



CHALMERS
UNIVERSITY OF TECHNOLOGY



UNIVERSITY OF GOTHENBURG

Applying Machine Learning to High-Dimensional Proteomics Datasets for Biomarker Discovery in Neurodegenerative Disorders

Master's thesis in Computer science and engineering

Christoffer Ivarsson Orrelid & Oscar Rosberg

Department of Computer Science and Engineering
CHALMERS UNIVERSITY OF TECHNOLOGY
UNIVERSITY OF GOTHENBURG
Gothenburg, Sweden 2024

MASTER'S THESIS 2024

Applying Machine Learning to High-Dimensional Proteomics Datasets for Biomarker Discovery in Neurodegenerative Disorders

Christoffer Ivarsson Orrelid & Oscar Rosberg



UNIVERSITY OF
GOTHENBURG



CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Computer Science and Engineering
CHALMERS UNIVERSITY OF TECHNOLOGY
UNIVERSITY OF GOTHENBURG
Gothenburg, Sweden 2024

Applying Machine Learning to High-Dimensional Proteomics Datasets for Biomarker
Discovery in Neurodegenerative Disorders
Christoffer Ivarsson Orreliid & Oscar Rosberg

© Christoffer Ivarsson Orreliid & Oscar Rosberg, 2024.

Supervisor: Lena Stempfle & Newton Mwai Kinyanjui, Department of Computer
Science and Engineering
Examiner: Fredrik Johansson, Department of Computer Science and Engineering

Master's Thesis 2024
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Typeset in L^AT_EX
Gothenburg, Sweden 2024

Applying Machine Learning to High-Dimensional Proteomics Datasets for Biomarker Discovery in Neurodegenerative Disorders
Christoffer Ivarsson Orreliid & Oscar Rosberg
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg

Abstract

Identifying biomarkers for Alzheimer’s Disease (AD), a progressive neurodegenerative disorder characterized by progressive cognitive decline is crucial for early diagnosis and treatment. This thesis explores proteomic abundances along the AD continuum using lumbar and ventricular cerebrospinal fluid (CSF) samples from patients with idiopathic normal pressure hydrocephalus (iNPH) to identify potential new biomarkers. Our study emphasizes the necessity of treating lumbar and ventricular CSF samples as separate datasets due to their distinct proteomic profiles.

Challenges such as handling high-dimensional data with missing values, small sample sizes and class imbalances were addressed through imputation, oversampling and k-fold cross-validation techniques. We discuss the presence and consequence of batch effect, a remnant of the mass spectrometry technique tandem mass tag. Comparative analysis through staging on existing biomarkers highlights the uniqueness of the dataset provided by Sahlgrenska University Hospital. Through machine learning and feature selection techniques, we propose eight protein and nine peptide biomarkers for distinguishing iNPH patients on the pathological AD spectra. One such biomarker shows relevance in both lumbar and ventricular CSF. Despite the study’s limited cohort size, our findings contribute insights into the proteomic analysis of neurodegenerative disorders.

Keywords: Alzheimer’s disease, neurodegenerative disorder, proteomics, mass spectrometry, high-dimensional data, biomarkers, machine learning, feature selection, staging

Acknowledgements

We would like to express our thanks to our supportive and helpful supervisors, Lena Stempfle and Newton Mwai Kinyanjui, who have supported us through this Master's thesis. Not only have they provided us with invaluable feedback, but they have also shared their never-ending knowledge, expertise, and positive outlook. We are also grateful to Sophia Weiner, who supported us with data, ideas, and a domain expert's opinion. Additionally, we would like to extend our gratitude to our examiner, Fredrik Johansson, for his valuable feedback and helpfulness. This project would not have been possible without them.

Christoffer Ivarsson Orrelid & Oscar Rosberg, 2024-08-21

Contents

List of Figures	xiii
List of Tables	xvii
List of Acronyms	xxi
1 Introduction	1
1.1 Problem	2
1.2 Ethical Considerations	2
2 Background	5
2.1 Bioinformatics	5
2.1.1 Biomarkers	5
2.1.2 Alzheimer’s Disease	6
2.1.3 Idiopathic Normal Pressure Hydrocephalus	7
2.1.4 Mass Spectrometry	7
2.1.4.1 Tandem Mass Tag	8
2.1.4.2 Batch Effect	8
2.2 Machine Learning	9
2.2.1 Missingness in Proteomics Data	9
2.2.2 Imputation	10
2.2.3 Models	12
2.2.3.1 Logistic Regression	12
2.2.3.2 Lasso	13
2.2.3.3 Random Forest	13
2.2.3.4 Extreme Gradient Boosting	15
2.2.4 Dimensionality Reduction	16
2.2.4.1 Feature Selection	16
2.2.4.2 Ensemble Feature Selection	17
2.2.4.3 Feature Extraction	18
2.2.5 Imbalanced Dataset	19
2.2.6 Evaluation	20
2.2.6.1 Accuracy	20
2.2.6.2 Recall	21
2.2.6.3 Precision	21
2.2.6.4 F_1 -score	21

2.2.6.5	Receiver Operating Characteristic & Area Under the Curve	22
2.2.6.6	Matthews Correlation Coefficient	22
3	Experimental Setup	23
3.1	The Bigger Picture	23
3.2	Data	24
3.2.1	Converting Multiclass to Binary Predictions	26
3.2.2	Data Exploration	27
3.2.3	Data Preprocessing	27
3.2.4	Missingness Mechanisms	28
3.3	Imputation	29
3.3.1	Multiple Imputation	29
3.3.2	Minimum Imputation	30
3.4	ComBat	30
3.5	Models	30
3.6	Feature Selection	31
3.7	Pipeline	32
4	Results	35
4.1	Data Exploration Results	35
4.1.1	Staging Biomarkers	35
4.1.2	Missingness Ratios in Lumbar and Ventricular datasets	38
4.1.3	Distribution of Analyte Mean Value Abundances	39
4.1.4	Batch Effect Results	40
4.1.5	Batch Effect on Predictive Results	41
4.2	Model Evaluation for Tissue Group Prediction	42
4.2.1	Ventricular protein modelling results	43
4.2.2	Lumbar protein modelling results	44
4.2.3	Ventricular peptide modelling results	45
4.2.4	Lumbar peptide modelling results	46
4.3	Proposed Biomarker Evaluation	47
4.3.1	Growth differentiation factor 8	48
4.3.2	Aspartate aminotransferase	48
4.3.3	Calcium/calmodulin-dependent protein kinase type II subunit gamma	48
4.3.4	Pituitary adenylate cyclase-activating polypeptide	49
4.3.5	Midkine	49
4.3.6	Neurogranin	49
4.3.7	Peptide biomarkers	50
5	Discussion	51
5.1	Data Preparation and Model Settings	51
5.2	Small Dataset and k -fold Validation	52
5.3	Model Performance and Ensemble Techniques	53
5.4	Feature Selection and Stability	54
5.5	Comparing Neurodegenerative Disorders	54

5.6	Limitations	55
5.7	Future work	56
6	Conclusion	57
	Bibliography	59
A	Appendix A	I
B	Appendix B	VII
C	Appendix C	XI

List of Figures

2.1	AD pathogenesis with respect to $A\beta$ plaques and tau tangles. Image inspiration from [20].	6
2.2	Simplified workflow of relative abundance quantification through MS. Image inspired from [20].	8
2.3	Batches for Tandem Mass Tag with the same reference.	9
2.4	Geometric interpretation of Lasso and Ridge regression. Image inspiration from [68].	14
2.5	High-level overview of XGBoosts cumulative process.	15
2.6	Ensemble feature selection through union. Image inspiration from [82]	18
2.7	Example of how PCA-plots can be used for data exploration.	19
2.8	Confusion matrix used for prediction visualization.	20
2.9	ROC & AUC, illustrating the performance of a binary classifier.	22
3.1	Overview of how missingness is introduced through batches.	28
3.2	Illustration representing the pipeline used for modelling.	33
4.1	Abundance distribution of proteins on tissue groupings and CSF sample type. The bars on the left in each figure are ventricular CSF, and those on the right are lumbar CSF. Blue bars represent abundance in the $A\beta^-T^-$ tissue group, orange in $A\beta^+T^-$ and green in $A\beta^+T^+$. Ideal staging biomarkers would include clear differences within the CSF sample type over the tissue groupings, which is somewhat seen in 4.1b lumbar CSF and 4.1d ventricular CSF. The caption of the subfigures corresponds to the protein description and gene symbol.	36

- 4.2 Missingness ratio in \tilde{D}_{PV} , \tilde{D}_{PL} , \tilde{D}_{PeV} and \tilde{D}_{PeL} . The vertical dashed lines denote the missingness ratio of 50%, and the horizontal line separates the proteins and peptides above and below this threshold. The proteins and peptides below the horizontal line are kept. There is no apparent difference in ratio when comparing lumbar and ventricular CSF samples in either the protein or peptide datasets, with fewer overall missing values in the protein datasets. When imputing missing values, features above a certain missingness ratio are discarded. Looking at the protein datasets, roughly 50% of all features are discarded if the max missingness ratio is set to 20%. The same missingness ratio on peptide data would discard roughly 70% of the features. Discarding all features with missing values would result in a protein dataset with 1201 features and a peptide dataset with 3887 features. 38
- 4.3 Figures 4.3a and 4.3c plots the protein and peptide PDF of the non-normalized datasets, and Figures 4.3b and 4.3d for the median-normalized datasets. The lumbar is represented by a blue plot, and the ventricular is represented by an orange plot. There is a noticeable difference in the distributions after normalization, particularly in the ventricular CSF, as both distributions approach zero mean. In both analytes, the lumbar distribution is more concentrated around the mean, while the ventricular distribution has a larger spread. 39
- 4.4 Four t-SNE plots of the \tilde{D}_{PeV} dataset with all features with missing values removed. Figures 4.4a and 4.4c are coloured by the TMT batch, while Figures 4.4b and 4.4d are coloured by tissue group. In Figures 4.4a and 4.4b, \tilde{D}_{PeV} has not undergone ComBat batch effect removal. Noticeable clusters in Figure 4.4a, as shown with red circles, indicate the presence of batch effect bias. After applying ComBat to \tilde{D}_{PeV} , Figure 4.4c shows increased entropy while retaining similar clustering patterns in the tissue group plot. 40
- 4.5 PCA plot on combined \tilde{D}_{PeL} and \tilde{D}_{PeV} with no missingness. The blue dots represent lumbar samples, and the orange dots represent ventricular samples. The clusters observed in the plot indicate a noticeable separation between the ventricular and lumbar samples, suggesting distinct patterns. This separation highlights the need to split the sample spaces into two different datasets for modelling. . . . 41
- 4.6 Illustrated are the sections for each dataset: \tilde{D}_{PV} in 4.2.1, \tilde{D}_{PL} in 4.2.2, \tilde{D}_{PeV} in 4.2.3 and \tilde{D}_{PeL} in 4.2.4. Each section follows the same format. 42
- 4.7 4.7a shows that the RF model performs well compared to the chance level, with 0.88 AUC. 4.7b shows a balanced and fairly strong performance. 4.7c shows that three features are stable through each k-fold and are considered biomarker candidates. 43

4.8	These three protein biomarker candidates are consistently extracted through feature selection in each k-fold. The subfigure captions depict the gene symbol. The proteins descriptions are: 4.8a - growth differentiation factor 8. 4.8b - aspartate aminotransferase, cytoplasmic. 4.8c - calcium/calmodulin-dependent protein kinase type II subunit gamma.	43
4.9	4.9a shows an XGB models ROC curve with 0.71 AUC. 4.9b highlights good predictions on the TP class, with the TN performance falling behind. 4.9c indicates that five features are selected across each k-fold. With $k = 5$, there are more features in the other bins, but can still be considered stable.	44
4.10	Plotted are three of the five protein biomarker candidates that are consistently extracted through feature selection in each k-fold from \tilde{D}_{PL} . Appendix C shows the full set of proteins. The subfigure captions depict the gene symbol. The protein descriptions are 4.10a - pituitary adenylate cyclase-activating polypeptide. 4.10b - midkine. 4.10c - neurogranin.	44
4.11	4.11a shows that the XGB model performs well compared to the chance level, with 0.88 AUC. 4.11b shows the strong predictions on the $A\beta^-T^-$ class, with worse results on $A\beta^+T^+$. Six biomarker candidates are found in the union datasets according to 4.11c.	45
4.12	Plotted are three of the six peptide biomarker candidates consistently extracted through feature selection in each k-fold. Appendix C shows the full set of peptides. The subfigure captions depict the peptide's position in their respective protein. Both 4.12b and 4.12c shows tissue group correlation on both ventricular and lumbar CSF.	45
4.13	4.13a shows a ROC curve and AUC slightly worse than for the other datasets. The same is true for 4.13b, where only half of the $A\beta^+T^+$ samples were correctly classified. Three biomarker candidates were found, as shown in 4.13c.	46
4.14	These three peptide biomarker candidates are consistently extracted through feature selection in each k-fold. The subfigure captions depict the peptide's position in their respective protein. Reading to plots, 4.14a seems to correlate to the tissue group for ventricular CSF, but no peptide has clear patterns for lumbar.	46
C.1	Proposed biomarkers from \tilde{D}_{PV}	XI
C.2	Proposed biomarkers from \tilde{D}_{PL}	XII
C.3	Proposed biomarkers from \tilde{D}_{PeV}	XIII
C.4	Proposed biomarkers from \tilde{D}_{PeL}	XIV

List of Tables

3.1	Descriptive statistics of demographic features categorized by tissue groups and divided into lumbar and ventricular CSF data. Rows 5-9 describe other clinical comorbid conditions of the patients where VCI is vascular cognitive impairment.	25
3.2	Description of experimental and demographic features.	25
3.3	Number of samples (lumbar and ventricular) divided into the tissue groups.	26
3.4	This table describes the number of missing values for each of the datasets. Within batches describe how many missing values there are when not combining the TMT sets. The following row describes the number of missing values when the TMT batches are combined. The next row describes how many missing values are introduced by removing outliers.	28
3.5	BayesSearchCV hyperparameters for each model.	31
4.1	Biomarker comparison between tissue groups on the ventricular protein dataset. The protein abundance of each of the tissue groups is expressed as mean \pm standard deviation. Statistical significance ($p < 0.05$) is highlighted with bold numbers. FABP3 protein abundance in group $A\beta^+T^+$ was found to be significantly different from those in group $A\beta^-T^-$	37
4.2	Biomarker comparison between tissue groups on the lumbar protein dataset. The protein abundance of each of the tissue groups is expressed as mean \pm standard deviation. Statistical significance ($p < 0.05$) is highlighted with bold numbers. YWHAG protein abundance in group $A\beta^+T^+$ was found to be significantly different from those in group $A\beta^-T^-$	37
4.3	Predicting TMT set based results.	41
4.4	95% confidence interval of the best model on protein ventricular data on accuracy, F_1 -score, AUC and MCC. The confidence interval is narrow on all scores, indicating a stable model.	43
4.5	Accuracy and AUC confidence scores indicate a tight fit. However, the F_1 -score and MCC spread highlights greater variance in the model predictions over multiple iterations. Each iteration's score is shown in Appendix B.	44

4.6	The table shows the 95% confidence interval on peptide ventricular data. The fit is fairly close for all metrics, with F_1 -score and MCC having a slightly broader spread.	45
4.7	The confidence interval on the scores all have a broad spread, indicating much variance in the results when training and evaluating the models. The MCC has an especially large variance.	46
4.8	Biomarker comparison between tissue groups on \tilde{D}_{PV} . The protein abundance of each of the tissue groups is expressed as mean \pm standard deviation. Statistical significance ($p < 0.05$) is highlighted with bold numbers. GOT1 protein abundance in group $A\beta^+T^+$ was found to be significantly different from those in both group $A\beta^+T^-$ and $A\beta^-T^-$	47
4.9	Biomarker comparison between tissue groups on \tilde{D}_{PL} . The protein abundance of each of the tissue groups is expressed as mean \pm standard deviation. Statistical significance ($p < 0.05$) is highlighted with bold numbers. No statistical significance was found among the tissue groups.	47
4.10	Biomarker comparison between tissue groups on \tilde{D}_{PeV} . The peptide abundance of each tissue group is expressed as mean \pm standard deviation. Statistical significance ($p < 0.05$) is highlighted with bold numbers. The $A\beta^+T^+$ tissue group showed statistically significant differences in P43652 [215-221] peptide abundance from both $A\beta^-T^-$ and $A\beta^+T^-$. Also, the $A\beta^+T^+$ tissue group was statistically significant from the $A\beta^+T^+$ tissue group in P49641 [277-283] peptide abundance.	47
4.11	Biomarker comparison between tissue groups on \tilde{D}_{PeL} . The peptide abundance of each tissue group is expressed as mean \pm standard deviation. Statistical significance ($p < 0.05$) is highlighted with bold numbers. The P23284 [159-165] peptide was found to have statistically significant differences in abundance in group $A\beta^+T^+$ compared to $A\beta^-T^-$ and $A\beta^+T^-$	48
A.1	Experiment results for all runs on \tilde{D}_{PV} binary classification task.	II
A.2	Experiment results for all runs on \tilde{D}_{PL} binary classification task.	III
A.3	Experiment results for all runs on \tilde{D}_{PeV} binary classification task.	IV
A.4	Experiment results for all runs on \tilde{D}_{PeL} binary classification task.	V
B.1	Confidence interval runs on best models on \tilde{D}_{PV} and \tilde{D}_{PL}	VIII
B.2	Confidence interval runs on best models on \tilde{D}_{PeV} and \tilde{D}_{PeL}	IX
C.1	Proposed protein biomarkers from ventricular CSF samples.	XI
C.2	Proposed protein biomarkers from lumbar CSF samples.	XII
C.3	Biomarker comparison between tissue groups on \tilde{D}_{PL} not present in Section 4.3. The protein abundance of each of the tissue groups is expressed as mean \pm standard deviation. Statistical significance ($p < 0.05$) is highlighted with bold numbers. No statistical significance was found among the tissue groups.	XII
C.4	Proposed peptide biomarkers from ventricular CSF samples.	XIII

C.5	Biomarker comparison between tissue groups on \tilde{D}_{PeV} not present in Section 4.3. The peptide abundance of each tissue group is expressed as mean \pm standard deviation. Statistical significance ($p < 0.05$) is highlighted with bold numbers. No statistical significance was found among the tissue groups.	XIII
C.6	Proposed peptide biomarkers from lumbar CSF samples.	XIV

List of Acronyms

$A\beta$	Beta-amyloid.
T	Tau protein.
AD	Alzheimer's Disease.
AUC	Area Under the Curve.
CN	Cognitive Normal.
CSF	Cerebrospinal Fluid.
FN	False Negative.
FP	False Positive.
IQR	Interquartile Range.
LR	Logistic Regression.
Lasso	Least Absolute Shrinkage Selection Operator.
MAR	Missing At Random.
MCAR	Missing Completely At Random.
MCC	Matthews Correlation Coefficient.
MCI	Mild Cognitive Impairment.
MICE	Multivariate Imputation by Chained Equation.
MI	Multiple Imputation.
ML	Machine Learning.
MNAR	Missing Not At Random.
MS	Mass Spectrometry.
NFTs	Neurofibrillary Tangles.
PCA	Principal Components Analysis.
PDF	Probability Density Function.
RFE	Recursive Feature Elimination.
RF	Random Forest.
ROC	Receiver Operating Characteristic.
SMOTE	Synthetic Minority Oversampling Technique.
SVM	Support Vector Machine.
TMT	Tandem Mass Tag.
TN	True Negative.
TP	True Positive.
XGBoost	eXtreme Gradient Boosting.
iNPH	idiopathic Normal Pressure Hydrocephalus.
t-SNE	t-Distributed Stochastic Neighbor Embedding.

1

Introduction

Alzheimer's Disease (AD) is a progressive neurodegenerative disorder, the most common form of dementia and places a substantial burden on caregivers and health-care systems worldwide [1]. The cause of disease progression remains unknown, and despite extensive research, there are no cures or effective drugs to prevent its pathogenesis. By examining protein and peptide levels in cerebrospinal fluid (CSF) samples across different stages of pathogenesis, we can analyze the AD continuum to identify observable patterns and the underlying cause of the disease. However, many neurodegenerative disorders share similar patterns, making them harder to distinguish from one another [2]. Furthermore, the presence of multiple disorders may distort previously established patterns for diagnosis, emphasizing the need for specialized observations for AD diagnosis in the presence of other neurodegenerative disorders.

Proteomics is a branch of molecular biology focusing on the large-scale study of proteins, their components, structure, identification, and quantification. Researching proteomics enables the understanding of how proteins interact with each other, how their levels change during disease progression and how they contribute to various biological processes [3]. The advancements of mass spectrometry (MS), a widely used technique in large-scale proteomics, have revolutionized the field, allowing for rapid quantification of tens of thousands of peptides and proteins from multiple samples simultaneously [4]. Analyzing these tens of thousands of variables is not a trivial task by hand. The advancement of MS has increased the influence of machine learning (ML) in the field, as more high-dimensional data on a larger scale than before is being produced. Utilizing ML in order to establish disease-associated proteins and peptides for diagnosis, staging, prognosis, and treatment is becoming common practice in proteomics [5].

These proteins and peptides are biological markers, known as biomarkers, and are measurable indicators of some biological state or condition, such as AD. Ideal biomarkers are characterized by their high specificity towards specific disease conditions. Many biomarkers are proteins expressed in several different diseases, though their expression levels vary among these diseases. By combining multiple biomarkers together, it is possible to identify the characteristics and pathogenesis of individual diseases [6].

The conventional method for diagnosing AD is through medical history, cognitive tests, neurological examinations and CSF biomarkers, resulting in a clinical diagnosis.

However, a definitive diagnosis of AD can only be made with certainty through a brain tissue biopsy, known as a pathological diagnosis [7]. Biomarkers used to give a pathological diagnosis of AD include beta-amyloid protein clumps, known as amyloid plaques, and neurofibrillary tangles—abnormal tau protein aggregates found in nerve cells [1]. As these biomarkers can only be found through invasive surgery in the brain, finding new biomarkers is crucial for faster, less invasive and more cost-efficient diagnosis, improved drug treatment research and pathogenesis understanding. Current proteomics AD research is focused on finding biomarkers in alternative biological samples, such as CSF, urine and blood.

CSF is a common biofluid for AD biomarker investigations due to its proximity to the pathology. Samples can be obtained from either the ventricular region, which is near the brain, or the lumbar area. MS can then process the samples to get the relative protein and peptide abundance. The resulting peptide and protein abundance may differ between these samples, even when processed simultaneously. Due to the invasive nature of ventriculostomy, access to ventricular CSF from living beings is rare, especially those with a pathological diagnosis. Therefore, the difference in AD biomarkers between CSF obtained via lumbar puncture and ventriculostomy remains unclear in current literature.

1.1 Problem

The following research questions are investigated in this thesis:

1. This project conducts an exploratory analysis of proteomic abundances along AD progression from lumbar and ventricular CSF. By investigating the characteristics and missingness in subgroups of the data, we will impute missing values, evaluate imputation strategies and reduce potential bias derived from batch effects. Further, we evaluate resampling and reweighing methods to handle the class imbalances. This exploratory analysis is done to improve the performance of the data on classification tasks and biomarker discovery and also to gain a deeper understanding of the data.
2. Using ML models, investigate biomarkers on a proteomic level using feature selection methods to predict patient’s pathological status and identify differences and similarities within the following subgroups:
 - (a) Proteins found in ventricular and lumbar CSF samples.
 - (b) Peptides found in ventricular and lumbar CSF samples.

1.2 Ethical Considerations

Ethical considerations are of great importance when it comes to omics data, particularly when it contains phenotypic information that can be traced back to individuals. One of the more pressing concerns is the issue of privacy and the potential for the data to be misused. However, any names or identifiers have been omitted as they

are considered personal data under the EU's General Data Protection Regulation. While the dataset provided by Sahlgrenska University Hospital consists of phenotypic data, it cannot be linked to particular individuals. The dataset consists of CSF samples from patients with idiopathic normal pressure hydrocephalus undergoing shunt surgery. Collecting ventricular CSF samples is considered an invasive operation, and for ethical reasons, it is not possible to collect them from healthy individuals who are not suffering from neurological conditions.

2

Background

The background chapter is divided into two main sections: bioinformatics and ML. The bioinformatics section introduces the concepts of biomarkers, AD, idiopathic normal pressure hydrocephalus (iNPH) and mass spectrometry to establish a foundational understanding of the problem domain. The ML section introduces missingness, imputation, feature selection, and model fitting, which are techniques common in the data science domain. These techniques form the backbone of the data preprocessing, feature extraction and model development stages.

2.1 Bioinformatics

The following foundational knowledge in bioinformatics is necessary to understand the project's problem and scope and provides domain-specific insights for missingness and bias patterns from batch effects. The biological aspects are somewhat simplified while remaining factually accurate.

2.1.1 Biomarkers

A biomarker refers to a measurable biological set of molecules, or pathogenic process, that can indicate the presence of a particular physiological or pathological disease [8]. These biomarkers may be proteins, peptides or any other kind of nucleic acid whose measurable level is an indicator of the stage of a given disease and can, therefore, be used to accurately predict the disease stage of a patient [9]. Furthermore, they are crucial in advancing medical research through earlier diagnosis and prognosis of disease treatment. Certain biomarkers may be able to identify the onset of a disease before physiological manifestations [10]. Simplified examples of what a biomarker could be is an increase in temperature indicating fever or how blood sugar level and insulin can indicate the presence of diabetes [11]. Biomarkers can serve different roles in disease diagnostics, prognosis or treatment strategies. An ideal diagnostic biomarker would allow immediate identification of a particular disease. However, combining several less specific biomarkers is often required to narrow down the potential disease spectrum [6].

2.1.2 Alzheimer's Disease

AD is an ageing-associated neurodegenerative disorder and is the leading cause of dementia. It is estimated to affect around 50 million people worldwide and is expected to increase to roughly 150 million in the year 2050 as life expectancy increases [12]. Although the disease is commonly observed in the elderly, it is not considered a standard element of ageing [13]. Some of the primary manifestations of AD include cognitive decline and progressive memory loss, leading to histopathological changes, such as degeneration of the hippocampus [14]. Due to the slow progression and initial signs, a common misconception is to assume that this degeneration is a normal part of ageing, something that prolongs an early prognosis [15]. Despite extensive research and drug development on AD, including recent breakthroughs in the field, the exact causes of the progression of the disease are still unknown, and there are no available cures or effective drugs for preventing the disease [16]. Therefore, early detection of AD is paramount to its management [17], as current treatment is focused on slowing down the pathogenesis progression. To improve the detection rate and treatment of AD, finding reliable biomarkers to help with early diagnosis and drug response is at the forefront of dementia-related research [18]. In 1984, it was discovered that the beta-amyloid ($A\beta$) peptide is associated with AD [19]. It is one of the primary studied biomarkers, together with tau protein, used in AD identification and treatment [20]. As $A\beta$ peptides increase, they aggregate and create amyloid plaques, followed by the formation of neurofibrillary tangles (NFTs) due to the increase in tau protein (T), as shown in figure 2.1.

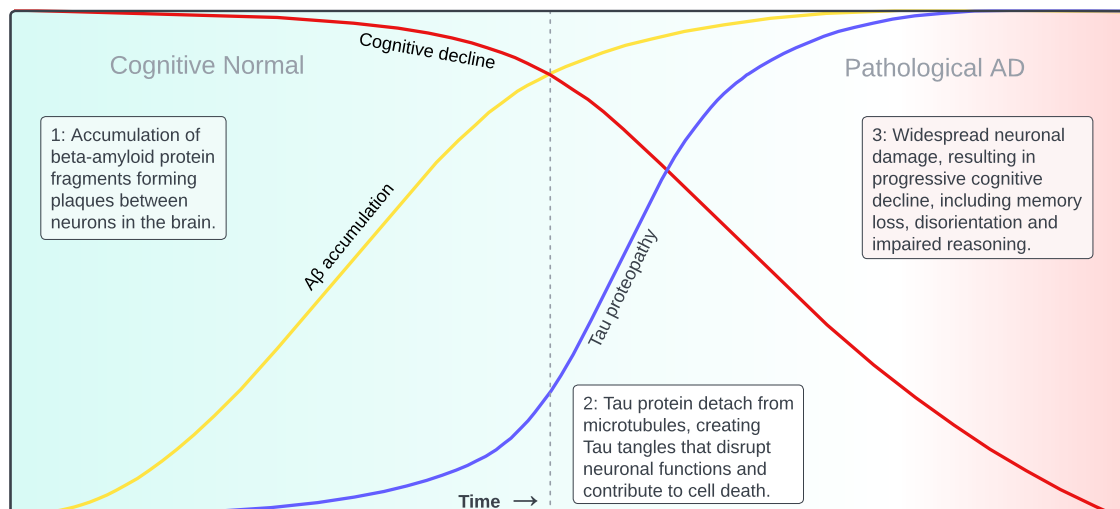


Figure 2.1: AD pathogenesis with respect to $A\beta$ plaques and tau tangles. Image inspiration from [20].

The formation of these biomarkers is used to categorize the disease stages [7]. The stages can be categorized as cognitive normal (CN), mild cognitive impairment (MCI) and AD. CN has no or low concentration of amyloid plaques and NFTs ($A\beta^-T^-$), MCI is recognized by an increase in amyloid plaques ($A\beta^+T^-$), and AD is identified by an increase in NFTs ($A\beta^+T^+$) [21]. This is a simplified version of staging, breaking the AD continuum into stages and examining singular biomarkers in each stage.

These biomarkers are not the only signs used to detect AD, with recent studies [17], [22], [23] proposing new biomarkers for neurodegenerative disorders.

2.1.3 Idiopathic Normal Pressure Hydrocephalus

iNPH is another neurodegenerative disorder, sharing some symptoms with AD, such as cognitive dysfunction [24]. Furthermore, patients with iNPH have a higher risk of developing AD, as the prevalence of AD is elevated in iNPH compared to the general population [25]. The similarities are further shown in the CSF. Biomarkers used for classification of the AD progression continuum, $A\beta$ and tau protein [26], seem to be prevalent in iNPH CSF samples as well, suggesting a connection between the two diseases [25]. The standard procedure of treatment of iNPH involves installing a CSF diversion shunt [27]. The shunt drains excess CSF from the cerebral ventricle to an extracerebral space so that the pressure on the brain is decreased [26]. Retrieving CSF samples from the lumbar is done a week before the surgery, a relatively non-invasive procedure that can be performed under spinal anaesthesia. Ventricular samples, however, are only collected during neurosurgery. During the installation of the CSF diversion shunt in iNPH patients, ventricular samples are procured. Thus, the ventricular CSF is never studied in healthy individuals [28]. The total peptide concentration in CSF can vary widely amongst individuals. This means that the composition of samples may differ, distorting the measurements of individual peptides [29]. The overall proteomic composition of samples from lumbar punctures and ventriculostomy are two common methods for extracting CSF samples. The overall proteomic composition of these different samples differs, as does the prevalence of AD biomarkers. A minority of peptides distribute equally between the two CSF spaces, while some are only detected in one of the samples. However, most were found in both samples using a mass spectrometer but differed in concentration [28].

2.1.4 Mass Spectrometry

The mass spectrometer is an important tool in the field of proteomics, and MS is the most widely used method in high-throughput proteomics [30]. It is an analytical device and process that can determine the masses of small molecules in a chemical compound, such as CSF, by separating molecular ions according to their mass-to-charge ratio (m/z). This is done by ionizing the compound, giving the content an electrical charge, allowing for separation and measurement [31]. In simpler terms, MS allows us to get the chemical composition from a CSF sample, which can be translated to a relative protein and peptide abundance, as shown in Figure 2.2.

