Incorporating multi-omics for Alzheimer's Disease predictions

James Lee[1]

Supervisor(s): Marcel Reinders[1]**, Timo Verlaan**[1]**, Roy Lardenoije**[1]**, Gerard Bouland**[1]

[1]**EEMCS, Delft University of Technology, The Netherlands**

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 20, 2025

Name of the student: James Lee
Final project course: CSE3000 Research Project
Thesis committee: Marcel Reinders, Timo Verlaan, Roy Lardenoije, Gerard Bouland, Ricardo Marroquim

Abstract

Background Currently, the relation between Amyloid-β (Aβ) and tau have been associated with Alzheimer's Disease (AD). However, the exact mechanism between their interactions have not been fully understood. With the fast-developing advances in the field of multi-omics technologies, raises the question whether using multiple biological measurements could help understand the intricate mechanism or help with the classification of AD. In this study, we looked at whether incorporating different omics, or a combination of multiple omics, will give a better prediction of AD.

Approach For our analysis we looked at the single-omics from the Religious Orders Study and Memory and Aging Project (ROSMAP), where we analyzed the Proteomics (LC-SRM), Metabolomics (Metabolon HD4), Epigenetics (ChIP Seq) and Gene Expression (RNA array). In our analysis, we used the Random Forest, Block Forest, *k*-Nearest Neighbor (*k*-NN) and Support Vector Machines (SVMs) models to determine the accuracies of classifying the cognitive diagnosis (cogdx).

Results The results suggest that there is no improvement in the classifications accuracies when combining the single-omics.

Conclusion It was concluded that the use of multi-omics did not improve the predictions of AD compared to using single-omics.

1 Introduction

Alzheimer's Disease (AD) is a complex age-related neurogenerative brain disease, that slowly destroys the memory and thinking skills of the individual. AD is currently the leading cause of dementia, with no cure available [14]. Currently, the leading hypothesis for AD is the "Amyloid Cascade Hypothesis", which states that increase of the protein Amyloid-β (Aβ) activates the tau pathology [10].

This relation between Aβ and tau is still not fully understood, existing work has used single-omics (The scientific field associated with measuring biological molecules in a high throughput way) [13] to see whether the correlation between Aβ and tau are prevelant in the data. A study by Kitani et al. [11] used Proteomics data from the Religious Orders Study and Memory and Aging Project (ROSMAP) [2] with integrative network analysis, concluded that they were able to identify potential Aβ interactions in the Proteomics data. Since the researchers focused solely on a single-omics approach, this raises the question whether incorporating different omics or a combination of multiple single-omics could provide deeper insights into the mechanisms of AD.

Development in multi-omics methods, raises the question whether combining different omics, such as proteomics and metabolomics, could provide better classifications of features or a deeper insight into which features are more prominent, than using single-omic data [17]. The use of multi-omics data has been seen in analysis of other diseases such as classifying tumors using a uncertainty-aware dynamic integration framework, which saw an accuracy of 98% for classifying turmors using the integration of DNA methylation, gene expression and miRNA expression [7] and for the large-scale benchmark study of survival prediction of cancer data [8].

There has been a study that did a comprehensive multi-omics analysis to predict AD [16]. The study looked at the results of using sparse generalized canonical correlation analysis (sGCCA) for predicting whether an individual had AD. The researchers looked at four datasets from the ROSMAP. Using the single-omics dataset Proteomics (LC-SRM), Epigenetics (DNA methylation), Genomic Variants (SNP Array) and Gene Expression (RNAseq from bulk brain), the study showed that incorporating the four datasets resulted in the accuracy of classifying clinical diagnosis at the time of death improved from about 0.60 for the single-omic predictions to 0.95 after integrating the four datasets. However, they only looked at the use of sparse generalized canonical correlation analysis, which uses a multivariate dimension reduction technique, and did not investigate the effect on other classifiers. Moreover, they only looked at the performance when all four single-omics were combined and have not considered any partial combination of the omics, where two or three omics were combined.

Here, we will look at whether incorporating different omics, or a combination of multiple omics, will be better for the prediction of AD. By comparing the results in [16] using different models for the evaluation of single-omic and multi-omics data and whether using a partial combination of single-omics would provide better predictions.

