A STUDY OF EFFECTIVENESS OF SIMULATED DATA IN CLASSIFYING ALZHEIMER'S DISEASE STATUS USING MRS PARAMETERS

A Thesis

presented to

the Faculty of the Graduate School

at the University of Missouri-Columbia

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

---

by

MARY YORK

Dr. Melissa Terpstra, Thesis Supervisor

MAY 2023

The undersigned, appointed by the dean of the Graduate School, have examined the thesis entitled

EFFECTIVENESS OF SIMULATED DATA IN CLASSIFYING ALZHEIMER'S DISEASE STATUS USING MRS PARAMETERS

presented by Mary York,

a candidate for the degree of Master of Science in Data Science and Analytics,

and hereby certify that, in their opinion, it is worthy of acceptance.

Professor Melissa Terpstra

Professor Lyndon Coghill

Professor Ilker Ersoy

DEDICATION

I owe many thanks to my loved ones. To my parents, thank you for always encouraging me to "try my hardest to be the smartest". Though unachievable, I'll never stop trying. To my sister Alicia, thank you for being the best academic role model I could ask for, your hard work has given me something to look up to and I don't think I'd be here without you. To my sister Erin, you are an integral part to my support system and I appreciate your rationale, encouragement, and reminders to take breaks throughout working on this project. To Grandma Brenda and Grandma Sandie, I get all my "smarts" from you, so I hope you both take as much credit for this work as I do. To Dana, I'm endlessly thankful for your undying support and encouragement, you have truly been a rock to lean on throughout this process. To Josh, you are and will always be my favorite brother in law. To Hallie, Abby, and Noah, thank you for always lending an open ear and extending an open arm. To Dr. Fernando Torralbo, you not only taught me how to properly balance academic work, but how to be kind and encouraging to colleagues along the way. I thank you, and hope to never lose sight of the lessons you've unknowingly taught me. Lastly, to my dog Tucker, if I manage to pass this defense I will buy you a toy to make up for all the fetch sessions that were cut short. I'll always strive to make you all proud.

ACKNOWLEDGEMENT

I would like to formally thank my committee members, including Dr. Melissa Terpstra, who has graciously allowed me to assist her in research.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

EFFECTIVENESS OF SIMULATED DATA IN CLASSIFYING ALZHEIMER'S DISEASE STATUS USING MRS PARAMETERS

Mary York

Dr. Melissa Terpstra, Thesis Supervisor

ABSTRACT

In vivo magnetic resonance spectroscopy, or MRS, has the potential to identify meaningful differences in the neurochemical profiles of patients with Alzheimer's disease (AD) relative to healthy controls, especially at ultra-high field (7 Tesla). Classification algorithms applied to such data could, theoretically, aid in disease diagnosis or act as an indicator of the effectiveness of a treatment. A common limitation when applying classification algorithms to such data is sample size, which arises from difficulty in recruitment of individuals with AD. Classification algorithms applied to small datasets may benefit from additional training data simulated from Bayesian networks that are learned via hill-climbing. Ultra-high field single voxel MRS and MRI data from Marjanska et al (2019) were used to explore the effect that including simulated data in the training of classification algorithms has on the ability to correctly classify AD status [2]. Three hill climbing methods, hill climbing scored via the Bayesian information criterion (BIC), hill climbing scored via the Akaike information criterion (AIC), and max-min hill climbing scored via BIC, were tested using the original data and data in which each Alzheimer's observation is doubled in search for the network that could produce the highest posterior predictive correlation across ascorbate, percent grey matter, and signal to noise ratio. The effect of including data simulated in the resulting manner was then characterized across three classification algorithms (Extreme Gradient Boosting, random forest, and support vector machine) and three data sets (original AD and control data from Marjanska et al (2019), Marjanska et al (2019) data reflected as principal components, and Marjanska et al (2019) data with additional controls from The Lifespan Human Connectome Project in Aging used in the training and simulation process) [3]. Variables used to predict AD status were mostly from MRS derived data, consisting of 14 water-referenced neurochemical concentrations, signal to noise ratio, and linewidth, with few

variables from MRI derived data, consisting of percent gray matter, percent white matter, percent cerebro-spinal fluid relative to the volume of interest. The strongest findings regarding structure learning arose from hill climbing scored via BIC using data in which every AD observation was over-sampled. The inclusion of data simulated from Bayesian networks whose structures were derived via this method did not widely lead to higher average sensitivity, specificity, or overall accuracy.

Chapter 1

INTRODUCTION

Alzheimer's disease is a neurological condition defined by the presence of amyloid B and tau in the brain and is largely associated with cognitive decline [4]. It is also the leading cause of dementia in the world [4]. With a growing elderly population, understanding the currently unknown etiology of this disease is crucial. Biomarkers pave the way for discoveries regarding disease onset, progression, prevention, diagnosis, and treatment. Commonly, diagnostic biomarkers for Alzheimer's require the use of neuroimaging techniques, like MRI and PET scans to observe atrophy, plaques, tangles, and other factors [5].

Magnetic resonance spectroscopy is a noninvasive technique for quantifying biochemical concentrations and changes in the brain. Generally speaking, MRS is feasible because distributions of electrons in an atom cause the nuclei of different molecules to experience differing magnetic fields [6]. Because hydrogen is the most abundant atom in living organisms, it is a clear candidate when applying MRS clinically [7]. After excitation via MR technology, radiofrequency signals are created which, after being decoded using Fourier transformation and subjected to spectral fitting, can be used to quantify metabolic concentrations [8]. A relatively recent advancement in the field of MR is the use of increasingly available 7 Tesla systems, which can more reliably detect, separate, and facilitate quantification of a larger number of metabolites [8]. Previously, if one wanted to quantify one of the harder to parse metabolites at a lower magnetic field, like ascorbate, they'd potentially need to sacrifice some of the other metabolites to do so in a process called editing or use methods

that are very sensitive to noise created if an individual moves, like the J-PRESS acquisition protocol [9].

Alzheimer related MRS research is seemingly becoming a more popular lens to view the disease. Most commonly, single voxel MRS is performed to quantify metabolite concentrations in one or two brain regions affected by AD, such as the posterior-cingulate cortex or the hippocampus [10]. Thus, differences or similarities in metabolite concentrations between Alzheimer's participants and controls can be observed and reported upon. The strongest risk factor of Alzheimer's disease is age, so controlling for age should be a relatively important aspect of this research [4]. Additionally, age commonly influences what fitting method is used, so if age is similar there are less likely to be differences due to changes in fit [8].

Currently, there are several MRS studies reporting on many brain region-specific differences between Alzheimer's participants and controls [10]. One study was able to report 88 percent sensitivity and 97 percent specificity in classifying Alzheimer's participants and controls with JMP's bootstrap forest method using 14 neurochemicals from 2 brain regions, the post-cingulate cortex (PCC) and the occipital cortex (OCC), leading to 28 total variables [2]. The PCC data from this publication will be used throughout this research. The PCC is a region well-documented to be impacted by Alzheimer's and is a part of the Default Mode Network, while the OCC is not known to experience substantial Alzheimer's or age-related changes [2]. It's worth noting that it was acknowledged in this publication that out of sample values would be lower than the reported results.

The paper mentioned above was able to compare means of chemicals reported on between Alzheimer's disease and controls and found a significant difference in average ascorbate in both the PCC and OCC regions, with both regions suggesting an increased ascorbate concentration in those with Alzheimer's [2]. This finding was replicated with a higher ascorbate concentration found in Alzheimer's participants compared to controls in the posteromedial cortex, another component of the default mode network, using J-PRESS acquisition at 3T [11]. In the PCC region, there have also been significant findings comparing means of myo-inositol and total choline, both of which exhibiting higher mean concentrations in Alzheimer's participants compared to controls [2]. The PCC also

has supporting documentation of differences between Alzheimer's participants and controls for MRI acquired parameters such as percent gray matter, percent white matter, and percent cerebral spinal fluid, so it's reasonable to conclude that combining MRI and MRS data may suit as a stronger biomarker than both independently, though combining these parameters in the past haven't influenced the ability to distinguish the two groups [2].

Given the statistically supported differences in MRS acquired data between AD and controls, there is potential to use this data in a classification algorithm which could support the use of MRS parameters as a biomarker for the disease, particularly a diagnostic biomarker. In addition, it could provide evidence to support research into when neurochemicals change in the disease course, which may give rise to preventative therapies. Another natural extension is that it could be used to track treatment response.

From a data science perspective, one of the most significant limitations of many MRS related Alzheimer datasets is the limited data sizes, with studies listed above having 16 AD participants and 20 AD participants respectively [2, 11]. With limited data, many issues can arise in disease studies, such as low probability of finding true effects, low positive predictive power, and an overly optimistic quantification of effect size [12]. Low positive predictive power is particularly meaningful in the realm of machine learning, as classification algorithms can struggle to find the variance between two groups, and potentially favor the group with the higher n value in an imbalanced data set [13].

Simulation may be a useful tool to increase the predictive power of classification models when it comes to determining one's AD status in a small MRS data set. Simulation can account for imbalance and increase gaussian variance of the dataset [14]. The increase of Gaussian variance can be beneficial to classification algorithms that rely on variance and distribution of provided variables. Classification models using simulated data can always be strengthened with new data, as well. Newly acquired data can also be used to test if simulated data is accurately capturing the truth of the data. For small sample sizes and simulation, it would probably be most telling if a classification method works for a given dataset by first using the "leave out one" method, which is primarily used for cross validation [15]. Essentially, all but one data point will be used for simulation and training to create a

robust sample size, while the one that is left out is then tested on.

Bayesian networks provide a means of simulation. A Bayesian network is a probabilistic graph model used to represent variables and their conditional dependencies [16]. In Bayesian networks, nodes characterize a random variable and lines characterize conditional probability for the corresponding random variable [16]. These relationships are of course determined using Bayes Theorem to integrate these variables to create a corresponding hypothesis of the causal relationship of any variable [16]. A benefit of using a Bayesian network for means of simulation is that there are no required assumptions of relationships in the data. If relationships are known though, they can be used to optimize the reliability of the network. Bayesian networks have previously been used in Alzheimer's research to characterize differences in default mode network connectivity for those at risk of the disease using MRI parameters [17].

A comparison study has previously reported that gradient boosted decision trees and random forest yield highest accuracy when compared to nine other classification models with results from 12 different data sets [18]. Random forest and GBDT are both non-linear classifiers, with a few key differences including the way trees are built, and sensitivity to overfitting. Specifically, GBDT are built sequentially, with each tree fitting to residuals from the previous tree, while in random forest, trees are built in parallel [19, 20]. A benefit of random forest over GBDT is the resilience to overfitting, despite the number of trees used in the model [20]. Other benefits to random forest include the ability to explain results, and tuning simplicity [19, 20]. Random forest classifications have resulted in the best accuracies in multiple imaging studies of neurological conditions, including Alzheimer's [21, 22].

A non-tree based classification algorithm that has previously supported Alzheimer's classification tasks is SVM, or support vector machine [23]. Support vector machines aim to find the closest data points between classes, and then seek to create a decision boundary between these points that best separates them and is further used for classification [24]. Though interpretability suffers when using non-linear SVM kernels, they may offer insight as to whether non-tree based methods are the best fit for separating such data.

Chapter 2

BACKGROUND

2.1 Sample Size Limitations

Small datasets are a frequent challenge in biomedical research where collecting many samples is often logistically difficult or cost prohibitive. Sensitivity, or the ability to accurately detect a true effect, is particularly limited in low sample size studies [12]. The focus of this project is to understand the suitability in using simulation data to train models for better accuracy and sensitivity in biomedical research where sample sizes are extremely limited using Alzheimer's disease as a test case.Specifically, the goal is to try to capture differences in neurochemistry between Alzheimer's individuals and healthy controls using low sample sizes (16 AD patients total). Thus, it is important to be aware of statistical complications arising from similar studies, including lower chances of finding true effects, low positive predictive power, an exaggerated estimate of the magnitude of the effect when a true effect is distinguished, and false positive results [12].

A common method to counteract limitations that arise from low sample size in machine learning is simulating data [25]. The goal of simulation models is to capture the underlying behavior of the data, then this data can be used to offer more training data into machine learning models.

2.2 Bayesian Networks

Bayesian Networks are a common means of simulation, that have several advantages over other methods of simulations for biological systems as experimental data and prior knowledge can converge to test specific hypotheses and map complex systems in a visually comprehensible manner [26, 27]. Bayesian networks are a probabilistic graphical structure displayed as a directed acyclic graph [26]. In these graphs, child nodes, or nodes with incoming edges, represent that there is a conditional dependency between that child node and its parent node [28]. An example of a Bayesian network can be found in 2.1 with the conditional dependencies of the Gaussian node 'H' shown in Figure 2.2. Worth noting, Bayesian networks can derive causal relationships from datasets containing continuous and discrete data, whose relationships are generally not well captured by other means of simulation and is of relevance for the data used in this research [28]. The structure of a Bayesian network can be learned, supplied by a user, or a mixture of both methods [28]. While there are benefits to supplying prior knowledge to Bayesian Networks in the form of probabilistic interactions, there are recent publications supporting the use of automated learning of Bayesian networks [29, 30]. In the scope of this research, heuristic structure learning is particularly useful as the causal relationships of the distributions of the neurochemicals are not particularly well documented, leaning itself to an automated process. Automated structure learning for Bayesian networks has been demonstrated to work well on smaller datasets, such as the UCI heart disease dataset with 300 observations, as well [28]. Despite their recent popularity in biological systems, historically, Bayesian networks have not been widely explored in the field of neuroscience [31]. Limited previous work has explored the functional connectivity between regions from MRI related data, paving way for potential use in MRS [31]. Even if classification models strengthened by MRS data simulated from Bayesian networks perform poorly, there is potential for discovering distinct interactions between available variables [31].

Bayesian networks are an incredibly computationally challenging problem (NP-hard), and therefore heuristic alternatives have been designed to help overcome these challenges. One promising and common approach to learning structure of a Bayesian network from the data is by using the

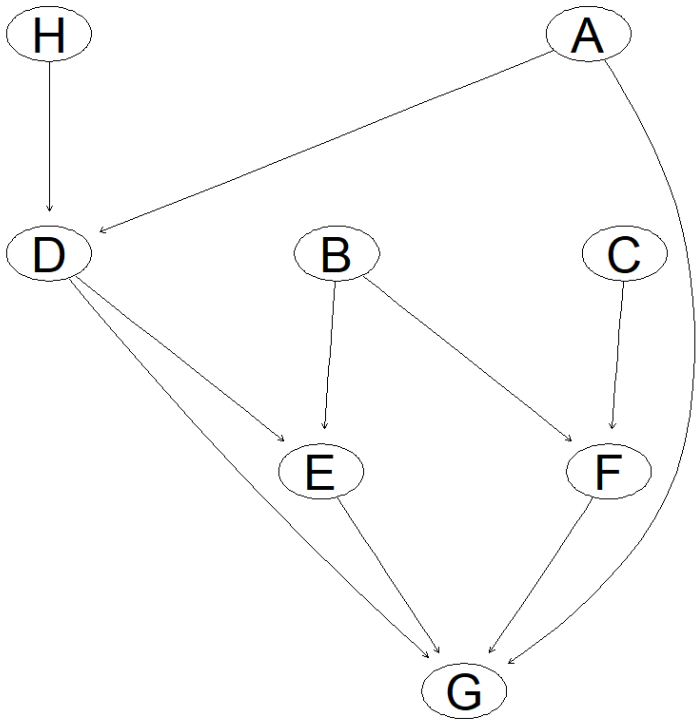

Figure 2.1: Bayesian network example in which nodes H,B,A, and C are root nodes (meaning they are without parents). G is a leaf node, as it does not have any child nodes. [1]

```
  Parameters of node D (conditional Gaussian distribution)

Conditional density: D | A + H
Coefficients:
                     0           1
(Intercept)   5.2919300  10.0410549
H             0.8817193   0.9834474
Standard deviation of the residuals:
        0         1
0.5099208 0.3073215
Discrete parents' configurations:
   A
0  a
1  b
```

Figure 2.2: Conditional distribution of node D are shown, which has two parent nodes, A and H. A is a categorical variable (whose values are 0 and 1 in this example) while H is a Gaussian variable. So, when fitting linear regressions from H to predict D, there is a linear regression fit per class of A. Coefficients represent the values to be multiplied by the predictor, while intercepts are the constants of the equation. Residuals describe the vertical distance between data points and the regression line. [1]

hill climbing algorithm, with previous research supporting good results [32, 33]. Hill climbing is a score-based learning criteria, which seeks to attain the best structure in accordance with a supplied function [34]. Common functions to use with the hill climbing algorithm include Bayesian Information Criteria (BIC), Akaike's Information Criteria (AIC), and Bayesian Dirichlet equivalence score (BDeu) [35].

