. Koch, C. M., Chiu, S. F., Akbarpour, M., Bharat, A., Ridge, K. M., Bartom, E. T., & Winter, D. R. (2018). A beginner's guide to analysis of RNA sequencing data. *American Journal of Respiratory Cell and Molecular Biology*, *59*(2), 145–157. https://doi.org/10.1165/rcmb.2017-0430tr

17. Law, C. W., Chen, Y., Shi, W., & Smyth, G. K. (2014). Voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, *15*(2). https://doi.org/10.1186/gb-2014-15-2-r29
18. Li, W. V., & Li, J. J. (2018). Modeling and analysis of RNA-seq data: a review from a statistical perspective. Quantitative biology (Beijing, China), 6(3), 195–209. https://doi.org/10.1007/s40484-018-0144-7
19. Love, M.I., Huber, W., Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15:550. 10.1186/s13059-014-0550-8
20. Love, M.I., Anders, S., Huber W. (2023) Analyzing RNA-seq data with DESeq2. Bioconductor. https://bioconductor.org/packages/devel/bioc/vignettes/DESeq2/inst/doc/DESeq2.html
21. Mayeux R, Stern Y. Epidemiology of Alzheimer disease. Cold Spring Harb Perspect Med. 2012 Aug 1;2(8):a006239. doi: 10.1101/cshperspect.a006239. PMID: 22908189; PMCID: PMC3405821.
22. Moura, D. A., & de Oliveira, J. R. (2021). What Do Machines Tell Us about Dementia? Machine Learning Applied to Aging, Dementia and Traumatic Brain Injury Study. https://doi.org/10.21203/rs.3.rs-840907/v1
23. Muzellec, B., Tele´nczuk, M. T., Cabeli, V., & Andreux, M. (2022). PyDESeq2: a python package for bulk RNA-seq differential expression analysis. BioRxiv, 2022.12.14.520412. https://doi.org/10.1101/2022.12.14.520412
24. Pedregosa *et al.*, Scikit-learn: Machine Learning in Python, (2011). JMLR 12, pp. 2825-2830
25. Pluto. (2021, October 18). DESeq2: *An overview of a popular RNA-seq analysis - pluto bioinformatics*. DESeq2: An Overview of a Popular RNA-Seq Analysis - Pluto Bioinformatics. https://pluto.bio/blog/deseq2-an-overview-of-popular-rna-seq-analysis-package
26. Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). Limma powers differential expression analyses for RNA-sequencing and Microarray Studies. *Nucleic Acids Research*, *43*(7). https://doi.org/10.1093/nar/gkv007
27. Sayers, E. W., Bolton, E. E., Brister, J. R., Canese, K., Chan, J., Comeau, D. C., Connor, R., Funk, K., Kelly, C., Kim, S., Madej, T., Marchler-Bauer, A., Lanczycki, C., Lathrop, S., Lu, Z., Thibaud-Nissen, F., Murphy, T., Phan, L., Skripchenko, Y., Tse, T., … Sherry, S. T. (2022). Database resources of the national center for biotechnology information. Nucleic acids research, 50(D1), D20–D26. https://doi.org/10.1093/nar/gkab1112
28. scikit-learn developers. (n.d.). *Sklearn.cluster.HDBSCAN*. scikit. https://scikit-learn.org/stable/modules/generated/sklearn.cluster.HDBSCAN.html
29. Skidmore, Z. (n.d.). *Differential expression with deseq2*. Griffith Lab. https://genviz.org/module-04-expression/0004/02/01/DifferentialExpression/
30. Smyth, G. K., Ritchie, M., Thorne, N., Wettenhall, J., Shi, W., & Hu, Y. (2002, December 2). Limma linear models for microarray and RNA-seq data User's Guide - bioconductor. https://bioconductor.org/packages/release/bioc/vignettes/limma/inst/doc/usersguide.pdf
31. *StatQuest: A gentle introduction to RNA-seq*. (2017). *YouTube*. Retrieved November 6, 2023, from https://youtu.be/tlf6wYJrwKY?feature=shared.

32. Teichman, G., Cohen, D., Ganon, O., Dunsky, N., Shani, S., Gingold, H., & Rechavi, O. (2023). RNALYSIS: Analyze your RNA sequencing data without writing a single line of code. *BMC Biology*, *21*(1). https://doi.org/10.1186/s12915-023-01574-6
33. Teichman, G. (2021). *Rnalysis.filtering.countfilter*. rnalysis.filtering.CountFilter - RNAlysis 3.10.1 documentation. https://guyteichman.github.io/RNAlysis/build/rnalysis.filtering.CountFilter.html
34. Twohig, D., Nielsen, H.M. α-synuclein in the pathophysiology of Alzheimer's disease. *Mol Neurodegeneration* 14, 23 (2019). https://doi.org/10.1186/s13024-019-0320-x
35. Volcano plot (statistics) https://en.wikipedia.org/wiki/Volcano_plot_(statistics)
36. World Health Organization. (2023, March 15). *Dementia*. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/dementia
37. Zhao, Y., Li, M.-C., Konaté, M. M., Chen, L., Das, B., Karlovich, C., Williams, P. M., Evrard, Y. A., Doroshow, J. H., & McShane, L. M. (2021). TPM, FPKM, or normalized counts? A comparative study of quantification measures for the analysis of RNA-seq data from the NCI patient-derived models repository. *Journal of Translational Medicine*, *19*(1). https://doi.org/10.1186/s12967-021-02936-w