2 Methodology

2.1 Datasets

The dataset we looked at, are from the Religious Orders Study and Memory and Aging Project (ROSMAP) [2]. The project looked at Catholic nuns, priests and brothers, from more than 40 groups across the United States and that agreed upon annual clinical evaluation and brain donation.

The datasets, came from Synapse[1] and we looked at the following single-omics: Proteomics (LC-SRM), Metabolomics (Metabolon HD4), Epigenetics (ChIP Seq) and Gene Expression (RNA array).

Proteomics is the set of proteins expressed by a cell, tissue or organism and the data has been collected by the use of Liquid Chromatography-Selected Reaction Monitoring (LC-SRM) [1]. Metabolomics is the set of molecules found as a result of metabolism, the data has been gathered by the use of the Metabolon LIMS system [13]. Epigenetics looks at how cells control gene activity without changing the DNA sequence with the data being gathered using Chromatin Immunoprecipitation Sequencing (ChIP-Seq) [12]. Lastly, Gene Expression measures the levels of mRNA, which have been gathered with an integrative network-based approach [19].

For our analysis we will be taking the "Final consensus cognitive diagnosis (cogdx)" as the targets of the predictors. The cogdx is the overall cognitive diagnosis, neurologists

[1] Synapse: https://www.synapse.org/Synapse:syn3219045

Target Values	Coding
CT	NCI: No cognitive impairment
CT	MCI: Mild cognitive impairment and No other cause of CI
CT	MCI: Mild cognitive impairment AND another cause of CI
AD	AD: Alzheimer's dementia and No other cause of CI
AD	AD: Alzheimer's dementia AND another cause of CI
-	Other dementia: Other primary cause of dementia

Table 1: The target values of the final consensus cognitive diagnosis (cogdx) values. Where the *control* (CT) values contains the individuals with Mild Cognitive Impairment (MCI), with or without other causes of Cognitive Impairment (CI), and individuals with No Cognitive Impairment (NCI). The *Alzheimer's Disease* (AD) values containing individuals with Alzheimer's Dementia (AD) with or without other causes of Cognitive Impairment (CI). Individuals with other type of dementia have been left out in the new assignment.

Single-Omic	Number of Samples	Number of Features	Number of Missing Values	Number of Features Containing Missing Values
Proteomics	147 (1209)	184 (184)	3 (375)	1 (48)
Epigenetics	147 (668)	26386 (26386)	0 (0)	0 (0)
Gene Expression	147 (668)	48805 (48805)	0 (0)	0 (0)
Metabolomics	147 (512)	1057 (1057)	49441 (171666)	576 (629)

Table 2: The number of samples and features for the different single-omics, including the number of missing values and the number of features that contain missing values, filtered on the samples that overlap in each single-omic. The numbers in bracket show the original values, before filtering.

with an expertise in dementia gave after reviewing all the available clinical data from a patient, after death.

The cogdx can contain six different values: No Cognitive Impairment (NCI); Mild Cognitive Impairment (MCI) and no other cause of Cognitive Impairment (CI); MCI and another cause of CI; Alzheimer's Dementia (AD) and no other cause of CI; and AD and another cause of CI; and other primary cause of dementia. However, for the analysis, we have combined the six different values to have the target be a binary classifier. We have done this by combining the values NCI and MCI into a single *control* (CT) target and combined the two Alzheimer's dementia labels into a single *Alzheimer's Disease* (AD) label. We dropped the individuals, which have another primary cause of dementia. The resulting target values with their original interpretation are shown in Table 1.

For the creation of the multi-omics dataset, we first combined the four different single-omics features into a single dataset keeping the samples that overlapped. The number of overlapping samples was 147 samples with the distribution of the target values being 102 for CT and 45 for AD based on Table 1. For the single-omics, we filtered them to only contain the 147 samples that are also present in the multi-omics dataset. The number of samples and features for each single-omic, with the number of missing values and the number of features containing missing values after filtering can be seen in Table 2. The number in brackets indicate the original numbers before filtering. The partial combinations have been created by combining the single-omics features and filtering them to contain the same 147 as in the multi-omics dataset.

2.2 Models

For our analysis, we will be taking 4 different models into consideration.

Firstly, we will be considering a k-Nearest Neighbors algorithm (k-NN), since it is a simple naive approach to see whether samples are similar to each other [5]. Secondly, we will also be considering, Support Vector Machines (SVMs) for our analysis, since they are useful for high dimensional datasets [5].