BIC has been demonstrated to perform better in the creation of network structure in comparison to AIC and BDeu and has been applied to studies of medical significance [36, 35]. In one case, networks derived from score-based hill-climbing approaches were compared to a predefined "gold standard" network [36]. In the case of networks derived from the Cleaveland Heart Database dataset and the Brest Cancer data set, BIC outperformed AIC by identifying correct edge placements and directions [36]. BIC and AIC are computed using a log-likelihood and regularization (penalty) component dependent on the resulting structure as well as the data [37]. In the bnlearn package in R, which was used in the scope of this work, BIC and AIC scores are calculated by 1) constructing contingency tables of data from nodes against configurations of data from parent nodes ,2) creating a probability table for each node, 3) determining the number of parameters, and 4) deriving log likelihood from the contingency table, which the penalty score is subtracted from [37]. The penalty score for BIC is particularly influenced by model complexity with more edges and parent nodes being deemed as more complex, meaning that networks with fewer edges and parent nodes will be penalized the most, especially in small datasets where the log likelihood is less influential of BIC scoring as a whole [38]. The penalty score for AIC does not increase with sample size and is less harsh of model complexity, which may be beneficial, as there is a better chance of interactions being found and utilized. In a study aimed to compare structure learning algorithms and scores used within them for breast cancer networks, AIC outperformed BIC in both hill-climbing and tabu search [39].

Max-Min Hill-Climbing, or MMHC, incorporates constraint based local learning in the hill-climbing algorithm [40]. First, the max-min parents and children algorithm is implemented, which uses conditional independence testing to determine candidate parents and children for all given variables in data, creating an underlying and undirected structure of a Bayesian network. [40]. After

this, hill-climbing searches through the given space with the goal of minimizing a given score, such as BIC or AIC [40]. The major difference between MMHC and standard hill-climbing, is that arcs are only permitted to add edges that were found by the max-min parents and children algorithm [41]. MMHC has been demonstrated to perform well in small sample sizes, drawing meaningful conclusions from a small-sample size prognosis evaluation of non-small-cell lung cancer [42]. Compared to MMHC, search and score methods of simple hill-climbing can be slow to converge, and complexity may limit ability to find an optimal structure [43].

Once a network structure is fitted via maximum likelihood, any number of rows can be simulated using the network. In general, smaller sample sizes can lead to fewer reliable simulations compared to larger sample sizes. Data for each row is populated sequentially, determined by its position in the network. Specifically, parent nodes are randomly sampled from their unconditional distribution, then their conditional child nodes are supplied via their conditional distributions from their parent nodes [44]. The process is repeated until all values are supplied. Simulated data then needs to be validated against original data, as doing so ensures that simulated data exhibits similar behavior to the original data. Without validation, there is a risk that models built on the simulated data will be capturing noise. There are many possible methods for validation, such as visualizations of distributions and correlation analysis of features. Posterior predicative correlation is a Bayesian network validation technique in which predictions are created from an arbitrary set of nodes using likelihood weighting to get Bayesian posterior estimates [37]. If the Bayesian network fits the data well, data generated from the models ought to be similar to observed data [45].

2.3 Classification Algorithms

A common machine learning mechanism for classification tasks related to bioinformatics is random forest [20]. This machine learning model has many benefits, including but not limited to the ability to overcome small sample sizes, handle imbalanced datasets, and ease of optimization due to few tuning parameters, and results are simple to explain [46]. The data for the Alzheimer's model being used in this study is from a small and imbalanced data set. Choosing a model robust

to challenges of the given data is a primary concern as over representation of a class often leads to over prediction of that class [47]. Many random forest packages in R have built in arguments that can automatically perform under sampling on the majority class to account for the imbalance or provide balanced class weights to be used in sampling [48]. Synthetic oversampling of the minority class has been demonstrated to support good accuracy in random forest models classifying cancer, supporting the use simulated data in classification over methods like random minority oversampling, in which random rows in the minority class are duplicated giving rise to issues of overftting [49]. Thus, simulated data may provide better results than random forests built in methods to address imbalance.

Random forest (RF) performs bagging on training data. This involves taking bootstrap samples of data (with replacement), fitting many decision trees in parallel, then aggregating the results from all the trees [46]. Bagging allows for the use of many training sets, and sampling with replacement allows a significant reduction in prediction variance [20]. Because sampling with replacement offers lower variance, this technique can be used to combat noisy data. Another notable ability of random forest is the capability to prevent overfitting when there are correlated variables, which is demonstrated in data to be used in this research as shown in Figure 2.3 [2]. Random forest relies on a specified or default ($\sqrt{p}$) mtry argument, which indicates the number of random variables to be sampled at each split [20]. If mtry is low, highly correlated variables are less likely to be sampled together for splits, reducing the negative impact of the correlations on the model, and reported feature importance from the model [23]. Retaining correlated features is also expected to aid in gathering a more complete picture of the mechanisms of the data [23].

Extreme gradient boosting (XGBoost or XGB) is another ensemble learning algorithm that classifies based on the congregation of outputs from single trees. XGB has been shown to perform closely with random forest, sometimes outperforming it, such as when predicting risk of breast cancer [50, 51].XGB has been applied to numerous classification tasks related to biological and health research. For example, XGB classifiers have been used to distinguish Alzheimer's status using cognitive measures, age, and gender [52]. Additionally, XGB has been applied to MRS and PET

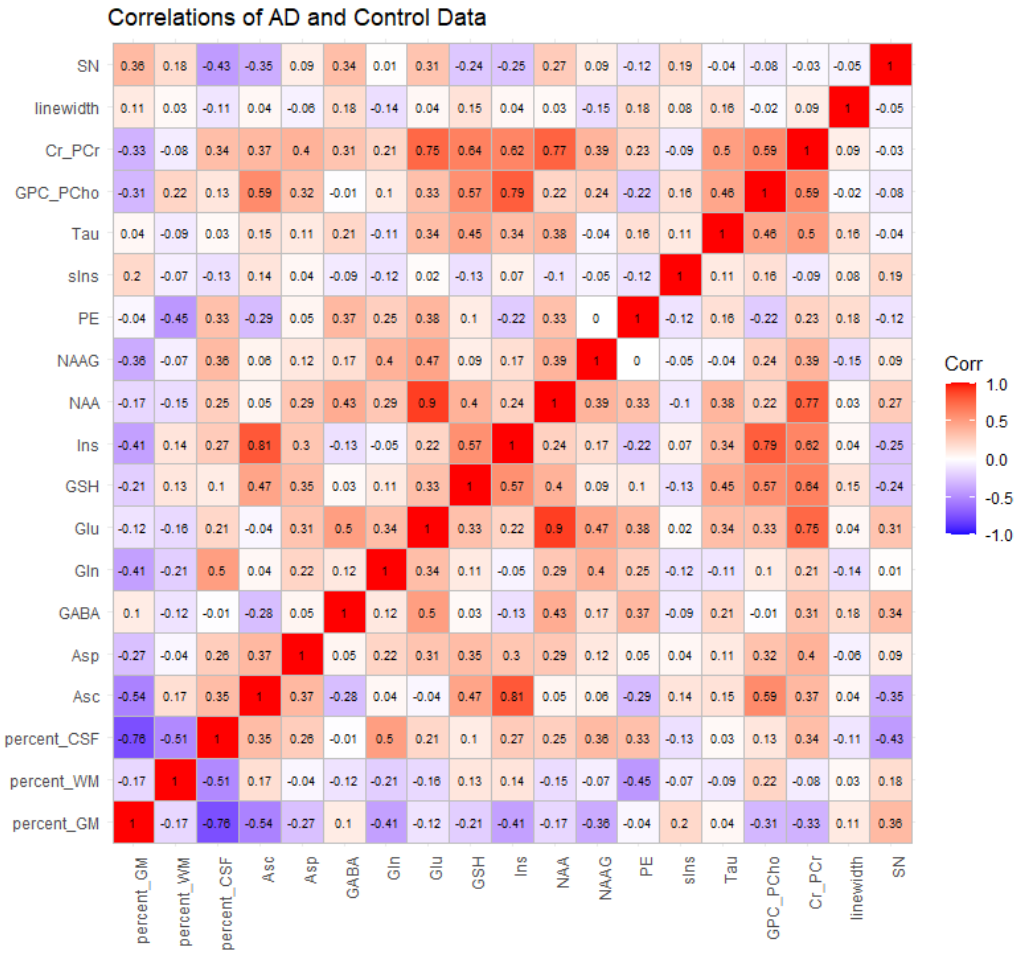

Figure 2.3: Correlations present in AD and Control data used in this study as well as the previous publication [2].

data with the goal to provide biomarkers to predict remission in depression, leading to models with 92% specificity and 77% overall accuracy [47].

XGB, as the name implies, uses boosting. In boosting algorithms, trees are formed sequentially, and aim to improve with each tree [53]. The model starts with the whole training set, and levels that are misclassified are given more weight in the production of the next tree [53]. In doing so, there's a distinct reduction in bias compared to random forest, supported by recent publications [54]. There are limitations that come along with using a more complex algorithm such as this, though, including harder to optimize hyperparameters as well as the risk of overfitting data, caused by the lack of variability in the training data [53] . Overfitting can be overcome in XGB by supplying a smaller number of columns to be tried at each split, reducing tree depth, among balancing other hyperparameters [53].

SVM, in contrast to XGBoost and random forest, does not rely on decision trees. Rather, the algorithm seeks to find a hyperplane that has the largest margin, or distance from the hyperplane to the closest data points from either class [55]. When data is not linearly separable, a kernel function is used to transform the data to become linearly separable in a feature space of high dimensions [55]. The radial basis kernel, or RBF kernel, has been demonstrated to be quite effective in maximizing accuracy when classifying disease data compared to other kernels while being easy to tune, requiring the selection of one added hyperparameter, gamma, as opposed to three with the polynomial kernel [56]. The equation for the RBF kernel is shown in in Equation 2.1,where xi and xj represent vectors in the input space and gamma determines how far the influence of training example reaches [57]. In detection of skin cancer, SVM models were able to outperform random forest and k-nearest neighbors' algorithms [58]. Additionally, SVM models are easily weighted, meaning penalties for misclassifying minority class observations can be higher than that of the majority class [59]. This is clearly applicable in biomedical data where class imbalance often arises. SVM classifiers have demonstrated capabilities of high accuracy when handling MRS data, as well, specifically in the classification of brain lesion using NAA, choline, myo-Inositol, creatine plus phosphocreatine, lipids, and lactate [60].

$$\mathrm{K}(x, x_i) = e^{-\gamma \|x - x_i\|^2} \tag{2.1}$$

In machine learning, to really estimate the performance of the model, an unseen test set needs to be applied. In the case of small sample sizes, this is costly, as the reduction of instances in training data effect the model's ability to distinguish the classes in the test set. A way to counteract this issue is to employ leave-one-out cross validation (LOOCV). In this cross-validation method, one instance is left out and treated as the test set, with the test instance changing each iteration until all observations have been predicted [15]. This has a clear advantage over holding out a larger proportion of data for testing, as the observations chosen in training can introduce bias. LOOCV thus gives a relatively unbiased estimate as to the accuracy of the method [61]. LOOCV has been supported in many recent publications of Alzheimer's research in which small datasets are common [62, 63, 64]. A common limitation is computational expense, which increases if steps like hyperparameter optimization need to take place within the cross validation.

Chapter 3

METHODS

3.1 Participants

3.1.1 AD and Control

Marjanska et al (2019) outlines recruitment for AD (Alzheimer's disease) and control participants [2]. In short, most AD participants were identified through a database of patients cared for at the Geriatric Research, Education and Clinical Center (GRECC) Memory Loss Clinic at the Minneapolis Veterans Affairs Medical Center. They were reviewed and diagnosed by a multidisciplinary panel using the NINCDS-ADRDA criteria. Participants met criteria for possible or probable AD, and no blood, cerebrospinal fluid (CSF) or brain amyloid biomarker was available for diagnosis of pathological AD. One AD patient was recruited from a community dementia community and their AD status was confirmed by the study neurologist. If an individual had a neurological diagnosis other than AD that could affect cognition, or other conditions that were thought to potentially interfere with the regions of interest in the brain, they were not recruited for the study. AD participants were deemed to have a mild to moderate case of dementia according to the Montreal Cognitive Assessment (MoCA) or Mini-Mental State Exam and also based on independence in basic activities of daily living. Participants with AD were required to meet the same criteria as control participants, with the exception of diagnosis of AD. Control participants were carefully screened to rule out presence of cognitive impairment, neurological diagnosis that could affect cognition, brain lesion or injury, major

psychiatric disorder, substance abuse, and serious systemic illness. Screening by a neuropsychologist and neurologist was based on medical history, MoCA scores >= 24, and neurological examination. Participants were excluded if they had conditions that could interfere with their ability to participate in MRI, such as claustrophobia.

3.1.2 Healthy Participants From Human Connectome Project on Aging

The Lifespan Human Connectome Project in Aging aims at providing detailed multimodal MRI and behavioral data from typically aging participants between the ages of 36-100+. Cognitively healthy participants from one location of the study, the University of Minnesota, were selected for the additional collection of MRS data to further understand biochemistry underlying aging. To qualify for the MRS portion of the study, individuals who were over 60 years of age had to have satisfactory scores in one to three tasks meant to gauge cognitive wellness: Montreal Cognitive Assessment (MOCA), Trail Making, and the Rey Auditory Verbal Learning Test (RAVLT) [65, 66]. Total score for MoCA had to be above 19. If the total MoCA score was higher than 22, this was all that was needed to deem someone cognitively healthy. If MoCa scores were above 19 but not 22 or higher, the participant's Trail Making Part A and Part B scores and RAVLT Learning Over Trials (LOT) score must have had a Z-score greater than -1.5, meaning the participants could not have scores less than 1.5 standard deviations below average scores. If those who deemed participants qualified had any hesitations of participants who passed these tests, the participants were further screened by a neuropsychologist. All participants over age 60 were examined by the study neurologist to rule out presence of cognitive impairment. Participants could not have a neurological diagnosis that could affect cognition, brain lesion or injury, major psychiatric disorder, substance abuse, or serious systemic illness. There are 61 healthy participants from the Lifespan Human Connectome Project on aging used in this project as additional controls when noted.