Appendix

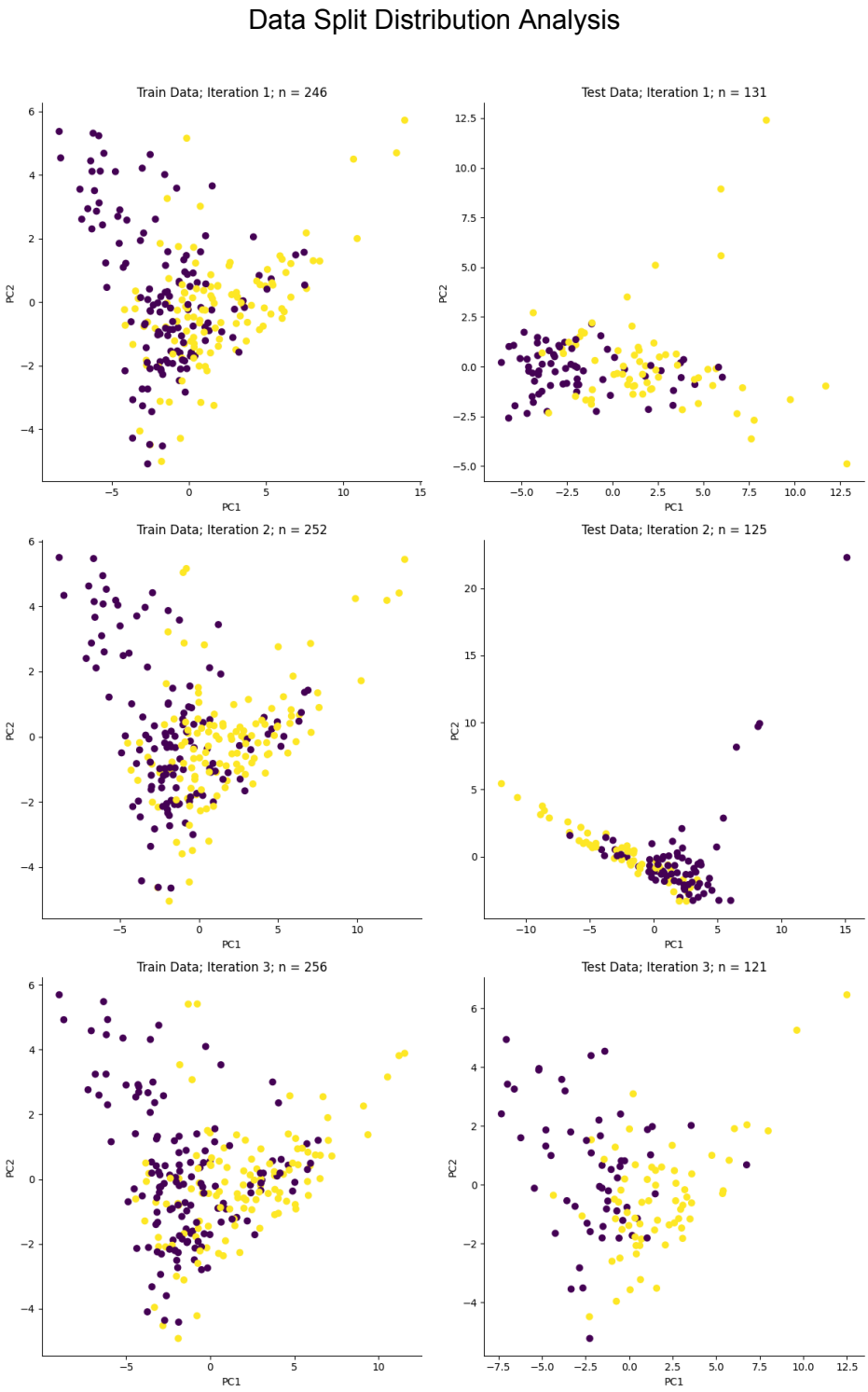
Incorporating Feedback

We received suggestions for a comprehensive literature review, defining our use case, and feature selection given our original goal to predict dementia in response to our project proposal. We completed a comprehensive literature review to familiarize ourselves with working with RNA sequencing and protein data and to see how others had analyzed this dataset. This informed the selection of models and clustering algorithms we experimented with. We did determine that the best use case for our analyses would be to explore more about the relationships within the data rather than be used in any sort of clinical setting because the samples were the result of an autopsy, and there would need to be more technological advancement to be able to use this to predict dementia for living patients. Because of this, we determined that a 'best' evaluation metric was not necessary, and we decided to report across evaluation metrics to see how model choice affected each metric.

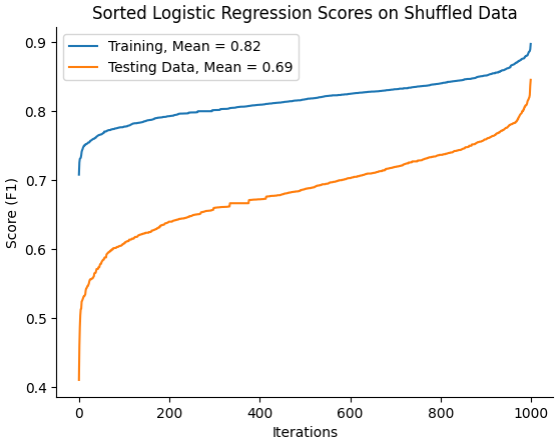
As a result of feedback from our advisor meeting, we noted that many of the model types and clustering algorithms were chosen due to their usage in similar studies.