However, while MS is a powerful technique, it has limitations. Sample preparation is a major part of the MS process and the step most prone to human and measuring errors. These errors cumulate, which can lead to two samples showing different compositions [32], despite being from the same individual. Consequently, comparing protein and peptide abundances across different MS experiments can be challenging due to potential biases and variability introduced by differences in experimental conditions. Multiple methods to combat these limitations have been developed, one of these being Tandem Mass Tag (TMT).

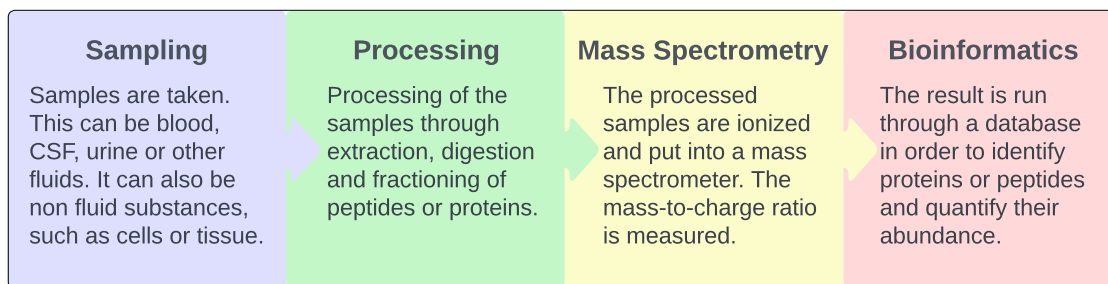


Figure 2.2: Simplified workflow of relative abundance quantification through MS. Image inspired from [20].

2.1.4.1 Tandem Mass Tag

TMT is an isobaric labelling strategy, allowing for parallel multiplexing, meaning multiple samples can be processed simultaneously through a mass spectrometer. This is one of the most frequently used techniques for quantifying relative protein and peptide abundance [33]. The process is similar to that of normal MS, with samples from up to 18 individuals being processed simultaneously with the least bias possible. The prepared samples are each tagged with a different isobaric chemical tag variant. Following the tagging, an equal quantity from each sample is pooled into one container that is run through the mass spectrometer. The first MS spectrum will show the peptide abundance across the entire pool, but a second MS spectrum can report the relative abundance from each sample in the pool based on the individual chemical tags [34]. The relative protein and peptide abundance is a comparative unit across all MS/MS runs, measuring the intensity compared to the reference sample used [35]. TMT is widely used and is an efficient method for unbiased quantification of proteins and peptides [36], is sensitive, accurate and easier to use than other methods [37]. Furthermore, TMT excels at reducing the number of missing values that can arise during multiplexing. The precision of quantifying the samples is particularly high [38] and allows for the analysis of multiple samples simultaneously. This increases the time cost efficiency. However, despite attempting to standardize and remove as much bias as possible from the sample preparations, new biases appear from the usage of TMT, namely batch effect [39].

2.1.4.2 Batch Effect

Batch effect refers to the variation in data that arises due to technical imperfections rather than biological differences. This means that when performing two TMT experiments on fluids from the same individuals, these two batches can result in fairly different MS spectra. These differences may arise from various sources, ranging from samples prepared by different people, locations or batches of preparation reagents [40]. It has been shown that the same run at different time points already introduces a batch effect [41]. A batch in this context is the combined samples of individuals used during a TMT run. Due to the constraint of sample capacity of TMT, when performing large-scale research of more than 18 individuals, multiple TMT batches must be combined [38]. Multiple papers discuss the effects and potential solutions for

dealing with batch effects [33] [36] [38] [42], one problem being the introduction of missing values. Within a single batch, missing values on a peptide level are usually in the $<1\%$ range, but when combining two batches, the concentration of missing values increases to roughly 20% [38]. The problem seems to be that despite its presence, not all peptides are displayed on the spectrum. Therefore, combining two or more batches with different sets of measured peptides results in a dataset with increased levels of missing data, yielding a higher missingness. Another problem is finding the relative abundance of peptides across batches. One common solution is to use the same reference sample mixed in with the individual samples in each TMT batch, as shown in Figure 2.3. The individual sample values can then be normalized based on the reference values participating in all batches if the reference value differs between batches.

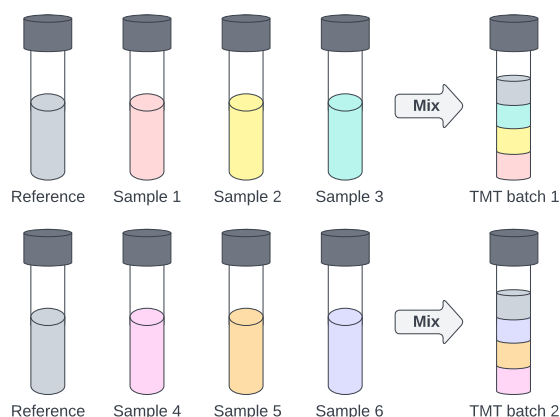


Figure 2.3: Batches for Tandem Mass Tag with the same reference.

2.2 Machine Learning

This section addresses key ML aspects relevant to this thesis, specifically managing high-dimensional proteomics data with missing values. It introduces commonly used models in bioinformatics and discusses methods for evaluating these models

2.2.1 Missingness in Proteomics Data

The missing data problem arises frequently in practice [43]. Missingness can result from measuring errors or inaccessibility of features or from sensitive information and non-responses during surveys. There are three main categories of missingness, depending on the underlying mechanics resulting in said missingness. In data that is missing completely at random (MCAR), the probability of data being missing is unrelated to both the observed and unobserved data, meaning that missingness occurs randomly without any systematic difference between observed and unobserved data. The conditions of MCAR is defined in 2.1, where R is a missing value matrix with the same dimensions as the dataset, with 0 if Y is observed, and 1 if it is missing and q is a vector of values that indicates the association between missingness in R and the dataset Y .

$$p(R|q) \tag{2.1}$$

For missing at random (MAR), the probability of data being missing only depends on the observed data, meaning that given the observed data, there should be no difference between observed and missing data. The conditions of MAR is defined in 2.2, where Y_o is the observed data.

$$p(R|Y_o, q) \tag{2.2}$$

For missing not at random (MNAR), the missing data is related to the missing values themselves and refers to missing data that is neither MCAR nor MAR [44]. Dealing with MNAR data usually requires more specialized techniques and knowledge within the domain of the data [45]. As mentioned in section 2.1.4.2, introducing missingness by combining data from multiple batches would result in a dataset with MNAR missingness. The conditions for MNAR is defined in 2.3, where Y_m is the missing data.

$$p(R|Y_o, Y_m, q) \tag{2.3}$$

Missingness within the domain of proteomics can appear due to multiple factors, including biological, technical and analytical. Some examples are sample abundance being below the instrument detection threshold, sample preparation errors, mislabeling of peptides during digestion, ionization problems or data analysis algorithms failing to recognize the peptides, or that the peptide is not in the database library [45].

MNAR typically occurs due to the abundance being below the threshold limit of the instrument, and values are typically missing across a feature [46]. MAR and MCAR are often missing across a sample or a batch and are typically due to technical issues such as co-elution of compounds [46].

There are multiple reasons for handling missing data. Missingness can increase the bias in the data and will lead to decreased efficiency [47]. This results in fewer conclusions being drawn from the data due to inadequate representation of the samples and reduced statistical power [45]. Furthermore, some ML models are unable to handle data with missingness while others can.

Some ways of handling missing data include deletion of other data that is affected by the missingness. However, it has been shown that this may introduce bias in analysis, especially when the missing data is not randomly distributed [48]. Instead of removing existing data in order to minimize missingness, imputation of missing data is often a better choice [44].

2.2.2 Imputation

Techniques for handling the missing data have a wide range of complexity, with the simpler being naïve imputation techniques. Zero imputation, where zero is imputed

to all missing values, is an example of a naïve imputation technique. Mean, median, constant and minimum imputation follow similar logic; however, imputing with a constant value can result in samples with too few differences and lead to poor inference of the actual values [45]. However, there are cases when naïve imputation performs well, particularly when values are MNAR [45]. A simple yet effective method is SampMin, where for each feature, all missing values are imputed with the minimum value found in the feature. Liu and Dongre [49] showed that SampMin performed well on TMT data compared to other methods for MNAR imputation, such as QRILC and MinProb. The reason may be that MNAR values are assumed to be introduced due to being below the detection threshold, and therefore SampMin imputes closer to that threshold [45].

Missingness can either be evenly spread throughout the data or more predominant in a minority of features. There seems to be no clear consensus on the level of missingness considered too low or too high to perform imputation on, neither on the entire dataset nor the feature level. Research has been done on different levels of missingness, ranging from as low as 5% to 90% [47], [50]. Furthermore, it is mentioned that results imputed on missingness higher than 40% should only be viewed as hypothetical [51], while others promote different imputation methods depending on the missingness level [45]. There are also recommendations of dropping either the sample containing missingness or the features themselves if there is >15% missingness [46].

A common yet more advanced imputation technique is Multivariate Imputation by Chained Equations (MICE) [52]. It is a multiple imputation method [53] and is often the choice for complex, incomplete datasets. MICE is performed through the following steps [54]:

1. The algorithm initializes the missing data with a simple imputation method, often mean imputation.
2. Then, one originally missing value, Y_j , is regressed on the other variables in the data.
3. The dependent value Y_j is regressed on the independent variables, then imputed based on the regression. This imputed value replaces the initial mean imputed value.
4. Step 2-4 is performed for each Y_j , consisting of one iteration.
5. Multiple iterations are performed until a stopping criterion is reached, often a certain number of iterations.
6. Multiple complete datasets are imputed, pooled and averaged to get a final imputation.

For MCAR and MAR data, there are many imputation techniques available. However, conventional imputation methods are unsuitable when the missing data is MNAR. There are other methods available, but these are complex and subject to multiple assumptions [55]. As proteomics data is often categorized as MNAR [46], [56], new methods for handling the missingness are continuously developed. Even techniques

such as MICE, which assumes MAR, often outperform more conventional methods when imputing on MNAR data, such as minimum imputation [50], [57]. Furthermore, imbalanced or high-dimensional datasets, something that is common in proteomics data, can result in biases or information loss during imputation [45].

2.2.3 Models

Machine learning models are used for predictive analytics, which are able to learn patterns and relationships in data. A recent study by Grueso and Viejo-Sobera [5] shows that the support vector machine (SVM) is the most common model within the field of proteomics. However, other studies show that it is often outperformed by random forest (RF) and Extreme Gradient Boosting (XGBoost) [17], [58]–[61] in AD prediction tasks. Furthermore, Logistic regression (LR) models have been used to predict whether individuals will experience a decline or maintain stability in their diagnostic status [21], [62]–[64]. Finally, the least absolute shrinkage and selection operator (Lasso) is a regularization method that can reduce the feature space, commonly used in ML proteomics [5], [17], [65].

The following section describes the fundamentals of four ML models: LR, Lasso, RF and XGBoost.

2.2.3.1 Logistic Regression

Logistic regression is a probabilistic classifier that makes use of supervised learning. The output is most commonly a binary discrete value such as yes/no or 0/1, but it can also be extended to provide predictions for more than two classes using a multinomial model. The logistic function is an S-shaped function that projects values from $[-\infty, \infty]$ to a range of $[0, 1]$.

The general logistic regression function for multiple predictors is defined as follows [66]:

$$p(Y = 1|X) = \frac{e^{\beta_0 + \beta_1 \cdot X_1 + \dots + \beta_p \cdot X_p}}{1 + e^{\beta_0 + \beta_1 \cdot X_1 + \dots + \beta_p \cdot X_p}} \quad (2.4)$$

It can further be rewritten as follows:

$$\frac{p(Y = 1|X)}{1 - p(Y = 1|X)} = e^{\beta_0 + \beta_1 \cdot X_1 + \dots + \beta_p \cdot X_p} \quad (2.5)$$

where the left-hand side is called the odds. It can take any value in the range $[0, \infty]$. However, by using the log odds or logits defined in eq. 2.6, the function can take any values in the range $[-\infty, \infty]$.

$$\log \left(\frac{p(Y = 1|X)}{1 - p(Y = 1|X)} \right) = \beta_0 + \beta_1 \cdot X_1 + \dots + \beta_p \cdot X_p \quad (2.6)$$

where $X = (X_1, \dots, X_p)$ are p predictors.

Maximum likelihood estimation estimates the coefficients $\beta_0, \beta_1, \dots, \beta_p$ from the logit transformation.

2.2.3.2 Lasso

Lasso is a method for regularizing linear models and feature selection. It penalizes the regression model's coefficients, bringing some of them to zero. The remaining non-zero coefficients are the selected features, increasing the model's interpretability by removing redundant features. The hyperparameter λ is used to tune the strength of the regularization. A large λ will shrink all the coefficients to exactly zero, while a λ equal to zero will give the least squares fit. The Lasso fits the following model [67]:

$$\mathbb{E}(y_i | \mathbf{x}_i) = \boldsymbol{\beta} \cdot \mathbf{x}_i, \quad (2.7)$$

given the response variables y_1, \dots, y_n and p -dimensional covariates $\mathbf{x}_1, \dots, \mathbf{x}_n$, by minimizing the following function:

$$\sum_{i=1}^n (y_i - \boldsymbol{\beta} \cdot \mathbf{x}_i)^2 + \lambda \sum_{j=1}^p |\beta_j| \quad (2.8)$$

where $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)$ is the regression coefficients. Not only is Lasso stable with its high consistency [67], but it is also commonly used for proteomics feature selection tasks. Reducing the high-dimensional data into fewer features with Lasso often results in better predictive performance than using the entire feature space [17].

As shown in Figure 2.4, the geometric interpretation of Lasso has sharp *corners* in the constraint. If the sum of squares *hits* one of these corners, the coefficient corresponding to that axis shrinks to zero. In two dimensions, the Lasso has the geometric shape of a diamond, and as the number of features increases, so does the number of *corners*. Hence, it is likely that some coefficients shrink, subsequently performing feature selection.

2.2.3.3 Random Forest

Random forests [69] are one of the most popular ML algorithms for both classification and regression tasks, not the least within the bioinformatics domain [70]. It is an ensemble technique wherein multiple decision trees are created. These tree-like data structures have internal nodes representing decisions about how to split the dataset into subsets. The final leaf nodes of these structures represent the class label prediction. Each decision tree is considered a weak classifier, performing only slightly better than random guessing. However, when combining their predictions, they usually result in a better model with higher accuracy.

During the training of decision trees, each node consists of a feature that splits the data into more homogeneous groups than before the split. Measuring the homogeneity can be done with different techniques, common measures of impurity being Gini and Entropy [71], with the functions being defined in 2.9 and 2.10:

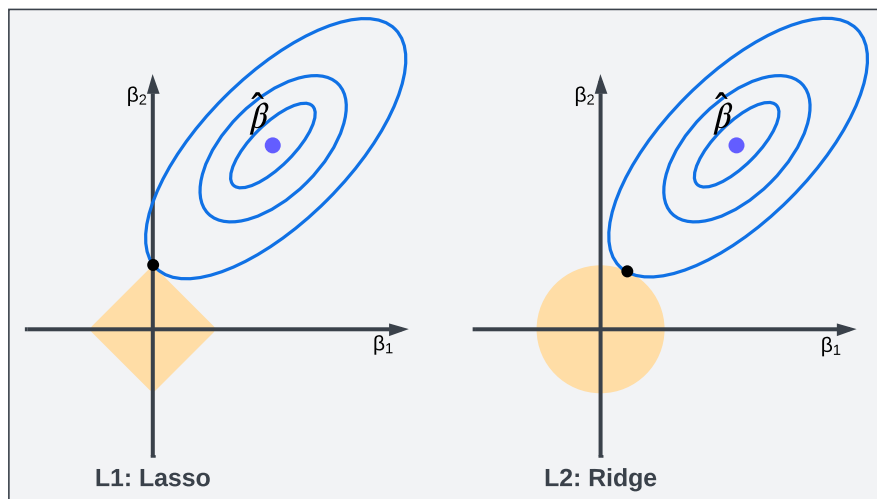


Figure 2.4: Geometric interpretation of Lasso and Ridge regression. Image inspiration from [68].

$$\text{Gini}(S) = 1 - \sum_{i=1}^c p_i^2 \quad (2.9)$$

$$\text{Entropy}(S) = - \sum_{i=1}^c p_i \log_2(p_i) \quad (2.10)$$

where $\text{Gini}(S)$ is the Gini score of the set S , $\text{Entropy}(S)$ represents the entropy of the set S , c is the number of classes in S , and p_i is the proportion of examples in S that belongs to class i . In a multi-class classification problem, one class should have a score of 1, while the other classes have a score of 0 in order to maximize the homogeneity of the dataset.

A single decision tree tends to overfit the data. One method RF utilizes to combat this is through bagging (Bootstrap Aggregating) [72]. Each tree is trained on a subset of the data through sampling with replacement, meaning not all data is shown to each decision tree. Furthermore, each tree only considers a subset of the features when performing the split, introducing some randomness into the tree ensemble. The estimation of each decision tree is taken into account through voting. The majority voted, also known as hard voted, classification is the final estimate of the model.

Recent works within the field of bioinformatics have seen an increase in the usage of the RF model [70]. The high interpretability and function to measure feature importance are important factors to the RF model's popularity in proteomics. Furthermore, the ability to generalize through bagging and random subsets, as well as being able to work on data with missing values, a variety of variables and high dimensional data enables it to outperform other models within the Alzheimer's Disease prediction domain [5], [58]–[60].

2.2.3.4 Extreme Gradient Boosting

XGBoost is another decision tree-based ensemble model, developed in 2016 by Chen and Guestrin [73], that has gained popularity in ML contests due to its flexibility, efficiency and portability in various tasks. It constructs an ensemble of decision trees sequentially, with each subsequent tree aiming to correct the errors of its predecessors.

Unlike traditional boosting methods such as AdaBoost, which adjusts the weights of training instances to focus on previously misclassified samples, gradient boosting models employ a gradient descent algorithm to minimize the loss function directly. This means that each new tree fits the gradient of the loss function with respect to the ensemble predictions made by the previous trees. By doing so, XGBoost directly addresses the residual errors left by the preceding trees, leading to more efficient and accurate model predictions. Using gradient descent enables XGBoost to optimize the model parameters principally, iteratively refining the ensemble predictions by reducing the loss function gradient [73]. Figure 2.5 shows a high-level overview of XGBoosts architecture. Summing up the individual contributions, f_k , of each individual tree results in a final prediction.

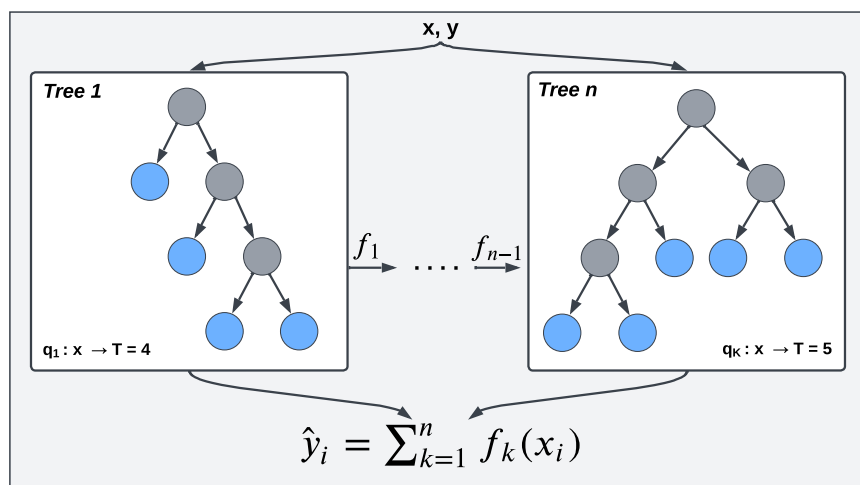


Figure 2.5: High-level overview of XGBoosts cumulative process.

One of the keys to XGBoosts performance is its use of Taylor expansion to approximate the loss function for the first and second-order derivatives, reducing computational complexity in finding the optimal split for each decision tree. This leads to a more efficient optimization, especially for large and complex datasets. XGBoost supports multiple loss functions, with softmax cross-entropy being the standard for multi-class classification tasks, as shown in eq. 2.11:

$$\text{Softmax Cross-Entropy} = -\frac{1}{n} \sum_{i=1}^n \sum_{j=1}^C y_{ij} \log(\hat{p}_{ij}). \quad (2.11)$$

Here, n is the number of samples, C is the number of classes, y_{ij} is an indicator variable and \hat{p}_{ij} is the predicted probability of sample i belonging to class j . In

order to control the complexity of the model, XGBoost incorporates a regularization term that penalizes large tree structures. This prevents complex decision trees, reducing the risk of overfitting on the training data. The mathematical formula for the regularization term in XGBoost can be seen in eq. 2.12:

$$\Omega(f_k) = \gamma T + \frac{1}{2} \lambda \|w\|_2^2. \quad (2.12)$$

Here, $\Omega(f_k)$ is the regularization term for the k -th tree, T is the number of leaves in the tree, and $\|w\|_2^2$ represents the squared L2 norm of the leaf weights. Furthermore, γ (the regularization parameter for tree complexity) and λ (the regularization parameter for the leaf weights) are hyperparameters that can be controlled in order to balance between complexity and performance to find the optimal trade-off between bias and variance. The objective function in XGBoost is a combination of the loss function and regularization term, as shown in eq. 2.13:

$$\text{Objective} = \sum_{i=1}^n \text{Softmax Cross-Entropy}(y_i, \hat{y}_i) + \Omega(f_k). \quad (2.13)$$

Here, Softmax Cross-Entropy is the loss function, y_i is the true label, \hat{y}_i the predicted value and $\Omega(f_k)$ the regularization term described in eq. 2.12. Finally, XGBoost is a combination of techniques resulting in a highly scalable tree-boosting system through accelerated parallel and distributed computing [73].

2.2.4 Dimensionality Reduction

The term "curse of dimensionality" refers to when the number of features p exceeds the number of observations n , and was initially coined by Bellman [74]. High-dimensional data refers to data that has many features or variables. However, many of these features may be redundant or unimportant for the task at hand. When these features are included in the analysis, they increase the search space, making it more difficult and time-consuming to identify patterns and relationships within the data. Furthermore, careful consideration of distance metrics has to be considered due to reasons such as increased sparsity in high dimensional data [75]. Including too many features can lead to overfitting, where the model is too complex and performs poorly on new data [76]. Therefore, it is important to identify and remove these redundant features to improve the efficiency and accuracy of the analysis. Two common approaches to address this issue are feature selection and feature extraction, further described below.

2.2.4.1 Feature Selection

The practice of feature selection offers a means to streamline datasets by reducing the number of features in order to reduce noise from redundant and irrelevant features. Feature selection can be used for various types of ML problems, such as supervised, unsupervised, and semi-supervised learning. Supervised feature selection is typically used on classification problems when labels are available, while unsupervised feature

selection is generally used for clustering tasks when labels are lacking [65]. Semi-supervised feature selection is often used when the data is partially labelled and typically relies on selecting the best features according to a similarity matrix [65].

There are three main categories of feature selection strategies: filter methods, embedded methods, and wrapper methods. Filter methods select features independent of the specific data modelling algorithm employed. Once the most relevant features are identified, they can further be used by any ML model [77]. Several common filter methods exist, such as Information gain, Chi-square and ReliefF [78]. Wrapper methods select features based on the performance of a given ML model by iteratively training models with different subsets of the data. The subsets are determined using a search strategy. By using an exhaustive search, the search space is $O(2^n)$. There are other search strategies, such as best-first and hill-climbing, that reduce the search space. Wrapper methods often obtain better-performing features since they are evaluated using real models [77]. However, they are also slower than filtering methods and require an independent validation sample and another modelling algorithm in order not to give biased results [77]. Embedded models are a trade-off between filter methods and wrapper methods. They select features during the modelling algorithm execution without further evaluation instead of first training multiple models and then selecting features based on the performance. They are, therefore, not as computationally expensive as wrapper methods.

In areas of bioinformatics, such as genomics and proteomics, the stability of the selected biomarkers is of significant importance. Different techniques often yield different features, and small data perturbations should ideally yield the same or a similar set of biomarkers. Otherwise, it would be difficult to trust that the selected biomarkers are the most predictive. However, selecting predictive yet stable biomarkers is not trivial [79].

Performing the feature selection stage correctly is important to reduce the risk of accidentally evaluating the models on the same data used for training, known as data leakage. It is common to use the entire dataset during feature selection, either intentionally or mistakenly [80]. This is especially common in small datasets that require cross-validation. Performing feature selection partly on test data can result in biases through data leakage.

2.2.4.2 Ensemble Feature Selection

Ensemble learning is based on the idea that multiple combined models yield a better result than just using a single model. Ensemble learning is commonly used in classification tasks but can also be extended to feature selection problems [81]. There are two types of feature selection ensembles. Homogeneous, which contains the same feature selection models but applied to different subsets of data, and heterogeneous, which uses a set of different feature selectors but on the same data and is more commonly used [82] than homogeneous ensembles. Finally, the feature subsets of each model in the ensemble have to be combined. Two common approaches to doing so are to take the union or the intersection of all feature subsets. Using the union as the combination method has been shown to yield lower errors than intersection [83].

Figure 2.6 shows the procedure of selecting features through an ensemble of selectors and combining them using the union.

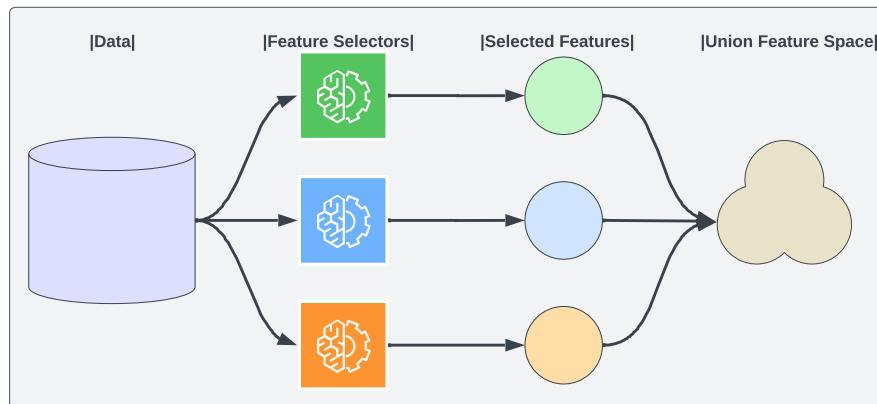


Figure 2.6: Ensemble feature selection through union. Image inspiration from [82]

A recent paper by Tandon, Levey, Lah, *et al.* [84] used an ensemble of two feature selectors, LR and SVM, with recursive feature elimination (RFE), a wrapper method that recursively eliminates features until a pre-defined number of features is reached. This was done to find the best predictive subset of proteins in classifying patients into three groups: asymptomatic Alzheimer’s disease, AD and control. Finally, the intersection of LR and SVM features was used for prediction.

2.2.4.3 Feature Extraction

Feature extraction is a technique to create new variables by combining existing ones. The aim is to reduce the number of features while still retaining the most important information. However, as the new features are created as a combination of the old ones, further analysis can be problematic since there is no physical meaning of the transformed features [65].

Principal Components Analysis (PCA) is a widely used feature extraction method. It reduces the number of data features by creating a new set of variables called principal components, which are linear combinations of the original features [85]. The first principal component has the largest possible variance. The second principal component is orthogonal to the first while still maximising the variance. The rest of the principal components are created similarly.

PCA can be utilized for data exploration by plotting the first two principal components. This can offer insights into how data preprocessing impacts the data. For example, it can help observe if any clusters are based on the target variable or different batches, suggesting batch effects. If PCA shows clustering on batches, the correlation of samples within the same batch is stronger than samples from different batches [86], indicating bias due to technical factors. An example is shown in Figure 2.7 where 2.7a illustrates how the classes cluster and 2.7b shows clustering on batches, implying batch effect.

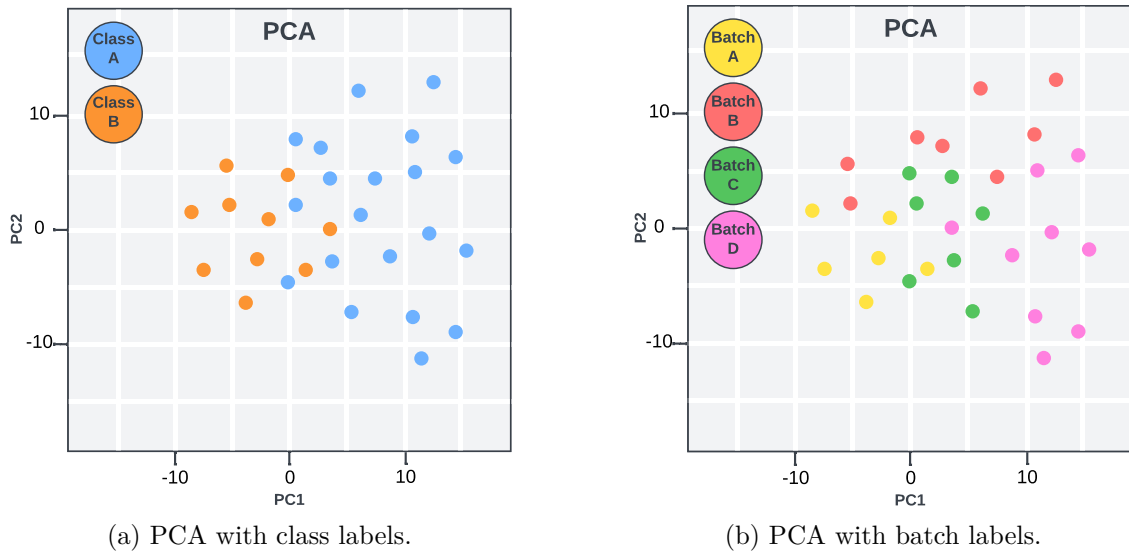


Figure 2.7: Example of how PCA-plots can be used for data exploration.

There are other techniques that can be used to visualize high-dimensional data in two dimensions, such as t-Distributed Stochastic Neighbor Embedding (t-SNE). It is a nonlinear method based on the stochastic neighbour embedding but using the Student-t distribution [87]. As shown by Wattenberg, Viégas, and Johnson [88], the overall structures and distances between the clusters generated by t-SNE are not always preserved and depend on the selection of the hyperparameter `perplexity`. This hyperparameter is related to how the algorithm balances between preserving the local and global structure of the data.

2.2.5 Imbalanced Dataset

A dataset is considered imbalanced if there is a significant difference in the number of samples in each class, which applies to binary and multi-class classification tasks [89]. Many classification algorithms aim to find the optimal decision boundary that effectively separates the classes. However, when the data suffers from imbalance, finding meaningful decision boundaries can be difficult, especially in a multi-class task [89].

Resampling techniques are used to rebalance the imbalanced sample space and are not dependent on the later classification method. Rebalancing can be done on both binary and multiclass data. The three most commonly used resampling methods are:

- **Under-sampling:** discards samples of the majority classes. Random under-sampling is a commonly used yet effective technique [81], which removes random instances of the majority class. Although this method has its advantages in some cases, one major disadvantage is the loss of potentially valuable information [90].
- **Over-sampling:** creates new samples of the minority classes. There are multiple methods, such as random over-sampling, which randomly selects samples and duplicates them, and Synthetic Minority Oversampling Technique

(SMOTE). This creates synthetic samples by randomly drawing a sample from the minority class, selecting its k nearest neighbours, and multiplying the difference of the feature vectors with a random number between 0 and 1 [91].