3.2 MR Acquisition

MR protocol for all participants used in this thesis is outlined in Marjanska et al [67]. Data was acquired using a 7T whole-body Siemens Magnetom scanner and a custom-built head coil. 7 T MRS allows for a larger number of neurochemicals to be detected reliably than is feasible at a lower magnetic field. Magnetization-prepared rapid gradient-echo (MPRAGE) images were used to identify the 8 mL regions of interest in the brain and quantify grey matter, white matter, and cerebrospinal fluid content using Freesurfer. Magnetic field (B0) homogeneity was optimized using FAST(EST)MAP [68]. Ultra-short echo time (8 ms) metabolite spectra (11 minute acquisition time) were processed in Matlab, where Eddy Current effects were corrected and frequency and phases aligned. Spectra were later fitted using LCModel,with the following neurochemicals recorded : Asc, aspartate (Asp), creatine (Cr), γ-aminobutyric acid (GABA), glucose (Glc), glutamine (Gln), glutamate (Glu), glutathione (GSH), glyc-erophosphorylcholine (GPC), lactate (Lac), mIns, NAA, NAAG, phosphocreatine (PCr), phosphorylcholine (PCho), phosphorylethanolamine (PE), scyllo-inositol (sIns), and taurine (Tau). Age specific macromolecules were used in the basis set as well [69]. Chemical quantification was accomplished using water signals (acquired using the same MRS protocol except without water suppression) corrected for gray matter, white matter, and cerebrospinal fluid content as an internal reference as outlined in equation 2 of Gussew et. al 2012 [70]. Only neurochemical concentrations that passed Cramer-Rao lower bound based quality control metrics were included in the models for the overarching study.

In total, data used in this research contain twenty variables collected from the posterior cingulate cortex (PCC) brain region : 14 neurochemical concentrations as mentioned above, percent gray matter, percent white matter, percent cerebrospinal fluid, signal to noise ratio, linewidth, and AD-status. It is challenging to recruit participants with AD for research, and even more challenging to enroll a cohort that does not have other complicating diseases and is at a particular disease stage (i.e., extant AD but not late stage). Thus, the sample size is small. With current focus on AD research, future cohorts are likely to be small, and the best possible data science approaches will be

needed to extract the largest possible extent of cohort characteristics. One limitation of the small n is the difficulty in drawing inferences outside of the sampled cohort. In the case of the data used here, the cohort was further limited to a specific geographic residence and care in a specific clinic.

3.3 Simulation via Bayesian Networks

3.3.1 Structure Learning

Structure learning refers to the process of learning the structure of a directed acyclic graph from the data itself. To find an optimum structure learning method three hill climbing algorithms were compared via the bnlearn package (version 4.8.1) within R programming (version 4.2.2): hill-climbing scored with Alkaline Information Criterion (aic-cg), hill-climbing scored with Bayesian Information Criterion (bic-cg), and max-min hill-climbing (mmhc), which employs the Max-Min Parents Children algorithm in the constraint phase to determine an undirected skeleton of structure to later be oriented via hill climbing scored with Bayesian Information Criterion [71, 1, 40] . Because models are to be fit via maximum likelihood estimation, both AIC and BIC offer appropriate scoring, as they are both based on log-likelihood and complexity as seen in Equations 3.1 and 3.2 [72, 73]. For training sets greater than or equal to seven observations, BIC will enforce a greater penalty for large models. While some past research has found AIC to perform better, particularly in small sample sizes where BIC can suffer due to the increased penalty for complexity, other studies have found that BIC scoring consistently outperforms AIC scoring in discovering underlying Bayesian structures [36, 74].

$$\mathrm{AIC} = \log \mathrm{L}\left(X_1, \ldots, X_v\right) - d \tag{3.1}$$

$$\mathrm{BIC} = \log \mathrm{L}\left(X_1, \ldots, X_v\right) - \frac{d}{2} \log n \tag{3.2}$$

3.3.2 Fitting

Once structure is determined via the optimal method from the above section, parameters of local distributions were estimated via the bn.fit function from bnlearn (version 4.8.1) [1]. Fitting was completed via the mle-cg method, which employs a maximum likelihood estimator for conditional probabilities when fitting discrete nodes, and a maximum likelihood estimator for least squares regression models for Gaussian nodes to create a conditional Gaussian Bayesian network. For continuous data, parameters take the form of regression coefficients while parameters for discrete nodes take the form of conditional probability tables. Distributions of continuous nodes are linearly dependent on continuous parents with parameters conditioned upon values from discrete parents. This method of fitting does not allow discrete nodes to have Gaussian parents, which allows the use of least squares regression to efficiently obtain maximum likelihood estimates for continuous nodes. Therefore for this study, AD status was a root node in every case of simulation, and probability of AD-status when simulating will be reflective of the class proportions in the observed data.

3.3.3 Generating Simulated Data

Simulated data was generated via forward sampling using the fitted network within the rbn function in bnlearn (version 4.8.1) [1]. Forward sampling is a stochastic process that occurs topologically, with root nodes being sampled first from their unconditional distributions [75]. This is followed by the sampling their child nodes by conditioning the child node's conditional probability distributions to values sampled from the parent node(s). This is repeated until each node has been sampled. The number of simulated observations is held constant throughout the experiment at 1,000 samples to ensure that simulated data characterize the original data well, and to limit computational expense when subsequently fitting classification models.

3.4 Comparing Model Performance With and Without Simulated Training Data

3.4.1 Generalized Workflow

The effect of including 1,000 rows of simulated data in the training of random forest, XGBoost and SVM, on predictive performance were evaluated using three derivations of data; original AD and control data, AD and control data reflected by principal components via the prcomp function in the stats package of R (version 4_4.2.2), and AD and control data with additional healthy non-AD participants from the Human Connectome Project on Aging acting as additional controls in training and simulation. Predictive performance was evaluated using training and testing subsets, encompassing 65 percent and 35 percent of the data respectively. Additionally, predictive performance was evaluated using leave-one-out cross-validation to prevent sampling bias from impacting model performance. Models were evaluated for accuracy, sensitivity, and specificity.

When models were evaluated using train-test splits, prior to splitting the data, seed state was set to ensure the same test sets are being used for predictions when using simulated enhanced data and original data in training. After splitting the data, hyperparameters were grid searched using three fold cross-validation, where all hyperparameter combinations are tested and the combination that results in the model that minimizes classification error is determined optimal. The optimal model was then used for classification on the test set. These steps were then repeated, with the exception of creating simulated data from the training data using methods mentioned above and appending it to the training data prior to grid searching. This process was repeated 100 times, and predictive performance on the test set from models that utilized simulated data were compared to models only utilizing original training data using the paired t-test.

The process was very similar for leave-one-out cross-validation, with the exception of training data consisting of all but one row of data. Given the 49 original AD and control data samples, there are 49 models used per run when utilizing LOOCV, thus the process was only repeated 10 times to

limit computational expense. These processes are outlined in Figure 3.1.

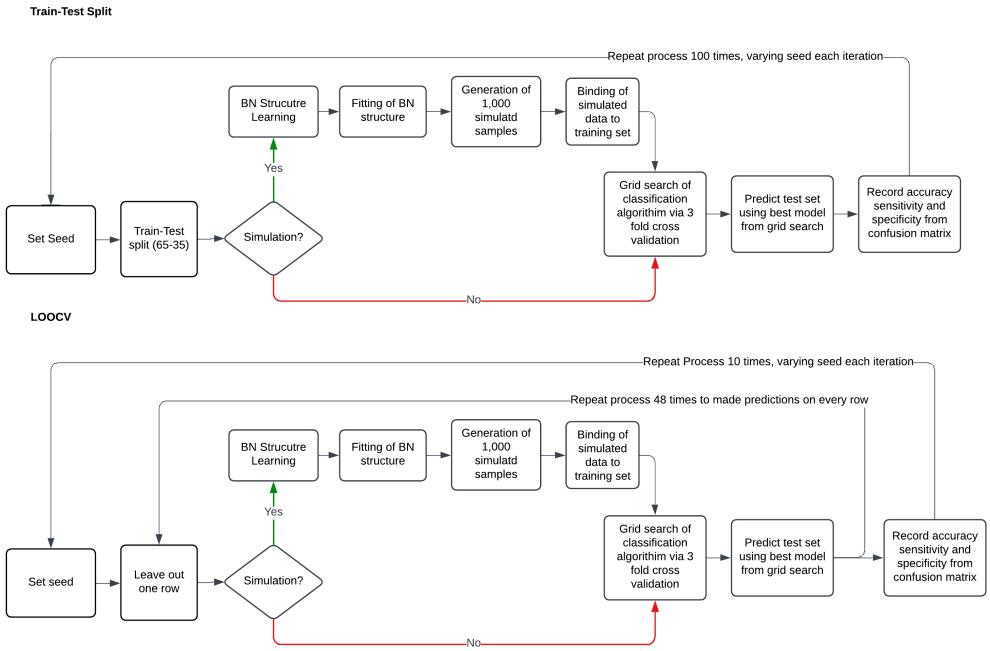

Figure 3.1: Generalized Workflow Flowcharts

3.4.2 Random Forest

Random forest has been applied to the same AD and control data in past publications, though using two brain regions rather than solely the post-cingulate cortex [2]. To create random forest models in this experiment, the randomForest package (version 4.7-1.1) in R (version 4.2.2) was utilized [71, 76]. In the past publication (Terpstra, 2019), Bootstrap Forest was implemented in JMP Pro ,though the impressive results have not proven to be reproducible using other implementations of random forest. Balanced class weights in training data were computed and supplied to the model, which provide prior probabilities and aid in the training of models with imbalanced data by allowing observations of the minority class a higher likelihood of being selected in bootstrap samples. The number of trees (ntree), number of randomly sampled columns to be used as candidates per split (mtry), and minimum size of terminal nodes (nodesize) were grid searched to find optimum values for these parameters via 3-fold cross validation for these parameters using the tune function from the e1071 package (version 1.7-12) in R (version 4.2.2) [77]. Possible values for ntree were 400 and

1000, with 1,000 trees being used for forests in the previously performed analysis of the AD and Control data. Using a higher number of trees can be related to better predictive performance, though this benefit can plateau at a given number of trees [13]. Using fewer trees is primarily beneficial to computation efficiency, especially when trying to find optimum values for other hyperparameters. Possible values for mtry were 3,4,5. Lastly, possible values for nodesize were 3,5, and 7, with higher nodesize resulting in better computational efficiency by creating shallow trees, which may lead to needing more trees. Classification error was used to deem the optimum model by 3-fold cross validated grid searching, which was then fitted using all the data and used to predict testing data. Variable importance was extracted from each model using varImportance in the randomForest package to offer model explainability via mean decrease of the Gini coefficient.

3.4.3 Support Vector Machines (SVM)

Support vector machines were executed via the svm function from the e1071 package (version 1.7-12) in R (version 4.2.2) [77]. A non-linear kernel, radial basis function, was used throughout SVM models in the experiment due to its demonstrated performance in classification tasks [78]. Training data are scaled with the scale argument within the svm function to ensure that all values have the same influence on the distance metric from the kernel matrix. Additionally, balanced class weights were supplied to the svm function, which offered higher penalty costs to the minority class (AD) to combat imbalance. SVM models were tuned via three fold cross validation grid search using tune.svm in the e1071 package [77]. Possible values for epsilon, or the insensitive loss function, were .1 and .01, with a smaller value implying steeper penalty for misclassification and therefore less support vectors. Possible values for cost of constraint violation, or cost, were 1,10,100, and 1000. The default cost value in the package used is 1, though low cost values may contribute to underfitting, while high cost can impose a high penalty for inseparable points and lead to overly complex models and overfitting. Possible values for gamma were .001,.01, and .1. Simply put, gamma represents how far influence from an observation reaches, with values that are too low creating support vectors that are unable to capture the shape of the data [79]. Inversely, when gamma values are too high, support vectors are

at risk of overfitting, as the influence of the support vectors only includes the vectors themselves [79]. Gamma and cost are often optimized together, as the negative effects of one being too high or low can often be overcome by the inverse of the other [79] . The best performing model from the cross validation based on classification error was trained on all of the training data and further used on the testing data.

3.4.4 XGBoost

XGBoost models were employed from the caret package (version 6.0-93) within R, with the method of the train function in this package defined as xgbTree [53, 80]. To limit computational expense, the number of trees per model was kept at the default of 500. To prevent overfitting, 70 percent of the training data was randomly sampled before growing trees, denoted by the subsample argument in the model. Additionally, the proportions of columns sampled by each tree was held constant at 70 percent. Balanced class weights from the training data were supplied to the model to prevent the effects of class imbalance by imposing a higher cost to misclassifications of the minority class. Similarly to random forest and SVM models, a three fold cross validation was performed to determine the best hyperparameter combination in a grid based on classification accuracy. Possible values for maximum depth of tree were 5,6, and 7, with a larger value creating complex models that are at risk of overfitting [53]. Possible values for eta, which denotes learning rate, were .2,.3, and .4, with the package default being 3 [53]. A higher eta value creates a more conservative algorithm that is less likely to overfit by shrinking feature weights after each boosting step, with lower values causing underfitting. Possible values for gamma, which is the minimum loss required to make a further partition on a leaf node, were 0 and .1, with a default of 0 and a higher value denoting a more conservative model. The best model resulting from the cross validation in terms of accuracy was used to perform predictions on the test set. Additionally, variable importance in the form of mean gain were kept from the XGBoost models in order to contribute to model explainability.

Chapter 4

RESULTS

4.1 Simulation via Bayesian Networks

10 fold cross validation, which was repeated 10 times, was used to find the best method of structure learning determined by average posterior predictive correlation for ascorbate, percent white matter, and signal to noise. The structure learning method that resulted in the highest posterior predictive correlation for ascorbate was hill climbing scored via BIC derived from data that contains over-sampled AD observations with a value of 0.8234, and was respectively followed by hill climbing scored via AIC created from over-sampled data (0.8097), max-min hill climbing using over-sampled AD data (0.7909), max-min hill climbing (0.7525), hill climbing scored via BIC (0.7116), then hill climbing scored via AIC (0.6678). The structure learning method that resulted in the highest posterior predictive correlation for percent white matter was also hill climbing scored via BIC derived from data that contained over-sampled AD observation with a value of .9996 and was followed respectively by max-min hill climbing using over-sampled AD data (0.9995), AIC scored hill climbing using over-sampled AD data (0.9992), hill climbing scored via BIC (0.9971), max-min hill climbing (0.9962), then hill climbing scored via AIC (0.9951). The structure learning method that resulted in the highest posterior predictive correlation for signal to noise ratio (SN) was again, hill climbing scored via BIC using data with over-sampled AD observations with a value of 0.5511, followed respectively by AIC scored hill climbing using over-sampled data (0.5126), hill climbing scored by BIC (0.3916),

hill climbing scored via AIC (0.3285), max-min hill climbing using over-sampled data (0.0573), then max-min hill climbing (-0.01904). Comparison of posterior predictive correlation across hill climbing variations for all variables can be found in Figure 4.3.