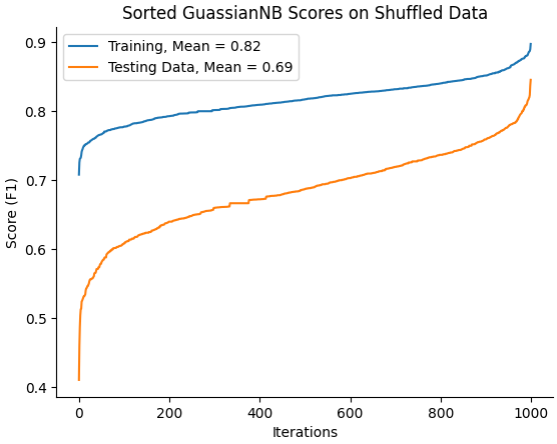
As a result of feedback from comments on our standup meeting, we have incorporated some ethical discussion within our "Broader Impacts" section of our report.

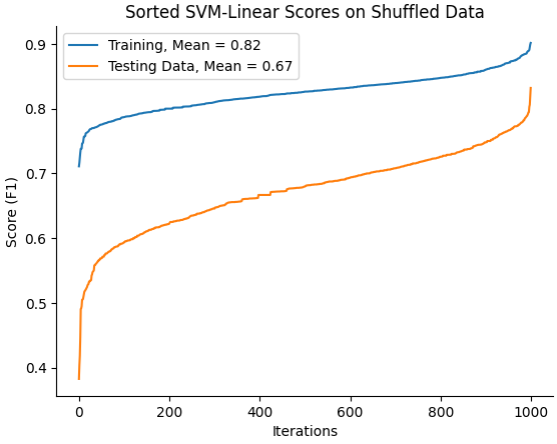
Gene Expression Analysis Supplementary Figures