- **Hybrid methods:** combines over-sampling by adding minority class instances to the sample space while removing samples from the majority class.

2.2.6 Evaluation

Datasets within proteomics commonly have a small sample size, often ranging between 50 and 300 samples, and are generally imbalanced. Therefore, utilizing the entire dataset and performing k -fold cross-validation can overfitting [80]. The data is partitioned into k equal-sized folds, each serving as a distinct validation set. The model is then trained and evaluated k times, with each fold taking turns as the validation set and the remaining data as the training set.

A set of evaluation metrics will be used to assess the performance of the selected ML models in each fold. In classification tasks, many metrics are based on the following four statistics: true positive (TP), true negative (TN), false positive (FP) and false negative (FN). TP and TN refer to situations where the model correctly classified a positive instance and a negative instance, respectively, while FP and FN refer to instances where the model incorrectly identified a negative instance as a positive and vice versa. These statistics can be visualized in a confusion matrix as in Figure 2.8.

		Predicted	
		Positive	Negative
Actual	Positive	TP	FN
	Negative	FP	TN

Figure 2.8: Confusion matrix used for prediction visualization.

2.2.6.1 Accuracy

Accuracy is a commonly used evaluation metric defined as the ratio of the number of correct predictions to the total number of predictions. This metric gives a score in the range of $[0, 1]$, where a score of 0 means that the model classified all instances wrongly and a score of 1 means that all instances were correctly classified. It is defined in eq. 2.14 [92]:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (2.14)$$

Accuracy can be a misleading metric if the classes that are to be predicted are imbalanced. For instance, imagine a classification task with 90 instances of class A, while class B only has 10 instances. By only predicting that all instances belong to class A, an accuracy of 90% would be achieved. When dealing with imbalanced datasets, balanced accuracy is a more appropriate metric defined as the arithmetic mean of sensitivity and specificity [93]. Accuracy can also be employed for multiclass tasks, calculated similarly to that in a binary task.

2.2.6.2 Recall

Recall, also known as sensitivity, measures the proportion of positive classified instances that are correctly classified. A score of 1 means that all instances of the positive class were correctly classified, while a low recall score indicates that the model misclassifies many positive instances. Recall is defined in eq. 2.15:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (2.15)$$

2.2.6.3 Precision

Precision, like accuracy and recall, has its best result at 1 and worst at 0. It is calculated by dividing the number of correct positive predictions by the total number of positive predictions and is defined in eq. 2.16.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (2.16)$$

2.2.6.4 F_1 -score

F_1 -score is the harmonic mean between precision and recall and ranges between $[0, 1]$. It gives equal importance to both recall and precision and is often used when classes are imbalanced. It is defined in eq. 2.17 [92]:

$$F_1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (2.17)$$

F_1 -score can also be used for multiclass tasks. Macro F_1 -score and Micro F_1 -score are popular approaches for multiclass problems. The Macro F_1 -score is calculated by taking the average precision and recall for each class and then calculating the harmonic mean of these averages [94]. The Micro F_1 -score calculates the score by globally counting the number of TP, FN and FP and then calculating the harmonic mean of the precision and recall as usual.

Macro F_1 -score is generally preferable when there is a class imbalance and all classes are of equal interest, while Micro F_1 -score doesn't take class sizes into consideration and is more similar to accuracy in that regard [94].

2.2.6.5 Receiver Operating Characteristic & Area Under the Curve

Receiver Operating Characteristic (ROC) and Area Under the Curve (AUC) measure the ability of a classifier to discriminate between positive and negative classes across various threshold settings. This allows the metric to capture the trade-off between true positive rate (sensitivity) and false positive rate (specificity), as well as making it robust against class imbalance. By plotting sensitivity against specificity, the ROC curve provides insights into the classifier's predictive power, with higher AUC indicating greater performance [92]. An example of a ROC and AUC curve can be seen in Figure 2.9. The area under the ROC curve is the AUC value and a value over 0.5 is better than a random classifier.

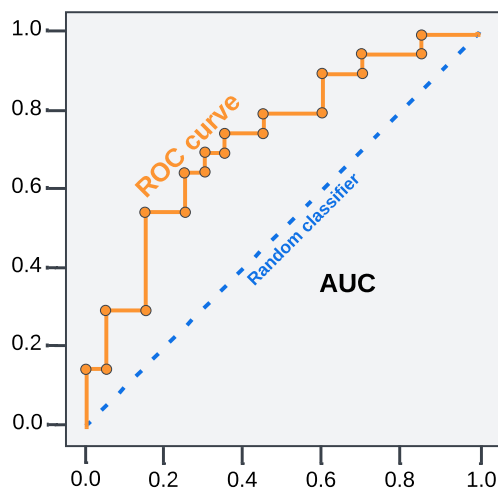


Figure 2.9: ROC & AUC, illustrating the performance of a binary classifier.

2.2.6.6 Matthews Correlation Coefficient

Matthews Correlation Coefficient (MCC) is a method unaffected by class imbalances. The MCC ranges between $[-1, 1]$ where a score of -1 equals a perfect misclassification, 0 is expected from random guessing, and 1 equals a perfect classification. It is defined as follows [95]:

$$\text{MCC} = \frac{\text{TP} \cdot \text{TN} - \text{FP} \cdot \text{FN}}{\sqrt{(\text{TP} + \text{FP}) \cdot (\text{TP} + \text{FN}) \cdot (\text{TN} + \text{FP}) \cdot (\text{TN} + \text{FN})}} \quad (2.18)$$

Several papers argue that the Matthews Correlation Coefficient generally provide more informative and truthful results than F_1 -score and balanced accuracy metrics, especially for classification tasks on imbalanced data [93], [95].

3

Experimental Setup

This chapter discusses the bigger picture of this project and the problems we aim to attack. It further introduces the datasets, the exploratory phase, and the preprocessing phase. It then discusses imputation methods, ML models, and feature selection techniques, which are implemented and applied to the datasets. Finally, it provides an overview of the pipeline that incorporates these techniques.

3.1 The Bigger Picture

This project's initial task is to use ML models to predict whether individuals in a dataset are healthy or sick. Healthy individuals in this dataset are classified as suffering from iNPH, while sick individuals suffer from both iNPH and AD. Furthermore, AD is diagnosed pathologically through a brain tissue biopsy, which means the presence of both amyloid plaques and tau tangles in affected individuals and the absence in healthy individuals. Therefore, in predicting the status of the pathology, we can find biomarkers that can distinguish the groups. These biomarkers can be used for further pathogenesis knowledge, improved drug development or more targeted treatments.

To do this, domain-specific knowledge needs to be considered. As we have missing values, the reasons for this missingness can help us impute more correctly. Due to the nature of TMT, it is reasonable to assume that missing values are missing because they are lower than the threshold required to be discovered through MS. Domain knowledge further introduces batch effect and possible ways to handle it. Additionally, the small sample size makes the potential discoveries harder to generalize. Therefore, it may be beneficial to utilize the entire dataset through cross-validation or possibly use augmented data.

Finally, we train the ML models to predict the status of the pathology on the dataset. The best-performing models are identified and analyzed to determine why they perform better in their predictions. These ML models will emphasise some proteins and peptides over others as more important for distinguishing the sick from the healthy. The proteins and peptides need to be consistent across the entire dataset to be considered as potential biomarkers. Biomarkers that always appear across cross-validation and have high feature importance will be presented as proposed biomarkers for future research.

3.2 Data

In this project, we have access to four datasets from a study conducted by Weiner, Junkkari, Sauer, *et al.* [26]. Their study aimed to identify prognostic CSF biomarkers to predict shunt responsiveness in iNPH patients. The datasets were generated using bottom-up proteomics. This involved digesting the proteins present in the CSF into peptides using Trypsin, a commonly used enzyme for this purpose. The peptides were analysed in an MS/MS instrument, and the MS/MS spectra matched to peptide sequences using SequestTM search engine with UniProtKB Swiss-Prot (TaxID = 9606, Homo sapiens) as database. Peptides were matched to proteins using Proteome Discoverer 2.5.0.400. The datasets are publicly available upon request.

All peptide abundances were first normalized to the reference channel (135N), which consists of the same sample and occupies the last channel of each TMT batch. As the abundance of peptides is known in the reference channel, normalizing the other samples based on the reference channel's result removes some of the theoretical batch effects. As aforementioned, the datasets were generated through bottom-up proteomics. This means that the relative peptide abundance is found and then used to reverse-engineer an assumed protein abundance in each sample.

The datasets contain protein and peptide abundances, respectively, where each of the datasets is median normalized and not normalized. Normalization is commonly used in TMT studies to remove some technical variations from sample preparation and TMT labelling [26]. Median normalization is calculated by dividing each protein abundance by the median protein abundance for each sample and is defined as follows [33]:

$$\tilde{X}_{ij} = \frac{X_{ij}}{\text{median}(X_i)} \quad (3.1)$$

where X_{ij} is the protein abundance of protein j in sample i and \tilde{X}_{ij} is the normalized protein abundance ratio.

To make it more clear, we will refer to the non-normalized datasets as D_{PL} , D_{PV} , D_{PeL} , and D_{PeV} where P and Pe indicate whether we refer to protein or peptide data and L and V implies lumbar and ventricular. The normalized datasets are represented with a tilde, for example, \tilde{D} .

Table 3.1: Descriptive statistics of demographic features categorized by tissue groups and divided into lumbar and ventricular CSF data. Rows 5-9 describe other clinical comorbid conditions of the patients where VCI is vascular cognitive impairment.

Description	$A\beta^-T^-$			$A\beta^+T^-$			$A\beta^+T^+$		
	Tot.	L	V	Tot.	L	V	Tot.	L	V
Mean Age at Biopsy	72.57	73.05	72.15	73.95	74.03	73.88	78.87	79.60	78.31
Min Age at Biopsy	53	53	53	59	64	59	64	64	64
Max Age at Biopsy	90	90	90	87	87	87	88	88	88
Std Dev of Age at Biopsy	8.10	7.96	8.28	5.70	5.27	6.10	6.23	6.52	6.20
VCI	4	2	2	0	0	0	0	0	0
AD	4	2	2	15	7	8	21	9	12
AD+VCI	4	2	2	4	2	2	2	1	1
Suspected AD	0	0	0	6	2	4	0	0	0
Genetic outlier	0	0	0	1	0	1	0	0	0
Male	54	25	29	45	19	26	14	7	7
Female	34	16	18	30	15	15	9	3	6

Table 3.1 describes the descriptive statistics of the dataset’s demographic features. The patients of the $A\beta^+T^+$ tissue group have a greater mean age than the other groups. Also, all patients in this group have a clinical AD diagnosis, and one of them also has VCI. Further, there are a few patients in the $A\beta^-T^-$ group that have a clinical AD or AD and VCI diagnosis, but no pathological lesions were found in the brain samples.

Table 3.2: Description of experimental and demographic features.

Feature	Description
names.channel	Combination of TMT Set and Pos.
Sample_Run_ID	Arbitrary ID (remapped to 1-106)
TMT Set	The set of samples tested simultaneously.
TMT Pos	The position in the set a sample had.
CSF_type	L (lumbar) or V (ventricular) CSF sample.
Diagnostic_classification_string3	iNPH + comorbid condition (AD or VCI).
Cortical_biopsy_grouping	0 (no AD), 1 (early stage), 2 (later stage).
Abeta_score	0-3 scale of brain lesions ($A\beta$ plaques).
Tau_score	0-3 scale of brain lesions (tau tangles).
Gender	Gender of sampled individual
Age_at_biopsy	Age of sampled individual

Table 3.2 describes the experimental and demographic features of the dataset. Where **Abeta_score** and **Tau_score** describe the number of pathological lesions, i.e. $A\beta$ plaques and tau-tangles that have been found in each brain sample and is an integer in the range of 0 to 3. A higher value means that more of the respective pathology has been found. These values have further been quantified into the feature **Cortical_biopsy_grouping** that describes the pathological stages of Alzheimer’s disease and is represented as an integer in the range of 0 to 2, where 0 indicates no

pathology ($A\beta^-T^-$), 1 is solely $A\beta$ pathology which represents an earlier disease stage ($A\beta^+T^-$). A value of 2 means both $A\beta$ pathology and tau pathology and represents a later disease stage ($A\beta^+T^+$). This is the target variable for classification.

The **TMT Set** (TMT batch) feature enumerates which batch each sample is from, and **TMT Pos** what position a sample had during in the mass spectrometer. Gender and age are also part of the dataset. Both features have been shown to influence risk factors and have distinct patterns in diagnosing AD [96], [97]. However, only protein and peptide features will be considered for this project’s predictive tasks.

What makes these datasets unique is the use of a pathological diagnosis rather than the more commonly employed clinical diagnosis. Clinical diagnosis identifies symptomatic signs of AD but can often misdiagnose other types of dementia. In contrast, pathological diagnosis is the conclusion from finding AD biomarkers through a brain biopsy. A clinical diagnosis is often classified in the CN, MCI and AD spectrum, roughly corresponding to the $A\beta^-T^-$, $A\beta^+T^-$ and $A\beta^+T^+$ pathogenesis of a pathological diagnosis. As aforementioned, the presence of proteins and peptides from ventricular CSF is fairly uncommon, as ventriculostomy is an invasive surgery not performed on healthy individuals. These two factors, combined with the presence of lumbar samples from the same individuals as the ventricular samples, make it an ideal dataset for biomarker-based staging across two different CSF sample spaces.

There are 186 samples, 85 of which are from lumbar CSF fluid and 101 from ventricular CSF fluid. The number of samples is unequal because some patients only attended one of the appointments. There are 2795 identified proteins and 18305 identified peptides from 15 TMT batches. As shown in Table 3.3, the sample sizes of the datasets are small while also imbalanced and high dimensional.

Table 3.3: Number of samples (lumbar and ventricular) divided into the tissue groups.

No. of subjects	Lumbar	Ventricular	Lumbar and ventricular
$A\beta^-T^-$	41	47	88
$A\beta^+T^-$	34	41	75
$A\beta^+T^+$	10	13	23
Total	85	101	186

3.2.1 Converting Multiclass to Binary Predictions

In Section 2.2.5, it was discussed that performing multiclass classification can be challenging, particularly when working with imbalanced datasets. Due to this difficulty and a suggestion from a domain expert, we have decided to exclude the $A\beta^+T^-$ tissue group during the pre-processing phase. As a result, this group will not be considered in the downstream prediction tasks, effectively converting the multiclass problem into binary prediction tasks.

3.2.2 Data Exploration

The data exploration phase is a critical step in understanding the datasets' characteristics, structure and correlations. This phase involves several key activities, including identification, summarization and visualization of patterns, anomalies and relevant information. Areas of interest will include the distribution of missingness and mean values according to different subgroups (sample space, analyte, batch, staging, etc).

Furthermore, this project aims to explore the distribution of already established biomarkers on these datasets. This is done by plotting biomarker-based staging of the disease spectra. Furthermore, comparisons are made by comparing the progression of staging between ventricular and lumbar CSF. The idea is to visualize differences and similarities in biomarker levels in the $A\beta^-T^-$, $A\beta^+T^-$ and $A\beta^+T^+$ tissue groups. After the modelling phase, if new potential biomarkers are found, this process will be revisited on these new biomarkers.

3.2.3 Data Preprocessing

Data cleaning is an important step that involves detecting and correcting errors and inconsistencies, thus improving the data quality for downstream analysis. We first identify what we consider erroneous measurements, namely elements equal to zero or infinity. For \tilde{D}_{PL} and \tilde{D}_{PV} , a total of 1239 values were equal to zero. For \tilde{D}_{PeL} and \tilde{D}_{PeV} , all values were within the threshold. Therefore, we suspect zero values were introduced when mapping peptides to proteins using Proteome Discoverer. We treated these values as NaN since they are below the cut-off threshold at 0.1.

The protein datasets contain a total of 282 infinity values, while the peptide datasets do not contain any. These values occur when dividing a positive protein abundance with a reference channel that equals zero and is treated as NaN, as there is no clear definition for them. Furthermore, biological data is often positively skewed. Therefore, the datasets were \log_2 -transformed to fulfil the requirements of a normal distribution. Thereafter, outliers of measurements within protein and peptides were removed if they were outside the lower or upper bounds defined in eqs. 3.2 and 3.3 respectively:

$$\text{Lower bound} = Q1 - 1.5 \cdot \text{IQR} \quad (3.2)$$

$$\text{Upper bound} = Q3 + 1.5 \cdot \text{IQR}, \quad (3.3)$$

where IQR is the abbreviation for the interquartile range. Other outlier methods, such as geometric mean-based methods, were considered, but the methods were kept simple to follow proteomics standard procedures [98]. The total number of missing values in the four normalized datasets can be seen in 3.4.

Table 3.4: This table describes the number of missing values for each of the datasets. Within batches describe how many missing values there are when not combining the TMT sets. The following row describes the number of missing values when the TMT batches are combined. The next row describes how many missing values are introduced by removing outliers.

Missingness	\tilde{D}_{PL}	\tilde{D}_{PV}	\tilde{D}_{PeL}	\tilde{D}_{PeV}
Within batches	0	0	22	45
Combined	76053	91046	687750	829928
Outliers	4304	6480	22828	37024
Total	80357	97526	710578	866952

3.2.4 Missingness Mechanisms

As mentioned in Section 2.1.4.2, TMT data generally has $< 1\%$ missing values within each batch, but when the batches are combined, more missingness is introduced due to the batches not capturing the same proteins and peptides. The same missingness mechanisms apply to our datasets, where there is no missingness within each batch in the protein datasets. When the batches are combined, the missingness increases to 32.01% and 32.25% in \tilde{D}_{PL} and \tilde{D}_{PV} , respectively. The peptide dataset \tilde{D}_{PeL} has 67 missing values total within each batch, and \tilde{D}_{PeV} has 45 missing values within each batch. When batches are combined, there is a missingness of 44.20% for \tilde{D}_{PeL} and 44.89% for \tilde{D}_{PeV} . There is no noticeable difference between missing variables between lumbar and ventricular CSF samples.

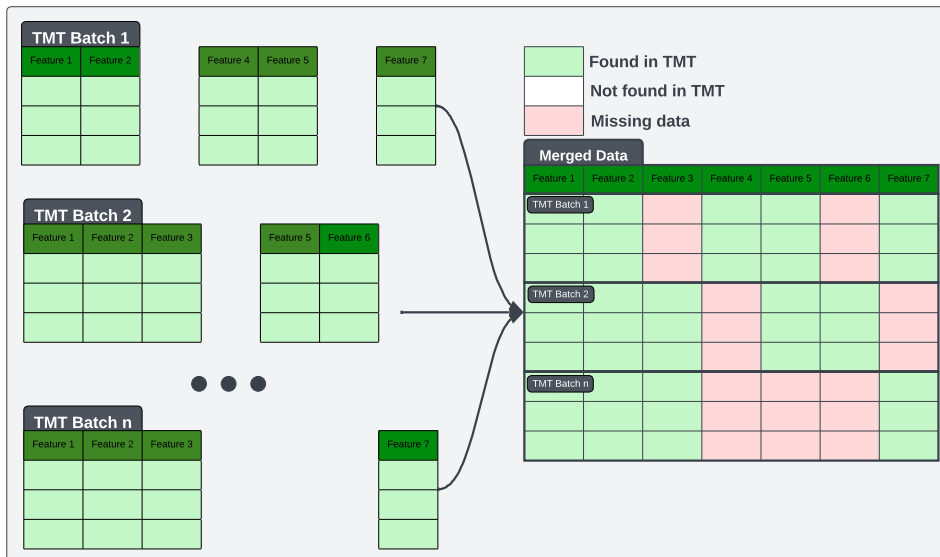


Figure 3.1: Overview of how missingness is introduced through batches.

Figure 3.1 shows how missingness is introduced through merging multiple batches with no missingness. Each batch has its own subset of features, and when these batches are combined, the union of features increase. This results in batches having features with no values, requiring either imputation of the feature in said batch or

removal of the entire feature itself. This missingness mechanism is suspected to be due to the peptide abundance in a TMT batch being below the visible threshold for the mass spectrometer.

3.3 Imputation

Many ML models, such as LR, SVM and linear regression, cannot handle missing values natively and thus depend on either imputing or removing the missing values. Therefore, the datasets are imputed using two imputation techniques, MICE and minimum imputation, and a comparison is made. As mentioned in Section 2.2.2, there is no consensus in the literature on how much missingness each feature should at most contain when imputing. Therefore, a maximum of 50% missing values for each protein or peptide was used, as imputing on more missingness tends to perform badly [45]. The remaining features were discarded for both imputation methods.

Utilizing a maximum missingness value of 50% removes ~ 8400 peptides from the dataset and ~ 950 proteins, and the remaining features with missingness undergo imputation. Minimum imputation is performed on all datasets. Furthermore, MICE is also attempted on the protein datasets, \tilde{D}_{PL} and \tilde{D}_{PV} . Finally, for exploratory purposes, two extremes were tested. Datasets where all missing values were minimum imputed, and datasets where all features with missingness were purged.

3.3.1 Multiple Imputation

To impute the missing values of the datasets, Scikit-learn `IterativeImputer` [99] was used with the `BayesianRidge` estimator. However, the `BayesianRidge` estimator was shown to utilise very large amounts of RAM. Over 230 GB in just a few iterations, compared to the `ExtraTreesRegressor` estimator, which used significantly less. However, the `BayesianRidge` estimator was much faster to execute. Therefore, an investigation into the source code of `IterativeImputer` was performed to determine why RAM usage was so high.

Listing 3.1: Removed code from `IterativeImputer` original implementation

```

798     estimator_triplet = _ImputerTriplet(
799         feat_idx, neighbor_feat_idx, estimator
800     )
801     self.imputation_sequence_.append(estimator_triplet)

```

In Listing 3.1, we can see a code snippet originating from the `fit_transform()` function of the class `IterativeImputer` on GitHub. The `estimator_triplet` is only used upon calling the `fit()` function and is not at any time called during `fit_transform()`. Therefore, we decided to remove this code, reducing the RAM usage significantly with the only drawback of being unable to refit the trained estimators' data.

Further, \tilde{D}_{PL} and \tilde{D}_{PV} were imputed five times independently with randomly drawn seeds for 30 iterations each. Thereafter, the five datasets were pooled into one by

averaging them column-wise. Due to MICE being very computationally heavy, it was not feasible to impute the peptide dataset using this method. Even allowing only a maximum of 20% missingness, there are still more than 6000 columns in both \tilde{D}_{PeL} and \tilde{D}_{PeV} respectively. Since the algorithm has a cubic time complexity with respect to the total number of columns, not only the columns containing missingness, the task is not feasible within this project.

3.3.2 Minimum Imputation

SampMin was chosen as a computationally efficient and effective imputation method for MNAR data. One of the reasons for the missing data is measuring errors during the MS phase. Despite the known existence of certain peptides, the instrument does not capture them as the minimum observable threshold is not reached. Therefore, it can be assumed that the overall peptide abundance in that batch is low yet still present in the biological sample. SampMin imputes missing values with the lowest observed value for each feature or column in the dataset, emulating the minimum observable threshold. There is no Python library with this functionality. Therefore, it was implemented by inheriting from Sklearn `SimpleImputer` and creating custom `fit()`, `transform()` and `fit_transform()` methods.

3.4 ComBat

ComBat is a commonly used batch effect correction method originally implemented by Johnson, Li, and Rabinovic [100]. It adjusts the data by estimating the location and scale using an Empirical Bayes method. ComBat has been shown to perform well on imbalanced data [39] and small datasets [100]. Behdenna, Colange, Haziza, *et al.* [101] has implemented a Python library called `pyComBat`, which was used to remove the batch effects on our datasets.

3.5 Models

In this project, XGBoost, LR and RF were used. LR and RF were implemented using the sci-kit learn library and XGBoost through the XGBoost library [102]. Hyperparameters were tuned using `BayesSearchCV` from the `scikit-optimize` library [103] utilizing Bayesian Optimization. It uses a surrogate model to represent the search space and is utilized to find parameters that try to maximize the scoring function passed as a hyperparameter. Other algorithms were also considered for choosing hyperparameters, such as `GridSearchCV` that exhaustively tries all combinations of hyperparameters and `RandomizedSearchCV` that randomly draws values for each parameter for a fixed number of times. Recent studies have demonstrated that random search outperforms grid search in efficiency. This is primarily because not all hyperparameters hold equal significance, and while grid search allocates time to exploring dimensions with minimal impact, random search focuses on more impactful dimensions [104]. The hyperparameter ranges for each model can be seen in table 3.5.

Table 3.5: BayesSearchCV hyperparameters for each model.

Model	Hyperparameter	Value
XGBoost	eta	Real(0.1, 0.5)
	max_depth	Integer(1, 20)
	n_jobs	-1
	n_estimators	Integer(50, 500)
	objective	multi:softmax
	num_classes	3
Logistic regression	penalty	elasticnet
	C	Real(0.000001, 100)
	solver	saga
	multi_class	multinomial
	max_iter	Integer(1000, 12000)
	n_jobs	-1
	l1_ratio	Real(0, 1)
Random forest	n_estimators	Integer(5, 500)
	max_depth	Integer(2, 50)
	n_jobs	-1
	min_samples_leaf	Integer(1,5)

3.6 Feature Selection

In high-dimensional statistics, the relationship between the number of variables p and the number of observations n is crucial. Traditional statistical methods are designed under the assumption that $n > p$. When $p \gg n$, these methods often fail or underperform [105]. Reducing the feature space p makes extracting meaningful insights from high-dimensional data possible.

Bolón-Canedo and Alonso-Betanzos [82] describes ensemble techniques for feature selection, an approach adopted for this project. The idea is that multiple feature selectors are given the training data, out of which the best k features are selected. Each of these feature subsets is then aggregated, either through some form of threshold method, ranking, intersection or union. The method employed in this study combined the selected features from four distinct models through the union aggregation. Intuitively, the intersection of the feature spaces seems the best choice. However, both the model evaluation and previous research suggest union outperforms intersection [82]. Four models were used for feature selection: Lasso, LR, RF and XGBoost. The Sklearn `RFE()` iteratively removes the m least important features from p until k features remain. Various values of k are examined during the modelling stage.

Moreover, the stability of this ensemble feature selection method is notable. When selecting features, especially biomarkers, selecting the same features deterministically is important for reproducibility and reliability. The stability of a feature selection algorithm can be seen as the robustness of the feature preferences it produces in training data drawn from the same distribution [79]. As feature selection is performed

on separate training data in each k-fold, more matching features across each k-fold would suggest higher stability.

3.7 Pipeline

Figure 3.2 shows the pipeline from data extracted through TMT until potential biomarkers and evaluation. Before entering the pipeline, the TMT data is normalized according to MS standards, which includes batch and individual normalization. Below is a brief description of every stage in the pipeline:

1. For some tests, the data will be binarized, as shown in 3.2.1.
2. Invalid values are replaced with NaN, as shown in 3.2.3.
3. The entire data is Log2 transformed to remove positive skew, as shown in 3.2.3.
4. Outliers are removed according to IQR, as shown in equation 3.2 and 3.3.
5. Features with missingness above a threshold are removed, as shown in 3.3.
6. Data is imputed as shown in 3.3.
7. Batch effect is removed through ComBat, as shown in 3.4.
8. Data is split into train and validation and run in 10-fold validation.
9. A subsample of features is selected, as shown in 3.6.
10. Synthetic SMOTE data is added, as shown in 2.2.5.
11. Train and validation data are scaled with StandardScaler mean variance.
12. Optimal hyperparameters are retrieved through Grid- and BayesSearchCV.
13. Models (as shown in 2.2.3) are trained on training data.
14. Data is validated through individual models and ensemble voting.
15. The feature importance through each fold is combined.
16. Results evaluated on metrics shown in 2.2.6.

The pipeline stages 1, 5, 6, 7, and 10 can be interpreted as adjustable settings rather than continuous stages. For step 1, models for both binarized and multiclass datasets were trained. Step 5 removed features based on different ratios of missing values, and step 6 explored multiple and minimum imputation. For steps 7 and 10, ComBat and SMOTE, multiple variations were explored. The settings and model results are shown in Appendix A.

According to Čuklina, Lee, Williams, *et al.* [86], batch effect correction methodologies should be performed before imputation to avoid introducing biases. ComBat does not work if there are missing values in the data; hence, this adjustment to the pipeline is needed. A possible solution would be to remove all features with missing values, which may lead to the loss of valuable quantitative information. ComBat works

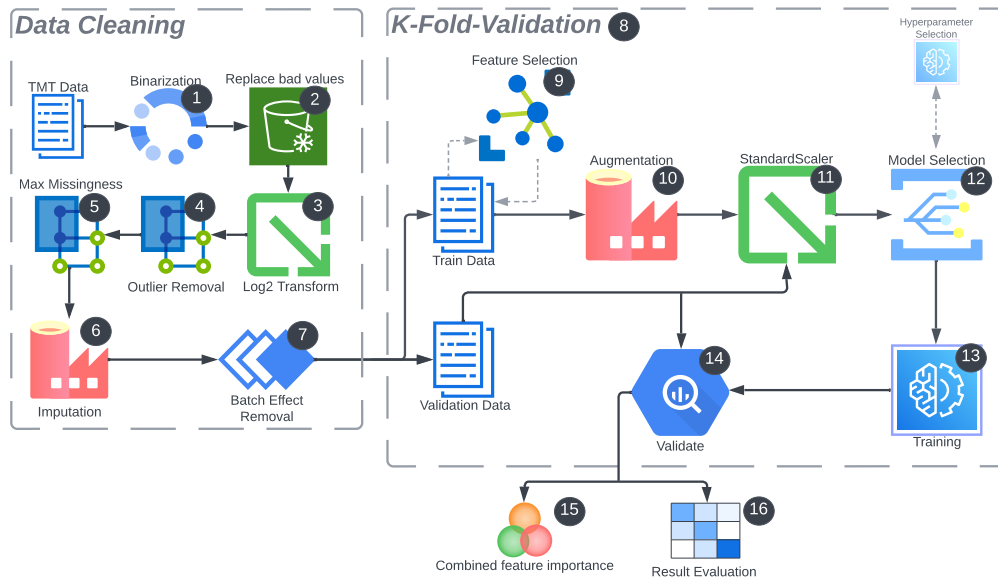


Figure 3.2: Illustration representing the pipeline used for modelling.

reasonably well after imputation, but a preferable alternative would be a software library that can perform batch correction before imputation [39].

Furthermore, performing the feature selection within each k-fold is important to reduce the data leakage risk. According to Demircioğlu [80], most papers using both cross-validation and feature selection describe utilizing the feature selection before the k-folds. This can introduce heavily biased results, as all data is used during selection. Samples used for feature selection should not be used during the evaluation of the results. While this paper is within the domain of radiomics, the problem with high dimensionality and small sample sizes is equally prevalent in proteomics.

4

Results

This chapter is divided into two parts: the exploratory analysis and the modelling results. In the exploratory analysis, we show that established biomarkers, visualized through staging, show little correlation with our dataset. Additionally, we investigate the ratios of missing data and the distributions of mean values, indicating differences in lumbar and ventricular CSF. We also employ PCA and t-SNE to visualize the data and identify the presence of batch effects. The second part includes the results gathered through the predictive models. We introduce plots showcasing metric results on binary data and how k-fold affects feature selection stability depending on the sample space size. Finally, we propose biomarkers with strong predictive power on the datasets and visualize them through staging.