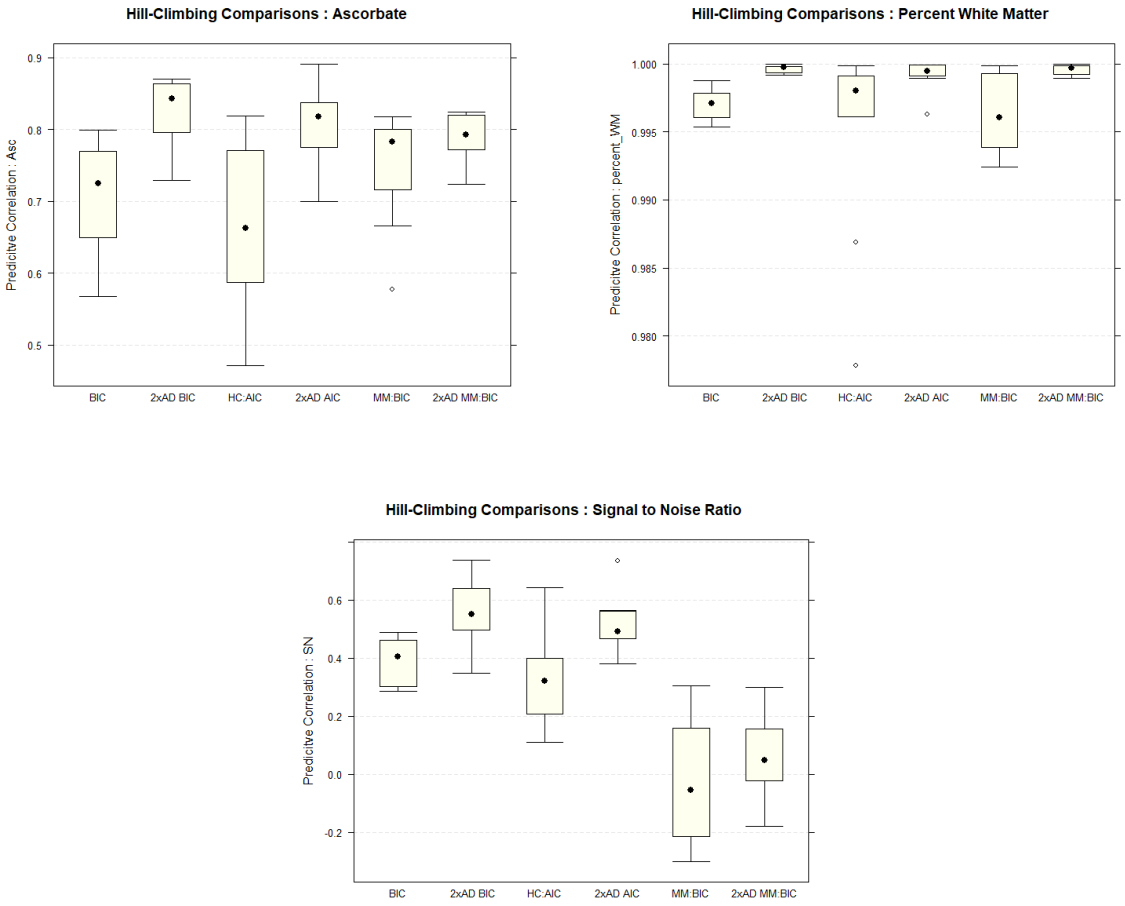

Figure 4.1: Hill climbing performance: posterior predictive correlation of ascorbate, percent white matter, and singal to noise ratio

Given the superior performance of Bayesian networks scored via BIC and fitted with over-sampled AD data, this was the method utilized for structure learning throughout the study. An example of a Bayesian network structure derived from hill climbing with BIC scoring using oversampled AD data can be found in Figure 4.2. Arc strength for the strongest half of arcs from this network are shown in Table 4.1, in which score refers to the change in score of the network (BIC) when that arc is removed, with a more negative score implying a stronger arc. The arc from percent gray matter to percent cerebrospinal fluid and the arc from percent white matter to percent cerebrospinal fluid are

the strongest arcs in the given network.

Distributions of simulated data for ascorbate, percent white matter, and SN can be found in Figure 4.3. Mean values and standard deviations for real and simulated values for all variables are shown in Table 4.2. Variables like ascorbate, glutamine, and GABA seem to be characterized well by simulated data, while variables like tau are poorly captured.

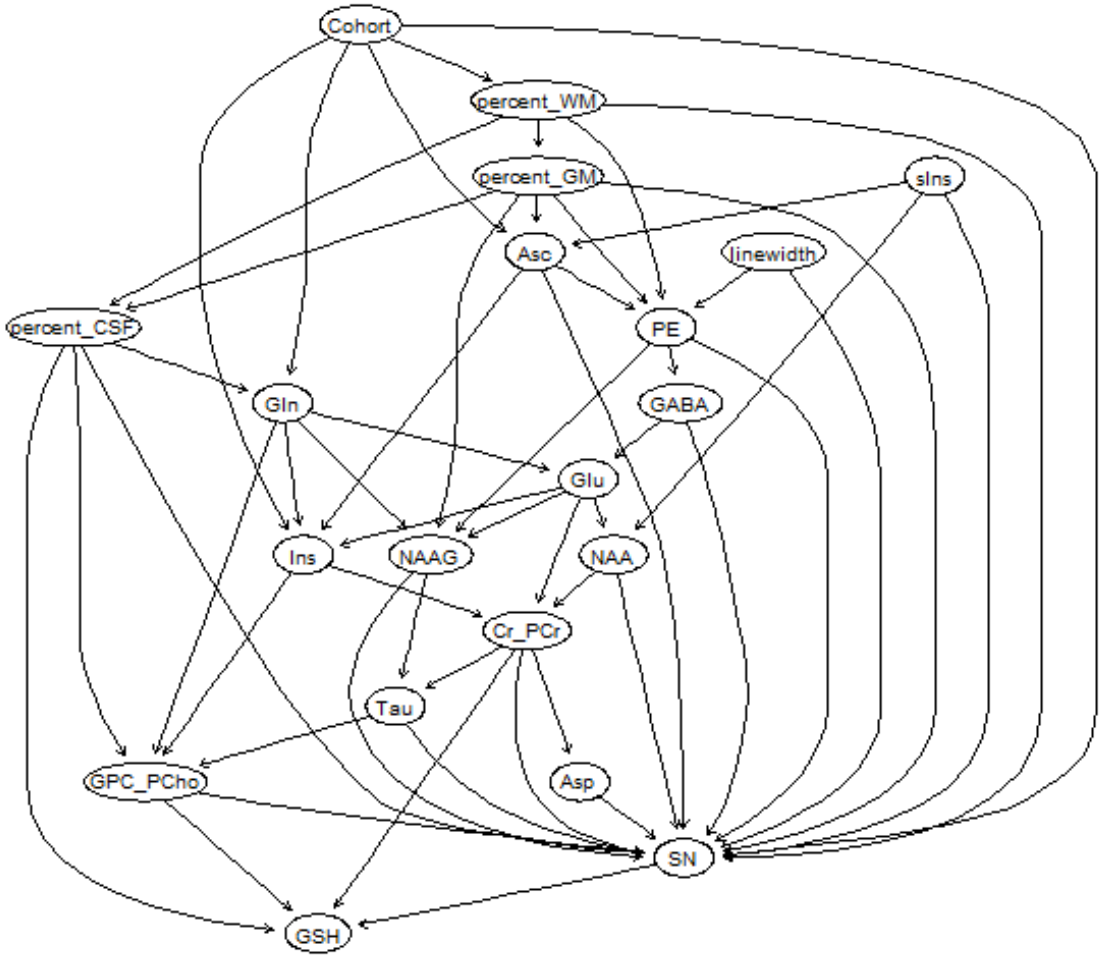

Figure 4.2: Bayesian Network dervied from Hill-Climbing scored via BIC and using AD and control data in which every AD case is doubled

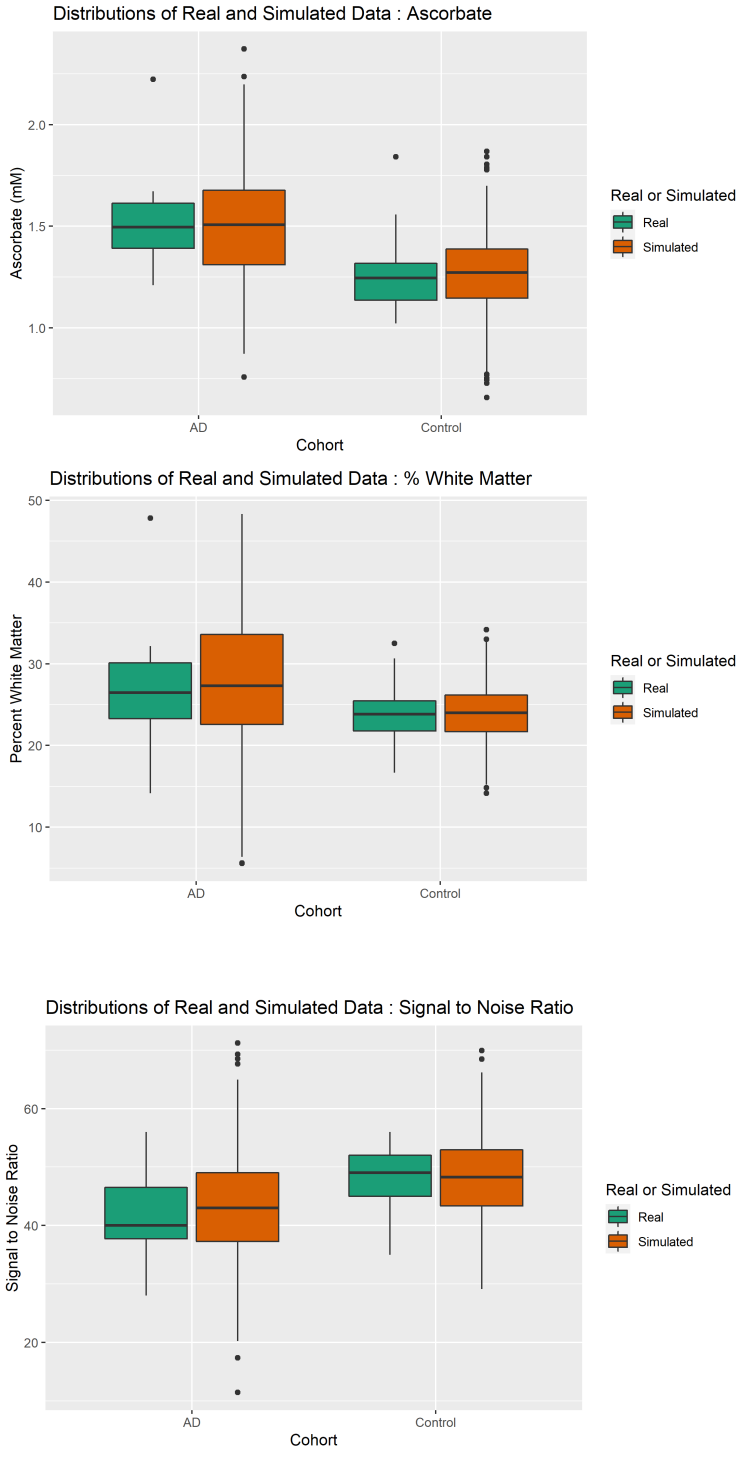

Figure 4.3: Distributions of real and simulated data for ascorbate, percent white matter, and SN

From	To	Strength
percent_GM	percent_CSF	-457.63
percent_WM	percent_CSF	-448.73
Cohort	SN	-113.98
sIns	SN	-113.07
Cr_PCr	SN	-77.63
NAA	SN	-77.34
Tau	SN	-71.68
GABA	SN	-69.4
percent_CSF	SN	-61.62
percent_GM	SN	-61.57
percent_WM	SN	-61.54
Glu	NAA	-58.47
GABA	GSH	-56.85
GSH	SN	-53.63
Ins	SN	-53.55
Cohort	GSH	-46.4
Glu	SN	-43.59
Tau	GSH	-38.66
Asc	Ins	-33.77
Ins	GPC_PCho	-31.05
Ins	Cr_PCr	-25.95
Glu	GSH	-23.97
GPC_PCho	SN	-23.06
Asp	SN	-21.4
Cohort	NAAG	-20.17
GPC_PCho	GSH	-19.93
Asp	NAAG	-16.47
NAAG	GSH	-15.29

Table 4.1: Arc strength, represented by change of BN score (BIC) after removal of defined arc

4.2 Model Evaluation: AD and Control Data

4.2.1 Random Forest

Train-Test Performance

Using a paired t-test, average differences of sensitivity, specificity, and accuracy for random forest models that do and do not use additional simulated training data are shown in Table 4.3. Average difference of specificity and accuracy did significantly differ in favor of modes that did not use simulated data under the Bonferroni corrected threshold for significance (p-value < 0.0028). Average difference of sensitivity between models that did not use simulated data and ones that did was 0.02 (p-value = 0.4). Mean difference of specificity was 0.06 (p-value = 8.42E-06). Mean difference of

Variable	Real AD	Simulated AD	Real Control	Simulated Control
percent_GM	42.02 σ 6.46	41.63 σ 6.73	46.75 σ 6.77	46.58 σ 6.71
percent_WM	26.65 σ 7.7	26.74 σ 7.68	23.88 σ 3.36	23.89 σ 3.28
percent_CSF	31.33 σ 7.75	31.62 σ 9.84	29.37 σ 8.15	29.53 σ 7.19
Asc	1.52 σ 0.23	1.53 σ 0.24	1.25 σ 0.17	1.25 σ 0.17
Asp	2.12 σ 0.36	2.08 σ 0.34	2.01 σ 0.35	2.02 σ 0.35
GABA	0.78 σ 0.24	0.81 σ 0.19	0.88 σ 0.16	0.87 σ 0.2
Gln	3.52 σ 0.56	3.53 σ 0.64	3.36 σ 0.39	3.36 σ 0.37
Glu	9.83 σ 1.12	9.92 σ 1.38	10.13 σ 0.89	10.1 σ 0.99
GSH	1.21 σ 0.14	1.21 σ 0.16	1.12 σ 0.14	1.12 σ 0.13
Ins	8.93 σ 1.72	9.01 σ 2.05	7.65 σ 0.82	7.61 σ 0.82
NAA	10.49 σ 1.19	10.69 σ 1.44	10.79 σ 0.99	10.73 σ 1.1
NAAG	1.02 σ 0.22	1.01 σ 0.21	0.99 σ 0.17	1 σ 0.19
PE	0.97 σ 0.22	1 σ 0.27	1.13 σ 0.24	1.11 σ 0.22
sIns	0.41 σ 0.18	0.43 σ 0.22	0.44 σ 0.25	0.43 σ 0.22
Tau	1.71 σ 0.31	1.84 σ 0.36	1.84 σ 0.3	1.77 σ 0.3
GPC_PCho	1.65 σ 0.24	1.65 σ 0.3	1.46 σ 0.16	1.46 σ 0.17
Cr_PCr	9.85 σ 1.22	9.98 σ 1.53	9.66 σ 0.93	9.57 σ 0.95
linewidth	8.43 σ 1.04	8.53 σ 1.09	8.49 σ 1.07	8.46 σ 1.07
SN	41.81 σ 7.53	43.11 σ 7.79	47.61 σ 6.13	46.84 σ 6.92

Table 4.2: Mean and standard deviation of real and simulated data grouped by AD status

overall accuracy was 0.05 (p-value = 8.67E-06).

Metric	Mean Difference (Model Not Using Simulated Data - Model Using Simulated Data)	p value
Sensitivity	0.02	0.36
Specificity	0.06	8.42E-06
Accuracy	0.05	8.67E-06

Table 4.3: AD and Control data RF train-test average performance comparison (paired t-test)

LOOCV Performance

Average sensitivity, specificity, and accuracy across random forest LOOCV runs that utilized simulated data and not are shown in Table 4.4 with p-values reported from t-tests. Average specificity and accuracy did significantly differ (p-value $<$ 0.0028) in favor of models that did not utilize simulated data. Average sensitivity for LOOCV runs that did and did not use additional simulated training data were 0.58 and 0.53 respectively (p-value = 0.08).Average specificity for LOOCV runs that do and do not use additional simulated training data were 0.76 and 0.90 respectively (p-value = 9.28E-07). Lastly, average accuracy for runs that did and did not use additional simulated training data were 0.70 and 0.78 respectively (p-value = 1.60E-4). Distributions of these performance metrics

are shown in Figure 4.4. Highest overall accuracy was found in LOOCV runs that did not utilize simulated data.