Moreover, we will be using the method of [8], that did a large-scale benchmark study of survival times using a multi-omics dataset of cancer datasets. It was concluded that the Cox model and Block Forest algorithm both show competitive performance. For our analysis we will use the Block Forest algorithm, since the Cox model is more specifically tailored for survival times and the Block Forest gives a more general approach. Additionally, Block Forest showed great results in the study that compared five different random forest variants developed for multi-omics covariate data analysis [9].

The Block Forest algorithm is an variant of the Random Forests algorithm [3]. The Block Forest algorithm randomly chooses which block of variables it should consider at each split. Afterwards, the algorithm uses the variables from the selected block to decide how to do the splitting. The Block Forest algorithm has been implemented as an extension from the Ranger implementation in C++ and R [18]. We will also be considering the (standard) Random Forests, to see whether the variation has a significant impact on the results.

2.3 Implementation Details

In this research we used the Delft AI Cluster (DAIC) from the TU Delft [6]. To use the Block Forest model, we used the R implementation with the rpy2[2] interface. This package makes it possible for R code to be used within Python. This decision has been made to ensure that we can run the model on the DAIC cluster, the parameters have been kept to their standard values for the analysis. For the Random Forests, *k*-Nearest Neighbors and Support Vector Machines. We used the implementations provided by the Scikit-learn[3] Python library. Specifically, the methods RandomForestClassifier, KNeighborsClassifier and SVC have been used, with their parameters kept at their default values.

Train/Test Split: The training and tests sets were created using a stratified 10-fold cross-validation, to ensure each fold preserves the percentage of samples for each class in the total dataset. Within each fold additional processing were done before running them on the models, this is done to avoid leaking information from the train data to the test data. Firstly, the training data has been normalized using the mean normalization in Formula 1, with x being the current cell value, μ being the mean of the samples in the feature and σ the standard deviation. An extra term ϵ has been added in the case that the standard deviation σ would be evaluated to zero. The ϵ was set to 10^{-100}. In addition, the test data has also been normalized by using the same mean μ and standard deviation σ from the training data.

$$x' = \frac{x - \mu}{\sigma + \epsilon} \tag{1}$$

Moreover, for each fold, the missing values have been filled by taking the mean values from every feature in the training data. The same mean values have also been used to fill in the test data. If the feature in the train data only contained missing values the feature was dropped in both the train and test data. This step only effected the Proteomics and Metabolomics datasets since the Epigenetics and Gene Expression datasets had no missing values.

Due to the imbalance in the target counts, with 102 samples being CT and 45 samples being AD, we also used the oversampling method Synthethic Minority Over-sampling Technique (SMOTE) to make the target counts in the train and test data in each fold more equal [4]. SMOTE creates synthetic samples of the minority class to balance the dataset.

Additionally, we used a feature selection technique for each fold to reduce the number of features for the models by calculating the ANOVA F-value of the train data and kept the top number of features equal to the number of rows in the train data. The ANOVA F-value looks at the variance between the features and chooses the features that are significantly different. The same features selected in the train data were also selected for the test data.

Evaluation Measures: Between different models in each fold, we trained the models on the same train data and determined the accuracy on the test data. Moreover, a confusion matrix has also been made by combining the individual confusion matrices from every fold on the test data. For the partial combinations of single-omics, we determined their mean accuracy on the test data across the 10-fold cross-validation.

We also kept track of the top 100 features selected in each fold by the ANOVA F-value and Random Forest, and determined which features occured the most in every fold. For the Epigenetics we used the CruzDB [4] package to determine the closest gene based on the start, end and chromosome information within the Epigenetics feature. CruzDB is an interface for the University of California, Santa Cruz (UCSC) genome browser, to retrieve information about genomes [15].

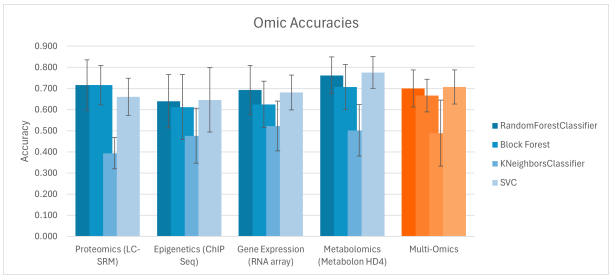


Figure 1: Barplot displaying the mean accuracies of classifying the cognitive diagnosis using different models, across the 10-fold stratified cross-validation for the different single-omics and the multi-omics dataset, where all 4 single-omics were combined. The error bar show the standard deviation present in the accuracy scores.