4.1 Data Exploration Results

The exploratory analysis is conducted to gain domain knowledge, particularly about the datasets. We focus on techniques in the proteomics field, such as staging, along with statistical methods, such as probability density function (PDF) plotting. Additionally, we apply dimensionality reduction techniques to the data and highlight batch effect and CSF differences.

4.1.1 Staging Biomarkers

When plotting the disease staging biomarkers, the least required data cleaning was performed to keep the data as pure as possible. In each dataset, invalid values were replaced with NaN and outliers were removed. Figure 4.1 shows staging of biomarkers common in neurodegenerative disorder research. These box plots show the distribution of protein abundance across tissue groupings and CSF sample types. The following proteins were provided by a domain expert at Sahlgrenska University Hospital biomarkers commonly elevated or decreased during neurodegenerative disorders: neurofilament light polypeptide (NEFL) [106]–[108], 14-3-3 protein gamma (YWHAG) [109], [110], neuronal pentraxin-2 (NPTX2) [111], [112] and fatty acid-binding protein - heart (FABP3) [113], [114], all used in numerous papers. Furthermore, amyloid-beta precursor protein (APP) and microtubule-associated protein tau (MAPT) are also plotted, as they are considered closely related biomarkers for AD. To further compare the tissue groups, Kruskal-Wallis tests were performed on the four proteins and peptides in both lumbar and ventricular CSF datasets.

4. Results

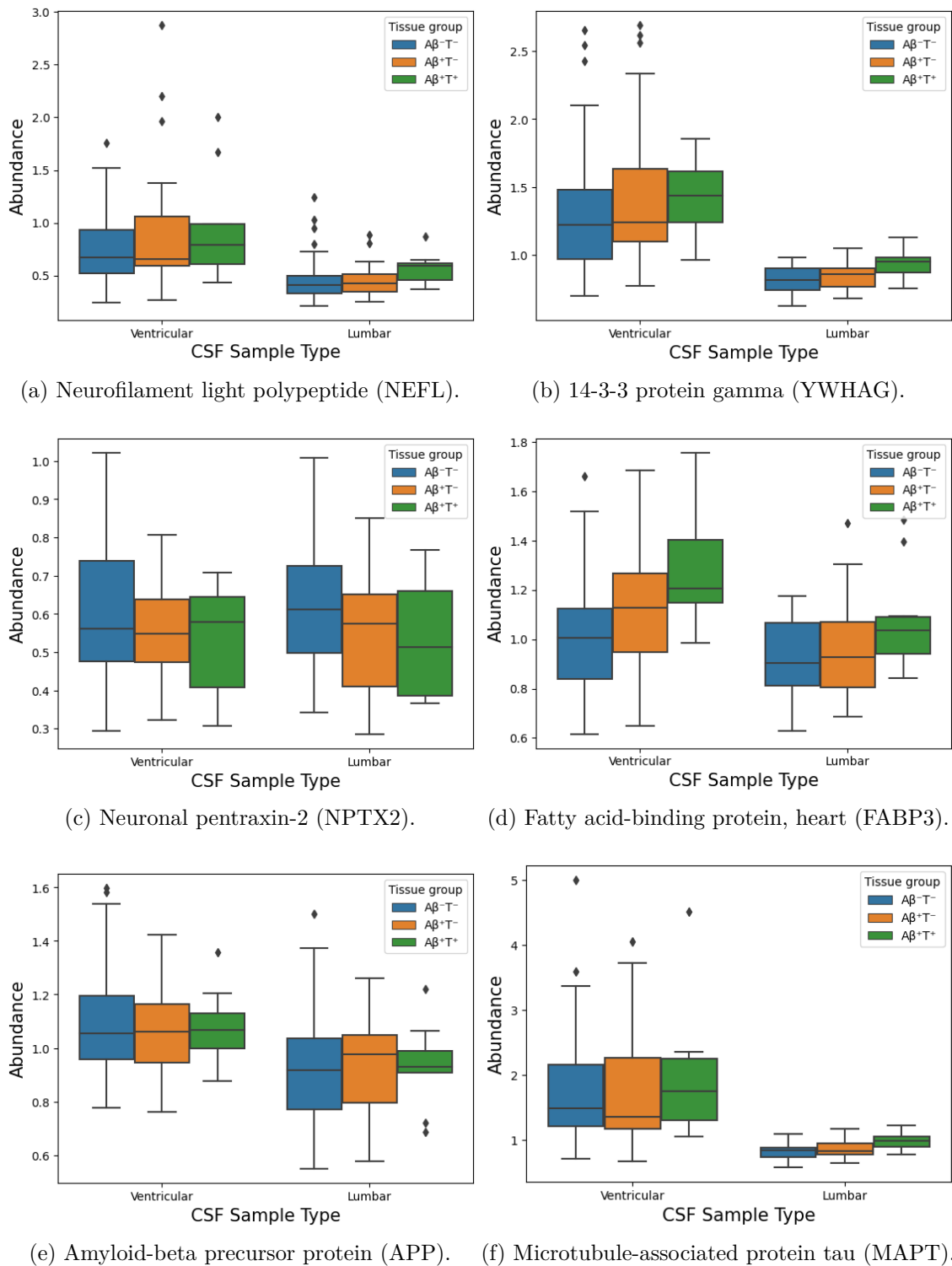


Figure 4.1: Abundance distribution of proteins on tissue groupings and CSF sample type. The bars on the left in each figure are ventricular CSF, and those on the right are lumbar CSF. Blue bars represent abundance in the $A\beta^-T^-$ tissue group, orange in $A\beta^+T^-$ and green in $A\beta^+T^+$. Ideal staging biomarkers would include clear differences within the CSF sample type over the tissue groupings, which is somewhat seen in 4.1b lumbar CSF and 4.1d ventricular CSF. The caption of the subfigures corresponds to the protein description and gene symbol.

It's a non-parametrical rank-based test that can be used to compare more than two different groups [115]. If statistical significance ($p < 0.05$) was achieved, post-hoc Dunn tests were performed to determine specific differences between pairs of tissue groups. Kruskal Wallis tests were performed using `scipy.stats` [116], and Dunn tests using `scikit-posthocs` [117].

Table 4.1: Biomarker comparison between tissue groups on the ventricular protein dataset. The protein abundance of each of the tissue groups is expressed as mean \pm standard deviation. Statistical significance ($p < 0.05$) is highlighted with bold numbers. FABP3 protein abundance in group $A\beta^+T^+$ was found to be significantly different from those in group $A\beta^-T^-$.

Protein	Kruskal Wallis	p	$A\beta^-T^-$	$A\beta^+T^-$	$A\beta^+T^+$	Post-hoc
NEFL	0.463	0.793	0.76 ± 0.35	0.87 ± 0.53	0.94 ± 0.49	-
YWHAG	2.451	0.294	1.31 ± 0.46	1.41 ± 0.50	1.42 ± 0.27	-
NPTX2	1.39	0.499	0.60 ± 0.19	0.55 ± 0.13	0.53 ± 0.14	-
FABP3	12.642	0.002	1.00 ± 0.22	1.11 ± 0.25	1.28 ± 0.21	$A\beta^-T^-$ and $A\beta^+T^+$, $p=0.002$

Table 4.2: Biomarker comparison between tissue groups on the lumbar protein dataset. The protein abundance of each of the tissue groups is expressed as mean \pm standard deviation. Statistical significance ($p < 0.05$) is highlighted with bold numbers. YWHAG protein abundance in group $A\beta^+T^+$ was found to be significantly different from those in group $A\beta^-T^-$.

Protein	Kruskal Wallis	p	$A\beta^-T^-$	$A\beta^+T^-$	$A\beta^+T^+$	Post-hoc
NEFL	4.739	0.094	0.47 ± 0.23	0.45 ± 0.15	0.56 ± 0.14	-
YWHAG	8.995	0.011	0.82 ± 0.10	0.84 ± 0.09	0.94 ± 0.10	$A\beta^-T^-$ and $A\beta^+T^+$, $p=0.008$
NPTX2	3.603	0.165	0.63 ± 0.17	0.55 ± 0.13	0.53 ± 0.15	-
FABP3	5.447	0.066	0.91 ± 0.15	0.95 ± 0.18	1.08 ± 0.20	-

4.1.2 Missingness Ratios in Lumbar and Ventricular datasets

Figure 4.2 shows the ratio of missingness for each protein and peptide in \tilde{D}_{PV} , \tilde{D}_{PL} , \tilde{D}_{PeV} and \tilde{D}_{PeL} . The missingness ratios in the non-normalized datasets are the same as their normalized counterparts. As aforementioned, almost no missing values exist in each individual TMT batch, but when combined, it introduces datasets with $\sim 35\%$ missing values for protein and $\sim 45\%$ for peptide.

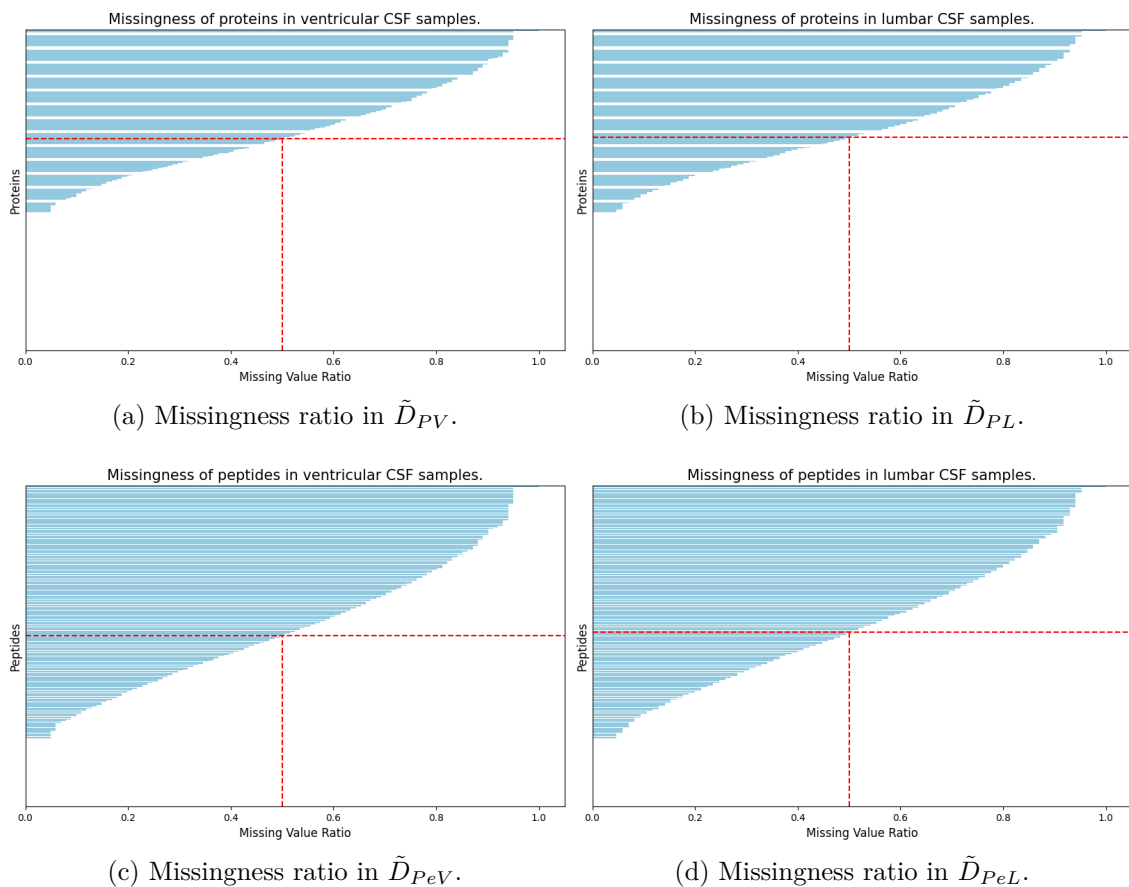
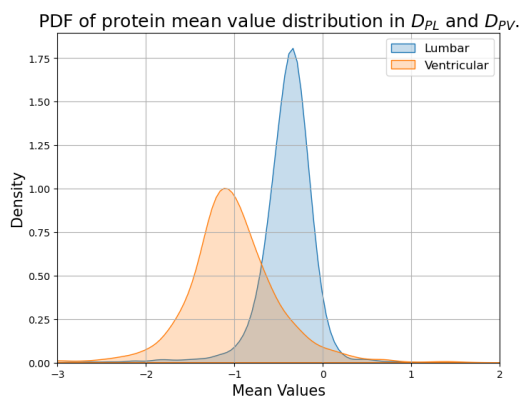


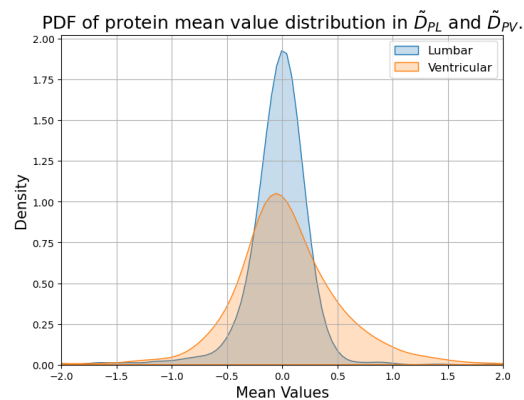
Figure 4.2: Missingness ratio in \tilde{D}_{PV} , \tilde{D}_{PL} , \tilde{D}_{PeV} and \tilde{D}_{PeL} . The vertical dashed lines denote the missingness ratio of 50%, and the horizontal line separates the proteins and peptides above and below this threshold. The proteins and peptides below the horizontal line are kept. There is no apparent difference in ratio when comparing lumbar and ventricular CSF samples in either the protein or peptide datasets, with fewer overall missing values in the protein datasets. When imputing missing values, features above a certain missingness ratio are discarded. Looking at the protein datasets, roughly 50% of all features are discarded if the max missingness ratio is set to 20%. The same missingness ratio on peptide data would discard roughly 70% of the features. Discarding all features with missing values would result in a protein dataset with 1201 features and a peptide dataset with 3887 features.

4.1.3 Distribution of Analyte Mean Value Abundances

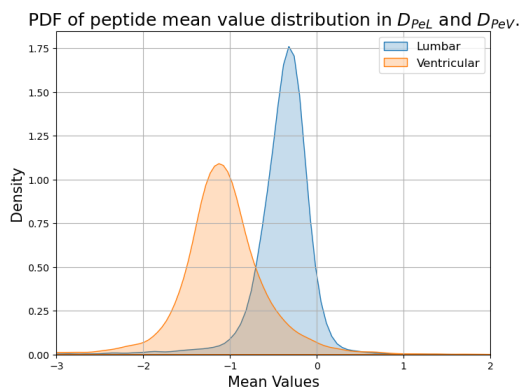
The proteomic profile differs between ventricular and lumbar CSF [28]. This variation also applies to the distribution of clinically relevant biomarkers within these CSF samples. However, the overall mean values of proteins and peptides should be relatively similar between the ventricular and lumbar CSF spaces. Furthermore, our analysis includes non-normalized and median-normalized datasets, as described in Section 3.2. In Figure 4.3, we present the PDF of the mean value distributions for proteins and peptides in lumbar and ventricular CSF. This comparison highlights the differences between non-normalized and median-normalized data. An assumption can be made that the sample spaces should have similar distributions, with ventricular samples having a higher abundance of brain-specific analytes and lumbar evenly distributed across the mean. Due to the distributions in Figure 4.3 and at the domain experts' suggestion, the modelling stage will utilize the median-normalized datasets.



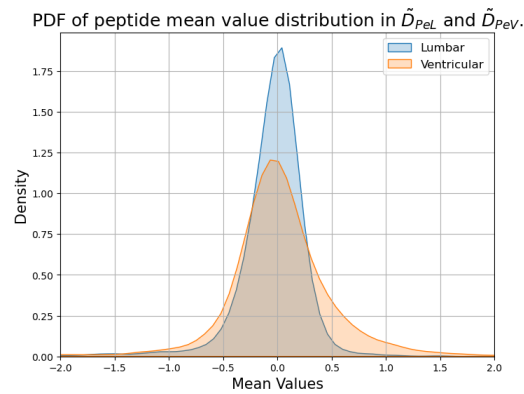
(a) Protein distribution, non-normalized.



(b) Protein distribution, median-normalized.



(c) Peptide distribution, non-normalized.



(d) Peptide distribution, median-normalized.

Figure 4.3: Figures 4.3a and 4.3c plots the protein and peptide PDF of the non-normalized datasets, and Figures 4.3b and 4.3d for the median-normalized datasets. The lumbar is represented by a blue plot, and the ventricular is represented by an orange plot. There is a noticeable difference in the distributions after normalization, particularly in the ventricular CSF, as both distributions approach zero mean. In both analytes, the lumbar distribution is more concentrated around the mean, while the ventricular distribution has a larger spread.

4.1.4 Batch Effect Results

From dimensionality reduction visualisations using t-SNE, we observe the presence of TMT batch clusters in the data. We compare the clusters on TMT batch and tissue groups before and after applying ComBat and see clear differences. However, the presence of batch effect does not necessarily correlate to worse clustering of tissue groups (see Figure 4.4).

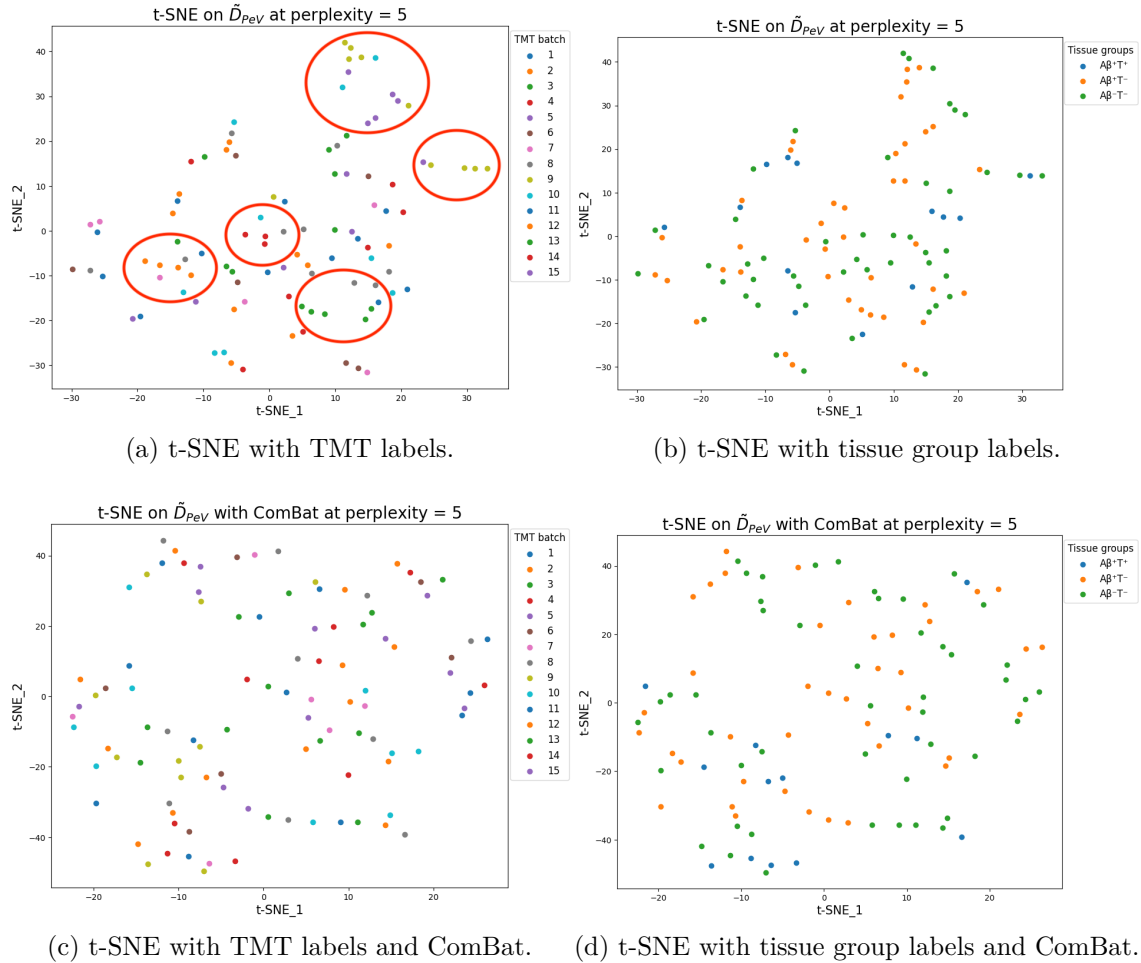


Figure 4.4: Four t-SNE plots of the \tilde{D}_{PeV} dataset with all features with missing values removed. Figures 4.4a and 4.4c are coloured by the TMT batch, while Figures 4.4b and 4.4d are coloured by tissue group. In Figures 4.4a and 4.4b, \tilde{D}_{PeV} has not undergone ComBat batch effect removal. Noticeable clusters in Figure 4.4a, as shown with red circles, indicate the presence of batch effect bias. After applying ComBat to \tilde{D}_{PeV} , Figure 4.4c shows increased entropy while retaining similar clustering patterns in the tissue group plot.

The results of PCA analysis further highlighted the differences in the two datasets \tilde{D}_{PeL} and \tilde{D}_{PeV} . Figure 4.5 shows the PCA plot of the first two principal components. It is evident that the datasets have different properties, as there are two distinguishable clusters, especially in the first principle component.

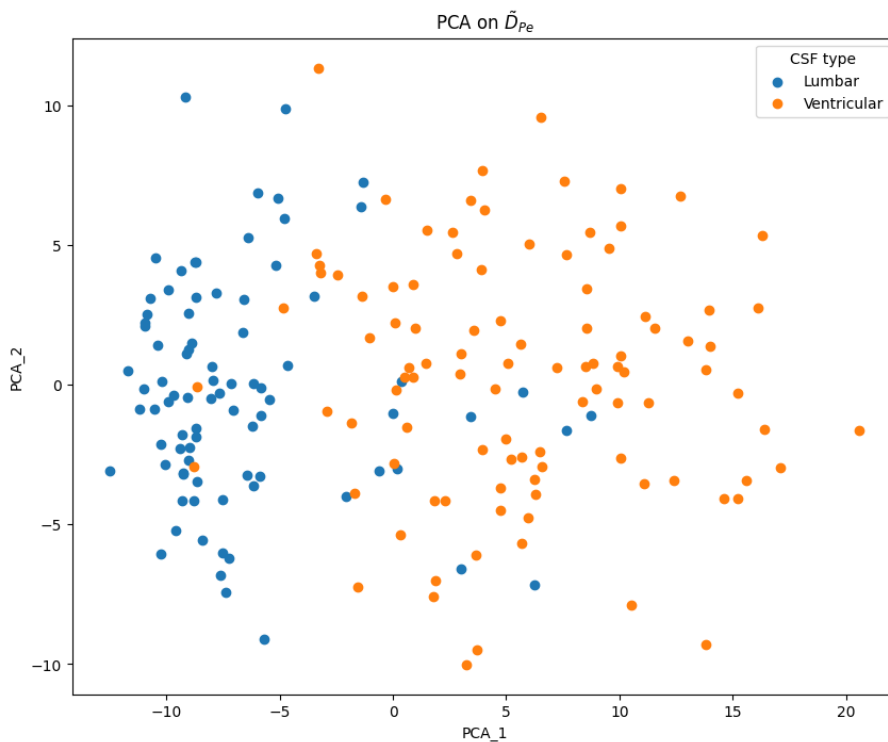


Figure 4.5: PCA plot on combined \tilde{D}_{PeL} and \tilde{D}_{PeV} with no missingness. The blue dots represent lumbar samples, and the orange dots represent ventricular samples. The clusters observed in the plot indicate a noticeable separation between the ventricular and lumbar samples, suggesting distinct patterns. This separation highlights the need to split the sample spaces into two different datasets for modelling.

4.1.5 Batch Effect on Predictive Results

We looked into how noticeable the batch effect is for ML models using a soft-voting ensemble of LR, RF and XGB. Table 4.3 presents the performance of models in predicting the TMT batch to which each sample belongs. Without using ComBat on the dataset, the models perform well when predicting which of the 15 batches a sample belongs to, with 55% and 77% accuracy. When ComBat is applied, the models perform worse, only reaching 1% and 11% accuracy. This holds true for both minimum and multiple imputations, further cementing the presence of batch effect and ComBat’s ability to remove it. However, as shown in Appendix A, ComBat does not appear to have a noticeable effect on predicting tissue groups, questioning the relevance batch effect removal.

Table 4.3: Predicting TMT set based results.

Dataset	Imputation	ComBat	SMOTE	Soft vote Acc
\tilde{D}_{PL}	MI	Off	On	55%
\tilde{D}_{PL}	MI	On	On	1%
\tilde{D}_{PL}	Minimum	Off	On	77%
\tilde{D}_{PL}	Minimum	On	On	11%

4.2 Model Evaluation for Tissue Group Prediction

After the exploratory analysis, we initiated the modelling stage. This section presents the results of the models applied to protein and peptide data in the ventricular and lumbar CSF subcategories, with the models performing relatively well for the downstream tasks of feature analysis. Each of the four sections in Figure 4.6 follows the same structure. In each section, we present the best-performing model’s ROC curve and confusion matrix. Additionally, the scoring metrics for this model are shown as a confidence interval. Finally, we show the stability of the feature selection process during k-fold ($k = 5$) cross-validation and introduce potential biomarkers, which are illustrated through staging.

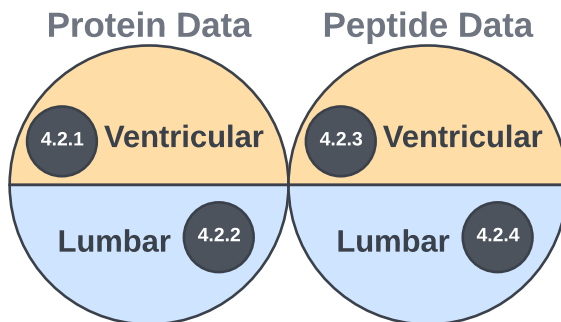


Figure 4.6: Illustrated are the sections for each dataset: \tilde{D}_{PV} in 4.2.1, \tilde{D}_{PL} in 4.2.2, \tilde{D}_{PeV} in 4.2.3 and \tilde{D}_{PeL} in 4.2.4. Each section follows the same format.

The best-performing models are evaluated based on the F1-score, AUC, and MCC metrics, with the model’s highest average score being highlighted. As previously mentioned, the multi-class data has been converted into a binary classification problem. Consequently, the CM follows the TP, TN, FP and FN structure in a 2x2 format.

We also present a summary table of results incorporating the 95% confidence interval for the metrics F_1 -score, accuracy, AUC and MCC. A frequency histogram is also included to illustrate the stability of the biomarker candidates. The x-axis indicates the number of folds where a feature appeared after feature selection. Features that are present in all five k-folds are considered biomarker candidates and will be presented through staging.

The considered biomarkers are extracted through ventricular or lumbar data, but the staging will include both. The hypothesis is that the sample space that extracted the biomarkers should show a clearer difference between the $A\beta^-T^-$ and $A\beta^+T^+$ tissue group than the sample space that did not. The full set of tests done on the binarized datasets is shown in Appendix A. In Appendix B, tables show the result of the best-performing models run 10 times to get a confidence interval on the scoring metrics. Finally, Appendix C highlights all biomarker candidates extracted through the models.

4.2.1 Ventricular protein modelling results

The best-performing model on \tilde{D}_{PV} had the following settings: all features with missingness were removed, data augmentation through SMOTE so that both classes had an equal amount of samples and feature selection through $\text{RFE}()$ until $k = 2$.

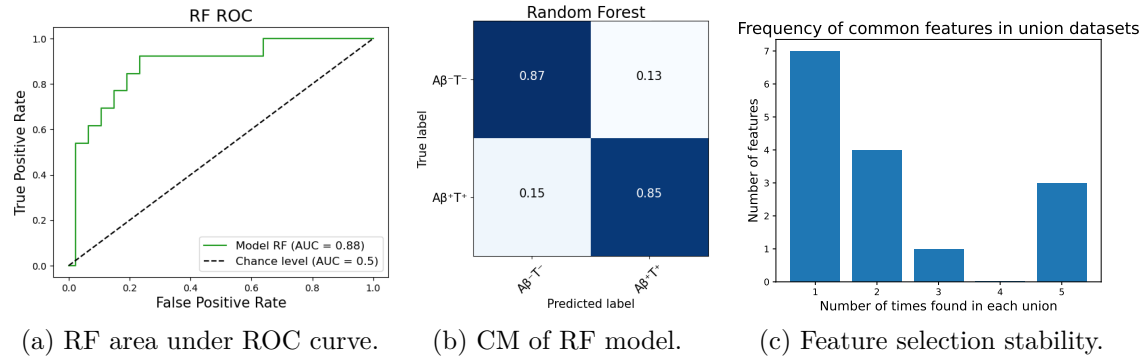


Figure 4.7: 4.7a shows that the RF model performs well compared to the chance level, with 0.88 AUC. 4.7b shows a balanced and fairly strong performance. 4.7c shows that three features are stable through each k-fold and are considered biomarker candidates.

Table 4.4: 95% confidence interval of the best model on protein ventricular data on accuracy, F_1 -score, AUC and MCC. The confidence interval is narrow on all scores, indicating a stable model.

Confidence	Accuracy	F_1 -score	AUC	MCC
Low	0.80	0.55	0.81	0.42
High	0.82	0.61	0.86	0.50

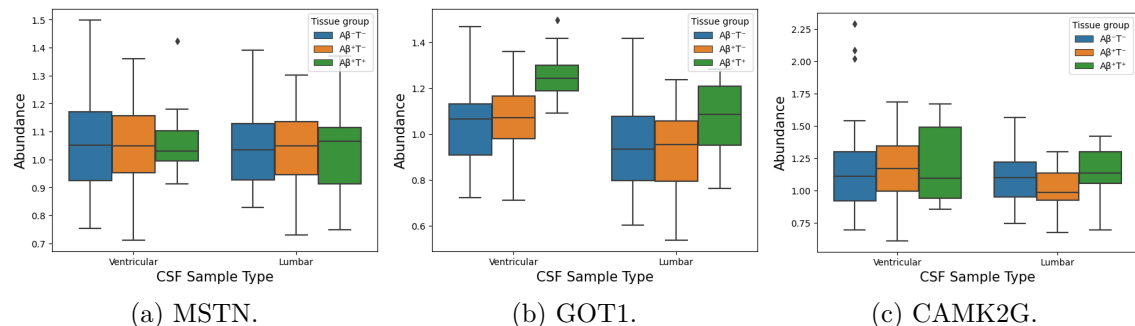
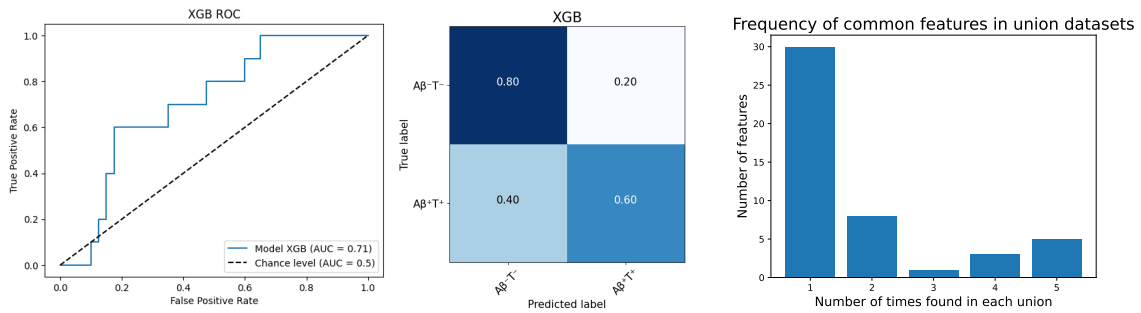


Figure 4.8: These three protein biomarker candidates are consistently extracted through feature selection in each k-fold. The subfigure captions depict the gene symbol. The proteins descriptions are: 4.8a - growth differentiation factor 8. 4.8b - aspartate aminotransferase, cytoplasmic. 4.8c - calcium/calmodulin-dependent protein kinase type II subunit gamma.