Metric	Mean Across LOOCV Runs Not Using Simulated Data	Mean Across LOOCV Runs Using Simulated Data	p value
Sensitivity	0.53	0.58	0.08
Specificity	0.90	0.76	9.28E-07
Accuracy	0.78	0.7	1.60E-04

Table 4.4: AD and Control data RF LOOCV average performance comparison (t-test)

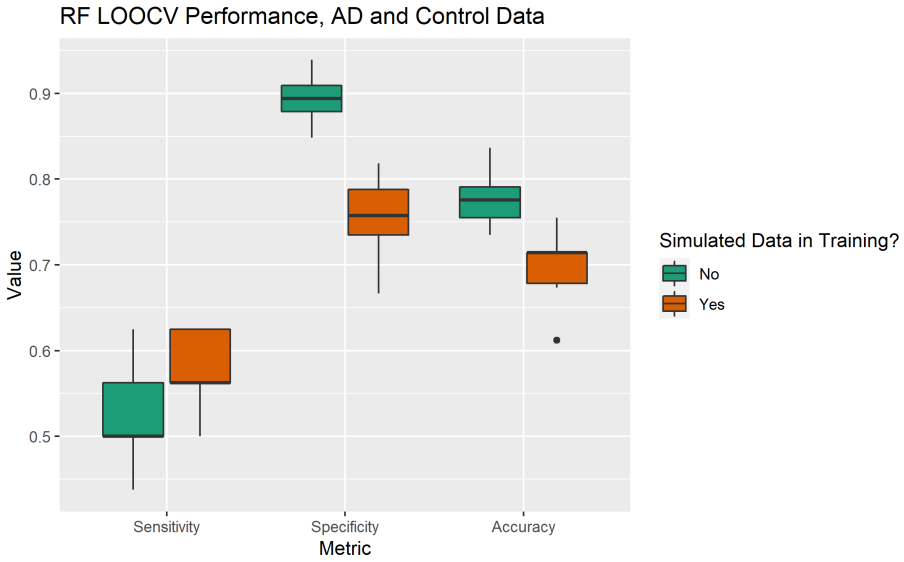

Figure 4.4: AD and Control data RF distributions of performance

Variable importance, represented by mean decrease in Gini coefficient, was recorded for each variable used in models within LOOCV runs. A comparison of average variable importance across runs that do and do not use additional simulated data in training are shown in Figure 4.5 . In order to make this comparison, values of mean decrease in Gini values were scaled for each model, as the value is partially dependent on the number of observations in the data. Ascorbate (Asc) and myo-Inositol (Ins) were the most important variables in both models utilizing simulated data and not. The variable with the largest difference of variable importance between models utilizing simulated data and not was percent white matter (percent-WM), which received a higher variable importance, on average, in models using simulated data.

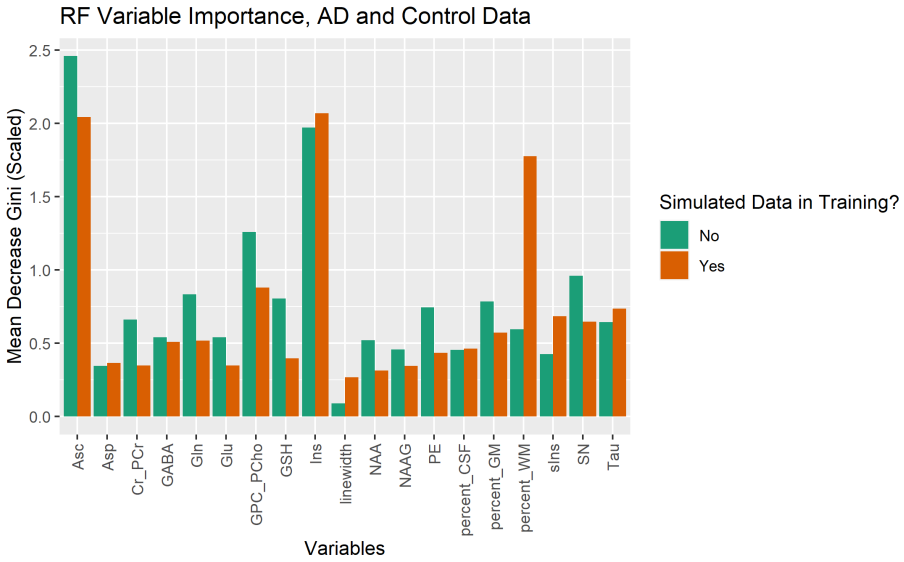

Figure 4.5: AD and Control data RF average variable importance (mean decrease Gini) across LOOCV runs

4.2.2 SVM

Train-Test Performance

Using a paired t-test, average differences of sensitivity, specificity, and accuracy for SVM models that do and do not use additional simulated training data are shown in Table 4.5. Average differences of specificity and accuracy were deemed significant (p-value < 0.0027) in favor of LOOCV runs that did not utilize simulated data. Mean difference of sensitivity between models that did not use simulated data and ones that did was 8.88E-18 (p-value = 1). Mean difference of specificity was 0.09 (p-value = 4.77E-09). Mean difference of overall accuracy was 0.06 (p-value = 2.52E-4).

Metric	Mean Difference (Model Not Using Simulated Data - Model Using Simulated Data)	p value
Sensitivity	8.88E-18	1
Specificity	0.09	4.77E-09
Accuracy	0.06	6.09E-08

Table 4.5: AD and Control data SVM train-test average performance comparison (paired t-test)

LOOCV Performance

Average sensitivity, specificity, and accuracy across SVM LOOCV runs that did and did not utilize simulated training data are shown in Table 4.6, with corresponding p-values from t-tests. Average specificity was the only metric to significantly differ (p-value < 0.0028) and favored runs that did not use simulated data. Average sensitivity for LOOCV runs that do and do not use additional simulated training data were 0.49 and 0.44 respectively (p-value = 0.3). Average specificity for LOOCV runs that did and did not use additional simulated training data were 0.78 and 0.92 respectively (p-value = 1.69E-4). Lastly, average accuracy for runs that did and did not use additional simulated training data were 0.69 and 0.76 (p = 5.52E-3). Distributions of these metrics are shown in Figure 4.6. Highest accuracy was recorded in a LOOCV run that did not utilize simulated data.

Metric	Mean Across LOOCV Runs Not Using Simulated Data	Mean Across LOOCV Runs Using Simulated Data	p value
Sensitivity	0.44	0.49	0.34
Specificity	0.92	0.78	1.69E-04
Accuracy	0.76	0.69	0.01

Table 4.6: AD and Control data SVM LOOCV average performance comparison (t-test)

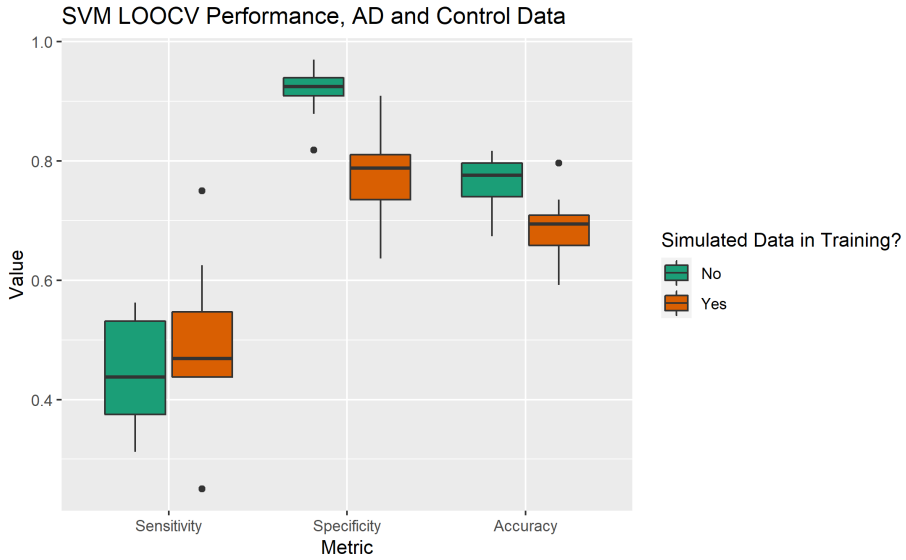

Figure 4.6: AD and Control data SVM distributions of performance

4.2.3 XGBoost

Train-Test Performance

Using a paired t-test, average difference of sensitivity, specificity, and accuracy for XGBoost models that do and do not use additional simulated training data are shown in Table 4.7. Average differences of specificity and accuracy were deemed significant in favor of models that did utilize simulated data (p-value < 0.0028). Mean difference of sensitivity between models that did not use simulated data and ones that did was -0.02 (p-value = 2). Mean difference of specificity was .06 (p-value = 1.11E-06). Mean difference of overall accuracy was .04 (p-value = 8.96E-4).

Metric	Mean Difference (Model Not Using Simulated Data - Model Using Simulated Data)	p value
Sensitivity	-0.02	0.24
Specificity	0.06	1.11E-06
Accuracy	0.04	8.96E-04

Table 4.7: AD and Control data XGBoost train-test average performance comparison (paired t-test)

LOOCV Performance

Average sensitivity, specificity, and accuracy across XGboost LOOCV runs that did and did not utilize simulated training data are shown in Table 4.8, with corresponding p-values from t-tests. Average sensitivity, specificity, and accuracy did not significantly differ (p-value > 0.0028). Average sensitivity for LOOCV runs that did and did not use additional simulated training data were .58 and .55 respectively (p-value = 0.2). Average specificity for LOOCV runs that did and did not utilize additional simulated training data were 0.79 and 0.82 respectively (p-value = 0.1). Average accuracy for runs that did and did not use additional simulated training data were 0.72 and 0.73 respectively (p-value =0.5). Distributions of these metrics are shown in Figure 4.7. Highest overall accuracy was achieved in a run that did not utilize simulated data.

Variable importance, represented by mean gain, was recorded for each variable used in models within XGBoost LOOCV runs. Gain represents the average gain of splits that use the variable. A comparison of average variable importance across runs that do and do not use additional simulated

Metric	Mean Across LOOCV Runs Not Using Simulated Data	Mean Across LOOCV Runs Using Simulated Data	p value
Sensitivity	0.55	0.58	0.18
Specificity	0.82	0.79	0.13
Accuracy	0.73	0.72	0.50

Table 4.8: AD and Control data XGBoost LOOCV average performance comparison (t-test)

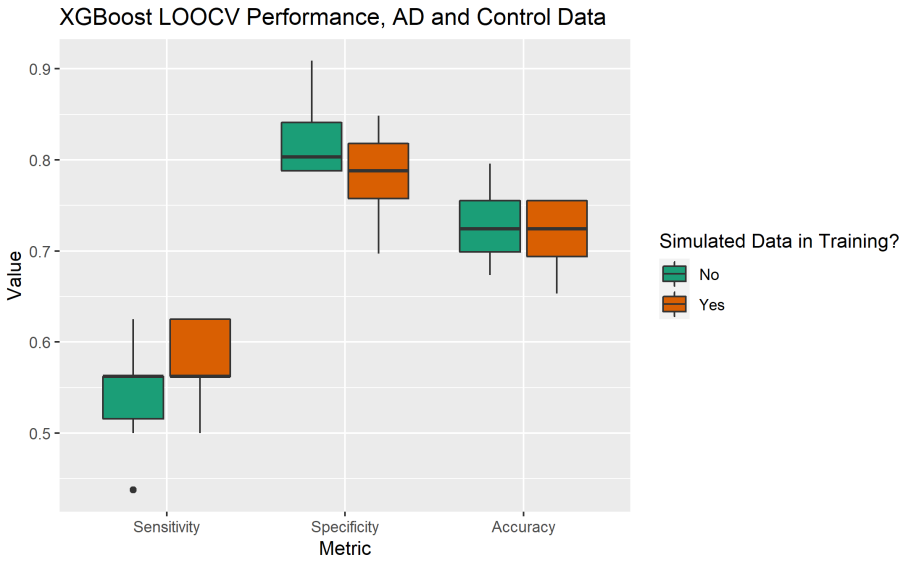

Figure 4.7: AD and Control data XGBoost distributions of performance

data in training are shown in Figure 4.8. Similar to random forest results, ascorbate (Asc) and myo-Inositol (Ins) were the most important variables in both models that utilized simulated data and not, with percent white-matter exhibiting the largest difference between the groups.

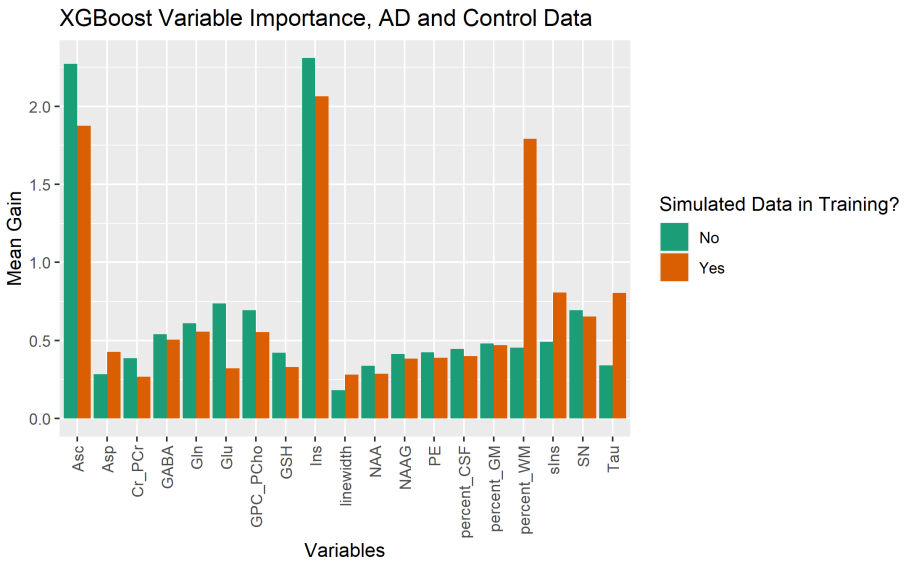

Figure 4.8: AD and control data XGBoost average variable importance (mean Gain) across LOOCV runs

4.3 Model Evaluation: Principal Component Data

4.3.1 Principal Component Data

Data were transformed to principal components via the "prcomp" function from the stats package in R. Principal components were able to capture 95 percent of the total variance in the AD and control data within 12 variables. These 12 variables were further used for simulation and modeling. Loadings, or the weight that each variable contributes to a principal component, of these 12 principal components are reflected in the Table 4.10.

4.3.2 Random Forest

Train-Test Performance

Using a paired t-test, average differences of sensitivity, specificity, and accuracy for random forest models that do and do not use additional simulated training data are shown in Table 4.9. Mean difference of specificity and accuracy did significantly differ for these groups (p-value $>$ .0028), favoring models that did not utilize simulated data. Average difference of sensitivity between models

that did not use simulated data and ones that did was -0.01 (p-value = 0.7) . Mean difference of specificity was 0.13 (p-value = 6.86E-14). Mean difference of overall accuracy was 0.08 (p-value = 3.34E-10).