3 Results

Method overview

In our analysis, we used the Random Forest, Block Forest, *k*-Nearest Neighbors and SVM classifiers to classify the cogdx values of AD patients from the ROSMAP dataset. After determining the number of overlapping samples for the multi-omics dataset, the single-omics were also filtered to also only contain the same overlapping samples.

Comparison of general accuracies show no improvement for multi-omics approach

Figure 1 displays the mean accuracy over the 10-fold cross-validation for the single-omics and multi-omics dataset. Additionally, the error bars show the standard deviation in the accuracies. The figure shows no improvement for the classification accuracy of cogdx for the multi-omics compared to the single-omics.

Comparison of Wilcoxon tests for *k*-Nearest Neighbors shows worse performance compared to the other models

Figure 1 shows that the *k*-Nearest Neighbors has the worst mean accuracy compared to the other models. However, to determine whether the difference is significant, we used the Wilcoxon test, since the samples are not assumed to be normally distributed. The null hypothesis (H_0) states that the two populations should be equal, while the alternative hypothesis

[2]RPY2: https://rpy2.github.io/

[3]Scikit-learn: https://scikit-learn.org/stable/

[4]CruzDB: https://github.com/brentp/cruzdb

(H_a) states that the two populations are not equal. The p-values from the Wilcoxon test were all significant ($p < 0.05$) for all of the omics and models compared to the k-Nearest Neighbors. Which means the *k*-Nearest Neigbhbor does not perform the same as the other models.

Wilcoxon test for Random Forest and Block Forest show the models perform the same

Comparing the Random Forest with the Block Forest model using the Wilcoxon test. The resulting p-values show that they are all not statistically significant ($p > 0.05$), with the p-values being [0.9375, 0.2382, 0.0937, 0.1250, 0.6640] respectively. Which means the model perform the same.

Confusion matrices show that *k*-Nearest Neighbors often miss-classifies the control label

Table 3 shows the confusion matrices of the different omics and models, where the matrices have been created by determining the individual confusion matrices on the test set in every fold of the 10-fold cross-validation and combining them, by adding the individual matrices together at the end. The table shows that for the Random Forest, Block Forest and SVM model, all classify the true CT label the most. The *k*-Nearest Neighbors does classify the AD labels the most, however it does misclassify the CT labels a lot.

UpSet plots for partial combination of single-omics classification of cogdx show uniform accuracies for the different models

Figure 2 shows the UpSet plots for the Random Forest, *k*-Nearest Neighbors and SVM. The UpSet plots show the mean accuracies for classifying the cogdx of the 10-fold cross-validation. The plots also indicate which single-omics were combined for the predictions. The figure shows that the accuracies are all quite uniform.

Feature importance of Random Forest and ANOVA F-Value show overlap in features

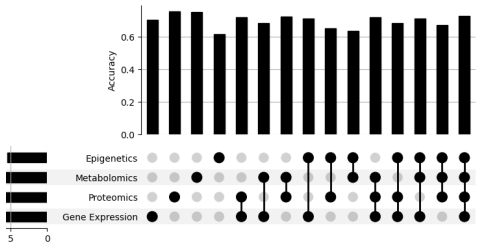
Table 4 show the top 10 most occurring features for the different omics from the Random Forest and ANOVA F-Value based on their feature importance. For the different omics we can see that there are some feature overlapping with what the Random Forest determined to be important with their corresponding ANOVA F-Value. As an example, for Proteomics we can see that both MLF2_2, bA and STX1A show importance for both the Random Forest and the ANOVA F-Value.

Feature importance of single-omics show relations with Alzheimer's Disease

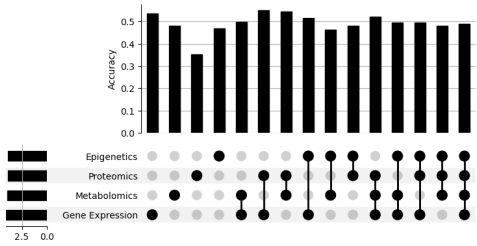
Taking the features from Table 4 in the context of AD, shows that some of the features do have a correlation with AD. As an example, for Proteomics we can see that bA (Amyloid-β) and tau_12E8_s262 have a correlation with AD, in the context of the "Amyloid Cascade Hypothesis".