4.2.2 Lumbar protein modelling results

The best-performing model on \tilde{D}_{PL} had the following settings: all features with missingness were removed, ComBat applied to the data to remove batch effect, data augmentation through SMOTE so that both classes had an equal amount of samples and feature selection through $RFE()$ until $k = 5$.



(a) XGB area under ROC curve. (b) CM of XGB model. (c) Feature times selection stability.

Figure 4.9: 4.9a shows an XGB models ROC curve with 0.71 AUC. 4.9b highlights good predictions on the TP class, with the TN performance falling behind. 4.9c indicates that five features are selected across each k-fold. With $k = 5$, there are more features in the other bins, but can still be considered stable.

Table 4.5: Accuracy and AUC confidence scores indicate a tight fit. However, the F_1 -score and MCC spread highlights greater variance in the model predictions over multiple iterations. Each iteration's score is shown in Appendix B.

Confidence	Accuracy	F_1 -score	AUC	MCC
Low	0.74	0.20	0.65	0.07
High	0.77	0.41	0.75	0.26

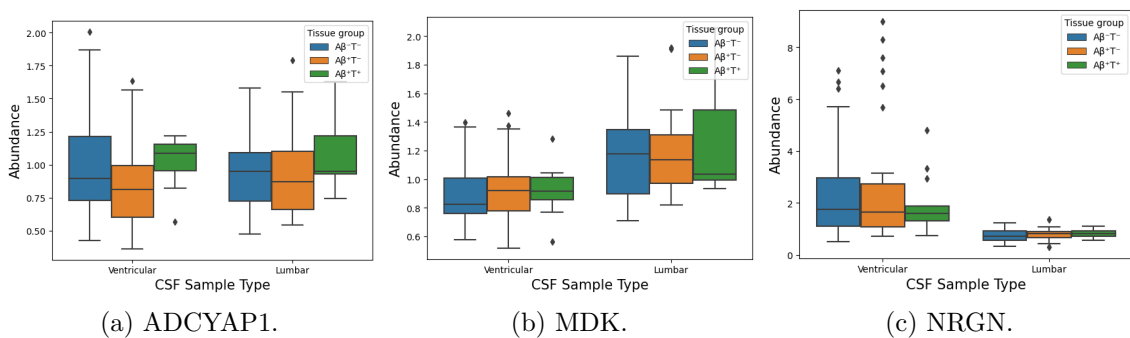


Figure 4.10: Plotted are three of the five protein biomarker candidates that are consistently extracted through feature selection in each k-fold from \tilde{D}_{PL} . Appendix C shows the full set of proteins. The subfigure captions depict the gene symbol. The protein descriptions are 4.10a - pituitary adenylate cyclase-activating polypeptide. 4.10b - midkine. 4.10c - neurogranin.

4.2.3 Ventricular peptide modelling results

The best-performing model on \tilde{D}_{PeV} had the following settings: all features with missingness were removed, data augmentation through SMOTE so that both classes had an equal amount of samples and feature selection through $RFE()$ until $k = 5$.

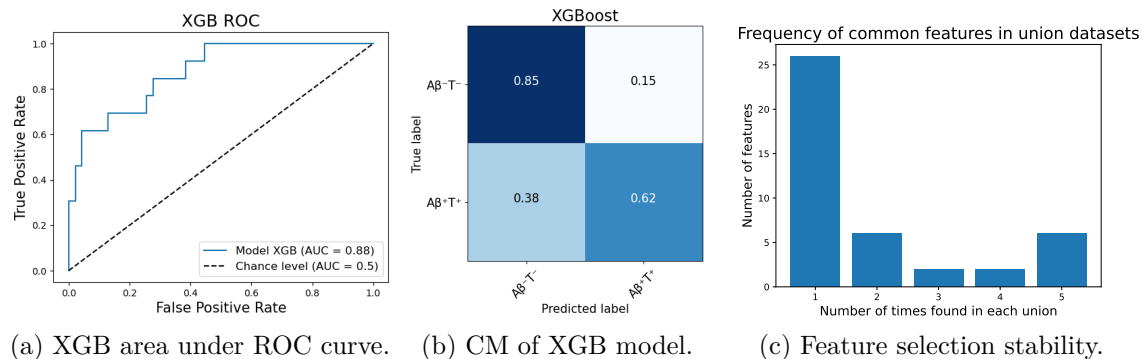


Figure 4.11: 4.11a shows that the XGB model performs well compared to the chance level, with 0.88 AUC. 4.11b shows the strong predictions on the $A\beta^-T^-$ class, with worse results on $A\beta^+T^+$. Six biomarker candidates are found in the union datasets according to 4.11c.

Table 4.6: The table shows the 95% confidence interval on peptide ventricular data. The fit is fairly close for all metrics, with F_1 -score and MCC having a slightly broader spread.

Confidence	Accuracy	F_1 -score	AUC	MCC
Low	0.75	0.46	0.76	0.30
High	0.81	0.59	0.85	0.48

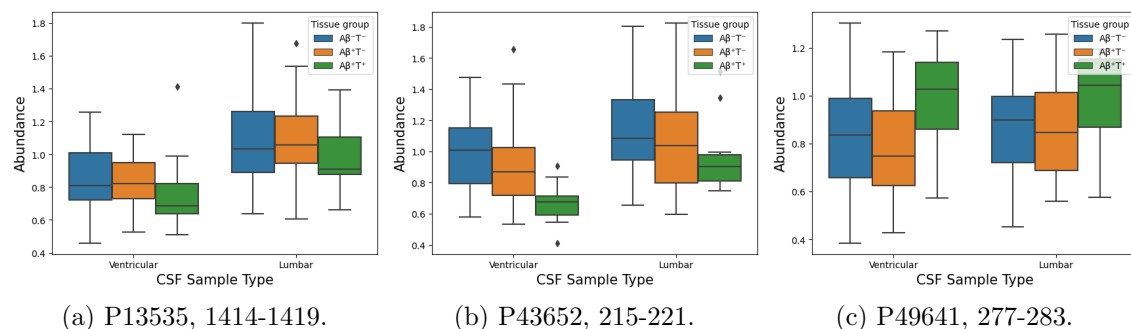


Figure 4.12: Plotted are three of the six peptide biomarker candidates consistently extracted through feature selection in each k-fold. Appendix C shows the full set of peptides. The subfigure captions depict the peptide's position in their respective protein. Both 4.12b and 4.12c shows tissue group correlation on both ventricular and lumbar CSF.

4.2.4 Lumbar peptide modelling results

The best-performing model on \tilde{D}_{PeL} had the following settings: all features with missingness were removed, ComBat applied to the data to remove batch effect, data augmentation through SMOTE so that both classes had an equal amount of samples and feature selection through $RFE()$ until $k = 2$.

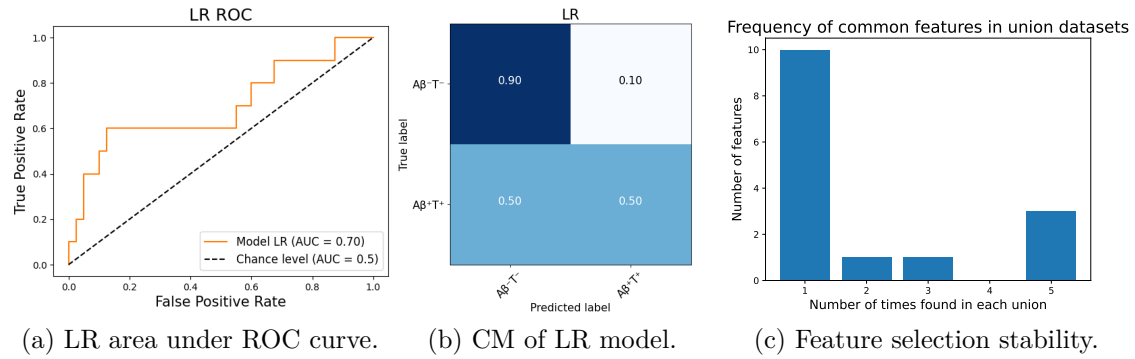


Figure 4.13: 4.13a shows a ROC curve and AUC slightly worse than for the other datasets. The same is true for 4.13b, where only half of the $A\beta^+T^+$ samples were correctly classified. Three biomarker candidates were found, as shown in 4.13c.

Table 4.7: The confidence interval on the scores all have a broad spread, indicating much variance in the results when training and evaluating the models. The MCC has an especially large variance.

Confidence	Accuracy	F_1 -score	AUC	MCC
Low	0.74	0.25	0.59	0.10
High	0.80	0.48	0.73	0.36

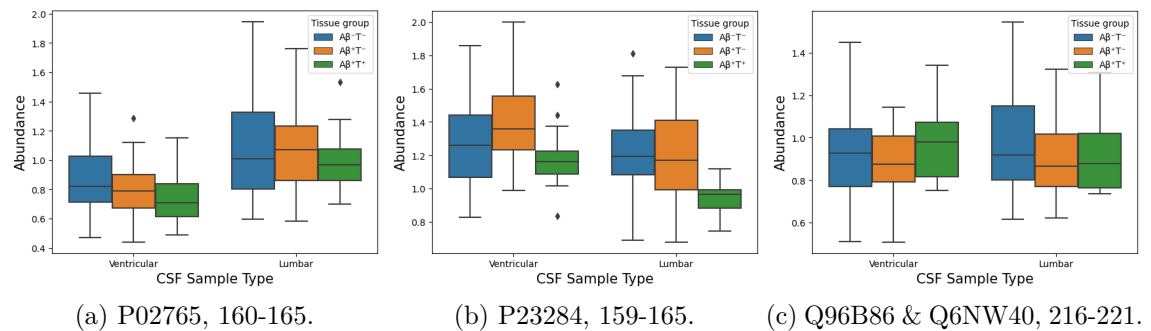


Figure 4.14: These three peptide biomarker candidates are consistently extracted through feature selection in each k-fold. The subfigure captions depict the peptide's position in their respective protein. Reading to plots, 4.14a seems to correlate to the tissue group for ventricular CSF, but no peptide has clear patterns for lumbar.

4.3 Proposed Biomarker Evaluation

This section explains our pipeline’s stable and predictive proteins and puts them in the context of results from previous literature. Similarly to Section 4.1.1, Kruskal-Wallis significance tests ($p < 0.05$) were carried out with additional Dunn tests for post-hoc analysis if statistical significance was determined with the Kruskal-Wallis test. The significance tests were performed independently on each of the datasets for the proteins or peptides that were considered important in the feature selection step in each dataset. Section 4.3.1, 4.3.2 and 4.3.3 considers biomarkers on \tilde{D}_{PV} while Section 4.3.4, 4.3.5 and 4.3.6 considers \tilde{D}_{PL} . Further significance tests on additional proteins and peptides are shown in Appendix C.

Table 4.8: Biomarker comparison between tissue groups on \tilde{D}_{PV} . The protein abundance of each of the tissue groups is expressed as mean \pm standard deviation. Statistical significance ($p < 0.05$) is highlighted with bold numbers. GOT1 protein abundance in group $A\beta^+T^+$ was found to be significantly different from those in both group $A\beta^+T^-$ and $A\beta^-T^-$.

Protein	Kruskal Wallis	p	$A\beta^-T^-$	$A\beta^+T^-$	$A\beta^+T^+$	Post-hoc
MSTN	0.031	0.984	1.06 ± 0.16	1.05 ± 0.16	1.07 ± 0.13	-
GOT1	20.247	0.00004	1.03 ± 0.16	1.07 ± 0.15	1.26 ± 0.11	$A\beta^-T^-$ and $A\beta^+T^+$, $p=0.00002$ $A\beta^+T^-$ and $A\beta^+T^+$, $p=0.0006$
CAMK2G	0.89	0.641	1.16 ± 0.33	1.17 ± 0.25	1.20 ± 0.28	-

Table 4.9: Biomarker comparison between tissue groups on \tilde{D}_{PL} . The protein abundance of each of the tissue groups is expressed as mean \pm standard deviation. Statistical significance ($p < 0.05$) is highlighted with bold numbers. No statistical significance was found among the tissue groups.

Protein	Kruskal Wallis	p	$A\beta^-T^-$	$A\beta^+T^-$	$A\beta^+T^+$	Post-hoc
ADCYAP1	3.02	0.221	0.93 ± 0.26	0.92 ± 0.31	1.10 ± 0.29	-
MDK	0.019	0.99	1.19 ± 0.31	1.17 ± 0.27	1.23 ± 0.37	-
NRGN	1.854	0.396	0.74 ± 0.24	0.79 ± 0.20	0.81 ± 0.17	-

Table 4.10: Biomarker comparison between tissue groups on \tilde{D}_{PeV} . The peptide abundance of each tissue group is expressed as mean \pm standard deviation. Statistical significance ($p < 0.05$) is highlighted with bold numbers. The $A\beta^+T^+$ tissue group showed statistically significant differences in P43652 [215-221] peptide abundance from both $A\beta^-T^-$ and $A\beta^+T^-$. Also, the $A\beta^+T^+$ tissue group was statistically significant from the $A\beta^+T^+$ tissue group in P49641 [277-283] peptide abundance.

Protein	Kruskal Wallis	p	$A\beta^-T^-$	$A\beta^+T^-$	$A\beta^+T^+$	Post-hoc
P13535..1414.1419.	4.199	0.123	0.85 ± 0.20	0.82 ± 0.14	0.76 ± 0.22	-
P43652..215.221.	20.429	0.00004	0.99 ± 0.22	0.90 ± 0.25	0.66 ± 0.12	$A\beta^-T^-$ and $A\beta^+T^+$ $p=0.00002$ $A\beta^+T^-$ and $A\beta^+T^+$ $p=0.004$
P49641..277.283.	10.07	0.007	0.82 ± 0.21	0.76 ± 0.19	0.98 ± 0.20	$A\beta^+T^-$ and $A\beta^+T^+$ $p=0.005$

4. Results

Table 4.11: Biomarker comparison between tissue groups on \tilde{D}_{PeL} . The peptide abundance of each tissue group is expressed as mean \pm standard deviation. Statistical significance ($p < 0.05$) is highlighted with bold numbers. The P23284 [159-165] peptide was found to have statistically significant differences in abundance in group $A\beta^+T^+$ compared to $A\beta^-T^-$ and $A\beta^+T^-$.

Peptide	Kruskal Wallis	p	$A\beta^-T^-$	$A\beta^+T^-$	$A\beta^+T^+$	Post-hoc
P02765..160.165.	0.472	0.79	1.10 ± 0.36	1.04 ± 0.27	1.00 ± 0.24	-
P23284..159.165.	10.987	0.004	1.20 ± 0.24	1.19 ± 0.28	0.95 ± 0.10	$A\beta^-T^-$ and $A\beta^+T^+$, $p=0.003$ $A\beta^+T^-$ and $A\beta^+T^+$, $p=0.01$
Q96B86..216.221...Q6NW40..216.221.	2.348	0.309	0.98 ± 0.24	0.88 ± 0.16	0.92 ± 0.18	-

4.3.1 Growth differentiation factor 8

Growth differentiation factor 8, also known as myostatin, is a protein that serves as a negative regulator of skeletal muscle mass [118]. Elevated levels of myostatin in humans have been linked to muscle wasting in humans [119]. In a study by Lin, Lin, and Hsiao [120], the researchers explored the association between myostatin and cognitive decline of AD in mice. Their findings suggested that elevated myostatin expression might play a role in initiating muscle atrophy and cognitive impairments. Myostatin was found to be an important and stable protein in both lumbar and ventricular CSF in our study. However, no significant differences between tissue groups were found during the Kruskal-Wallis tests, as seen in Table 4.8 and Table 4.9. Figure 4.8a shows the boxplots of myostatin abundances in \tilde{D}_{PV} and \tilde{D}_{PL} .

4.3.2 Aspartate aminotransferase

The presence of aspartate aminotransferase (AST) and alanine aminotransferase (ALT) enzymes can be used to diagnose people with tissue injury, especially tissue of the heart and the liver [121]. A few papers have investigated the effect of AST and ALT enzymes on AD. The research paper by Nho, Kueider-Paisley, Ahmad, *et al.* [122] showed that elevated AST to ALT ratios were associated with the diagnosis of AD. Lu, Pike, Selvin, *et al.* [123] concluded that low levels of aminotransferase, particularly ALT, were linked to a higher long-term risk of dementia. AST was concluded to be a stable ventricular CSF biomarker in our feature extraction pipeline. The Kruskal-Wallis test, followed by the posthoc Dunn test, performed on \tilde{D}_{PV} , showed that there were significant differences in the GOT1 abundance in $A\beta^+T^+$ tissue group compared to both $A\beta^-T^-$ and $A\beta^+T^-$ as seen in Table 4.8. Observing Figure 4.8b we can see that the abundance is elevated in the $A\beta^+T^+$ tissue group compared to the other groups.

4.3.3 Calcium/calmodulin-dependent protein kinase type II subunit gamma

Calcium/calmodulin-dependent protein kinase type II subunit gamma (CaMKII) plays an important role in synaptic plasticity and memory formation and has also been proposed to be a tau kinase [124]. Tau kinases are enzymes that phosphorylate the tau protein. An increased phosphorylation is associated with tau aggregation,

an important hallmark of AD [125]. CaMKII is known to be dysregulated in AD hippocampus, but further research must be conducted to determine how the dysregulation occurs [124]. No significant differences were found between tissue groups for the CaMKII protein abundance.

4.3.4 Pituitary adenylate cyclase-activating polypeptide

There have not been many studies regarding the role of pituitary adenylate cyclase-activating polypeptide in AD. However, a study by Han, Liang, Baxter, *et al.* [126] which included a cohort of RNA with 34 AD and 14 CN patients with a validation cohort of 12 AD and 11 CN patients, suggested that this protein is reduced in patients with AD. ADCYAP1 was found to be one of the stable proteins during feature selection in the \tilde{D}_{PL} dataset. In Figure 4.10a, we can see that the median abundance is elevated in the $A\beta^+T^-$ tissue group. However, no statistically significant differences were found between the tissue groups as seen in Table 4.9.

4.3.5 Midkine

Midkine is a protein which is highly expressed in the development of the human embryo and has been shown to be upregulated in the central nervous system upon several conditions such as multiple sclerosis, brain injury and cancer [127]. It has also been shown to be significantly higher in sera of patients with AD ($n = 36$) compared to control ($n = 32$) in a study from 2005 [128]. Midkine is believed to be upregulated in AD patients to prevent $A\beta$ peptide cell death due to it binding the $A\beta$ peptide and therefore neutralizes the cytotoxic activity [129]. Figure 4.10b shows the boxplot of midkine abundance. There were no statistically significant differences in midkine abundance between the tissue groups.

4.3.6 Neurogranin

Neurogranin is a protein known for its association with synaptic plasticity and long-term potentiation, characterized by the persistent enhancement of synaptic strength. This modulation is mediated by the calcium- and calmodulin-signaling pathways [130]. It has also been proposed as a potential biomarker of diseases such as AD and Parkinson's disease [130]. Another research paper by Wellington, Paterson, Portelius, *et al.* [131] concluded that the concentration of lumbar CSF was higher in AD samples than in control samples and samples of other neurodegenerative diseases in a cohort of 331 patients. No statistically significant differences were found between tissue groups.

The domain expert from Sahlgrenska University Hospital found neurogranin particularly interesting because it is considered a well-established AD biomarker. The other proteins and peptides are less well-known in the context of AD research.

4.3.7 Peptide biomarkers

Kruskal Wallis tests and posthoc tests on biomarkers from feature selection on \tilde{D}_{PeL} and \tilde{D}_{PeV} are seen in Table 4.11 and Table 4.10. In \tilde{D}_{PeL} , there were significant differences in peptide abundance between the $A\beta^+T^+$ tissue group and the other tissue groups in the P23284 [159-165] peptide. In \tilde{D}_{PeV} , significant differences between tissue group abundances were found in P43652 [215-221] and P49641 [277-283]. Furthermore, Appendix C shows Kruskal Wallis tests on additional peptides from the feature selection step in the pipeline.

5

Discussion

This chapter includes an in-depth analysis of the study’s methodologies, findings, and implications. It delves into the challenges encountered during data preparation and model configuration, addressing issues such as high-dimensional data handling and missing value imputation. The chapter also discusses the impact of dataset characteristics, including small sample size and class imbalance, on model performance. The effectiveness of validation techniques, ensemble methods, and feature selection strategies are examined. Additionally, we compare our results with existing literature, highlighting the uniqueness of our approach and dataset. Finally, we acknowledge the study’s limitations and propose avenues for future research.

5.1 Data Preparation and Model Settings

High-dimensional data: A recurring theme in this project is the challenge of handling high-dimensional data with significant levels of missingness. There is no universally accepted approach for managing missing values, especially in scenarios where $p \gg n$ and the dataset is relatively small. Striking a balance between removing features with excessive missing data and imputing these missing values is crucial. On one hand, removing features can lead to the loss of potentially valuable patterns and predictive information. On the other hand, imputation can introduce biases by distorting the distribution of features. Domain-specific considerations are essential in making these decisions. For example, a good biomarker for diagnosis only found in every other individual may be worse than a decent one found in all, favouring stable biomarkers without missingness.

Missingness in data: When the missing data is MNAR, domain knowledge should help guide the choice of the imputation method. When performing MS, only features above a certain threshold are recorded. Consequently, imputing missing values with the minimum observed value can minimize bias. We applied MI to the protein datasets on features with 20% or less missingness and minimum imputation on both protein and peptide datasets with 20% and 50% missing values, respectively, based on Kong, Hui, Peng, *et al.* [45]. Additionally, we evaluated the impact of removing features with any missing data. Our results, detailed in the best-performing models and Appendix A, indicate that any imputation generally resulted in worse performance than using only features without missing data. The only exception is on \tilde{D}_{PV} with 20% missing values imputed through MI, resulting in fairly strong models.

However, this appears to be an outlier rather than the norm, as removing all features exhibiting any form of missing values resulted in stronger models.

Insights from lumbar and ventricular data: The initial datasets in this project included both lumbar and ventricular CSF data. The PDF and PCA plots reveal distinct differences in the proteomic profiles in these samples. The PDF indicates that ventricular CSF samples have a higher standard deviation in protein and peptide levels, whereas lumbar CSF samples are more concentrated around the mean. This observation supports that the analytes near the brain are more abundant in ventricular CSF and dilute as they travel down the spine to the lumbar region, where they mix with new analytes [28]. The PCA plot in Figure 4.5 further illustrates this distinction, as the first principal component clearly separates the CSF samples into two distinct clusters. These findings highlight the necessity of treating lumbar and ventricular CSF samples as separate datasets, necessitating different data preparation and modelling processes. However, the significance tests show poor scores on both Kruskal-Wallis, with only a few proteins and peptides being significantly different. One of these, GOT1 seen in Table 4.8, is significantly different both between pathogenesis classes $A\beta^-T^-$ and $A\beta^+T^+$ and also between $A\beta^+T^-$ and $A\beta^+T^+$.

Batch effect: Furthermore, we visualized the batch effect in the datasets using t-SNE and ML models predicting batches. The t-SNE plots in Figure 4.4 clearly illustrate the clustering of TMT batches. Comparing the clustering before and after applying ComBat demonstrates the impact, with higher batch entropy observed post-ComBat while maintaining similar tissue group clustering. This suggests that ComBat positively affects the data preparation process for modelling. This observation is further supported by Table 4.3, which shows the effect of ComBat on batch prediction. However, in most cases, the application of ComBat has minor, negligible or even negative effects on the result. This leads to the possible conclusion that the presence of batches in the data has less impact than initially hypothesized.

Small patient cohort: Finally, the small size of the dataset presents a significant challenge. ML models typically perform better with larger quantities of data, but these datasets are small and imbalanced. Under-sampling and hybrid methods become impractical, as efficient use of available data is preferred. Consequently, we applied the over-sampling technique SMOTE, experimentally determining the number of synthesized observations. The best-performing models used SMOTE to balance the minority class and match the number of samples in the majority class. As shown in Appendix A, over-sampling through SMOTE is the most effective strategy for enhancing ML model performance.

5.2 Small Dataset and k -fold Validation

In addition to the challenges of over-sampling, small datasets present further difficulties. The protein and peptide datasets have a minority class sample size of 10 and 13, respectively. Splitting these small datasets into training and testing sets can introduce biases, especially when the test set contains few and potentially non-diverse samples, and different splits can greatly impact the models' performance. To

mitigate this, we applied cross-validation, allowing the entire dataset to be used for validation. Both five-fold and ten-fold cross-validation were evaluated, with five-fold cross-validation providing slightly higher consistency. In both five and ten folds, the confidence intervals are wide yet slightly more narrow in the five-fold setting.

However, despite the higher consistency offered by five-fold cross-validation, there remains variability in scoring metrics across the folds. This is particularly evident when analysing F_1 -score and MCC, where a single misdiagnosed instance in the minority class can significantly skew the results. It is common to observe excellent performance in three or four folds, sometimes perfect predictions, but poor results in the others, often worse than random guessing. This issue is more pronounced in small, noisy datasets, where each prediction heavily influences the final outcome.

The impact of this uncertainty is also reflected in the 95% confidence interval tables in Section 4.2. While a few models produced relatively stable intervals, most exhibited a wide confidence interval, indicating variability in the predicted scoring metrics. A plausible hypothesis is that the nature of the folds affects model performance, resulting in unstable models and suggesting the presence of potential outliers in the dataset.

Additionally, the introduction of k-fold cross-validation complicates the feature selection process. When data is initially split into training and testing sets, feature selection is performed only on the training set. It is, therefore, crucial to perform feature selection within each k-fold to reduce the sample space, not before the data is split in each k-fold. If not, the models risk overfitting the data due to data leakage. This is a common pitfall when working with high-dimensional, small sample-size datasets.

5.3 Model Performance and Ensemble Techniques

For all tests, we used three models and two ensemble methods to predict the tissue group on the data: XGB, LR, and RF, and ensembles of these based on soft and hard voting. We initially assumed that XGB would outperform the other models due to its strengths in handling high-dimensional data. However, as the feature space was reduced, all models performed similarly. As shown in Section 4.2, each model performed better than the others, at least once, on the datasets. While individual models sometimes underperformed, ensembles somewhat mitigated this issue.

Despite rarely being the best-performing model, both ensemble techniques demonstrated higher consistency than the individual models. We theorize that when one individual model underperforms, the other two compensate for it within the ensemble. Therefore, we believe the ensemble methods would have better reproducibility and more narrow confidence intervals, although this theory has not been explicitly tested. A potential downside to ensemble models is the lack of interpretability, as a singular model is always easier to understand than a combination of multiple.

5.4 Feature Selection and Stability

Identifying potential biomarkers is crucial, so we must reduce the feature space of the data. Patterns observed in Appendix A suggest that significantly reducing the feature space is more beneficial than retaining more features for model training. We conducted experiments on this topic, ranging from the entire dataset to selecting the most important feature in each fold. Applying the scikit-learn `RFE()` on the four feature-selecting models, we iteratively reduced k before taking the union of the models' feature sets. The intersection of the feature sets sometimes reduced k to zero, while the union approach allowed each model to contribute its own strengths to the ensemble. Additionally, the stability of the feature sets was higher with the union approach than with the intersection approach.

According to a domain expert at Sahlgrenska University Hospital, a good biomarker should be clear, disease-specific, and particularly stable across samples. Extracting features separately in each k-fold ensures that a biomarker is stable only if present in each fold. If a strong feature is selected in one fold but not in others, it may emphasize outlier samples in the dataset. Therefore, the proposed biomarkers in this thesis have all been selected in all k-folds, ensuring their applicability to the entire dataset. Furthermore, the protein neurogranin was particularly interesting to the expert and can be considered an established biomarker in AD research.

It should be noted that most of the proposed proteins and the previously established biomarkers do not perform well during a Kruskal-Wallis test. Only one proposed protein and three proposed peptide biomarkers are statistically significant on the data, as shown in Tables 4.8 and 4.11, and Appendix C.5. The same is true for previously established biomarkers, having an even lower statistical significance, as shown in Tables 4.1 and 4.2. We still believe the proposed biomarkers hold value, as the combination of multiple proteins and peptides may work as a good biomarker, something that a Kruskal-Wallis test does not take into consideration.

5.5 Comparing Neurodegenerative Disorders

Exploring previous research on similar topics, that of predicting CN or AD from CSF samples using ML models, it is not uncommon to see results with 99% accuracy. However, it is important to note that our task is different: we predict pathological diagnosis based on tissue groups, which is far less common when utilizing CSF samples. Additionally, AD prediction datasets typically compare healthy individuals to those with AD. In our dataset, the negative group suffers from iNPH, another neurodegenerative disorder. Thus, our task involves differentiating AD from other neurodegenerative disorders. As shown in Table 3.1, the *healthy* group includes individuals with iNPH and other clinically diagnosed diseases.

Instead of predicting CN, MCI and AD, we are predicting $A\beta^{-}T^{-}$, $A\beta^{+}T^{-}$ and $A\beta^{+}T^{+}$. Initially, this predictive task seemed challenging, as the scoring metrics for the multi-class problem barely outperformed random guessing. This suggests there may be an overlap between iNPH patients with tissue groups $A\beta^{-}T^{-}$ and $A\beta^{+}T^{-}$,

complicating the predictive tasks. Removing the $A\beta^+T^-$ tissue group from the data significantly improved the performance of the models. Furthermore, established biomarkers common in neurodegenerative disorders did not perform well on this dataset. In Figure 4.1, only one of the biomarkers showed any correlation to the tissue group. Given that all samples in the dataset are from individuals with some disorder, this result is unsurprising. Therefore, we propose that different biomarkers are needed to predict pathological tissue groups for AD in a cohort with other neurodegenerative disorders.

Finally, the proteins and peptides identified in Section 4.2 and Appendix C are biomarkers we propose for this purpose. These proposed protein and peptide biomarkers differ between lumbar and ventricular CSF, further emphasizing the difference in CSF composition. However, the MSTN gene was found in both samples, suggesting some similarities. In Section 4.3, we demonstrate that some of our proposed biomarkers have been studied in the context of neurodegeneration, particularly in AD. There may be benefits in incorporating these proteins and peptides into the current biomarker set, differentiating iNPH patients from those with both iNPH and AD.

5.6 Limitations

The small cohort of the dataset can result in potential biases, overfitting the few minority-class samples present. A larger cohort, especially with more minority-class samples, would minimize the need for synthesizing data through SMOTE. However, a larger cohort risk increases both batch effect and the introduction of missing values due to adding TMT batches.

Furthermore, the lack of good results when predicting three classes made multi-class classification non-trivial, and we decided to focus on binary classification. While patterns within the multi-class dataset may exist, our findings suggest otherwise.