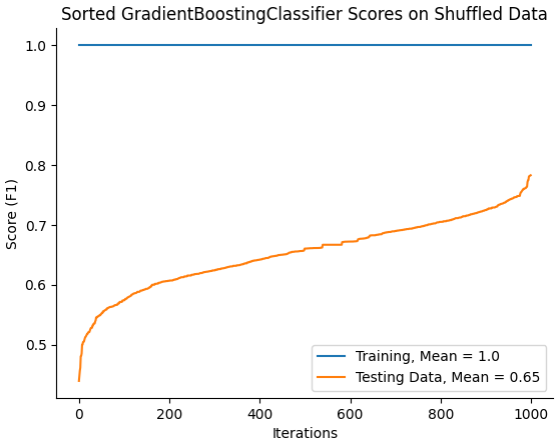


Base Model Scores on Bootstrapped Data Splits

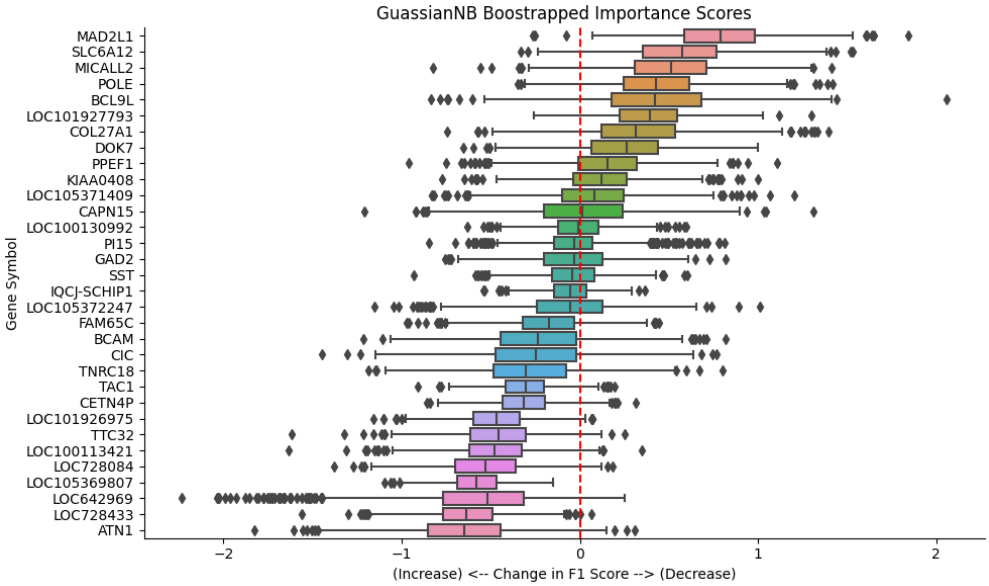


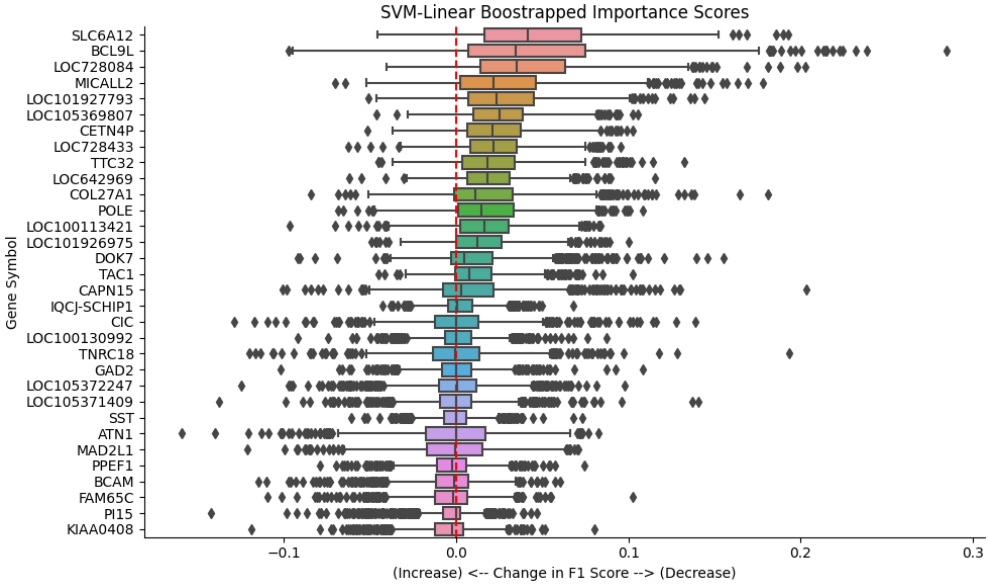


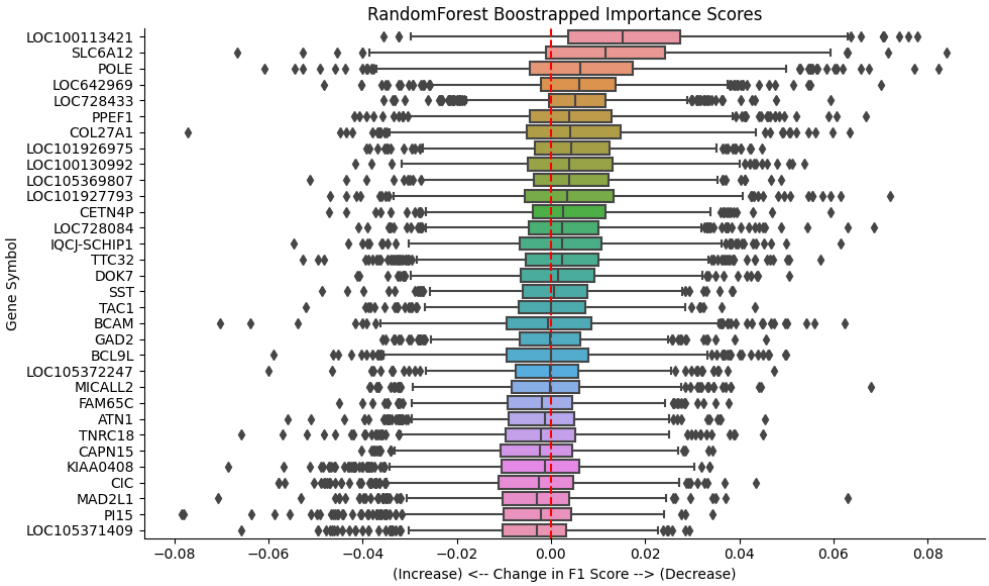


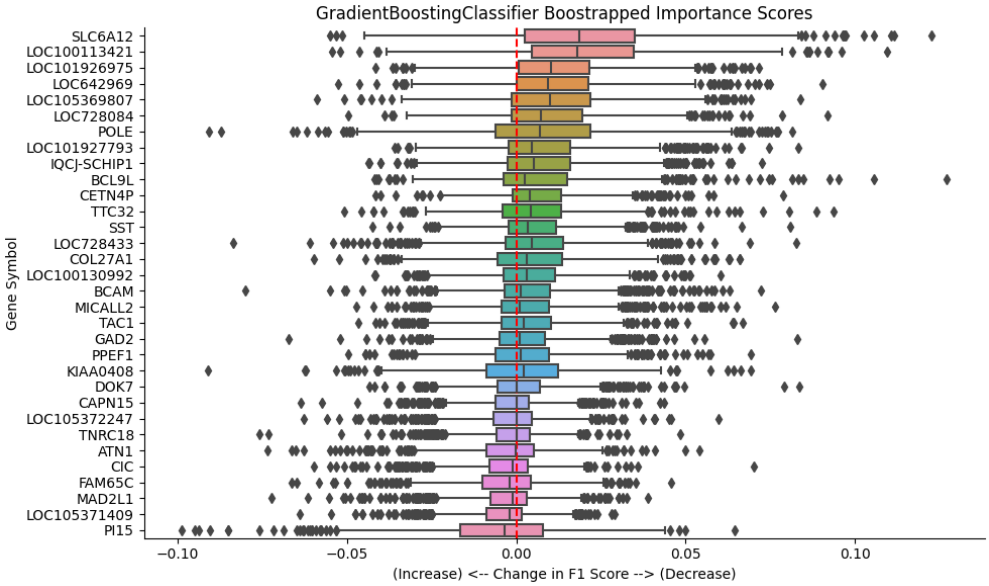


Base Model Importance Scores

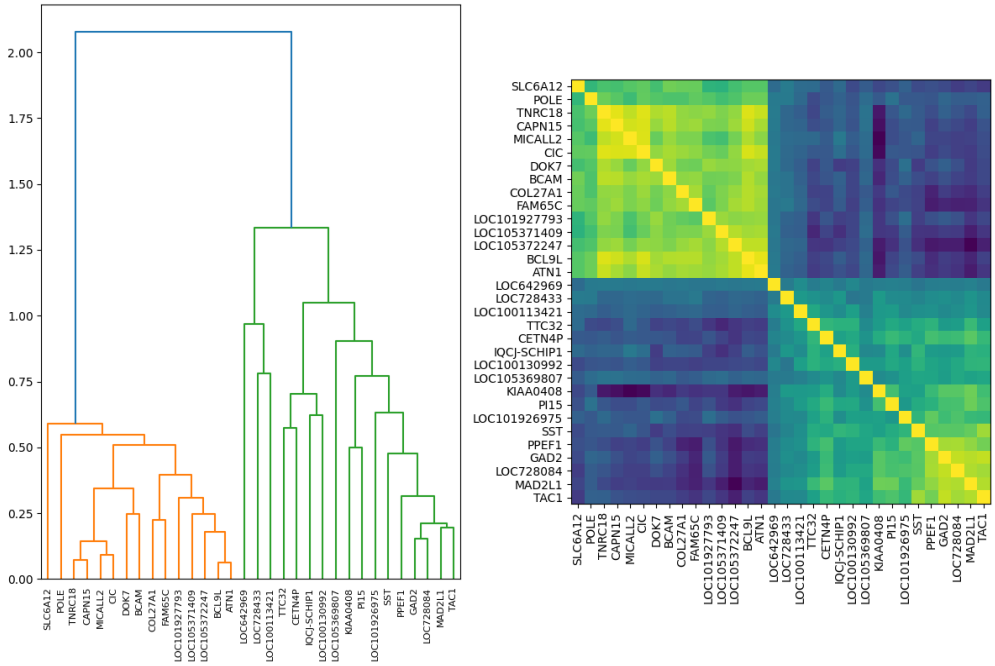




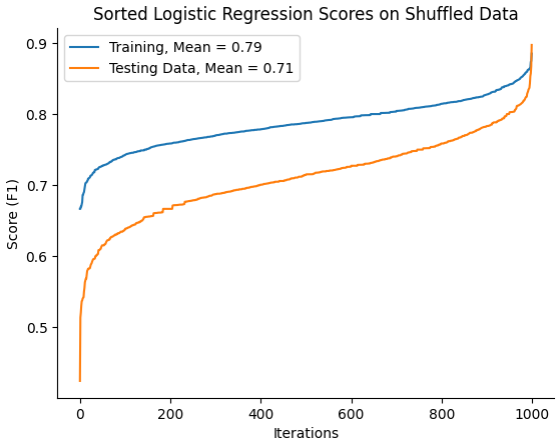

RandomForest Boostrapped Importance Scores
LOC100113421
SLC6A12
POLE
LOC642969
LOC728433
PPEF1
COL27A1
LOC101926975
LOC100130992
LOC105369807
LOC101927793
CETN4P
LOC728084
IQCJ-SCHIP1
TTC32
DOK7
SST
TAC1
BCAM
GAD2
BCL9L
LOC105372247
MICALL2