Multi-omics feature importance show a strong contribution from Epigenetics features

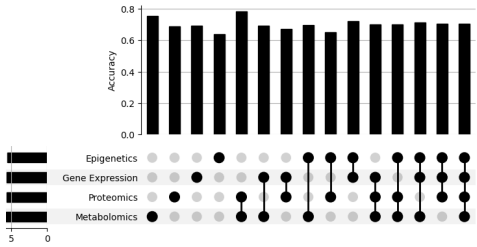
The feature importance for multi-omics in Table 4, also shows a strong preference for features from the Epigenetics dataset.



(a) UpSet plot displaying the mean accuracies for the Random Forest model on partial-omics



(b) UpSet plot displaying the mean accuracies for the *k*-Nearest Neighbor model on partial-omics



(c) UpSet plot displaying the mean accuracies for the Support Vector Machine (SVM) model on partial-omics

Figure 2: UpSet plots for the different models, with the bar displaying the mean accuracies of classifying the cognitive diagnosis across the 10-fold stratified cross-validation, for the different combinations of single-omics (partial-omics).

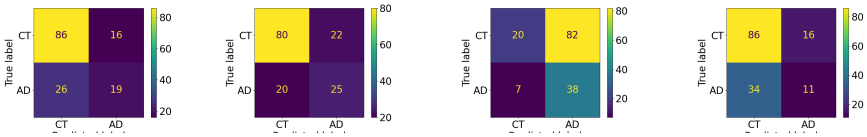
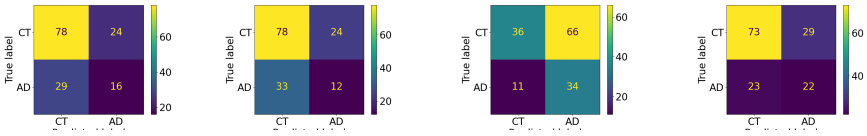
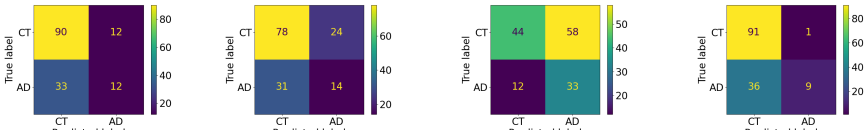
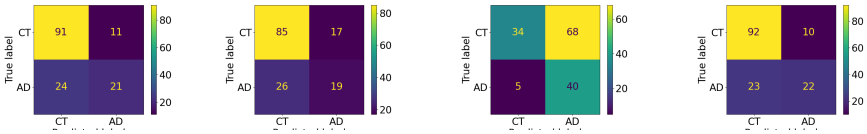
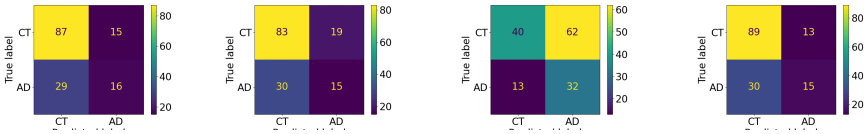
	Random Forest	Block Forest	k-Nearest Neighbors	SVM
Proteomics				
Epigenetics				
Gene Expression				
Metabolomics				
Multi-Omics				

Table 3: Confusion matrices showing the true labels against the predicted labels, for the cognitive diagnosis *control* (CT) and *Alzheimer's Disease* (AD) values, for the different models and omics across the 10-fold stratified cross-validation. The vertical axis of the confusion matrices contain the true labels and the horizontal axis contain the predicted labels.