Metric	Mean Difference (Model Not Using Simulated Data - Model Using Simulated Data)	p value
Sensitivity	-0.01	0.73
Specificity	0.13	6.86E-14
Accuracy	0.08	3.34E-10

Table 4.9: PC data RF train-test average performance comparison (paired t-test)

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12
percent_GM	-0.25	0.17	-0.36	-0.13	-0.17	-0.24	0.07	0.19	0.33	-0.26	0.04	-0.15
percent_WM	-0.04	-0.24	-0.26	0.34	0.46	0.24	0.01	0.09	-0.31	0.31	0.18	0.14
percent_CSF	0.24	0.01	0.49	-0.11	-0.15	0.05	-0.07	-0.23	-0.09	0.02	-0.16	0.04
Asc	0.25	-0.4	0.03	-0.01	-0.11	0.16	0.07	-0.16	0.04	0.02	-0.06	-0.27
Asp	0.23	-0.03	0	0.12	-0.18	0.12	0.77	-0.14	0.1	-0.18	0.08	0.41
GABA	0.1	0.37	-0.15	-0.01	0.14	0.19	-0.07	-0.39	-0.35	-0.52	0.26	-0.21
Gln	0.17	0.16	0.34	0.21	-0.03	0.15	0.09	0.63	-0.23	-0.23	-0.28	-0.26
Glu	0.31	0.33	-0.13	0.12	-0.03	-0.02	-0.05	0.02	0.1	0.24	0.05	-0.09
GSH	0.3	-0.12	-0.17	-0.19	0.23	-0.1	0.2	0.42	0.12	-0.07	0.21	-0.05
Ins	0.31	-0.31	-0.11	-0.03	-0.05	-0.03	-0.11	-0.19	0.12	-0.06	0.04	-0.3
NAA	0.31	0.3	-0.11	0.07	0.06	-0.08	0.01	-0.07	0.17	0.39	-0.22	-0.14
NAAG	0.21	0.13	0.15	0.38	-0.02	0.02	-0.45	0.09	0.34	-0.2	0.25	0.48
PE	0.09	0.35	0.16	-0.39	0.04	0.08	0.13	0.11	-0.19	0.27	0.46	-0.02
sIns	-0.02	-0.06	-0.19	0.03	-0.75	0.27	-0.14	0.15	-0.17	0.24	0.29	0
Tau	0.21	0.03	-0.29	-0.33	-0.08	-0.26	-0.14	0.04	-0.44	-0.03	-0.4	0.47
GPC_PCho	0.31	-0.24	-0.2	0.08	-0.09	-0.08	-0.14	0.16	-0.21	-0.29	0.1	-0.05
Cr_PCr	0.4	0.08	-0.13	-0.04	0.07	-0.09	-0.04	-0.08	0.16	0.07	0.01	-0.04
linewidth	0.01	0.04	-0.18	-0.37	0.11	0.76	-0.13	0.09	0.29	-0.09	-0.27	0.13
SN	-0.06	0.26	-0.31	0.44	-0.13	0.12	0.15	-0.1	-0.07	0.03	-0.31	-0.11

Table 4.10: Variable loadings from principal components created from AD and control data.

LOOCV Performance

Average sensitivity, specificity, and accuracy across random forest LOOCV runs that did and did not utilize simulated training data are shown in Table 4.11 with corresponding p-values from t-tests. Average specificity, and accuracy did significantly differ, in favor of models that did not utilize simulated data (p-value < 0.0028). Average sensitivity for LOOCV runs that do and do not use additional simulated training data were 0.6 and 0.53 respectively (p-value = .03). Average specificity for LOOCV runs that do and do not use additional simulated training data were 0.9 and 0.78 respectively (p-value of 4.50E-06) . Lastly, average accuracy for runs that did and did not use additional simulated training data were 0.80 and 0.70 respectively (p-value = 2.11E-07). Distributions of these metrics are shown in Figure 4.9. Highest value for accuracy was found in LOOCV runs that did not utilize simulated data.

Metric	Mean Across LOOCV Runs Not Using Simulated Data	Mean Across LOOCV Runs Using Simulated Data	p value
Sensitivity	0.6	0.53	0.03
Specificity	0.9	0.78	4.50E-06
Accuracy	0.80	0.70	2.11E-07

Table 4.11: PC data RF LOOCV average performance comparison (t-test)

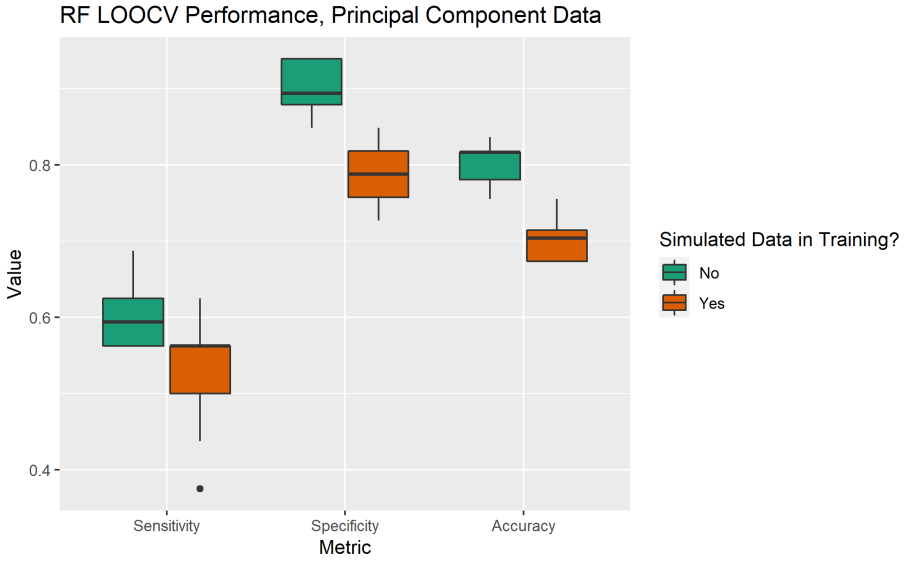

Figure 4.9: PC data RF distributions of performance

Variable importance, represented by mean decrease of the Gini coefficient, was recorded for

each variable used in random forest models within LOOCV runs. A comparison of average variable importance across runs that do and do not use additional simulated data in training are shown in Figure 4.10. PC2 was the most important variable in models utilizing simulated data and not. Variable weights for PC2 can be found in Table 4.10, which shows that ascorbate (Asc) had the highest contributing weight, followed respectively by GABA, glutamate (Glu), phosphorylethanolamine (PE) , and N-acetylaspartate (NAA).

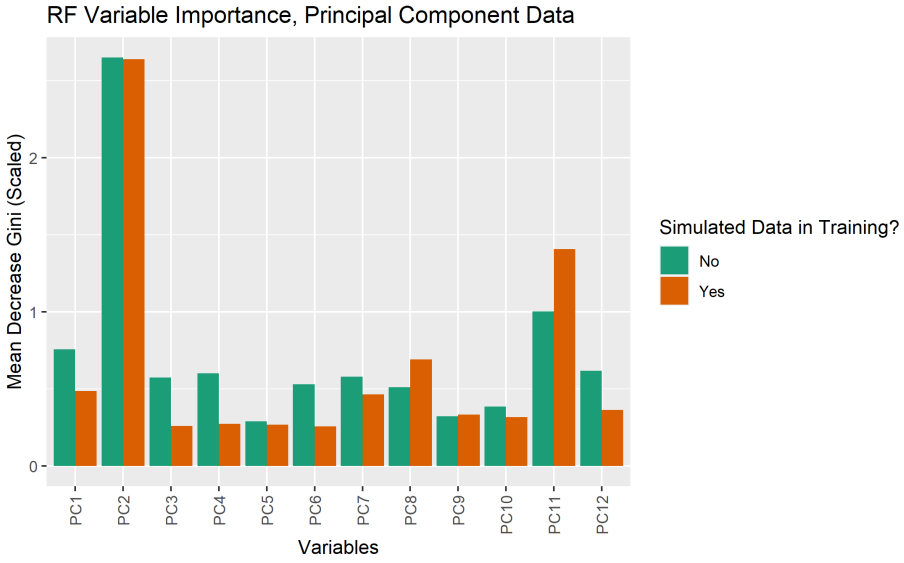

Figure 4.10: PC data RF average variable importance (mean decrease Gini) across LOOCV runs

4.3.3 SVM

Train-Test Performance

Using a paired t-test, average differences of sensitivity, specificity, and accuracy for SVM models that do and do not use additional simulated training data are shown in Table 4.12. Mean difference for specificity and accuracy did significantly differ for these groups ($p < 0.0028$), favoring models that do not utilize simulated data. Average difference of sensitivity between models that did not use simulated data and ones that did was 0.004 (p-value of 0.9) . Mean difference of specificity was 0.06 (p-value = 2.51E-3). Mean difference of overall accuracy was 0.04 (p-value = 2.52E-4).

Metric	Mean Difference (Model Not Using Simulated Data - Model Using Simulated Data)	p value
Sensitivity	0.004	0.91
Specificity	0.059	2.50E-03
Accuracy	0.042	2.52E-04

Table 4.12: PC data SVM train-test average performance comparison (paired t-test)

LOOCV Performance

Average sensitivity, specificity, and accuracy across SVM LOOCV runs that did and did not utilize additional simulated training data are recorded in Table 4.13, with corresponding p-values from t-tests. Average sensitivity, specificity, and accuracy did not significantly differ (p-value $>$ 0.0028). Average sensitivity for LOOCV runs that did and did not use additional simulated training data were 0.41 and 0.36 respectively (p-value = 0.3). Average specificity for LOOCV runs that did and did not use additional simulated training data were 0.81 and 0.82 respectively (p-value = 0.8) . Average accuracy for runs that did and did not use additional simulated training data were 0.68 and 0.67 respectively (p-value = 0.8). Distributions of these performance metrics are shown in Figure 4.11. Highest value of overall accuracy was found in a LOOCV run that did utilize simulated data.

Metric	Mean Across LOOCV Runs Not Using Simulated Data	Mean Across LOOCV Runs Using Simulated Data	p value
Sensitivity	0.37	0.41	0.34
Specificity	0.82	0.81	0.79
Accuracy	0.67	0.68	0.78

Table 4.13: PC data SVM LOOCV average performance comparison (t-test)

4.3.4 XGBoost

Train-Test Performance

Using a paired t-test, average differences of sensitivity, specificity, and accuracy for XGBoost models that did and did not use additional simulated training data are shown in Table 4.14. Mean difference for sensitivity, specificity, and accuracy did not significantly differ ($p < 0.0028$). Average difference of sensitivity between models that did not use simulated data and ones that did was 0.02 (p-value = 0.28) . Mean difference of specificity was 0.03 (p-value = 0.04). Mean difference of overall

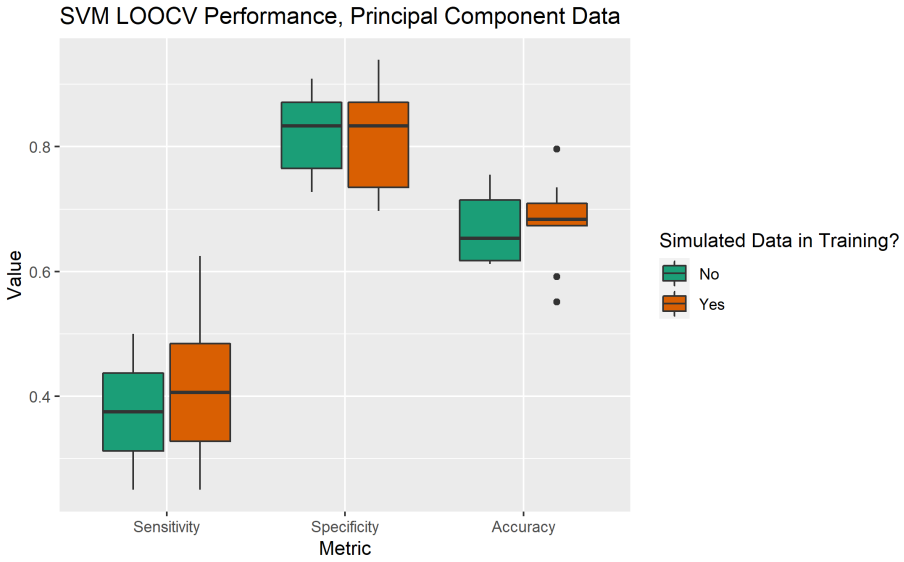

Figure 4.11: PC data SVM distributions of performance

accuracy was 0.03 (p-value = 0.01).

Metric	Mean Difference (Model Not Using Simulated Data - Model Using Simulated Data)	p value
Sensitivity	0.02	0.28
Specificity	0.03	0.04
Accuracy	0.03	0.01

Table 4.14: PC data XGBoost train-test average performance comparison (paired t-test)

LOOCV Performance

Average sensitivity, specificity, and accuracy across XGBoost LOOCV runs that did and did not utilize simulated data in model training are shown in Table 4.15 with corresponding p-values from t-tests. Average sensitivity, specificity, and accuracy did not significantly differ (p-value > 0.0028). Average sensitivity for LOOCV runs that did and did not use additional simulated training data were 0.56 and 0.61 respectively (p-value = 0.27). Average specificity for LOOCV runs that did and did not use additional simulated training data were 0.77 and 0.82 respectively (p-value = 0.04) . Lastly, average accuracy for runs that did and did not use additional simulated training data were 0.72 and 0.75 respectively (p-value = 0.02). Distributions of these performance metrics are shown in Figure 4.12. Highest value of overall accuracy was found in LOOCV runs that did not utilize

simulated data.

Metric	Mean Across LOOCV Runs Not Using Simulated Data	Mean Across LOOCV Runs Using Simulated Data	p value
Sensitivity	0.61	0.56	0.27
Specificity	0.82	0.77	0.04
Accuracy	0.75	0.72	0.02

Table 4.15: PC data XGBoost LOOCV average performance comparison (t-test)

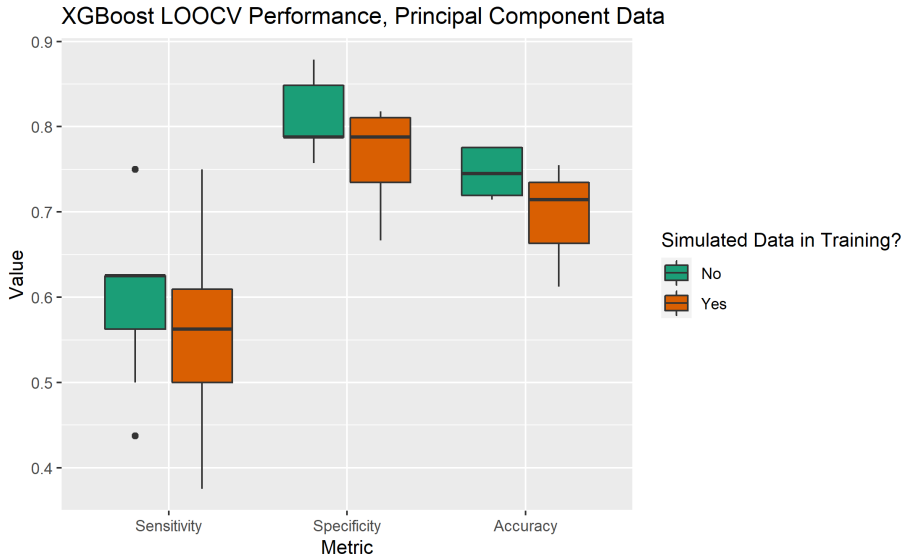

Figure 4.12: PC data XGBoost distributions of performance

Variable importance, represented by mean gain, was recorded for each variable used in XGBoost models within LOOCV runs. A comparison of average variable importance across runs that do and do not use additional simulated data in training are shown in Figure 4.13. Similar to random forest models trained and tested with principal component data, PC2 was the variable with highest average variable importance. There was a relatively large increase in variable importance of PC11, which is heavily influenced by phosphorylethanolamine (PE), Tau, and signal to noise ratio (SN), when utilizing simulated data.