FAM65C
ATN1
TNRC18
CAPN15
KIAA0408
CIC
MAD2L1
PI15
LOC105371409
Gene Symbol
−0.08
−0.06
−0.04
−0.02
0.00
0.02
0.04
0.06
0.08
(Increase) <-- Change in F1 Score --> (Decrease)

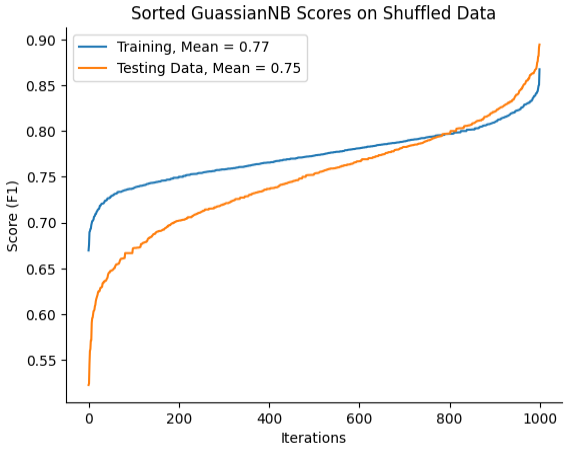

GradientBoostingClassifier Boostrapped Importance Scores
SLC6A12
LOC100113421
LOC101926975
LOC642969
LOC105369807
LOC728084
POLE
LOC101927793
IQCJ-SCHIP1
BCL9L
CETN4P
TTC32
SST
LOC728433
COL27A1
LOC100130992
BCAM
MICALL2
TAC1
GAD2
PPEF1
KIAA0408
DOK7
CAPN15
LOC105372247
TNRC18
ATN1
CIC
FAM65C
MAD2L1
LOC105371409
PI15
Gene Symbol
−0.10
−0.05
0.00
0.05
0.10
(Increase) <-- Change in F1 Score --> (Decrease)