As is common in projects of this nature, the scope expanded beyond initial expectations, resulting in time constraints. Consequently, we could not perform certain tests we had intended, e.g., Rubin’s Rules exploration of the MI data, Kruskal–Wallis tests on the proposed biomarkers and wider and deeper testing setups for the binary classification predictions.

Finally, due to our limited knowledge of proteomics, we evaluate our proposed biomarkers primarily by comparing them with existing research. A more robust approach would involve extensively consulting domain experts to theorise our findings’ validity and significance.

5.7 Future work

The dataset is fairly unique in proteomics, including pathological diagnosis of AD on lumbar and ventricular CSF from iNPH patients, allowing for numerous avenues to explore. Firstly, adding multimodal data, such as brain scans of the cohort, would enable more comprehensive comparisons along the AD continuum. Secondly, longitudinal studies should be conducted on the same patient group to examine lumbar CSF samples and track biomarker changes over time. This could provide insights into the progression of AD and the role of identified biomarkers in different disease stages. Thirdly, future studies should aim to include larger and more diverse cohorts. Increasing the sample size and diversity can help generalize the identified biomarkers and validate them for new cohorts. Fourthly, there are avenues of adding a binary class to the data indicating ventricular or lumbar CSF, treating them as one dataset. Despite the distinct proteomic profiles, the PCA only indicated one dimension of great difference. Finally, experimenting with different preprocessing and imputation methods and exploring various ML models, such as the domain popular SVM, on the same dataset could enhance the robustness and accuracy of biomarker identification.

6

Conclusion

In this project, we set out to conduct an exploratory analysis of proteomic abundances along the progression of AD in lumbar and ventricular CSF. This exploration was done on a dataset unique in its composition, consisting of iNPH patients, some pathologically diagnosed on the AD continuum. We faced significant challenges, including handling high-dimensional data with missing values, small sample sizes, and class imbalances. We applied imputation and oversampling techniques, attempting to enhance model performance, and five-fold cross-validation allowed the entire dataset to be validated.

Due to their distinct proteomic profiles, we noticed the necessity of treating lumbar and ventricular CSF samples as separate datasets. Our results also suggest that removing all features with missing values provided stronger models than imputing the missingness. Additionally, despite the presence of batch effect, it barely affects the resulting models. We also noticed that no model outperformed the others in all situations, and all suffered from inconsistent scoring metrics' across folds. Ensemble models performed slightly worse yet were more consistent in the scoring metrics confidence intervals. Furthermore, feature selection was crucial for identifying stable biomarkers and ensuring their reliability across all folds.

Comparative analysis of existing biomarkers highlighted the uniqueness of our dataset. Correlation to traditional biomarkers on the AD continuum was lacking when visualizing the staging, suggesting that new proteins and peptides must be considered when the control group suffers from iNPH. We propose eight protein and nine peptide biomarkers to help differentiate iNPH patients on the pathological AD spectra. One such biomarker shows promise in both lumbar and ventricular CSF.

Despite the study's limitations, including the small dataset size and challenges in multi-class classification, our findings contribute insights into the proteomic analysis of neurodegenerative disorders. Future research should expand the cohort size, evaluate our proposed biomarkers, incorporate multimodal data, and conduct longitudinal studies to validate and build on our findings.

Bibliography

- [1] A. Association, “2023 Alzheimer’s disease facts and figures,” en, *Alzheimer’s & Dementia*, vol. 19, no. 4, pp. 1598–1695, Apr. 2023, ISSN: 1552-5260, 1552-5279. DOI: 10.1002/alz.13016. [Online]. Available: <https://alz-journals.onlinelibrary.wiley.com/doi/10.1002/alz.13016> (visited on 02/27/2024).
- [2] S. Johnson, M. Suárez-Calvet, I. Suridjan, *et al.*, “Identifying clinically useful biomarkers in neurodegenerative disease through a collaborative approach: The NeuroToolKit,” *Alzheimer’s Research & Therapy*, vol. 15, Jan. 2023. DOI: 10.1186/s13195-023-01168-y.
- [3] S. D. Patterson and R. H. Aebersold, “Proteomics: The first decade and beyond,” en, *Nature Genetics*, vol. 33, no. S3, pp. 311–323, Mar. 2003, ISSN: 1061-4036, 1546-1718. DOI: 10.1038/ng1106. [Online]. Available: <https://www.nature.com/articles/ng1106z> (visited on 04/23/2024).
- [4] E. J. Dupree, M. Jayathirtha, H. Yorkey, M. Mihasan, B. A. Petre, and C. C. Darie, “A Critical Review of Bottom-Up Proteomics: The Good, the Bad, and the Future of This Field,” en, *Proteomes*, vol. 8, no. 3, p. 14, Sep. 2020, Number: 3 Publisher: Multidisciplinary Digital Publishing Institute, ISSN: 2227-7382. DOI: 10.3390/proteomes8030014. [Online]. Available: <https://www.mdpi.com/2227-7382/8/3/14> (visited on 03/28/2024).
- [5] S. Grueso and R. Viejo-Sobera, “Machine learning methods for predicting progression from mild cognitive impairment to Alzheimer’s disease dementia: A systematic review,” en, *Alzheimer’s Research & Therapy*, vol. 13, no. 1, p. 162, Dec. 2021, ISSN: 1758-9193. DOI: 10.1186/s13195-021-00900-w. [Online]. Available: <https://alzres.biomedcentral.com/articles/10.1186/s13195-021-00900-w> (visited on 03/28/2024).
- [6] Q.-Y. He and J.-F. Chiu, “Proteomics in biomarker discovery and drug development,” en, *Journal of Cellular Biochemistry*, vol. 89, no. 5, pp. 868–886, Aug. 2003, ISSN: 0730-2312, 1097-4644. DOI: 10.1002/jcb.10576. [Online]. Available: <https://onlinelibrary.wiley.com/doi/10.1002/jcb.10576> (visited on 02/22/2024).
- [7] M. A. DeTure and D. W. Dickson, “The neuropathological diagnosis of Alzheimer’s disease,” en, *Molecular Neurodegeneration*, vol. 14, no. 1, p. 32, Dec. 2019, ISSN: 1750-1326. DOI: 10.1186/s13024-019-0333-5. [Online]. Available: <https://molecularneurodegeneration.biomedcentral.com/articles/10.1186/s13024-019-0333-5> (visited on 02/27/2024).
- [8] K. Blennow, H. Hampel, M. Weiner, and H. Zetterberg, “Cerebrospinal fluid and plasma biomarkers in Alzheimer disease,” en, *Nature Reviews Neurology*, vol. 6, no. 3, pp. 131–144, Mar. 2010, ISSN: 1759-4758, 1759-4766. DOI: 10.1038/nrneuro1.2010.4. [Online]. Available: <https://www.nature.com/articles/nrneuro1.2010.4> (visited on 01/15/2024).
- [9] C. Christin, H. C. J. Hoefsloot, A. K. Smilde, *et al.*, “A Critical Assessment of Feature Selection Methods for Biomarker Discovery in Clinical Proteomics *,” English, *Molecular & Cellular Proteomics*, vol. 12, no. 1, pp. 263–276, Jan. 2013, Publisher: Elsevier, ISSN: 1535-9476, 1535-9484. DOI: 10.1074/mcp.M112.022566. [Online]. Available: [https://www.mcponline.org/article/S1535-9476\(20\)33445-9/abstract](https://www.mcponline.org/article/S1535-9476(20)33445-9/abstract) (visited on 01/31/2024).
- [10] J. Silberring and P. Ciborowski, “Biomarker discovery and clinical proteomics,” en, *TrAC Trends in Analytical Chemistry*, vol. 29, no. 2, pp. 128–140, Feb. 2010, ISSN: 01659936. DOI: 10.1016/j.trac.2009.11.007. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S016599360900288X> (visited on 02/22/2024).

- [11] D. Cagney, J. Sul, R. Y. Huang, K. L. Ligon, P. Y. Wen, and B. M. Alexander, “BEST (Biomarkers, EndpointS, and other Tools) Resource,” en, *Neuro-oncology*, vol. 20, no. 9, pp. 1162–1172, 2018, Oxford University Press US.
- [12] World Health Organization, *Risk reduction of cognitive decline and dementia: WHO guidelines*, en. Geneva: World Health Organization, 2019, Section: xiii, 78 p., ISBN: 978-92-4-155054-3. [Online]. Available: <https://iris.who.int/handle/10665/312180> (visited on 02/27/2024).
- [13] R. H. Swerdlow, “Is aging part of Alzheimer’s disease, or is Alzheimer’s disease part of aging?” en, *Neurobiology of Aging*, vol. 28, no. 10, pp. 1465–1480, Oct. 2007, ISSN: 01974580. DOI: 10.1016/j.neurobiolaging.2006.06.021. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S0197458006002351> (visited on 02/27/2024).
- [14] P. Scheltens, K. Blennow, M. M. Breteler, *et al.*, “Alzheimer’s disease,” *The Lancet*, vol. 388, no. 10043, pp. 505–517, 2016, Elsevier. DOI: [https://doi.org/10.1016/S0140-6736\(15\)01124-1](https://doi.org/10.1016/S0140-6736(15)01124-1).
- [15] D. Zafeiris, S. Rutella, and G. R. Ball, “An Artificial Neural Network Integrated Pipeline for Biomarker Discovery Using Alzheimer’s Disease as a Case Study,” en, *Computational and Structural Biotechnology Journal*, vol. 16, pp. 77–87, 2018, ISSN: 20010370. DOI: 10.1016/j.csbj.2018.02.001. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S2001037017300843> (visited on 11/24/2023).
- [16] S. Asif and T. Khan, “A Machine Learning Model to Predict the Onset of Alzheimer Disease using Potential Cerebrospinal Fluid (CSF) Biomarkers,” en, *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 12, 2017, ISSN: 21565570, 2158107X. DOI: 10.14569/IJACSA.2017.081216. [Online]. Available: <http://thesai.org/Publications/ViewPaper?Volume=8&Issue=12&Code=ijacsa&SerialNo=16> (visited on 11/30/2023).
- [17] H. Alamro, M. A. Thafar, S. Albaradei, T. Gojobori, M. Essack, and X. Gao, “Exploiting machine learning models to identify novel Alzheimer’s disease biomarkers and potential targets,” en, *Scientific Reports*, vol. 13, no. 1, p. 4979, Mar. 2023, ISSN: 2045-2322. DOI: 10.1038/s41598-023-30904-5. [Online]. Available: <https://www.nature.com/articles/s41598-023-30904-5> (visited on 11/24/2023).
- [18] A. J. Bayer, “The role of biomarkers and imaging in the clinical diagnosis of dementia,” en, *Age and Ageing*, vol. 47, no. 5, pp. 641–643, Sep. 2018, ISSN: 0002-0729, 1468-2834. DOI: 10.1093/ageing/afy004. [Online]. Available: <https://academic.oup.com/ageing/article/47/5/641/4843988> (visited on 02/27/2024).
- [19] G. G. Glenner and C. W. Wong, “Alzheimer’s disease: Initial report of the purification and characterization of a novel cerebrovascular amyloid protein,” en, *Biochemical and Biophysical Research Communications*, vol. 120, no. 3, pp. 885–890, May 1984, ISSN: 0006291X. DOI: 10.1016/S0006-291X(84)80190-4. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S0006291X84801904> (visited on 02/27/2024).
- [20] B. Bai, D. Vanderwall, Y. Li, *et al.*, “Proteomic landscape of Alzheimer’s Disease: Novel insights into pathogenesis and biomarker discovery,” en, *Molecular Neurodegeneration*, vol. 16, no. 1, p. 55, Aug. 2021, ISSN: 1750-1326. DOI: 10.1186/s13024-021-00474-z. [Online]. Available: <https://moleculareuro-degeneration.biomedcentral.com/articles/10.1186/s13024-021-00474-z> (visited on 11/19/2023).
- [21] H. V. Dansson, L. Stempfle, H. Egilsdóttir, *et al.*, “Predicting progression and cognitive decline in amyloid-positive patients with Alzheimer’s disease,” en, *Alzheimer’s Research & Therapy*, vol. 13, no. 1, p. 151, Sep. 2021, ISSN: 1758-9193. DOI: 10.1186/s13195-021-00886-5. [Online]. Available: <https://alzres.biomedcentral.com/articles/10.1186/s13195-021-00886-5> (visited on 11/26/2023).
- [22] Y. Wang, Y. Sun, Y. Wang, *et al.*, “Identification of novel diagnostic panel for mild cognitive impairment and Alzheimer’s disease: Findings based on urine proteomics and machine learning,” en, *Alzheimer’s Research & Therapy*, vol. 15, no. 1, p. 191, Nov. 2023, ISSN: 1758-9193. DOI: 10.1186/s13195-023-01324-4. [Online]. Available: <https://doi.org/10.1186/s13195-023-01324-4> (visited on 01/31/2024).
- [23] C.-H. Chang, C.-H. Lin, and H.-Y. Lane, “Machine Learning and Novel Biomarkers for the Diagnosis of Alzheimer’s Disease,” en, *International Journal of Molecular Sciences*, vol. 22,

- no. 5, p. 2761, Mar. 2021, ISSN: 1422-0067. DOI: 10.3390/ijms22052761. [Online]. Available: <https://www.mdpi.com/1422-0067/22/5/2761> (visited on 11/24/2023).
- [24] J. M Das and M. C. Biagioni, “Normal Pressure Hydrocephalus,” eng, in *StatPearls*, Treasure Island (FL): StatPearls Publishing, 2024. [Online]. Available: <http://www.ncbi.nlm.nih.gov/books/NBK542247/> (visited on 02/14/2024).
- [25] D. Cabral, T. G. Beach, L. Vedders, *et al.*, “Frequency of Alzheimer’s disease pathology at autopsy in patients with clinical normal pressure hydrocephalus,” en, *Alzheimer’s & Dementia*, vol. 7, no. 5, pp. 509–513, Sep. 2011, ISSN: 1552-5260, 1552-5279. DOI: 10.1016/j.jalz.2010.12.008. [Online]. Available: <https://alz-journals.onlinelibrary.wiley.com/doi/10.1016/j.jalz.2010.12.008> (visited on 02/28/2024).
- [26] S. Weiner, A. Junkkari, M. Sauer, *et al.*, “Novel cerebrospinal fluid biomarkers correlating with shunt responsiveness in patients with idiopathic normal pressure hydrocephalus,” en, *Fluids and Barriers of the CNS*, vol. 20, no. 1, p. 40, Jun. 2023, ISSN: 2045-8118. DOI: 10.1186/s12987-023-00440-5. [Online]. Available: <https://fluidsbarrierscns.biomedcentral.com/articles/10.1186/s12987-023-00440-5> (visited on 11/24/2023).
- [27] A. Gontsarova, D. Richardson, A. M. Methley, K. Tsang, R. Pearce, and C. Carswell, “Shunting for idiopathic normal pressure hydrocephalus,” en, *Cochrane Database of Systematic Reviews*, vol. 2022, no. 3, Cochrane Dementia and Cognitive Improvement Group, Ed., Mar. 2022, ISSN: 14651858. DOI: 10.1002/14651858.CD014923. [Online]. Available: <http://doi.wiley.com/10.1002/14651858.CD014923> (visited on 02/28/2024).
- [28] N. Rostgaard, M. H. Olsen, M. Ottenheim, *et al.*, “Differential proteomic profile of lumbar and ventricular cerebrospinal fluid,” en, *Fluids and Barriers of the CNS*, vol. 20, no. 1, p. 6, Jan. 2023, ISSN: 2045-8118. DOI: 10.1186/s12987-022-00405-0. [Online]. Available: <https://fluidsbarrierscns.biomedcentral.com/articles/10.1186/s12987-022-00405-0> (visited on 11/30/2023).
- [29] K. Blennow, P. Fredman, A. Wallin, *et al.*, “Protein analysis in cerebrospinal fluid: II. Reference values derived from healthy individuals 18-88 years of age,” *European neurology*, vol. 33, no. 2, pp. 129–133, 1993, S. Karger AG Basel, Switzerland.
- [30] M. Ahram and E. F. Petricoin, “Proteomics Discovery of Disease Biomarkers,” en, *Biomarker Insights*, vol. 3, BMI.S689, Jan. 2008, ISSN: 1177-2719, 1177-2719. DOI: 10.4137/BMI.S689. [Online]. Available: <http://journals.sagepub.com/doi/10.4137/BMI.S689> (visited on 11/19/2023).
- [31] G. Siuzdak, *Mass Spectrometry for Biotechnology*, en. Elsevier, Feb. 1996, Google-Books-ID: muoMxirJdKkC, ISBN: 978-0-08-053584-5.
- [32] G. Lubec and L. Afjehi-Sadat, “Limitations and Pitfalls in Protein Identification by Mass Spectrometry,” en, *Chemical Reviews*, vol. 107, no. 8, pp. 3568–3584, Aug. 2007, ISSN: 0009-2665, 1520-6890. DOI: 10.1021/cr068213f. [Online]. Available: <https://pubs.acs.org/doi/10.1021/cr068213f> (visited on 02/13/2024).
- [33] S. Weiner, M. Sauer, P. J. Visser, *et al.*, “Optimized sample preparation and data analysis for TMT proteomic analysis of cerebrospinal fluid applied to the identification of Alzheimer’s disease biomarkers,” en, *Clinical Proteomics*, vol. 19, no. 1, p. 13, Dec. 2022, ISSN: 1542-6416, 1559-0275. DOI: 10.1186/s12014-022-09354-0. [Online]. Available: <https://clinicalproteomicsjournal.biomedcentral.com/articles/10.1186/s12014-022-09354-0> (visited on 11/19/2023).
- [34] L. Zhang and J. E. Elias, “Relative Protein Quantification Using Tandem Mass Tag Mass Spectrometry,” in *Proteomics*, L. Comai, J. E. Katz, and P. Mallick, Eds., vol. 1550, Series Title: Methods in Molecular Biology, New York, NY: Springer New York, 2017, pp. 185–198, ISBN: 978-1-4939-6745-2. DOI: 10.1007/978-1-4939-6747-6_14. [Online]. Available: http://link.springer.com/10.1007/978-1-4939-6747-6_14 (visited on 02/14/2024).
- [35] G. Zhang, B. M. Ueberheide, S. Waldemarson, *et al.*, “Protein Quantitation Using Mass Spectrometry,” *Methods in molecular biology (Clifton, N.J.)*, vol. 673, pp. 211–222, 2010, ISSN: 1064-3745. DOI: 10.1007/978-1-60761-842-3_13. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3758905/> (visited on 03/28/2024).
- [36] M. B. de Geus, S. N. Leslie, T. Lam, *et al.*, “Mass spectrometry in cerebrospinal fluid uncovers association of glycolysis biomarkers with Alzheimer’s disease in a large clinical sample,” en, *Scientific Reports*, vol. 13, no. 1, p. 22406, Dec. 2023, Number: 1 Publisher: Nature

- Publishing Group, ISSN: 2045-2322. DOI: 10.1038/s41598-023-49440-3. [Online]. Available: <https://www.nature.com/articles/s41598-023-49440-3> (visited on 02/07/2024).
- [37] A. Thompson, J. Schäfer, K. Kuhn, *et al.*, “Tandem Mass Tags: A Novel Quantification Strategy for Comparative Analysis of Complex Protein Mixtures by MS/MS,” en, *Analytical Chemistry*, vol. 75, no. 8, pp. 1895–1904, Apr. 2003, ISSN: 0003-2700, 1520-6882. DOI: 10.1021/ac0262560. [Online]. Available: <https://pubs.acs.org/doi/10.1021/ac0262560> (visited on 01/28/2024).
- [38] A. Brenes, J. Hukelmann, D. Bensaddek, and A. I. Lamond, “Multibatch TMT Reveals False Positives, Batch Effects and Missing Values,” en, *Molecular & Cellular Proteomics*, vol. 18, no. 10, pp. 1967–1980, Oct. 2019, ISSN: 15359476. DOI: 10.1074/mcp.RA119.001472. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S1535947620315255> (visited on 02/15/2024).
- [39] S.-X. Phua, K.-P. Lim, and W. W.-B. Goh, “Perspectives for better batch effect correction in mass-spectrometry-based proteomics,” en, *Computational and Structural Biotechnology Journal*, vol. 20, pp. 4369–4375, 2022, ISSN: 20010370. DOI: 10.1016/j.csbj.2022.08.022. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S2001037022003567> (visited on 05/14/2024).
- [40] M. Sprang, M. A. Andrade-Navarro, and J.-F. Fontaine, “Batch effect detection and correction in RNA-seq data using machine-learning-based automated assessment of quality,” *BMC Bioinformatics*, vol. 23, no. 6, p. 279, Jul. 2022, ISSN: 1471-2105. DOI: 10.1186/s12859-022-04775-y. [Online]. Available: <https://doi.org/10.1186/s12859-022-04775-y> (visited on 02/15/2024).
- [41] N. Altman, “Batches and blocks, sample pools and subsamples in the design and analysis of gene expression studies,” in *Batch effects and noise in microarray experiments: sources and solutions*, A. Scherer, Ed., Wiley Online Library, 2009, pp. 33–50, ISBN: 978-0-470-74138-2 978-0-470-68598-3. DOI: 10.1002/9780470685983.
- [42] X. Wang, “Statistical Assessment of QC Metrics on Raw LC-MS/MS Data,” in *Proteomics*, L. Comai, J. E. Katz, and P. Mallick, Eds., vol. 1550, Series Title: Methods in Molecular Biology, New York, NY: Springer New York, 2017, pp. 325–337, ISBN: 978-1-4939-6745-2. DOI: 10.1007/978-1-4939-6747-6_22. [Online]. Available: http://link.springer.com/10.1007/978-1-4939-6747-6_22 (visited on 02/15/2024).
- [43] D. B. Rubin, “Inference and missing data,” *Biometrika*, vol. 63, no. 3, pp. 581–592, 1976.
- [44] T. Emmanuel, T. Maupong, D. Mpoeleng, T. Semong, B. Mphago, and O. Tabona, “A survey on missing data in machine learning,” en, *Journal of Big Data*, vol. 8, no. 1, p. 140, Oct. 2021, ISSN: 2196-1115. DOI: 10.1186/s40537-021-00516-9. [Online]. Available: <https://journalofbigdata.springeropen.com/articles/10.1186/s40537-021-00516-9> (visited on 04/02/2024).
- [45] W. Kong, H. W. H. Hui, H. Peng, and W. W. B. Goh, “Dealing with missing values in proteomics data,” en, *PROTEOMICS*, vol. 22, no. 23-24, p. 2200092, 2022, ISSN: 1615-9861. DOI: 10.1002/pmic.202200092. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1002/pmic.2022-00092> (visited on 01/24/2024).
- [46] W. W. B. Goh, H. W. H. Hui, and L. Wong, “How missing value imputation is confounded with batch effects and what you can do about it,” *Drug Discovery Today*, vol. 28, no. 9, p. 103661, Sep. 2023, ISSN: 1359-6446. DOI: 10.1016/j.drudis.2023.103661. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1359644623001770> (visited on 08/01/2024).
- [47] P. Madley-Dowd, R. Hughes, K. Tilling, and J. Heron, “The proportion of missing data should not be used to guide decisions on multiple imputation,” en, *Journal of Clinical Epidemiology*, vol. 110, pp. 63–73, Jun. 2019, ISSN: 08954356. DOI: 10.1016/j.jclinepi.2019.02.016. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S0895435618308710> (visited on 01/29/2024).
- [48] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data*, en. John Wiley & Sons, Apr. 2019, Google-Books-ID: BemMDwAAQBAJ, ISBN: 978-0-470-52679-8.
- [49] M. Liu and A. Dongre, “Proper imputation of missing values in proteomics datasets for differential expression analysis,” *Briefings in Bioinformatics*, vol. 22, no. 3, bbaa112, May

- 2021, ISSN: 1477-4054. DOI: 10.1093/bib/bbaa112. [Online]. Available: <https://doi.org/10.1093/bib/bbaa112> (visited on 05/02/2024).
- [50] J. Hyuk Lee and J. C. Huber Jr., “Evaluation of Multiple Imputation with Large Proportions of Missing Data: How Much Is Too Much?” en, *Iranian Journal of Public Health*, Jul. 2021, ISSN: 2251-6093, 2251-6085. DOI: 10.18502/ijph.v50i7.6626. [Online]. Available: <https://publish.kne-publishing.com/index.php/ijph/article/view/6626> (visited on 01/29/2024).
- [51] J. C. Jakobsen, C. Gluud, J. Wetterslev, and P. Winkel, “When and how should multiple imputation be used for handling missing data in randomised clinical trials – a practical guide with flowcharts,” en, *BMC Medical Research Methodology*, vol. 17, no. 1, p. 162, Dec. 2017, ISSN: 1471-2288. DOI: 10.1186/s12874-017-0442-1. [Online]. Available: <https://bmcmredsmethodol.biomedcentral.com/articles/10.1186/s12874-017-0442-1> (visited on 04/17/2024).
- [52] S. V. Buuren and K. Groothuis-Oudshoorn, “**mice** : Multivariate Imputation by Chained Equations in R,” en, *Journal of Statistical Software*, vol. 45, no. 3, 2011, ISSN: 1548-7660. DOI: 10.18637/jss.v045.i03. [Online]. Available: <http://www.jstatsoft.org/v45/i03/> (visited on 01/21/2024).
- [53] D. B. Rubin, “Multiple Imputation after 18+ Years,” en, *Journal of the American Statistical Association*, vol. 91, no. 434, pp. 473–489, Jun. 1996, ISSN: 0162-1459, 1537-274X. DOI: 10.1080/01621459.1996.10476908. [Online]. Available: <http://www.tandfonline.com/doi/abs/10.1080/01621459.1996.10476908> (visited on 04/19/2024).
- [54] M. J. Azur, E. A. Stuart, C. Frangakis, and P. J. Leaf, “Multiple imputation by chained equations: What is it and how does it work?” en, *International Journal of Methods in Psychiatric Research*, vol. 20, no. 1, pp. 40–49, 2011, ISSN: 1557-0657. DOI: 10.1002/mpr.329. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1002/mpr.329> (visited on 04/19/2024).
- [55] M. C. M. De Goeij, M. Van Diepen, K. J. Jager, G. Tripepi, C. Zoccali, and F. W. Dekker, “Multiple imputation: Dealing with missing data,” en, *Nephrology Dialysis Transplantation*, vol. 28, no. 10, pp. 2415–2420, Oct. 2013, ISSN: 0931-0509, 1460-2385. DOI: 10.1093/ndt/gft221. [Online]. Available: <https://academic.oup.com/ndt/article-lookup/doi/10.1093/ndt/gft221> (visited on 04/16/2024).
- [56] M. L. Gardner and M. A. Freitas, “Multiple Imputation Approaches Applied to the Missing Value Problem in Bottom-Up Proteomics,” en, *International Journal of Molecular Sciences*, vol. 22, no. 17, p. 9650, Sep. 2021, ISSN: 1422-0067. DOI: 10.3390/ijms22179650. [Online]. Available: <https://www.mdpi.com/1422-0067/22/17/9650> (visited on 04/16/2024).
- [57] L. M. Bramer, J. Irvahn, P. D. Piehowski, K. D. Rodland, and B.-J. M. Webb-Robertson, “A Review of Imputation Strategies for Isobaric Labeling-Based Shotgun Proteomics,” en, *Journal of Proteome Research*, vol. 20, no. 1, pp. 1–13, Jan. 2021, ISSN: 1535-3893, 1535-3907. DOI: 10.1021/acs.jproteome.0c00123. [Online]. Available: <https://pubs.acs.org/doi/10.1021/acs.jproteome.0c00123> (visited on 04/17/2024).
- [58] C. Kavitha, V. Mani, S. R. Srividhya, O. I. Khalaf, and C. A. Tavera Romero, “Early-Stage Alzheimer’s Disease Prediction Using Machine Learning Models,” en, *Frontiers in Public Health*, vol. 10, p. 853 294, Mar. 2022, ISSN: 2296-2565. DOI: 10.3389/fpubh.2022.853294. [Online]. Available: <https://www.frontiersin.org/articles/10.3389/fpubh.2022.853294/full> (visited on 03/28/2024).
- [59] X. Tang and J. Liu, “Comparing different algorithms for the course of Alzheimer’s disease using machine learning,” en, *Annals of Palliative Medicine*, vol. 10, no. 9, pp. 9715–9724, Sep. 2021, ISSN: 22245820, 22245839. DOI: 10.21037/apm-21-2013. [Online]. Available: <https://apm.amegroups.com/article/view/76527/html> (visited on 03/28/2024).
- [60] M. Bari Antor, A. H. M. S. Jamil, M. Mamtaz, *et al.*, “A Comparative Analysis of Machine Learning Algorithms to Predict Alzheimer’s Disease,” en, *Journal of Healthcare Engineering*, vol. 2021, J. Kang, Ed., pp. 1–12, Jul. 2021, ISSN: 2040-2309, 2040-2295. DOI: 10.1155/2021/9917919. [Online]. Available: <https://www.hindawi.com/journals/jhe/2021/9917919/> (visited on 03/28/2024).
- [61] J. F. Beltrán, B. M. Wahba, N. Hose, D. Shasha, R. P. Kline, and For the Alzheimer’s Disease Neuroimaging Initiative, “Inexpensive, non-invasive biomarkers predict Alzheimer

- transition using machine learning analysis of the Alzheimer’s Disease Neuroimaging (ADNI) database,” en, *PLOS ONE*, vol. 15, no. 7, S. D. Ginsberg, Ed., e0235663, Jul. 2020, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0235663. [Online]. Available: <https://dx.plos.org/10.1371/journal.pone.0235663> (visited on 05/27/2024).
- [62] T. Tanaka, R. Lavery, V. Varma, *et al.*, “Plasma proteomic signatures predict dementia and cognitive impairment,” en, *Alzheimer’s & Dementia: Translational Research & Clinical Interventions*, vol. 6, no. 1, e12018, Jan. 2020, ISSN: 2352-8737, 2352-8737. DOI: 10.1002/trc2.12018. [Online]. Available: <https://alz-journals.onlinelibrary.wiley.com/doi/10.1002/trc2.12018> (visited on 05/27/2024).
- [63] J. L. Shaffer, J. R. Petrella, F. C. Sheldon, *et al.*, “Predicting Cognitive Decline in Subjects at Risk for Alzheimer Disease by Using Combined Cerebrospinal Fluid, MR Imaging, and PET Biomarkers,” en, *Radiology*, vol. 266, no. 2, pp. 583–591, Feb. 2013, ISSN: 0033-8419, 1527-1315. DOI: 10.1148/radiol.12120010. [Online]. Available: <http://pubs.rsna.org/doi/10.1148/radiol.12120010> (visited on 05/27/2024).
- [64] T. A. Pascoal, J. Therriault, S. Mathotaarachchi, *et al.*, “Topographical distribution of Abeta predicts progression to dementia in Abeta positive mild cognitive impairment,” en, *Alzheimer’s & Dementia: Diagnosis, Assessment & Disease Monitoring*, vol. 12, no. 1, Jan. 2020, ISSN: 2352-8729, 2352-8729. DOI: 10.1002/dad2.12037. [Online]. Available: <https://onlinelibrary.wiley.com/doi/10.1002/dad2.12037> (visited on 05/27/2024).
- [65] S. Wang, J. Tang, and H. Liu, “Feature Selection,” in Jan. 2016, pp. 1–9. DOI: 10.1007/978-1-4899-7502-7_101-1.
- [66] G. James, D. Witten, T. Hastie, R. Tibshirani, and J. Taylor, *An Introduction to Statistical Learning: with Applications in Python* (Springer Texts in Statistics), en. Cham: Springer International Publishing, 2023, ISBN: 978-3-031-38746-3 978-3-031-38747-0. DOI: 10.1007/978-3-031-38747-0. [Online]. Available: <https://link.springer.com/10.1007/978-3-031-38747-0> (visited on 05/28/2024).
- [67] S. Chatterjee, *Assumptionless consistency of the Lasso*, Jun. 2014. [Online]. Available: <http://arxiv.org/abs/1303.5817> (visited on 04/16/2024).
- [68] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning* (Springer Series in Statistics). New York, NY: Springer, 2009, ISBN: 978-0-387-84857-0 978-0-387-84858-7. DOI: 10.1007/978-0-387-84858-7. [Online]. Available: <http://link.springer.com/10.1007/978-0-387-84858-7> (visited on 05/27/2024).
- [69] L. Breiman, “Random Forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001, ISSN: 1573-0565. DOI: 10.1023/A:1010933404324. [Online]. Available: <https://doi.org/10.1023/A:1010933404324>.
- [70] Y. Qi, “Random Forest for Bioinformatics,” in *Ensemble Machine Learning: Methods and Applications*, C. Zhang and Y. Ma, Eds., New York, NY: Springer New York, 2012, pp. 307–323, ISBN: 978-1-4419-9326-7. DOI: 10.1007/978-1-4419-9326-7_11. [Online]. Available: https://doi.org/10.1007/978-1-4419-9326-7_11.
- [71] S. Boumerdassi, É. Renault, and P. Mühlethaler, Eds., *Machine Learning for Networking: Second IFIP TC 6 International Conference, MLN 2019, Paris, France, December 3–5, 2019, Revised Selected Papers* (Lecture Notes in Computer Science), en. Cham: Springer International Publishing, 2020, vol. 12081, ISBN: 978-3-030-45777-8 978-3-030-45778-5. DOI: 10.1007/978-3-030-45778-5. [Online]. Available: <https://link.springer.com/10.1007/978-3-030-45778-5> (visited on 06/25/2024).
- [72] L. Breiman, “Bagging predictors,” en, *Machine Learning*, vol. 24, no. 2, pp. 123–140, Aug. 1996, ISSN: 0885-6125, 1573-0565. DOI: 10.1007/BF00058655. [Online]. Available: <http://link.springer.com/10.1007/BF00058655> (visited on 03/06/2024).
- [73] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” en, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco California USA: ACM, Aug. 2016, pp. 785–794, ISBN: 978-1-4503-4232-2. DOI: 10.1145/2939672.2939785. [Online]. Available: <https://dl.acm.org/doi/10.1145/2939672.2939785> (visited on 03/06/2024).
- [74] R. Bellman, *Dynamic programming*, en. Princeton, NJ: Princeton Univ. Pr, 1984, ISBN: 978-0-691-07951-6.