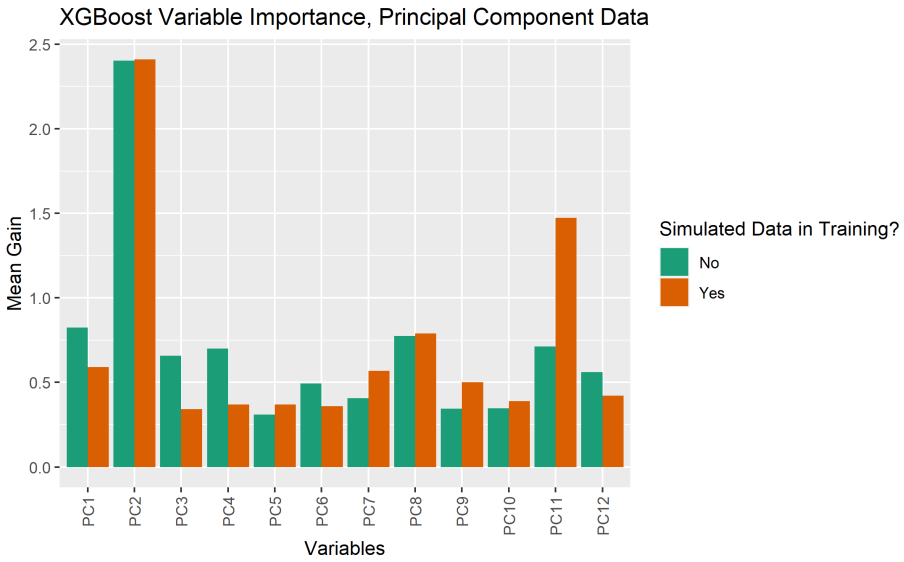

Figure 4.13: PC data XGBoost average variable importance (mean Gain) across LOOCV runs

4.4 Model Evaluation: AD and Control Data with Added Controls in Training and Simulation

4.4.1 Random Forest

Train-Test Performance

A Bayesian network that's structure was derived from AD and control data with additional controls from the Human Connectome Project-Aging is reported in Figure 4.14. This network contained more arcs that the network derived from AD and control data only, and is thus more

complex. Using a paired t-test, average differences of sensitivity, specificity, and accuracy for random forest models that did and did not use additional simulated training data are shown in Table 4.16. Mean difference sensitivity, specificity, and accuracy did significantly differ for these groups, with specificity and accuracy in favor of models that did not utilize simulated data. Average difference of sensitivity between models that did not use simulated data and ones that did was -0.20 (p-value =3.30E-13) . Mean difference of specificity was 0.19 (p-value = 1.64E-18). Mean difference of overall accuracy was 0.07 (p-value = 2.22E-06).

Metric	Mean Difference (Model Not Using Simulated Data - Model Using Simulated Data)	p value
Sensitivity	-0.20	3.30E-13
Specificity	0.19	1.64E-18
Accuracy	0.07	2.22E-06

Table 4.16: AD and Control Data With additional controls for training and/or simulation from HCPA random forest performance comparison on test set (paired t-tests)

LOOCV Performance

Average sensitivity, specificity, and accuracy across random forest LOOCV runs that did and did not utilize simulated data in model training are shown in Table 4.17 with corresponding p-values from t-tests. Average sensitivity and specificity did significantly differ, in favor of models not utilizing simualted data (p-value < 0.0028). Average sensitivity for LOOCV runs that did and did not use additional simulated training data were 0.61 and 0.43 respectively (p-value = 2.3E-4). Average specificity for LOOCV runs that did and did not use additional simulated training data were 0.73 and 0.89 respectively (p-value = 5.96E-05) . Lastly, average accuracy for runs that did and did not use additional simulated training data were 0.69 and 0.74 respectively (p-value = 0.03). Distributions of these performance metrics are shown in Figure 4.15. Models not utilizing simulated data had noticeably smaller distributions for sensitivity, specificity, and accuracy. Highest overall accuracy was achieved by a model utilizing simulated training data.

Variable importance, represented by mean decrease in the Gini impurity index, was recorded for each variable used in random forest models within LOOCV runs. A comparison of average variable importance across runs is shown in Figure 4.16. Linewidth notably displayed higher importance in

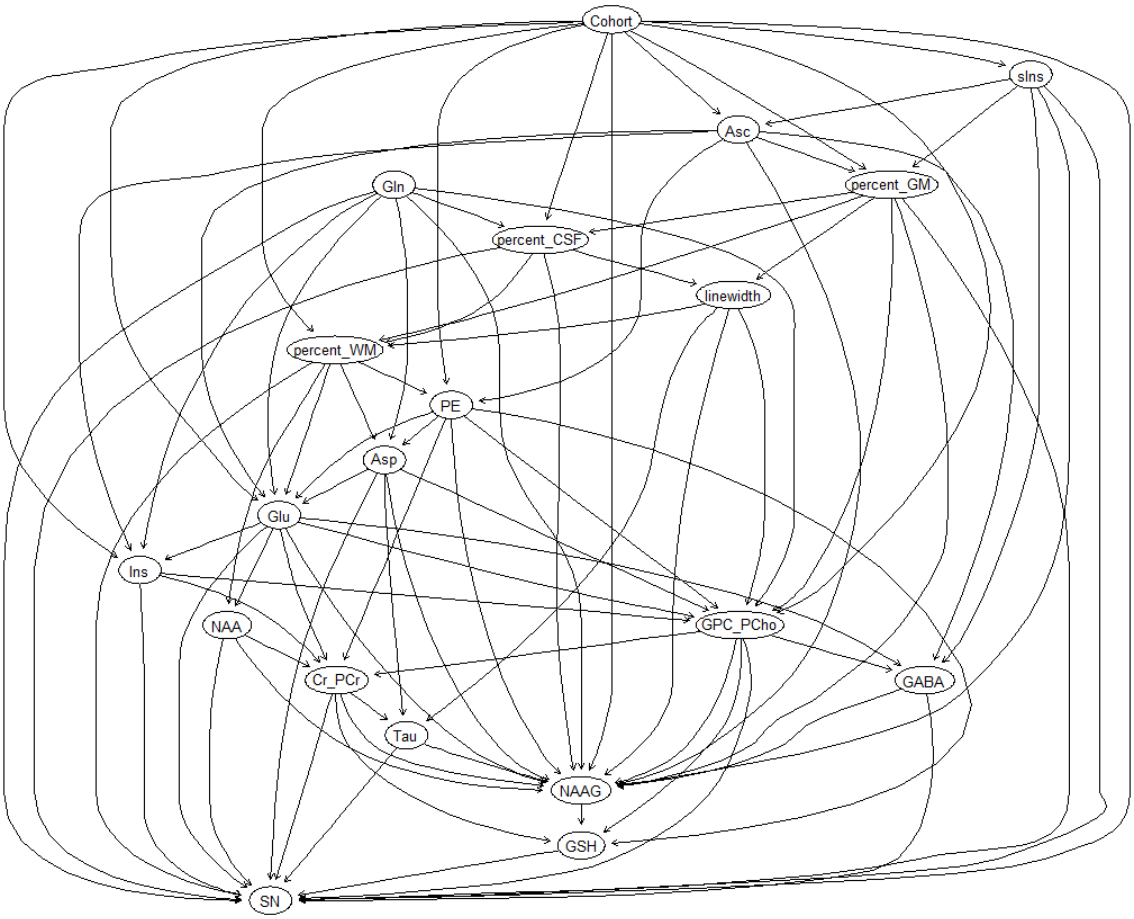

Figure 4.14: Bayesian Network derived from AD and Control Data with additional controls from HCPA

Metric	Mean Across LOOCV Runs Not Using Simulated Data	Mean Across LOOCV Runs Using Simulated Data	p value
Sensitivity	0.43	0.61	2.25E-04
Specificity	0.90	0.73	5.96E-05
Accuracy	0.74	0.69	0.03

Table 4.17: AD and Control data with additional HCPA controls RF LOOCV average performance comparison (t-test)

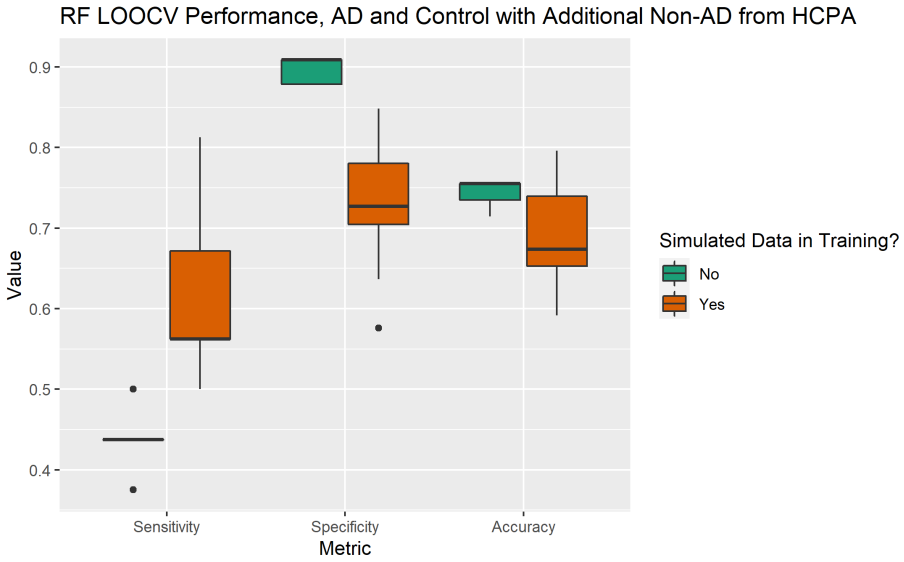

Figure 4.15: Random forest LOOCV Performance for AD and Control Data with additional controls from HCPA

models used in LOOCV using AD and control data with additional controls from HCPA compared to random forest only utilizing AD and control data. High variable importance was also observed in phosphorylethanolamine (PE), and the three MRI derived parameters.

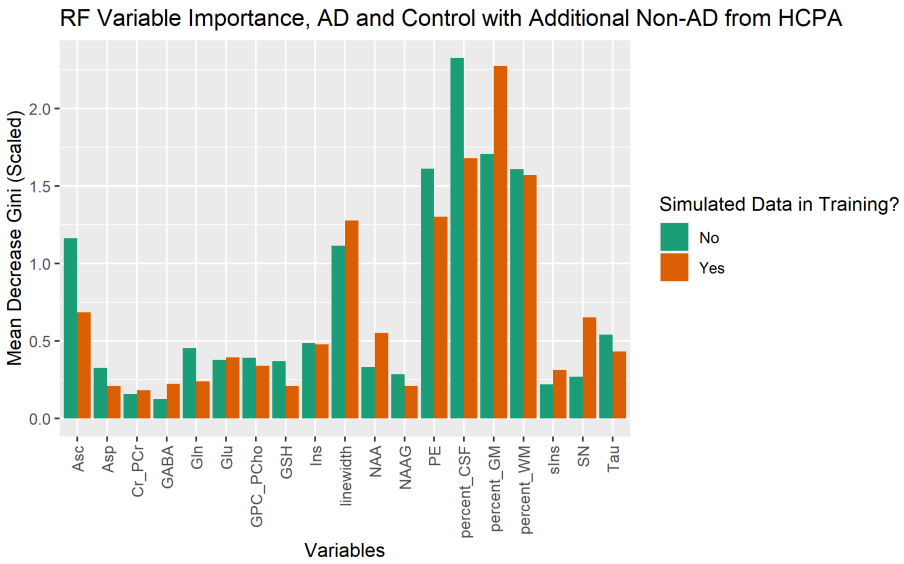

Figure 4.16: AD and Control data with additional HCPA controls RF average variable importance (mean decrease Gini) across LOOCV runs

4.4.2 SVM

Train-Test Performance

Using a paired t-test, average differences of sensitivity, specificity, and accuracy for SVM models that do and do not use additional simulated training data are shown in Table 4.18. Mean difference for specificity and accuracy did not significantly differ for these groups ($p > 0.0028$). Average difference for sensitivity between models that did not use simulated data and ones that did was 0.05 (p-value of 0.06) . Mean difference of specificity was 0.02 (p-value = .1). Mean difference of overall accuracy was 0.03 (p-value = 4.1E-3).

Metric	Mean Difference (Model Not Using Simulated Data - Model Using Simulated Data)	p value
Sensitivity	0.05	0.06
Specificity	0.02	0.10
Accuracy	0.03	4.13E-03

Table 4.18: AD and Control data with additional HCPA controls SVM train-test average performance comparison (paired t-test)

LOOCV Performance

Average sensitivity, specificity, and accuracy across SVM LOOCV runs that did and did not utilize simulated data in model training are shown in Table 4.19 with corresponding p-values from t-tests. Average specificity and accuracy did significantly differ (p-value < 0.0028). Average sensitivity for LOOCV runs that did and did not use additional simulated training data were 0.41 and 0.44 respectively (p-value = 0.5). Average specificity for LOOCV runs that did and did not use additional simulated training data were 0.81 and 0.92 respectively (p-value = 3.02E-06) . Lastly, average accuracy for runs that did and did not use additional simulated training data were 0.68 and 0.76 respectively (p-value = 3.55E-4). Distributions of these metrics are shown in Figure 4.17. Distributions of sensitivity between LOOCV runs utilizing similar data and not are quite similar. Highest overall accuracy was achieved by a model that did not use simulated data.

Metric	Mean Across LOOCV Runs Not Using Simulated Data	Mean Across LOOCV Runs Using Simulated Data	p value
Sensitivity	0.44	0.41	0.57
Specificity	0.92	0.81	3.02E-06
Accuracy	0.76	0.68	3.55E-04

Table 4.19: AD and Control data with additional HCPA controls SVM LOOCV average performance comparison (t-test)

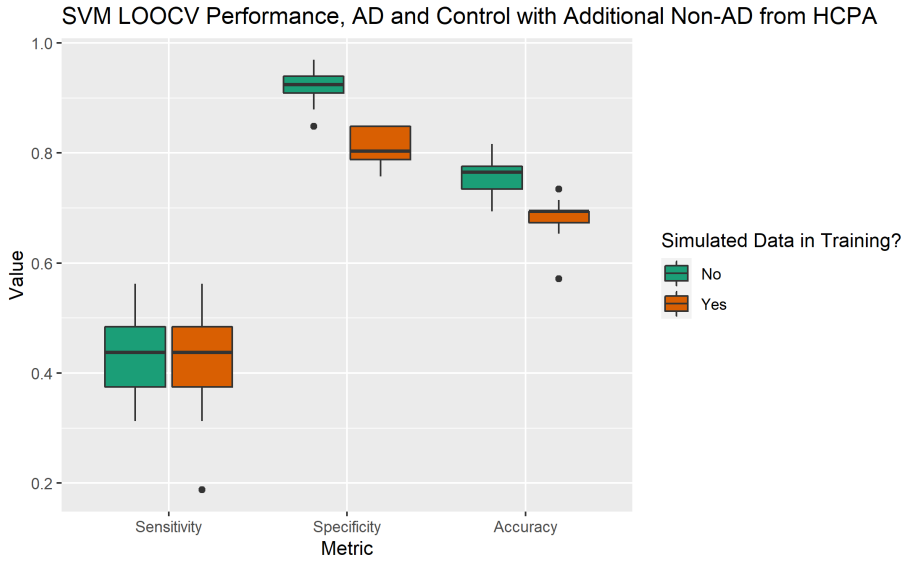

Figure 4.17: AD and Control data with additional HCPA controls SVM distributions of performance

4.4.3 XGBoost

Train-Test Performance

Using a paired t-test, average difference of sensitivity, specificity, and accuracy for XGBoost models that do and do not use additional simulated training data are shown in Table 4.20. Mean difference for specificity and accuracy did significantly differ for these groups ($p < 0.0027$). Average difference of sensitivity between models utilizing simulated data and not was -0.17 (p-value = 1.85E-14) . Mean difference of specificity was 0.20 (p-value = 3.65E-25). Mean difference of overall accuracy was .08 (p-value = 2.21E-11).