Multicollinearity Analysis

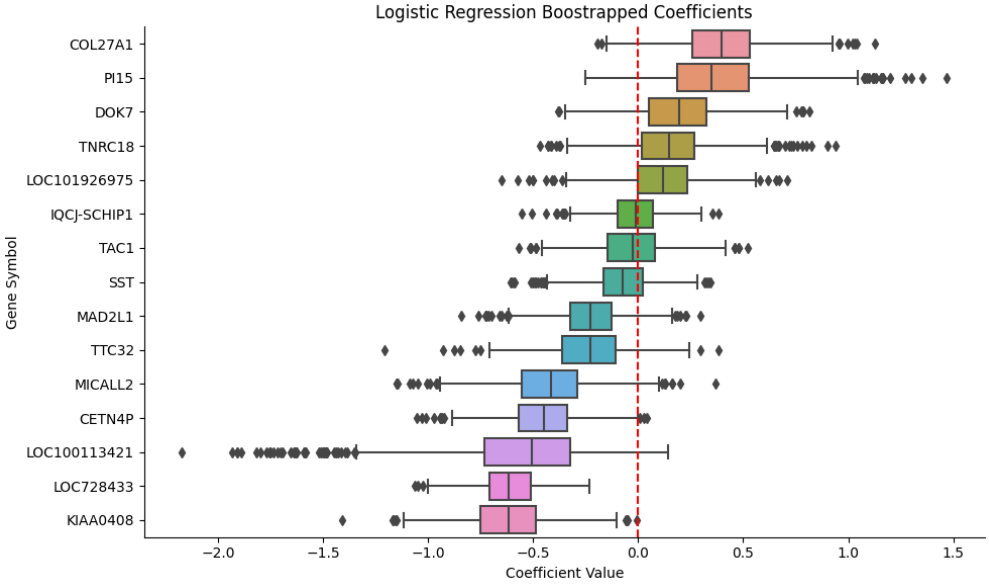


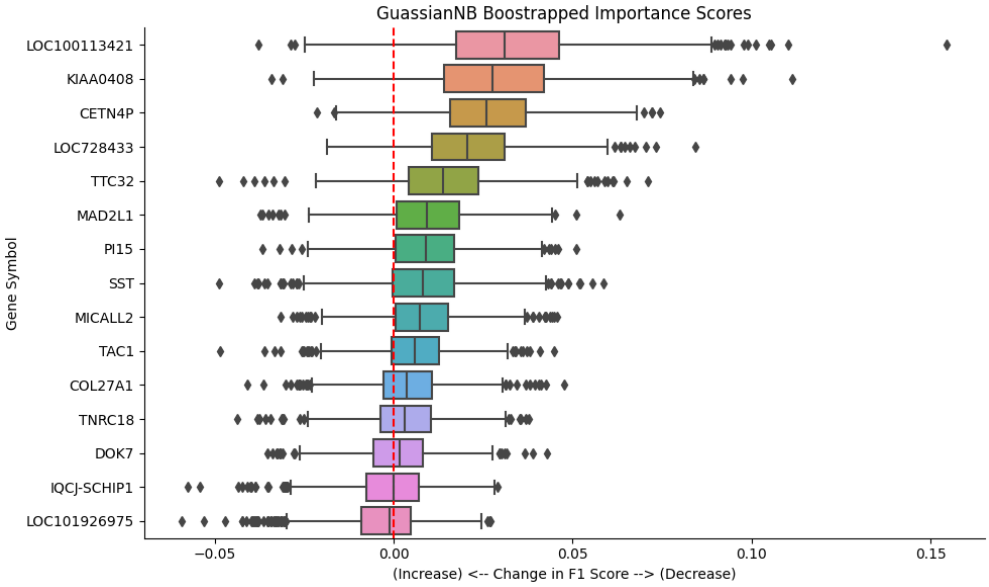
Reduced Model Scores on Bootstrapped Data Splits

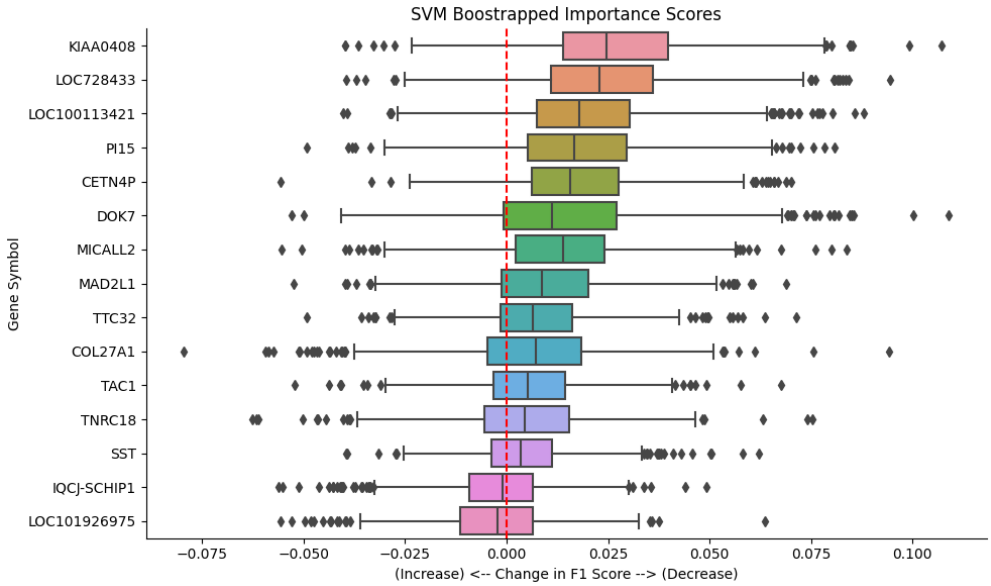




Reduced Model Importance Scores






SVM Boostrapped Importance Scores
KIAA0408
LOC728433
LOC100113421
PI15
CETN4P
DOK7
MICALL2
MAD2L1
TTC32
COL27A1
TAC1
TNRC18
SST
IQCJ-SCHIP1
LOC101926975
Gene Symbol
−0.075
−0.050
−0.025
0.000
0.025
0.050
0.075
0.100
(Increase) <-- Change in F1 Score --> (Decrease)