Protemics		Gene Expression		Epigenetics	
Random Forest	ANOVA F-Value	Random Forest	ANOVA F-Value	Random Forest	ANOVA F-Value
MLF2_2	MLF2_2	HS.246177	WDR64	MYO1D	SH3PXD2B
bA	bA	MCF2L2.1	HS.246177	MPP2	MPP2
STX7	VGF	WDR64	MCF2L2.1	SGF29	MYO1D
STX5	STX1A	GSDMC	LOC148915	SH3PXD2B	RBFOX3
STX1A	tau_12E8_s262	LOC654085	BMI1	S100A6	MIR29A
SNAP25_3	AMPD2_2	HS.542293	KIAA0090	RBFOX3	VAPA
VGF	LDHB_1	LOC644783	C1ORF2	MARF1	MIR1302-7
SNAP25_7	PLXNB1	HS.552393	FRAS1	SRSF6	LOC102724163
AMPD2_2	IGFBP5	LOC148915	HS.529514	LOC102724163	DCAF8
LDHB_1	STX7	LOC339240	FAM83H	TTC34	POLR1A

Metabolomics		Multi-Omics	
Random Forest	ANOVA F-Value	Random Forest	ANOVA F-Value
glycerate	glycerophospho-ethanolamine	HS.246177	RBFOX3
glycerophospho-ethanolamine	glycerophosphoryl-choline (GPC)	MCF2L2.1	WDR64
O-sulfo-L-tyrosine	glycerate	RBFOX3	SH3PXD2B
1,2-dipalmitoyl-GPG (16:0/16:0)	glutamate	KIF5A.1	MARF1
tryptophan betaine	1-stearoyl-2-oleoyl-GPE (18:0/18:1)	WDR64	glycerophosphoryl-choline (GPC)
glycerophosphoryl-choline (GPC)	12-HHTrE	POLR2A	glycerophospho-ethanolamine
homocarnosine	12-HETE	SH3PXD2B	VAPA
12-HHTrE	homoarginine	GSDMC	LARP4
putrescine	X - 23739	TMEM2	MCF2L2.1
1-stearoyl-2-oleoyl-GPE (18:0/18:1)	2-aminoadipate	glycerophospho-ethanolamine	BMI1

Table 4: Tables showing the top 10 features for the different single-omics and multi-omics dataset. The features have been determined by noting the top 100 features for the Random Forest and ANOVA F-Value across each of the 10-fold cross-validation and seeing which features appeared the most across the 10-folds. The Epigenetics features show the closest gene from their peaks based on the University of California, Santa Cruz (UCSC) genome database.

Since the RBFOX3, SH3PXD2B, RBFOX3, MARF1, VAPA and LARP4 all originate from the Epigenetics dataset.

4 Discussion

In this paper, we looked at whether incorporating multiple single-omics could improve the the accuracy of predicting Alzheimer's Disease (AD) compared to only using single-omics. For our analysis, we have used the datasets from the Religious Orders Study and Memory and Aging Project (ROSMAP), were we mostly looked at the omics: Proteomics (LC-SRM), Metabolomics (Metabolon HD4), Epigenetics (ChIP Seq) and Gene Expression (RNA array).

For the analysis, we looked at the accuracies and confusion matrices for the Random Forest, Block Forest, *k*-Nearest Neighbor (*k*-NN) and Support Vector Machines (SVMs). Additionally, we looked at the most occurring important features and looked at the upset plots of partial-omics, where just 2 or 3 single-omics were combined.

In our analysis, we compared the general accuracies of the single-omics against the multi-omics dataset, and saw that there were no improvement in the multi-omics approach. Comparing our results, with the results from Vacher et al. [16]. They showed an improvement in the accuracy for the use of multi-omics, while we saw no improvement in our analysis.

This different could be explained by the difference in classification we made for the "cogdx" labels in Table 1. Where they combined the Alzheimer's Dementia (AD) and Mild Cognitive Impairment (MCI) cases together and left the No Cognitive Impairment (NCI) cases separate, while we combined the MCI and NCI labels, and kept the AD samples separate. Moreover, the difference in accuracies could be explained by the different single-omics we used in our analysis. While we both use Proteomics (LC-SRM), the other single-omics were different since they used Epigenetics (DNA methylation array), Genomic Variants (SNP Array) and Gene Expression (RNAseq from bulk brain) to maximize the number of overlapping samples, to 455 samples. The researchers also used different pre-processing techniques for each single-omic, while we used a more general pipeline for processing and for reducing our feature space. Furthermore, the difference could be explained by the different models we used since Vacher et al. [16] used space generalized canonical correlation analysis (sGCCA), while we used more standard models Random Forest, *k*-Nearest Neighbor (*k*-NN) and Support Vector Machines (SVMs).