- [75] C. C. Aggarwal, A. Hinneburg, and D. A. Keim, "On the Surprising Behavior of Distance Metrics in High Dimensional Space," en, in *Database Theory — ICDT 2001*, G. Goos, J. Hartmanis, J. Van Leeuwen, J. Van Den Bussche, and V. Vianu, Eds., vol. 1973, Series Title: Lecture Notes in Computer Science, Berlin, Heidelberg: Springer Berlin Heidelberg, 2001, pp. 420–434, ISBN: 978-3-540-41456-8 978-3-540-44503-6. DOI: 10.1007/3-540-44503-X_27. [Online]. Available: http://link.springer.com/10.1007/3-540-44503-X_27 (visited on 05/18/2024).
- [76] J. Lever, M. Krzywinski, and N. Altman, "Model selection and overfitting," en, *Nature Methods*, vol. 13, no. 9, pp. 703–704, Sep. 2016, ISSN: 1548-7091, 1548-7105. DOI: 10.1038/nmeth.3968. [Online]. Available: <https://www.nature.com/articles/nmeth.3968> (visited on 05/14/2024).
- [77] A. Jovic, K. Brkić, and N. Bogunovic, *A review of feature selection methods with applications*. May 2015, Pages: 1205. DOI: 10.1109/MIPRO.2015.7160458.
- [78] R. J. Urbanowicz, R. S. Olson, P. Schmitt, M. Meeker, and J. H. Moore, "Benchmarking relief-based feature selection methods for bioinformatics data mining," *Journal of Biomedical Informatics*, vol. 85, pp. 168–188, Sep. 2018, ISSN: 1532-0464. DOI: 10.1016/j.jbi.2018.07.015. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S15320464183-01412> (visited on 04/22/2024).
- [79] A. Kalousis, J. Prados, and M. Hilario, "Stability of feature selection algorithms: A study on high-dimensional spaces," en, *Knowledge and Information Systems*, vol. 12, no. 1, pp. 95–116, May 2007, ISSN: 0219-1377, 0219-3116. DOI: 10.1007/s10115-006-0040-8. [Online]. Available: <https://link.springer.com/10.1007/s10115-006-0040-8> (visited on 04/19/2024).
- [80] A. Demircioğlu, "Measuring the bias of incorrect application of feature selection when using cross-validation in radiomics," *Insights into Imaging*, vol. 12, no. 1, p. 172, Nov. 2021, ISSN: 1869-4101. DOI: 10.1186/s13244-021-01115-1. [Online]. Available: <https://doi.org/10.1186/s13244-021-01115-1> (visited on 04/25/2024).
- [81] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, "Learning from class-imbalanced data: Review of methods and applications," en, *Expert Systems with Applications*, vol. 73, pp. 220–239, May 2017, ISSN: 09574174. DOI: 10.1016/j.eswa.2016.12.035. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S0957417416307175> (visited on 01/30/2024).
- [82] V. Bolón-Canedo and A. Alonso-Betanzos, "Ensembles for feature selection: A review and future trends," en, *Information Fusion*, vol. 52, pp. 1–12, Dec. 2019, ISSN: 15662535. DOI: 10.1016/j.inffus.2018.11.008. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S1566253518303440> (visited on 03/18/2024).
- [83] D. Álvarez-Estévez, N. Sánchez-Marño, A. Alonso-Betanzos, and V. Moret-Bonillo, "Reducing dimensionality in a database of sleep EEG arousals," en, *Expert Systems with Applications*, vol. 38, no. 6, pp. 7746–7754, Jun. 2011, ISSN: 09574174. DOI: 10.1016/j.eswa.2010.12.134. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S0957417410014971> (visited on 04/23/2024).
- [84] R. Tandon, A. I. Levey, J. J. Lah, N. T. Seyfried, and C. S. Mitchell, "Machine Learning Selection of Most Predictive Brain Proteins Suggests Role of Sugar Metabolism in Alzheimer's Disease," en, *Journal of Alzheimer's Disease*, vol. 92, no. 2, pp. 411–424, Jan. 2023, Publisher: IOS Press, ISSN: 1387-2877. DOI: 10.3233/JAD-220683. [Online]. Available: <https://content.iospress.com/articles/journal-of-alzheimers-disease/jad220683> (visited on 03/18/2024).
- [85] H. Abdi and L. J. Williams, "Principal component analysis," en, *WIREs Computational Statistics*, vol. 2, no. 4, pp. 433–459, 2010, ISSN: 1939-0068. DOI: 10.1002/wics.101. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1002/wics.101> (visited on 05/15/2024).
- [86] J. Čuklina, C. H. Lee, E. G. Williams, *et al.*, "Diagnostics and correction of batch effects in large-scale proteomic studies: A tutorial," en, *Molecular Systems Biology*, vol. 17, no. 8, e10240, Aug. 2021, ISSN: 1744-4292, 1744-4292. DOI: 10.15252/msb.202110240. [Online]. Available: <https://www.embopress.org/doi/10.15252/msb.202110240> (visited on 05/13/2024).

- [87] Laurens van der Maaten and Geoffrey Hinton, “Visualizing data using t-SNE.,” *Journal of machine learning research*, vol. 9, no. 11, 2008.
- [88] M. Wattenberg, F. Viégas, and I. Johnson, “How to Use t-SNE Effectively,” *Distill*, vol. 1, no. 10, 10.23915/distill.00002, Oct. 2016. DOI: 10.23915/distill.00002. [Online]. Available: <http://distill.pub/2016/misread-tsne> (visited on 05/16/2024).
- [89] T. M. Khoshgoftaar, C. Seiffert, J. V. Hulse, A. Napolitano, and A. Folleco, “Learning with limited minority class data,” en, in *Sixth International Conference on Machine Learning and Applications (ICMLA 2007)*, Cincinnati, OH, USA: IEEE, Dec. 2007, pp. 348–353, ISBN: 978-0-7695-3069-7. DOI: 10.1109/ICMLA.2007.76. [Online]. Available: <http://ieeexplore.ieee.org/document/4457255/> (visited on 02/16/2024).
- [90] M. R. Longadge, S. S. Dongre, and D. L. Malik, “Class Imbalance Problem in Data Mining: Review,” en, vol. 2, no. 1, 2013.
- [91] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique,” en, *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, Jun. 2002, ISSN: 1076-9757. DOI: 10.1613/jair.953. [Online]. Available: <https://www.jair.org/index.php/jair/article/view/10302> (visited on 04/30/2024).
- [92] Z. Vujovic, “Classification Model Evaluation Metrics,” *International Journal of Advanced Computer Science and Applications*, vol. Volume 12, pp. 599–606, Jul. 2021. DOI: 10.14569/IJACSA.2021.0120670.
- [93] D. Chicco, N. Tötsch, and G. Jurman, “The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation,” en, *BioData Mining*, vol. 14, no. 1, p. 13, Feb. 2021, ISSN: 1756-0381. DOI: 10.1186/s13040-021-00244-z. [Online]. Available: <https://doi.org/10.1186/s13040-021-00244-z> (visited on 05/27/2024).
- [94] M. Grandini, E. Bagli, and G. Visani, *Metrics for Multi-Class Classification: An Overview*, en, arXiv:2008.05756 [cs, stat], Aug. 2020. [Online]. Available: <http://arxiv.org/abs/2008.05756> (visited on 05/28/2024).
- [95] D. Chicco and G. Jurman, “The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation,” en, *BMC Genomics*, vol. 21, no. 1, p. 6, Jan. 2020, ISSN: 1471-2164. DOI: 10.1186/s12864-019-6413-7. [Online]. Available: <https://doi.org/10.1186/s12864-019-6413-7> (visited on 05/27/2024).
- [96] J. Dukart, K. Mueller, A. Villringer, *et al.*, “Relationship between imaging biomarkers, age, progression and symptom severity in Alzheimer’s disease,” en, *NeuroImage: Clinical*, vol. 3, pp. 84–94, 2013, ISSN: 22131582. DOI: 10.1016/j.nicl.2013.07.005. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S2213158213000922> (visited on 05/18/2024).
- [97] M. Mielke, P. Vemuri, and W. Rocca, “Clinical epidemiology of Alzheimers disease: Assessing sex and gender differences,” en, *Clinical Epidemiology*, p. 37, Jan. 2014, ISSN: 1179-1349. DOI: 10.2147/CLEP.S37929. [Online]. Available: <http://www.dovepress.com/clinical-epidemiology-of-alzheimers-disease-assessing-sex-and-gender-peer-reviewed-article-CLEP> (visited on 05/18/2024).
- [98] F. Erhard and R. Zimmer, “Detecting outlier peptides in quantitative high-throughput mass spectrometry data,” en, *Journal of Proteomics*, vol. 75, no. 11, pp. 3230–3239, Jun. 2012, ISSN: 18743919. DOI: 10.1016/j.jprot.2012.03.032. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S1874391912001650> (visited on 06/25/2024).
- [99] *Sklearn.impute.IterativeImputer*, en. [Online]. Available: <https://scikit-learn/stable/modules/generated/sklearn.impute.IterativeImputer.html> (visited on 04/05/2024).
- [100] W. E. Johnson, C. Li, and A. Rabinovic, “Adjusting batch effects in microarray expression data using empirical Bayes methods,” *Biostatistics*, vol. 8, no. 1, pp. 118–127, Jan. 2007, ISSN: 1465-4644. DOI: 10.1093/biostatistics/kxj037. [Online]. Available: <https://doi.org/10.1093/biostatistics/kxj037> (visited on 03/14/2024).
- [101] A. Behdenna, M. Colange, J. Haziza, *et al.*, “pyComBat, a Python tool for batch effects correction in high-throughput molecular data using empirical Bayes methods,” en, *BMC Bioinformatics*, vol. 24, no. 1, p. 459, Dec. 2023, ISSN: 1471-2105. DOI: 10.1186/s12859-023-05578-5. [Online]. Available: <https://doi.org/10.1186/s12859-023-05578-5> (visited on 03/14/2024).

- [102] *Python Package Introduction — xgboost 2.0.3 documentation*. [Online]. Available: https://xgboost.readthedocs.io/en/stable/python/python_intro.html (visited on 05/19/2024).
- [103] *Skopt.BayesSearchCV — scikit-optimize 0.8.1 documentation*. [Online]. Available: <https://scikit-optimize.github.io/stable/modules/generated/skopt.BayesSearchCV.html> (visited on 05/19/2024).
- [104] J. Bergstra and Y. Bengio, “Random Search for Hyper-Parameter Optimization,” en,
- [105] N. Altman and M. Krzywinski, “The curse (s) of dimensionality,” en, *Nature Methods*, vol. 15, no. 6, pp. 399–400, Jun. 2018. DOI: <https://doi.org/10.1038/s41592-018-0019-x>.
- [106] M. M. Mielke, J. A. Syrjanen, K. Blennow, *et al.*, “Plasma and CSF neurofilament light,” *Neurology*, vol. 93, no. 3, e252–e260, Jul. 2019, ISSN: 0028-3878. DOI: [10.1212/WNL.00000000000007767](https://doi.org/10.1212/WNL.00000000000007767). [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6656645/> (visited on 05/29/2024).
- [107] K. Dhiman, V. B. Gupta, V. L. Villemagne, *et al.*, “Cerebrospinal fluid neurofilament light concentration predicts brain atrophy and cognition in Alzheimer’s disease,” en, *Alzheimer’s & Dementia: Diagnosis, Assessment & Disease Monitoring*, vol. 12, no. 1, e12005, 2020, ISSN: 2352-8729. DOI: [10.1002/dad2.12005](https://doi.org/10.1002/dad2.12005). [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1002/dad2.12005> (visited on 05/29/2024).
- [108] G. Giacomucci, S. Mazzeo, S. Bagnoli, *et al.*, “Plasma neurofilament light chain as a biomarker of Alzheimer’s disease in Subjective Cognitive Decline and Mild Cognitive Impairment,” *Journal of Neurology*, vol. 269, no. 8, pp. 4270–4280, 2022, ISSN: 0340-5354. DOI: [10.1007/s00415-022-11055-5](https://doi.org/10.1007/s00415-022-11055-5). [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9293849/> (visited on 05/29/2024).
- [109] G. Sathe, C. H. Na, S. Renuse, *et al.*, “Quantitative Proteomic Profiling of Cerebrospinal Fluid to Identify Candidate Biomarkers for Alzheimer’s Disease,” eng, *Proteomics. Clinical Applications*, vol. 13, no. 4, e1800105, Jul. 2019, ISSN: 1862-8354. DOI: [10.1002/prca.201800105](https://doi.org/10.1002/prca.201800105).
- [110] Q.-Q. Tao, X. Cai, Y.-Y. Xue, *et al.*, “Alzheimer’s disease early diagnostic and staging biomarkers revealed by large-scale cerebrospinal fluid and serum proteomic profiling,” *The Innovation*, vol. 5, no. 1, p. 100544, Jan. 2024, ISSN: 2666-6758. DOI: [10.1016/j.xinn.2023.100544](https://doi.org/10.1016/j.xinn.2023.100544). [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2666675823001728> (visited on 06/03/2024).
- [111] A. Soldan, S. Oh, T. Ryu, *et al.*, “NPTX2 in Cerebrospinal Fluid Predicts the Progression From Normal Cognition to Mild Cognitive Impairment,” eng, *Annals of Neurology*, vol. 94, no. 4, pp. 620–631, Oct. 2023, ISSN: 1531-8249. DOI: [10.1002/ana.26725](https://doi.org/10.1002/ana.26725).
- [112] M.-F. Xiao, D. Xu, M. T. Craig, *et al.*, “NPTX2 and cognitive dysfunction in Alzheimer’s Disease,” *eLife*, vol. 6, e23798, ISSN: 2050-084X. DOI: [10.7554/eLife.23798](https://doi.org/10.7554/eLife.23798). [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5404919/> (visited on 05/29/2024).
- [113] F. N. Sepe, D. Chiasserini, and L. Parnetti, “Role of FABP3 as Biomarker in Alzheimer’s Disease and Synucleinopathies,” *Future Neurology*, vol. 13, no. 4, pp. 199–207, Jul. 2018, ISSN: 1479-6708. DOI: [10.2217/fnl-2018-0003](https://doi.org/10.2217/fnl-2018-0003). [Online]. Available: <https://doi.org/10.2217/fnl-2018-0003> (visited on 05/29/2024).
- [114] M. Dulewicz, A. Kulczyńska-Przybik, A. Słowik, R. Borawska, and B. Mroczo, “Fatty Acid Binding Protein 3 (FABP3) and Apolipoprotein E4 (ApoE4) as Lipid Metabolism-Related Biomarkers of Alzheimer’s Disease,” eng, *Journal of Clinical Medicine*, vol. 10, no. 14, p. 3009, Jul. 2021, ISSN: 2077-0383. DOI: [10.3390/jcm10143009](https://doi.org/10.3390/jcm10143009).
- [115] E. Ostertagová, O. Ostertag, and J. Kováč, “Methodology and Application of the Kruskal-Wallis Test,” en, *Applied Mechanics and Materials*, vol. 611, pp. 115–120, Aug. 2014, ISSN: 1662-7482. DOI: [10.4028/www.scientific.net/AMM.611.115](https://doi.org/10.4028/www.scientific.net/AMM.611.115). [Online]. Available: <https://www.scientific.net/AMM.611.115> (visited on 08/13/2024).
- [116] *Statistical functions (scipy.stats) — SciPy v1.14.0 Manual*. [Online]. Available: <https://docs.scipy.org/doc/scipy/reference/stats.html> (visited on 08/17/2024).
- [117] *Scikit-posthocs — scikit-posthocs 0.7.0 documentation*. [Online]. Available: <https://scikit-posthocs.readthedocs.io/en/latest/index.html> (visited on 08/17/2024).

- [118] S.-J. Lee and A. C. McPherron, "Regulation of myostatin activity and muscle growth," *Proceedings of the National Academy of Sciences*, vol. 98, no. 16, pp. 9306–9311, Jul. 2001, Publisher: Proceedings of the National Academy of Sciences. DOI: 10.1073/pnas.151270098. [Online]. Available: <https://www.pnas.org/doi/full/10.1073/pnas.151270098> (visited on 06/03/2024).
- [119] M. Sharma, B. Langley, J. Bass, and R. Kambadur, "Myostatin in Muscle Growth and Repair," en-US, *Exercise and Sport Sciences Reviews*, vol. 29, no. 4, p. 155, Oct. 2001, ISSN: 0091-6331. [Online]. Available: https://journals.lww.com/acsm-essr/fulltext/2001/10000/Myostatin_in_Muscle_Growth_and_Repair.4.aspx (visited on 06/04/2024).
- [120] Y.-S. Lin, F.-Y. Lin, and Y.-H. Hsiao, "Myostatin Is Associated With Cognitive Decline in an Animal Model of Alzheimer's Disease," eng, *Molecular Neurobiology*, vol. 56, no. 3, pp. 1984–1991, Mar. 2019, ISSN: 1559-1182. DOI: 10.1007/s12035-018-1201-y.
- [121] X.-J. Huang, Y.-K. Choi, H.-S. Im, O. Yarimaga, E. Yoon, and H.-S. Kim, "Aspartate Aminotransferase (AST/GOT) and Alanine Aminotransferase (ALT/GPT) Detection Techniques," *Sensors (Basel, Switzerland)*, vol. 6, no. 7, pp. 756–782, Jul. 2006, ISSN: 1424-8220. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3894536/> (visited on 06/03/2024).
- [122] K. Nho, A. Kueider-Paisley, S. Ahmad, *et al.*, "Association of Altered Liver Enzymes With Alzheimer Disease Diagnosis, Cognition, Neuroimaging Measures, and Cerebrospinal Fluid Biomarkers," *JAMA Network Open*, vol. 2, no. 7, e197978, Jul. 2019, ISSN: 2574-3805. DOI: 10.1001/jamanetworkopen.2019.7978. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6669786/> (visited on 06/03/2024).
- [123] Y. Lu, J. R. Pike, E. Selvin, *et al.*, "Low liver enzymes and risk of dementia. The Atherosclerosis Risk in Communities (ARIC) Study," *Journal of Alzheimer's disease : JAD*, vol. 79, no. 4, pp. 1775–1784, 2021, ISSN: 1387-2877. DOI: 10.3233/JAD-201241. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8679120/> (visited on 06/03/2024).
- [124] A. Ghosh and K. P. Giese, "Calcium/calmodulin-dependent kinase II and Alzheimer's disease," *Molecular Brain*, vol. 8, p. 78, Nov. 2015, ISSN: 1756-6606. DOI: 10.1186/s13041-015-0166-2. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4657223/> (visited on 06/03/2024).
- [125] Y. Xia, S. Prokop, and B. I. Giasson, "'Don't Phos Over Tau': Recent developments in clinical biomarkers and therapies targeting tau phosphorylation in Alzheimer's disease and other tauopathies," *Molecular Neurodegeneration*, vol. 16, no. 1, p. 37, Jun. 2021, ISSN: 1750-1326. DOI: 10.1186/s13024-021-00460-5. [Online]. Available: <https://doi.org/10.1186/s13024-021-00460-5> (visited on 06/03/2024).
- [126] P. Han, W. Liang, L. C. Baxter, *et al.*, "Pituitary adenylate cyclase-activating polypeptide is reduced in Alzheimer disease," *Neurology*, vol. 82, no. 19, pp. 1724–1728, May 2014, ISSN: 0028-3878. DOI: 10.1212/WNL.0000000000000417. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4032204/> (visited on 06/04/2024).
- [127] E. E. Neumaier, V. Rothhammer, and M. Linnerbauer, "The role of midkine in health and disease," English, *Frontiers in Immunology*, vol. 14, Nov. 2023, Publisher: Frontiers, ISSN: 1664-3224. DOI: 10.3389/fimmu.2023.1310094. [Online]. Available: <https://www.frontiersin.org/journals/immunology/articles/10.3389/fimmu.2023.1310094/full> (visited on 06/05/2024).
- [128] R. Salama, H. Muramatsu, E. Shimizu, *et al.*, "Increased midkine levels in sera from patients with Alzheimer's disease," *Progress in neuro-psychopharmacology & biological psychiatry*, vol. 29, pp. 611–6, Jun. 2005. DOI: 10.1016/j.pnpbp.2005.01.018.
- [129] E. Shimizu and D. Matsuzawa, "Midkine in Psychiatric and Neurodegenerative Diseases," en, in *Midkine: From Embryogenesis to Pathogenesis and Therapy*, M. Ergüven, T. Muramatsu, and A. Bilir, Eds., Dordrecht: Springer Netherlands, 2012, pp. 165–170, ISBN: 978-94-007-4234-5. DOI: 10.1007/978-94-007-4234-5_14. [Online]. Available: https://doi.org/10.1007/978-94-007-4234-5_14 (visited on 06/05/2024).
- [130] Y. Xiang, J. Xin, W. Le, and Y. Yang, "Neurogranin: A Potential Biomarker of Neurological and Mental Diseases," *Frontiers in Aging Neuroscience*, vol. 12, p. 584743, Oct. 2020, ISSN: 1663-4365. DOI: 10.3389/fnagi.2020.584743. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7573493/> (visited on 06/04/2024).

- [131] H. Wellington, R. W. Paterson, E. Portelius, *et al.*, “Increased CSF neurogranin concentration is specific to Alzheimer disease,” *Neurology*, vol. 86, no. 9, pp. 829–835, Mar. 2016, ISSN: 0028-3878. DOI: 10.1212/WNL.0000000000002423. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4793782/> (visited on 06/04/2024).

A

Appendix A

The following 4 tables (A.1-A.4) include the results for all the binary classification runs with multiple combinations of missingness thresholds, imputation methods, size of union feature selectors, models and evaluation metrics.

Models				XGB				LR				RF				Soft votes				Hard votes				Tot f.			
CSF	k Union	Misc.	Imp.	CombBat	Ang.	Acc	F1	AUC	MCC	Acc	F1	AUC	MCC	Acc	F1	AUC	MCC	Acc	F1	AUC	MCC	Acc	F1	AUC	MCC	ShH f.	ShH f.
V	1	0	x	OH	OH	0.83	0.47	0.83	0.57	0.83	0.95	0.87	0.46	0.80	0.90	0.90	0.38	0.83	0.58	0.89	0.48	0.82	0.19	0.89	0.48	5	2
V	2	0	x	OH	OH	0.80	0.45	0.85	0.35	0.78	0.43	0.79	0.31	0.80	0.40	0.82	0.31	0.78	0.38	0.84	0.27	0.78	0.38	0.84	0.27	16	3
V	5	0	x	OH	OH	0.75	0.29	0.75	0.15	0.78	0.65	0.76	0.41	0.80	0.40	0.84	0.31	0.77	0.36	0.85	0.23	0.78	0.32	0.85	0.23	43	7
V	10	0	x	OH	OH	0.78	0.43	0.79	0.31	0.82	0.66	0.82	0.45	0.75	0.12	0.75	0.02	0.78	0.32	0.83	0.23	0.78	0.32	0.83	0.23	82	14
V	1	20	Min	OH	OH	0.77	0.30	0.73	0.19	0.82	0.52	0.84	0.42	0.83	0.50	0.82	0.44	0.80	0.40	0.83	0.31	0.82	0.48	0.83	0.39	6	2
V	2	20	Min	OH	OH	0.77	0.22	0.58	0.13	0.70	0.25	0.59	0.06	0.77	0.13	0.67	0.06	0.78	0.32	0.62	0.23	0.77	0.22	0.62	0.13	21	3
V	5	20	Min	OH	OH	0.73	0.27	0.74	0.12	0.70	0.19	0.61	0.03	0.70	0.10	0.67	-0.07	0.72	0.11	0.72	-0.04	0.72	0.19	0.72	0.03	52	6
V	10	20	Min	OH	OH	0.78	0.38	0.78	0.27	0.67	0.09	0.61	-0.19	0.75	0.00	0.77	-0.10	0.75	0.00	0.74	-0.10	0.75	0.00	0.74	-0.10	100	11
V	1	50	Min	OH	OH	0.80	0.54	0.66	0.41	0.75	0.40	0.71	0.24	0.77	0.36	0.73	0.23	0.75	0.35	0.72	0.20	0.78	0.43	0.72	0.31	9	1
V	2	50	Min	OH	OH	0.73	0.11	0.62	-0.01	0.77	0.42	0.67	0.27	0.78	0.13	0.62	0.13	0.72	0.19	0.63	0.03	0.77	0.13	0.63	0.06	23	2
V	5	50	Min	OH	OH	0.73	0.11	0.73	-0.01	0.72	0.26	0.53	0.09	0.75	0.21	0.61	0.09	0.75	0.12	0.64	0.02	0.75	0.21	0.64	0.09	52	6
V	10	50	Min	OH	OH	0.75	0.21	0.76	0.09	0.70	0.18	0.62	0.01	0.78	0.00	0.67	0.00	0.75	0.12	0.69	0.02	0.77	0.13	0.69	0.06	109	11
V	1	0	x	On	OH	0.78	0.43	0.81	0.31	0.80	0.40	0.79	0.31	0.82	0.42	0.86	0.36	0.80	0.45	0.85	0.35	0.78	0.32	0.85	0.23	7	2
V	2	0	x	On	OH	0.65	0.22	0.62	-0.00	0.70	0.36	0.60	0.16	0.67	0.00	0.57	-0.19	0.72	0.27	0.62	0.19	0.68	0.17	0.62	-0.02	18	2
V	5	0	x	On	OH	0.80	0.25	0.76	0.25	0.75	0.29	0.76	0.15	0.78	0.00	0.77	0.00	0.78	0.24	0.77	0.18	0.80	0.14	0.77	0.25	41	6
V	10	0	x	On	OH	0.78	0.32	0.79	0.23	0.73	0.38	0.78	0.21	0.82	0.42	0.73	0.36	0.83	0.44	0.80	0.43	0.82	0.42	0.80	0.36	84	13
V	1	0	x	OH	SMOTE	0.78	0.52	0.81	0.38	0.77	0.61	0.85	0.50	0.78	0.55	0.80	0.41	0.78	0.52	0.83	0.38	0.80	0.57	0.83	0.44	8	1
V	2	0	x	OH	SMOTE	0.83	0.58	0.83	0.48	0.75	0.52	0.81	0.36	0.85	0.64	0.85	0.55	0.82	0.56	0.83	0.45	0.83	0.62	0.83	0.51	14	3
V	5	0	x	OH	SMOTE	0.78	0.58	0.81	0.45	0.82	0.56	0.83	0.45	0.80	0.54	0.89	0.41	0.85	0.64	0.87	0.55	0.82	0.59	0.87	0.48	38	8
V	10	0	x	OH	SMOTE	0.77	0.46	0.79	0.31	0.77	0.56	0.73	0.23	0.80	0.45	0.80	0.35	0.77	0.46	0.80	0.31	0.80	0.50	0.80	0.38	79	13
V	1	0	x	On	SMOTE	0.65	0.32	0.63	0.10	0.67	0.41	0.68	0.21	0.65	0.16	0.64	-0.06	0.65	0.28	0.66	0.05	0.65	0.28	0.66	0.05	6	1
V	2	0	x	On	SMOTE	0.68	0.34	0.64	0.14	0.68	0.34	0.72	0.14	0.75	0.48	0.68	0.32	0.70	0.40	0.67	0.21	0.72	0.41	0.67	0.23	18	2
V	5	0	x	On	SMOTE	0.87	0.71	0.80	0.63	0.80	0.69	0.86	0.61	0.83	0.62	0.88	0.51	0.87	0.69	0.90	0.61	0.87	0.69	0.90	0.61	40	8
V	10	0	x	On	SMOTE	0.78	0.48	0.80	0.34	0.80	0.54	0.82	0.41	0.82	0.56	0.79	0.45	0.77	0.46	0.80	0.31	0.82	0.56	0.80	0.45	80	13
V	1	20	MICE	OH	OH	0.67	0.09	0.62	-0.11	0.72	0.19	0.73	0.03	0.73	0.27	0.67	0.12	0.70	0.10	0.69	-0.07	0.72	0.19	0.69	0.03	10	1
V	2	20	MICE	OH	OH	0.75	0.40	0.75	0.24	0.77	0.46	0.78	0.31	0.72	0.32	0.68	0.14	0.75	0.40	0.75	0.24	0.75	0.40	0.75	0.24	19	3
V	5	20	MICE	OH	OH	0.72	0.11	0.68	-0.04	0.63	0.15	0.63	-0.08	0.70	0.10	0.61	-0.07	0.68	0.10	0.62	-0.09	0.68	0.10	0.62	-0.09	52	6
V	10	20	MICE	OH	OH	0.72	0.19	0.68	0.03	0.73	0.33	0.69	0.17	0.78	0.13	0.68	0.13	0.73	0.20	0.71	0.06	0.75	0.21	0.71	0.09	97	12
V	1	20	MICE	On	OH	0.77	0.30	0.73	0.19	0.82	0.52	0.84	0.42	0.83	0.50	0.82	0.44	0.80	0.40	0.83	0.31	0.82	0.48	0.83	0.39	6	2
V	2	20	MICE	On	OH	0.77	0.22	0.68	0.13	0.70	0.25	0.59	0.06	0.77	0.13	0.67	0.06	0.78	0.32	0.62	0.23	0.77	0.22	0.62	0.13	21	3
V	5	20	MICE	On	OH	0.73	0.27	0.74	0.12	0.72	0.19	0.61	0.03	0.70	0.10	0.67	-0.07	0.72	0.11	0.72	-0.04	0.72	0.19	0.72	0.03	52	6
V	10	20	MICE	On	OH	0.78	0.38	0.78	0.27	0.67	0.00	0.61	-0.19	0.75	0.00	0.77	-0.10	0.75	0.00	0.74	-0.10	0.75	0.00	0.74	-0.10	100	11
V	1	20	MICE	OH	SMOTE	0.77	0.53	0.73	0.39	0.83	0.64	0.80	0.54	0.82	0.59	0.77	0.48	0.80	0.57	0.78	0.44	0.82	0.59	0.78	0.48	7	1
V	2	20	MICE	OH	SMOTE	0.77	0.46	0.76	0.31	0.72	0.41	0.76	0.23	0.77	0.46	0.76	0.31	0.80	0.54	0.76	0.44	0.75	0.44	0.76	0.28	19	3
V	5	20	MICE	OH	SMOTE	0.83	0.62	0.84	0.51	0.73	0.38	0.70	0.21	0.82	0.56	0.81	0.45	0.78	0.52	0.82	0.38	0.80	0.54	0.82	0.41	53	6
V	10	20	MICE	OH	SMOTE	0.75	0.40	0.82	0.24	0.65	0.22	0.69	-0.00	0.77	0.36	0.79	0.23	0.72	0.26	0.75	0.09	0.75	0.35	0.75	0.20	96	11
V	1	20	MICE	On	SMOTE	0.58	0.24	0.49	-0.03	0.62	0.26	0.53	0.01	0.62	0.26	0.51	0.01	0.65	0.28	0.51	0.05	0.62	0.30	0.51	0.06	13	1
V	2	20	MICE	On	SMOTE	0.67	0.09	0.55	-0.11	0.77	0.42	0.73	0.27	0.70	0.10	0.57	-0.07	0.72	0.19	0.63	0.03	0.72	0.19	0.63	0.03	27	2
V	5	20	MICE	On	SMOTE	0.68	0.30	0.60	0.09	0.73	0.38	0.77	0.21	0.75	0.35	0.67	0.20	0.70	0.25	0.66	0.06	0.72	0.32	0.66	0.14	56	5
V	10	20	MICE	On	SMOTE	0.68	0.30	0.58	0.09	0.73	0.33	0.65	0.17	0.77	0.36	0.70	0.23	0.72	0.22	0.64	0.14	0.72	0.32	0.64	0.14	101	10

Table A.1: Experiment results for all runs on \tilde{D}_{PV} binary classification task.