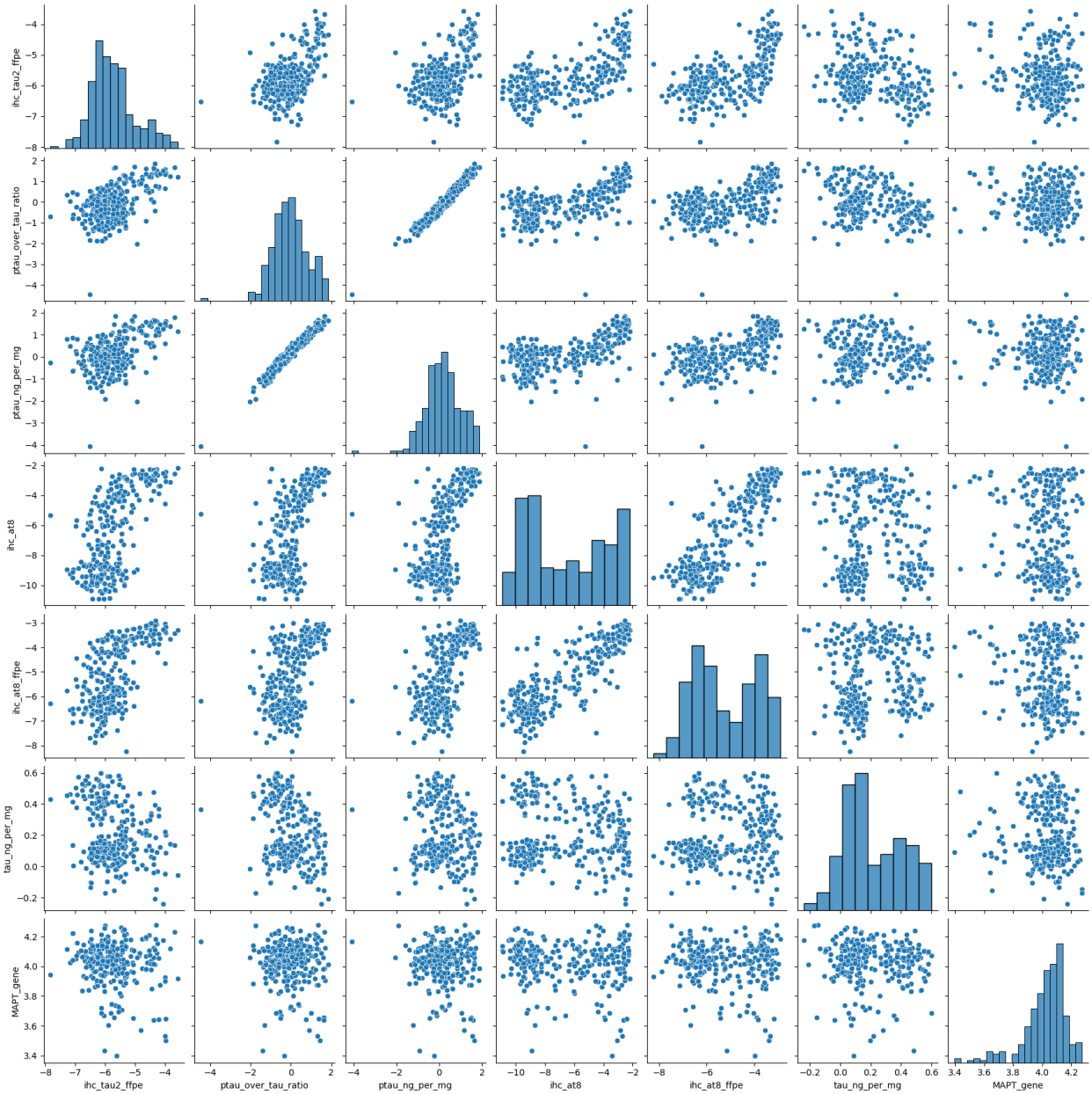
Protein Features Dictionary

ihc_a_syn : α-synuclein (ihc)
ihc_tau2_ffpe : Tau2 (ihc-ffpe)
ihc_at8_ffpe : AT8 (anti-tau antibody) (ihc-ffpe)
ihc_at8 : AT8 (anti-tau antibody) (ihc)
ihc_ptdp_43_ffpe : phospho-TDP43 (TAR DNA binding protein) (ihc-ffpe)
ihc_a_beta_ffpe : Amyloid β (ihc-ffpe)
ihc_a_beta : Amyloid β (ihc)
ihc_iba1_ffpe : Ionized calcium binding adaptor molecule 1 (ihc-ffpe)
ihc_gfap_ffpe : glial fibrillary acidic protein (ihc-ffpe)
ptau_ng_per_mg : Phosphorylated Tau (conc)
vegf_pg_per_mg : VEGF (conc)
ab42_over_ab40_ratio : Amyloid β (ratio)
tnf_a_pg_per_mg : TNF alpha
tau_ng_per_mg : Tau (conc)
il_10_pg_per_mg : Interleukin 10 (conc)
isoprostane_pg_per_mg : Isoprostane (conc)
il_6_pg_per_mg : Interleukin 6 (conc)
il_1b_pg_per_mg : Interleukin 1b (conc)
ptau_over_tau_ratio : Phosphorylated Tau (ratio)
il_4_pg_per_mg : Interleukin 4 (conc)
rantes_pg_per_mg : RANTES (conc)
ab40_pg_per_mg : Amyloid β 40 (conc)
a_syn_pg_per_mg : α-synuclein (conc)
ifn_g_pg_per_mg : Interferon gamma (conc)
mcp_1_pg_per_mg : Monocyte Chemoattractant Protein-1 (conc)
bdnf_pg_per_mg : Brain-derived neurotrophic factor (conc)
mip_1a_pg_per_mg : Macrophage Inflammatory Protein-1 Alpha (conc)
il_7_pg_per_mg : Interleukin 7 (conc)
ab42_pg_per_mg : Amyloid β 42 (conc)