For our analysis, we also looked at the confusion matrices for the different omics and saw that *k*-Nearest Neighbors misclassified the *control* (CT) label a lot, compared to the other models. This observation could be the results of the oversampling procedure using Synthetic Minority Over-Sampling Technique (SMOTE), since SMOTE creates synthetic samples from the features space between the the minority classes, which can result in the AD groupings created by *k*-Nearest Neighbors to overlap with the CT samples.

We have also looked at the mean accuracies of partial combinations of single-omics, where two or three omics were combined. The plots show that the accuracies were all quite uniform, which could indicate that there is no additional predictive ability to combine multiple single-omics. However, this could also be the result of the feature selection process to make the number of features equal to the number of rows, which could have greatly reduced the feature space and predictive ability of the models.

Furthermore, we determined the feature importance for the Random Forest model and ANOVA F-Value used for feature selection and saw that there are features that overlapped that were important for both metrics. However, this could the consequence of the implementation, since we first used the ANOVA F-Value for feature selection and then used the selected features for the Random Forest, so this could have had an effect on the selected features.

Lastly, the feature importance also showed that features with a relation with AD were also seemed as important predictive features for the Random Forest and ANOVA F-Value.

However, there is room for further exploration whether there is a correlation of the features importance between the different single-omics. As an example, whether the top proteins found for Proteomics have a relation with the top genes found in Gene Expression.

Moreover, our work did not consider all the different single-omics available from the ROSMAP dataset. This could also give room for exploration how the different single-omics are correlated with each other.

5 Responsible Research

Ethical considerations: Firstly for our research, we have used machine learning models to evaluate the accuracy of predicting Alzheimer's Disease (AD). Machine learning models are known to be quite resource intensive so we were given access to the Delft AI Cluster (DAIC) [6] to gain extra computation power to evaluate our models. However, as Aritificial Intelligence (AI) models and machine learning models become more prevelant, the demand to use these type of clusters and the need for computational power, requires more and more energy to supply these machines.

Data Sensitivity: Secondly, the data we used come from the Religious Orders Study and Memory and Aging Project (ROSMAP) [2]. These kind of datasets contain sensitive personal medical information, which do have to be used with permission and care. The participants of the study did give permission for the use of the data. So the biggest discussion is about the privacy and security of the data, since we are working with medical data. The platform Synapse (https://www.synapse.org/), were we gathered the data had us, first requests access to the datasets and explain that we wanted to use it for scientific research purposes. However, as the bioinformatics field grows, there would be more demand for data and raises the question how we should ethically handle, safely store and use these sensitive information.

Reproducibility: Additionally, it is important that the work we do and present are reproducible. Not only is it an integrable part of scientific research, it also gives the opportunity to verify results and avoid any false conclusions made. Moreover, it gives other's the opportunity to build on existing work and gives more transparancy in the work and results

we achieve. For our analysis, since the data and the models we used in the study are publically available it should give an individual the possibility of reproducing the study with the same models.

Biases: Lastly, in our work we used relatively simple models, where there results can be explanatory. However, a lot of models, and more complex models can be seen as black-boxes, which mean the reasoning behind how the models come to their results are unexplanatory. This could raise the issue of the models becoming overtrusted than the physician or clinician, which can lead to issues of medical malpractice.

To summarise, the use of machine learning and other AI models can help the progress in developed to solve medical problems. However, these development does require us to reflect on the use and safety of the data collected and the interpretability of the results, the models give us.

References

[1] Victor P. Andreev, Vladislav A. Petyuk, Heather M. Brewer, Yuliya V. Karpievitch, Fang Xie, Jennifer Clarke, David Camp, Richard D. Smith, Andrew P. Lieberman, Roger L. Albin, Zafar Nawaz, Jimmy El Hokayem, and Amanda J. Myers. Label-free quantitative lc-ms proteomics of alzheimer's disease and normally aged human brains. *Journal of Proteome Research*, 11(6):3053–3067, may 2012.

[2] David A. Bennett. Overview and findings from the religious orders study. *Current Alzheimer Research*, 9(6):628–645, 2012.

[3] L Breiman. Random forests. *Machine Learning*, 45:5–32, 10 2001.

[4] Nitesh Chawla, Kevin Bowyer, Lawrence Hall, and W. Kegelmeyer. Smote: Synthetic minority oversampling technique. *J. Artif. Intell. Res. (JAIR)*, 16:321–357, 06 2002.