Model:		XGB				LR				RF				Soft vote				Hard vote					
CSF	k Union	Miss.	Imp.	ComBat	Aug.	Acc	F1	AUC	MCC	Acc	F1	AUC	MCC	Acc	F1	AUC	MCC	Acc	F1	AUC	MCC	Tot. f.	Stabl. f.
L	1	0	x	Off	Off	0.73	0.36	0.72	0.19	0.65	0.18	0.68	-0.04	0.78	0.42	0.70	0.16	0.76	0.40	0.70	0.25	12	1
L	2	0	x	Off	Off	0.73	0.13	0.62	-0.03	0.76	0.40	0.70	0.25	0.73	0.13	0.62	0.13	0.76	0.14	0.66	0.04	18	3
L	5	0	x	Off	Off	0.75	0.24	0.62	0.09	0.71	0.12	0.66	-0.05	0.78	0.15	0.42	0.09	0.76	0.14	0.54	0.09	49	6
L	10	0	x	Off	Off	0.76	0.14	0.77	0.04	0.73	0.13	0.62	-0.03	0.76	0.00	0.59	-0.10	0.75	0.00	0.68	-0.12	104	11
L	1	20	Min	Off	Off	0.75	0.13	0.62	0.00	0.71	0.21	0.68	0.03	0.75	0.13	0.60	0.00	0.75	0.24	0.62	0.09	11	1
L	2	20	Min	Off	Off	0.73	0.00	0.67	-0.14	0.73	0.00	0.50	-0.14	0.76	0.00	0.64	-0.10	0.73	0.00	0.67	-0.14	22	3
L	5	20	Min	Off	Off	0.78	0.15	0.83	0.09	0.69	0.11	0.65	-0.08	0.76	0.00	0.70	-0.10	0.75	0.00	0.76	-0.10	56	6
L	10	20	Min	Off	Off	0.80	0.29	0.80	0.22	0.75	0.00	0.63	-0.12	0.75	0.00	0.67	-0.12	0.75	0.00	0.77	-0.10	112	11
L	1	50	Min	Off	Off	0.69	0.20	0.50	0.00	0.76	0.25	0.61	0.13	0.75	0.24	0.50	0.09	0.76	0.25	0.53	0.13	11	1
L	2	50	Min	Off	Off	0.76	0.00	0.60	-0.10	0.67	0.11	0.63	-0.10	0.73	0.13	0.64	-0.03	0.69	0.11	0.62	-0.08	20	2
L	5	50	Min	Off	Off	0.73	0.22	0.70	0.06	0.67	0.11	0.56	-0.10	0.75	0.00	0.52	-0.12	0.73	0.00	0.60	-0.14	59	5
L	10	50	Min	Off	Off	0.73	0.00	0.66	-0.14	0.69	0.11	0.59	-0.08	0.78	0.00	0.58	-0.07	0.73	0.00	0.59	-0.14	114	11
L	1	0	x	On	Off	0.78	0.42	0.62	0.29	0.76	0.33	0.58	0.19	0.82	0.40	0.67	0.33	0.78	0.42	0.66	0.29	10	1
L	2	0	x	On	Off	0.78	0.42	0.67	0.29	0.74	0.13	0.58	0.00	0.80	0.29	0.72	0.22	0.74	0.13	0.72	0.00	18	3
L	5	0	x	On	Off	0.80	0.29	0.68	0.22	0.78	0.35	0.79	0.23	0.74	0.24	0.64	0.09	0.78	0.27	0.73	0.17	48	6
L	10	0	x	On	Off	0.78	0.35	0.72	0.23	0.66	0.26	0.64	0.05	0.78	0.15	0.66	0.08	0.74	0.24	0.64	0.09	97	12
L	1	0	x	Off	SMOTE	0.65	0.31	0.66	0.09	0.67	0.41	0.66	0.23	0.75	0.38	0.70	0.22	0.67	0.37	0.69	0.17	10	1
L	2	0	x	Off	SMOTE	0.69	0.33	0.63	0.14	0.69	0.33	0.62	0.14	0.73	0.36	0.57	0.10	0.67	0.26	0.59	0.05	20	2
L	5	0	x	Off	SMOTE	0.71	0.12	0.68	-0.05	0.69	0.11	0.69	-0.08	0.75	0.00	0.70	-0.12	0.76	0.25	0.67	-0.14	52	5
L	10	0	x	Off	SMOTE	0.76	0.25	0.56	0.13	0.73	0.00	0.51	-0.14	0.73	0.13	0.54	-0.03	0.76	0.14	0.57	0.04	105	11
L	1	0	x	On	SMOTE	0.72	0.46	0.71	0.30	0.74	0.48	0.77	0.33	0.84	0.36	0.72	0.15	0.72	0.46	0.77	0.30	7	2
L	2	0	x	On	SMOTE	0.78	0.35	0.73	0.23	0.82	0.53	0.70	0.42	0.76	0.25	0.71	0.12	0.80	0.44	0.74	0.33	15	3
L	5	0	x	On	SMOTE	0.72	0.13	0.69	-0.03	0.78	0.27	0.77	0.17	0.78	0.27	0.73	0.17	0.80	0.29	0.77	0.22	46	6
L	10	0	x	On	SMOTE	0.76	0.25	0.58	0.12	0.72	0.13	0.74	-0.03	0.78	0.27	0.62	0.17	0.74	0.13	0.65	0.00	99	12

Table A.4: Experiment results for all runs on \tilde{D}_{PeL} binary classification task.

B

Appendix B

This section shows the tests run to confirm the confidence intervals on the best-performing models. Each model was run on the same settings ten times, and the confidence intervals were calculated from all scoring metrics.

B. Appendix B

Methods										XGB			LR			RF			Soft			Hard					
CSF	k Union	Mis.	Imp.	ComBat	Aug.	Acc	F1	AUC	MCC	Acc	F1	AUC	MCC	Acc	F1	AUC	MCC	Acc	F1	AUC	MCC	Tot. f.	Stbl. f.				
V	2	0	x	OH	SMOTE	0.80	0.50	0.73	0.38	0.73	0.43	0.82	0.26	0.82	0.59	0.83	0.48	0.82	0.56	0.85	0.45	0.80	14	3			
V	2	0	x	OH	SMOTE	0.80	0.57	0.83	0.44	0.78	0.58	0.75	0.45	0.82	0.59	0.83	0.44	0.82	0.59	0.84	0.48	0.80	14	3			
V	2	0	x	OH	SMOTE	0.80	0.54	0.77	0.41	0.77	0.53	0.83	0.39	0.78	0.48	0.82	0.34	0.78	0.55	0.83	0.41	0.83	17	3			
V	2	0	x	OH	SMOTE	0.78	0.55	0.85	0.41	0.75	0.55	0.88	0.40	0.82	0.59	0.86	0.48	0.80	0.57	0.86	0.44	0.82	14	3			
V	2	0	x	OH	SMOTE	0.83	0.58	0.85	0.48	0.77	0.56	0.82	0.42	0.83	0.64	0.88	0.54	0.83	0.64	0.87	0.54	0.83	16	3			
V	2	0	x	OH	SMOTE	0.72	0.41	0.75	0.23	0.72	0.48	0.77	0.31	0.78	0.52	0.76	0.38	0.75	0.48	0.78	0.32	0.75	16	3			
V	2	0	x	OH	SMOTE	0.80	0.57	0.85	0.44	0.72	0.48	0.78	0.31	0.82	0.62	0.82	0.51	0.82	0.62	0.85	0.51	0.82	17	3			
V	2	0	x	OH	SMOTE	0.78	0.52	0.80	0.38	0.80	0.60	0.81	0.48	0.82	0.59	0.87	0.48	0.82	0.59	0.85	0.48	0.80	16	3			
V	2	0	x	OH	SMOTE	0.85	0.64	0.81	0.55	0.80	0.54	0.76	0.41	0.83	0.62	0.85	0.51	0.82	0.56	0.84	0.45	0.82	15	3			
V	2	0	x	OH	SMOTE	0.82	0.62	0.82	0.51	0.75	0.55	0.77	0.40	0.80	0.57	0.86	0.44	0.82	0.62	0.85	0.51	0.82	18	3			
Conf. Low (95%)						0.77	0.50	0.78	0.36	0.74	0.40	0.77	0.34	0.80	0.55	0.81	0.42	0.79	0.55	0.82	0.41	0.79	0.53	0.82	0.39		
Conf. High (95%)						0.82	0.60	0.84	0.49	0.78	0.57	0.83	0.43	0.82	0.61	0.86	0.50	0.82	0.61	0.86	0.61	0.86	0.50	0.83	0.62	0.86	0.50
L	5	0	x	On	SMOTE	0.78	0.42	0.74	0.29	0.82	0.61	0.81	0.50	0.74	0.13	0.71	0.00	0.76	0.33	0.78	0.19	0.78	0.35	0.78	0.23	41	5
L	5	0	x	On	SMOTE	0.76	0.00	0.64	-0.10	0.70	0.00	0.74	-0.17	0.74	0.00	0.70	-0.13	0.74	0.00	0.68	-0.13	0.74	0.00	0.68	-0.13	45	5
L	5	0	x	On	SMOTE	0.76	0.14	0.75	0.04	0.72	0.13	0.67	-0.03	0.80	0.17	0.71	0.15	0.74	0.24	0.76	0.09	0.78	0.15	0.76	0.08	47	5
L	5	0	x	On	SMOTE	0.74	0.24	0.53	0.09	0.72	0.22	0.68	0.05	0.66	0.19	0.59	-0.02	0.74	0.24	0.62	0.09	0.70	0.21	0.62	0.03	52	5
L	5	0	x	On	SMOTE	0.76	0.40	0.78	0.25	0.74	0.38	0.78	0.22	0.74	0.13	0.74	0.00	0.76	0.33	0.78	0.19	0.76	0.33	0.78	0.19	48	6
L	5	0	x	On	SMOTE	0.74	0.38	0.71	0.22	0.72	0.13	0.54	-0.03	0.76	0.25	0.68	0.12	0.74	0.24	0.67	0.09	0.74	0.24	0.67	0.09	44	5
L	5	0	x	On	SMOTE	0.74	0.32	0.71	0.16	0.82	0.61	0.78	0.50	0.80	0.29	0.79	0.22	0.76	0.25	0.81	0.12	0.78	0.27	0.81	0.17	40	5
L	5	0	x	On	SMOTE	0.78	0.48	0.78	0.34	0.74	0.32	0.78	0.16	0.72	0.13	0.64	-0.03	0.80	0.44	0.74	0.33	0.78	0.35	0.74	0.23	48	5
L	5	0	x	On	SMOTE	0.72	0.36	0.68	0.19	0.74	0.43	0.70	0.27	0.74	0.24	0.75	0.09	0.72	0.36	0.72	0.19	0.76	0.40	0.72	0.25	46	5
L	5	0	x	On	SMOTE	0.76	0.33	0.69	0.19	0.80	0.44	0.76	0.33	0.78	0.35	0.76	0.23	0.80	0.38	0.76	0.28	0.82	0.47	0.76	0.37	44	5
Conf. Low (95%)						0.74	0.20	0.65	0.07	0.72	0.18	0.67	0.02	0.72	0.12	0.66	-0.02	0.74	0.19	0.69	0.05	0.74	0.18	0.69	0.05		
Conf. High (95%)						0.77	0.41	0.75	0.26	0.78	0.47	0.78	0.34	0.78	0.26	0.75	0.15	0.77	0.37	0.77	0.37	0.77	0.23	0.79	0.37	0.77	0.25

Table B.1: Confidence interval runs on best models on \tilde{D}_{PV} and \tilde{D}_{PL} .

Methods										XGB			LR			RF			Soft			Hard		
CSF	k Union	Mis.	Imp.	ComBat	Aug.	Acc	F1	AUC	MCC	Acc	F1	AUC	MCC	Acc	F1	AUC	MCC	Acc	F1	AUC	MCC	Tot. f.	Stbl. f.	
V	5	0	x	OH	SMOTE	0.75	0.40	0.75	0.24	0.77	0.30	0.76	0.19	0.75	0.40	0.74	0.24	0.73	0.38	0.77	0.21	0.77	47	6
V	5	0	x	OH	SMOTE	0.80	0.57	0.86	0.44	0.73	0.50	0.78	0.34	0.78	0.52	0.84	0.38	0.77	0.50	0.83	0.35	0.77	42	6
V	5	0	x	OH	SMOTE	0.83	0.64	0.88	0.54	0.82	0.62	0.89	0.51	0.83	0.62	0.86	0.51	0.83	0.67	0.89	0.57	0.83	41	6
V	5	0	x	OH	SMOTE	0.87	0.69	0.89	0.61	0.78	0.52	0.71	0.38	0.87	0.64	0.81	0.57	0.82	0.52	0.83	0.42	0.85	45	6
V	5	0	x	OH	SMOTE	0.73	0.43	0.79	0.26	0.75	0.48	0.79	0.32	0.80	0.57	0.83	0.44	0.83	0.64	0.81	0.54	0.78	45	6
V	5	0	x	OH	SMOTE	0.77	0.50	0.73	0.35	0.78	0.52	0.83	0.38	0.77	0.53	0.80	0.39	0.78	0.58	0.82	0.45	0.80	43	6
V	5	0	x	OH	SMOTE	0.77	0.53	0.81	0.39	0.78	0.52	0.85	0.38	0.78	0.52	0.82	0.38	0.80	0.60	0.84	0.48	0.80	40	6
V	5	0	x	OH	SMOTE	0.72	0.41	0.71	0.23	0.70	0.36	0.67	0.16	0.73	0.33	0.71	0.17	0.70	0.36	0.71	0.16	0.70	48	6
V	5	0	x	OH	SMOTE	0.82	0.56	0.83	0.45	0.78	0.52	0.82	0.38	0.77	0.42	0.82	0.27	0.83	0.62	0.82	0.51	0.80	45	6
V	5	0	x	OH	SMOTE	0.75	0.52	0.80	0.36	0.78	0.52	0.80	0.38	0.78	0.55	0.82	0.41	0.77	0.53	0.86	0.39	0.78	46	6
						0.75	0.46	0.76	0.30	0.74	0.42	0.75	0.27	0.76	0.44	0.77	0.29	0.75	0.47	0.78	0.31	0.76	0.47	0.78
Conf. Low (95%)						0.81	0.59	0.85	0.48	0.79	0.55	0.85	0.41	0.81	0.58	0.84	0.46	0.82	0.61	0.85	0.50	0.82	0.60	0.85
Conf. High (95%)																								
L	2	0	x	On	SMOTE	0.78	0.27	0.85	0.17	0.80	0.44	0.78	0.23	0.80	0.44	0.80	0.23	0.80	0.44	0.82	0.23	0.80	19	3
L	2	0	x	On	SMOTE	0.74	0.38	0.76	0.22	0.76	0.45	0.69	0.30	0.84	0.56	0.81	0.46	0.80	0.50	0.77	0.38	0.78	19	3
L	2	0	x	On	SMOTE	0.76	0.25	0.85	0.12	0.78	0.35	0.60	0.23	0.76	0.25	0.81	0.12	0.78	0.27	0.82	0.17	0.78	10	2
L	2	0	x	On	SMOTE	0.72	0.22	0.72	0.05	0.76	0.14	0.57	0.04	0.72	0.22	0.67	0.05	0.74	0.13	0.70	0.00	0.68	16	3
L	2	0	x	On	SMOTE	0.74	0.24	0.69	0.09	0.70	0.12	0.50	-0.06	0.76	0.14	0.73	0.04	0.72	0.22	0.68	0.05	0.76	18	2
L	2	0	x	On	SMOTE	0.86	0.70	0.81	0.62	0.84	0.64	0.78	0.54	0.84	0.56	0.81	0.46	0.86	0.67	0.84	0.58	0.84	12	3
L	2	0	x	On	SMOTE	0.68	0.20	0.56	0.00	0.76	0.25	0.68	0.12	0.74	0.32	0.60	0.16	0.70	0.21	0.68	0.03	0.70	19	2
L	2	0	x	On	SMOTE	0.76	0.40	0.86	0.25	0.80	0.50	0.78	0.38	0.82	0.53	0.86	0.42	0.82	0.53	0.85	0.42	0.82	16	3
L	2	0	x	On	SMOTE	0.80	0.50	0.76	0.38	0.76	0.45	0.73	0.30	0.80	0.44	0.78	0.33	0.76	0.45	0.76	0.30	0.78	18	3
L	2	0	x	On	SMOTE	0.78	0.35	0.73	0.23	0.72	0.30	0.62	0.13	0.80	0.38	0.67	0.28	0.74	0.24	0.68	0.09	0.80	15	2
						0.73	0.24	0.67	0.08	0.74	0.25	0.59	0.10	0.76	0.28	0.67	0.15	0.74	0.24	0.66	0.09	0.74	0.22	0.66
Conf. Low (95%)						0.80	0.46	0.80	0.34	0.80	0.48	0.73	0.26	0.82	0.49	0.80	0.37	0.81	0.48	0.80	0.37	0.81	0.48	0.80
Conf. High (95%)																								

Table B.2: Confidence interval runs on best models on \tilde{D}_{PeV} and \tilde{D}_{PeL} .

C

Appendix C

The following figures show proposed biomarkers based on the feature selection process.

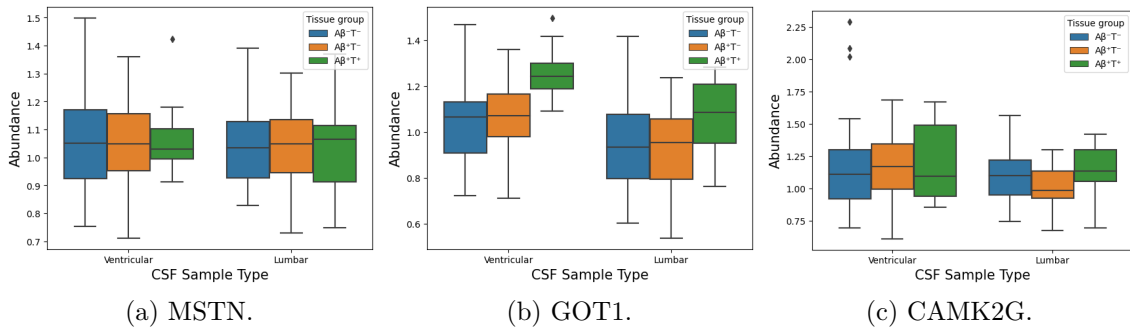


Figure C.1: Proposed biomarkers from \tilde{D}_{PV} .

Table C.1: Proposed protein biomarkers from ventricular CSF samples.

Accession	Description	Gene
O14793	Growth differentiation factor 8	MSTN
P17174	Aspartate aminotransferase, cytoplasmic	GOT1
Q13555	Calcium/calmodulin-dependent protein kinase type II	CAMK2G

C. Appendix C

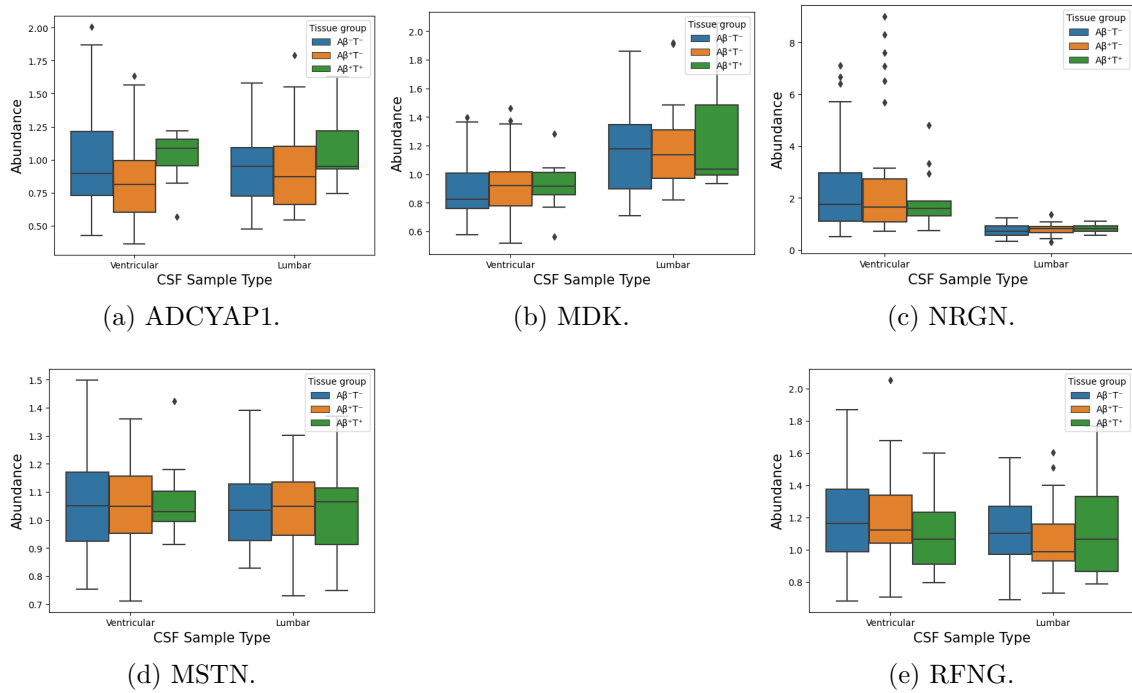


Figure C.2: Proposed biomarkers from \tilde{D}_{PL} .

Table C.2: Proposed protein biomarkers from lumbar CSF samples.

Accession	Description	Gene
P18509	Pituitary adenylate cyclase-activating polypeptide	ADCYAP1
P21741	Midkine	MDK
Q92686	Neurogranin	NRGN
O14793	Growth/differentiation factor 8	MSTN
Q9Y644	Beta-1,3-N-acetylglucosaminyltransferase radical fringe	RFNG

Table C.3: Biomarker comparison between tissue groups on \tilde{D}_{PL} not present in Section 4.3. The protein abundance of each of the tissue groups is expressed as mean \pm standard deviation. Statistical significance ($p < 0.05$) is highlighted with bold numbers. No statistical significance was found among the tissue groups.

Protein	Kruskal Wallis	p	$A\beta^-T^-$	$A\beta^+T^-$	$A\beta^+T^+$	Post-hoc
O14793	0.132	0.93624	1.05 \pm 0.13	1.03 \pm 0.15	1.05 \pm 0.19	-
Q9Y644	1.915	0.38381	1.11 \pm 0.22	1.05 \pm 0.20	1.13 \pm 0.30	-

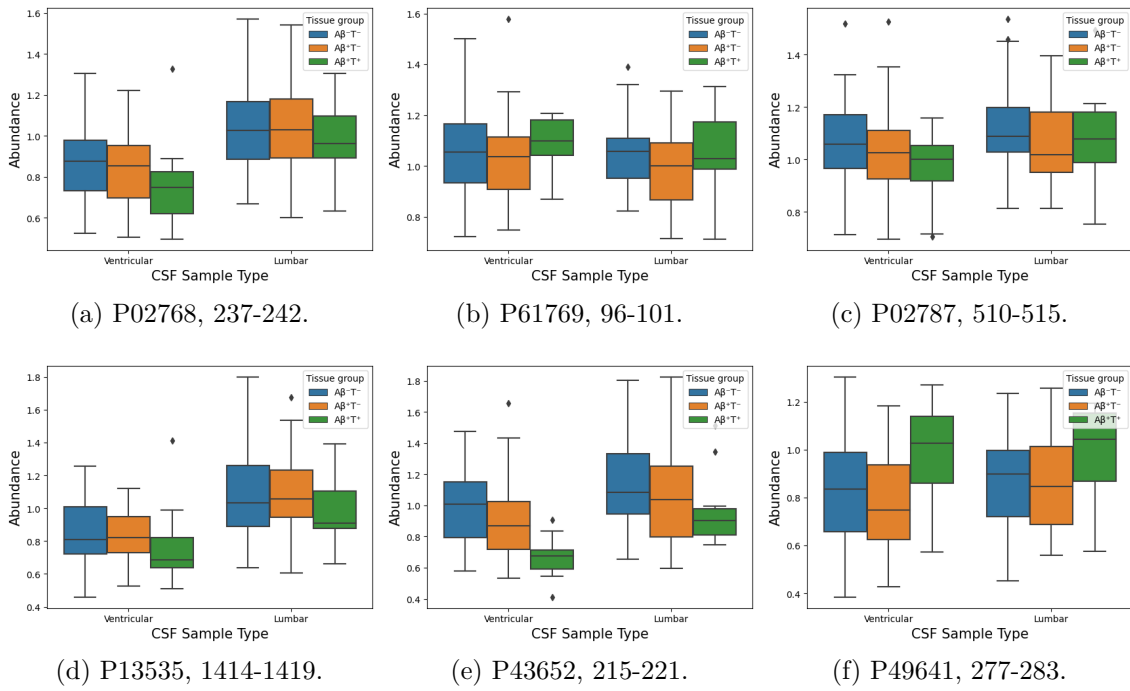
Figure C.3: Proposed biomarkers from \tilde{D}_{PeV} .

Table C.4: Proposed peptide biomarkers from ventricular CSF samples.

Sequence	Modifications	Protein and position
[K].AWAVAR.[LG]	1xTMTpro [N-Term]	P02768 [237-242]
[K].DEYACR.[V]	1xCarbamidomethyl [C5]; 1xTMTpro [N-Term]	P61769 [96-101]
[K].DSSLCK.[L]	1xCarbamidomethyl [C5]; 1xTMTpro [K6]; 1xTMTpro [N-Term]	P02787 [510-515]
[K].CASLEK.[T]	1xCarbamidomethyl [C1]; 1xTMTpro [K6]; 1xTMTpro [N-Term]	P13535 [1414-1419]
[K].AFSSYQK.[H]	1xTMTpro [K7]; 1xTMTpro [N-Term]	P43652 [215-221]
[R].NLGATPR.[S]	1xTMTpro [N-Term]	P49641 [277-283]

Table C.5: Biomarker comparison between tissue groups on \tilde{D}_{PeV} not present in Section 4.3. The peptide abundance of each tissue group is expressed as mean \pm standard deviation. Statistical significance ($p < 0.05$) is highlighted with bold numbers. No statistical significance was found among the tissue groups.

Peptide	Kruskal Wallis	p	$A\beta^-T^-$	$A\beta^+T^-$	$A\beta^+T^+$	Post-hoc
P02768..237.242.	5.432	0.066	0.88 ± 0.18	0.84 ± 0.17	0.76 ± 0.20	-
P61769..96.101.	2.19	0.334	1.05 ± 0.16	1.03 ± 0.16	1.09 ± 0.10	-
P02787..510.515.	5.204	0.074	1.08 ± 0.15	1.03 ± 0.17	0.96 ± 0.13	-

C. Appendix C

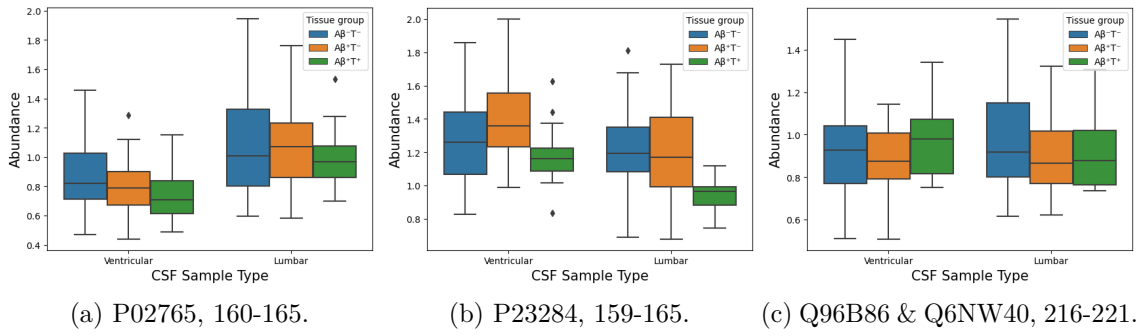


Figure C.4: Proposed biomarkers from \tilde{D}_{PeL} .

Table C.6: Proposed peptide biomarkers from lumbar CSF samples.

Sequence	Modifications	Protein and position
[R].VVHAAK.[A]	1xTMTpro [K6]; 1xTMTpro [N-Term]	P02765 [160-165]
[K].TAWLDGK.[H]	1xTMTpro [K7]; 1xTMTpro [N-Term]	P23284 [159-165]
[K].ITIIFK.[NA]	1xTMTpro [K6]; 1xTMTpro [N-Term]	Q96B86 & Q6NW40 [216-221]