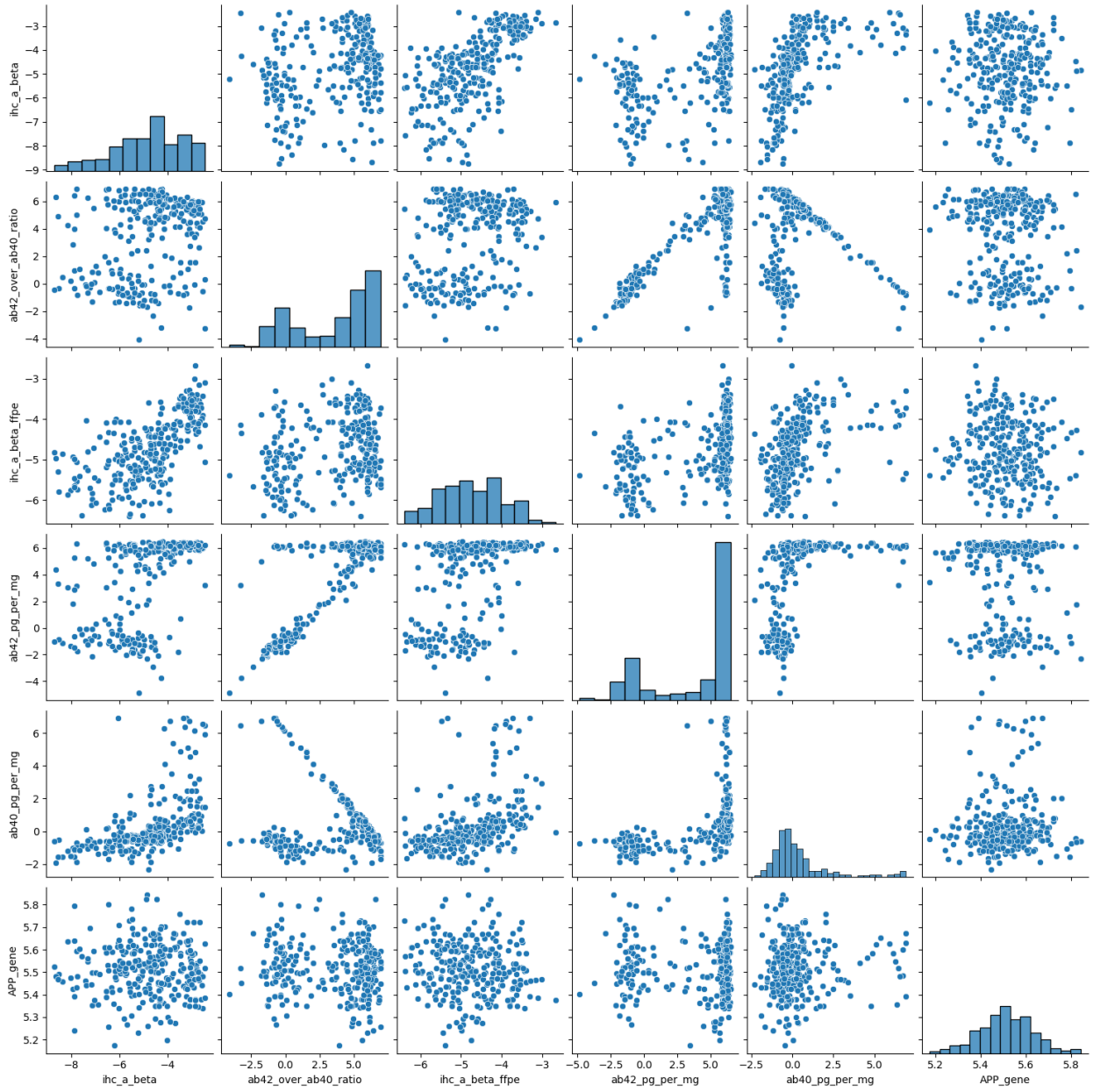
Gene expression to protein expression correlation analysis

Exploratory analysis was done to see if there was any correlation between gene expression and protein features for Tau proteins (MAPT gene) and Amyloid proteins (APP gene). Data was log transformed and a SPLOM plot was created to evaluate potential correlations between gene expression and protein expression

MAPT/Tau SPLOM



APP/Amyloid SPLOM



The rightmost column on each SPLOM shows the scatterplots with the gene expression on the x-axis and respective protein expressions on the y-axis. No correlations are seen between gene expression and protein expression. This could be due to multiple factors, most relevant being that protein expression quantification for these types of analyses are usually done through mass spectrometry. There are likely too many confounding variables when measuring protein expression through other means such as blood concentration and IHC.