[5] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, sep. 1995.

[6] Delft AI Cluster (DAIC). The Delft AI Cluster (DAIC), RRID:SCR_025091, 2024.

[7] Ling Du, Chaoyi Liu, Ran Wei, and Jinmiao Chen. Uncertainty-aware dynamic integration for multi-omics classification of tumors. *Journal of Cancer Research and Clinical Oncology*, 149, 08 2022.

[8] Moritz Herrmann, Philipp Probst, Roman Hornung, Vindi Jurinovic, and Anne-Laure Boulesteix. Large-scale benchmark study of survival prediction methods using multi-omics data. *Briefings in Bioinformatics*, 22(3):bbaa167, 08 2020.

[9] Roman Hornung and Marvin N. Wright. Block forests: random forests for blocks of clinical and omics covariate data. *BMC Bioinformatics*, 20(1), jun. 2019.

[10] Kasper P Kepp, Nikolaos K Robakis, Poul F Høilund-Carlsen, Stefano L Sensi, and Bryce Vissel. The amyloid cascade hypothesis: an updated critical review. *Brain*, 146(10):3969–3990, 05 2023.

[11] Akihiro Kitani, Yusuke Matsui, and Alzheimer's Disease Neuroimaging Initiative. Integrative network analysis reveals novel moderators of aβ-tau interaction in alzheimer's disease. *Alzheimer's Research & Therapy*, 17(1):70, 2025.

[12] Hans-Ulrich Klein, Cristin McCabe, Elizabeta Gjoneska, Sarah E. Sullivan, Belinda J. Kaskow, Anna Tang, Robert V. Smith, Jishu Xu, Andreas R. Pfenning, Bradley E. Bernstein, Alexander Meissner, Julie A. Schneider, Sara Mostafavi, Li-Huei Tsai, Tracy L. Young-Pearse, David A. Bennett, and Philip L. De Jager. Epigenome-wide study uncovers large-scale changes in histone acetylation driven by tau pathology in aging and alzheimer's human brains. *Nature Neuroscience*, 22(1):37–46, dec 2018.

[13] Christine M. Micheel, Sharly J. Nass, and Gilbert S. Omenn, editors. *Evolution of Translational Omics: Lessons Learned and the Path Forward*. National Academies Press (US), Washington (DC), March 2012.

[14] National Insitute on Aging (NIH). Alzheimer's disease fact sheet. https://www.nia.nih.gov/health/alzheimers-and-dementia/alzheimers-disease-fact-sheet, 2023.

[15] Brent S. Pedersen, Ivana V. Yang, and Subhajyoti De. Cruzdb: software for annotation of genomic intervals with ucsc genome-browser database. *Bioinformatics*, 29(23):3003–3006, 09 2013.

[16] Michael Vacher, Rodrigo Canovas, Simon M. Laws, and James D. Doecke. A comprehensive multi-omics analysis reveals unique signatures to predict alzheimer's disease. *Frontiers in Bioinformatics*, 4, jun. 2024.

[17] Katy Vandereyken, Alejandro Sifrim, Bernard Thienpont, and Thierry Voet. Methods and applications for single-cell and spatial multi-omics. *Nature Reviews Genetics*, 24(8):494–515, mar. 2023.

[18] Marvin Wright and Andreas Ziegler. ranger: A fast implementation of random forests for high dimensional data in c++ and r. *Journal of Statistical Software*, 77, 08 2015.

[19] Bin Zhang, Chris Gaiteri, Liviu-Gabriel Bodea, Zhi Wang, Joshua McElwee, Alexei A. Podtelezhnikov, Chunsheng Zhang, Tao Xie, Linh Tran, Radu Dobrin, Eugene Fluder, Bruce Clurman, Stacey Melquist, Manikandan Narayanan, Christine Suver, Hardik Shah, Milind Mahajan, Tammy Gillis, Jayalakshmi Mysore, Marcy E. MacDonald, John R. Lamb, David A. Bennett, Cliona Molony, David J. Stone, Vilmundur Gudnason, Amanda J. Myers, Eric E. Schadt, Harald Neumann, Jun Zhu, and Valur Emilsson. Integrated systems approach identifies genetic nodes and networks in late-onset alzheimer's disease. *Cell*, 153(3):707–720, 2013.