Metric	Mean Difference (Model Not Using Simulated Data - Model Using Simulated Data)	p value
Sensitivity	-0.17	1.85E-14
Specificity	0.20	3.65E-25
Accuracy	0.09	2.21E-11

Table 4.20: AD and Control data with additional HCPA controls XGBoost train-test average performance comparison (paired t-test)

LOOCV Performance

Average sensitivity, specificity, and accuracy across XGBoost LOOCV runs that did and did not utilize simulated data in model training are shown in Table 4.21 with corresponding p-values from t-tests. Average specificity, and accuracy did significantly differ (p-value < 0.0028). Average sensitivity for LOOCV runs that did and did not use additional simulated training data were 0.74 and 0.53 respectively (p-value = 1.86E-05). Average specificity for LOOCV runs that did and did not use additional simulated training data were 0.63 and 0.86 respectively (p-value = 3.03E-07). Average accuracy for runs that did and did not use additional simulated training data were 0.67 and 0.75 respectively (p-value = 5.37E-4). Distributions of these metrics are shown in Figure 4.18. LOOCV sensitivity for runs that utilized additional simulated training data performed with higher sensitivity, but lower specificity compared to runs that did not use simulated data. Highest overall accuracy was achieved by a model not utilizing simulated data.

Variable importance, represented by mean gain, was recorded for each variable used in

Metric	Mean Across LOOCV Runs Not Using Simulated Data	Mean Across LOOCV Runs Using Simulated Data	p value
Sensitivity	0.53	0.74	1.86E-05
Specificity	0.86	0.63	3.03E-07
Accuracy	0.75	0.67	5.37E-04

Table 4.21: AD and Control data with additional HCPA controls XGBoost LOOCV average performance comparison (t-test)

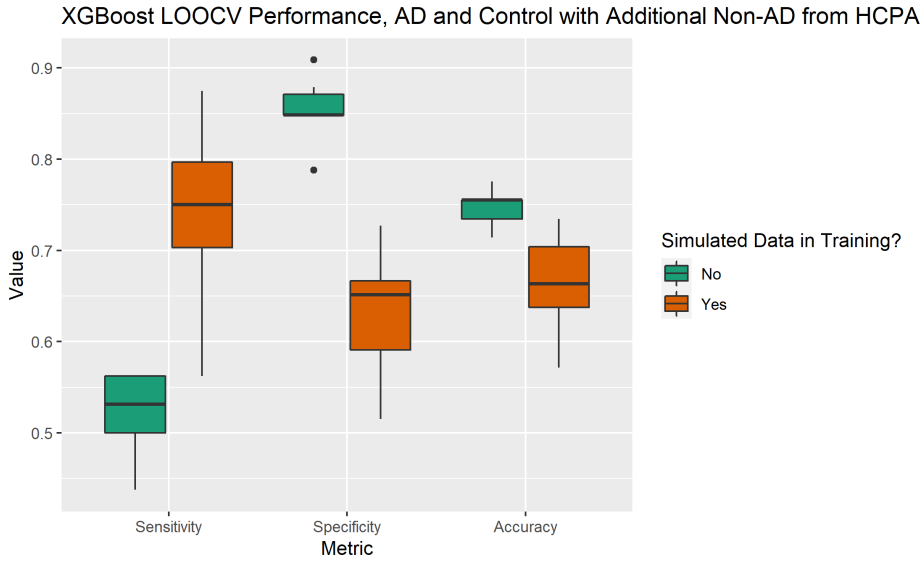

Figure 4.18: AD and Control data with additional HCPA controls XGBoost distributions of performance

XGBoost models within LOOCV runs. A comparison of average variable importance across runs that do and do not use additional simulated data in training are shown in Figure 4.19. Similar to random forest models trained and tested with AD and control data with additional controls for training and simulation from HCPA, high variable importance was found in MRI derived parameters and phosphorylethanolamine (PE).

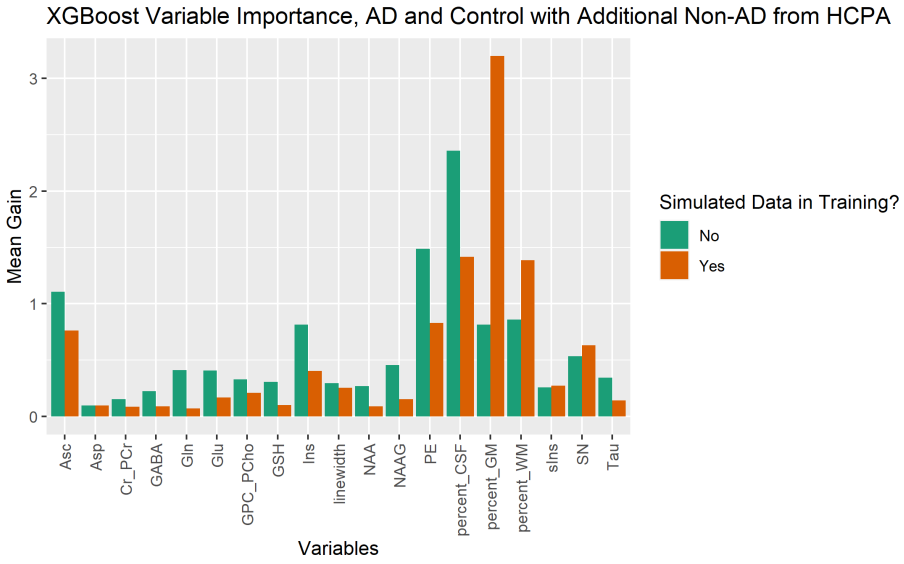

Figure 4.19: AD and Control with additional HCPA controls data XGBoost average variable importance (mean Gain) across LOOCV runs

4.5 Summary Figures

To provide a means of uncovering overall trends, figures are provided. Figure 4.20 shows average sensitivity across LOOCV runs that did and did not utilize simulated training data across every model type and data combination. Figure 4.21 shows average specificity across LOOCV runs that did and did not utilize simulated training data across every model type and data combination. Figure 4.22 shows average accuracy across LOOCV runs that did and did not utilize simulated training data across every model type and data combination. Figure 4.23 shows average difference of sensitivity, specificity, and accuracy from models that did not utilize simulated data and ones that did. In Figure 4.23 the test sets encompassed 35% of training data.

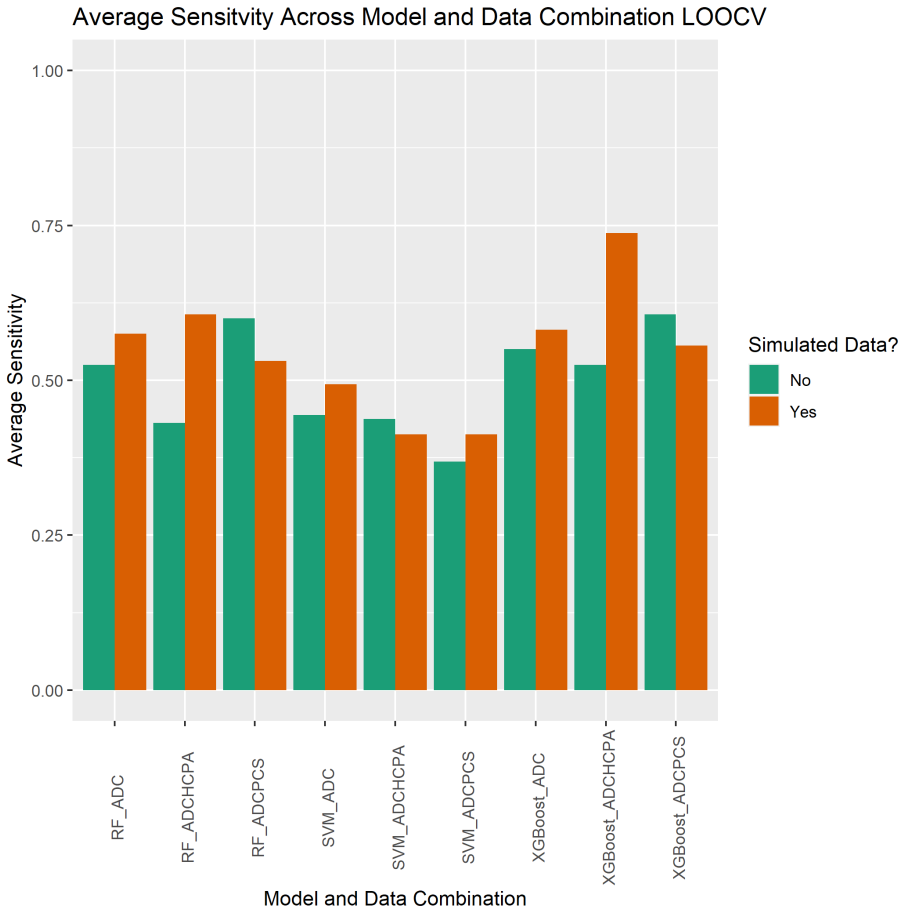

Figure 4.20: Average sensitivity of LOOCV runs that did and did not utilize simulated data for each model and data combination. Significance of these differences are recorded in the corresponding area of this chapter. AD = AD and control data, ADPCS = AD and control data reflected as principal components, ADCHCPA = AD and control with additional controls from HCPA in training and simulation

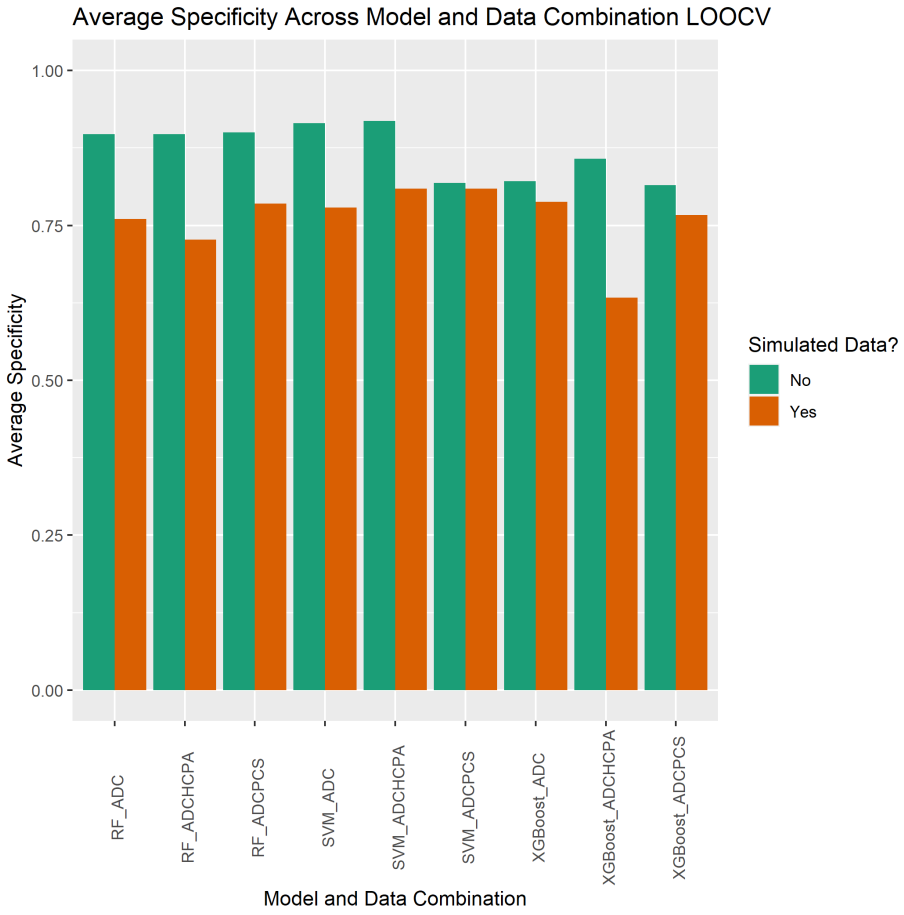

Figure 4.21: Average specificity of LOOCV runs that did and did not utilize simulated data for each model type and data combination. Significance of these differences are recorded in the corresponding area of this chapter. AD = AD and control data, ADPCS = AD and control data reflected as principal components, ADCHCPA = AD and control with additional controls from HCPA in training and simulation

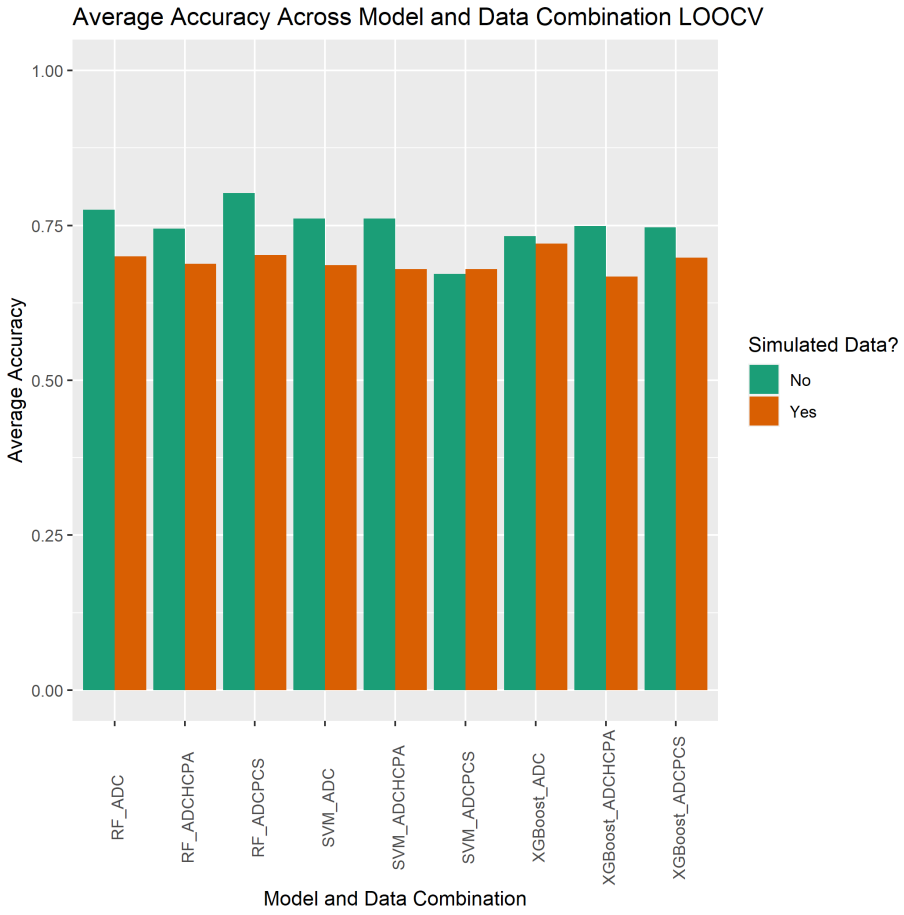

Figure 4.22: Average specificity of LOOCV runs that did and did not utilize simulated data for each model type and data combination. Significance of these differences are recorded in the corresponding area of this chapter.AD = AD and control data, ADPCS = AD and control data reflected as principal components, ADCHCPA = AD and control with additional controls from HCPA in training and simulation

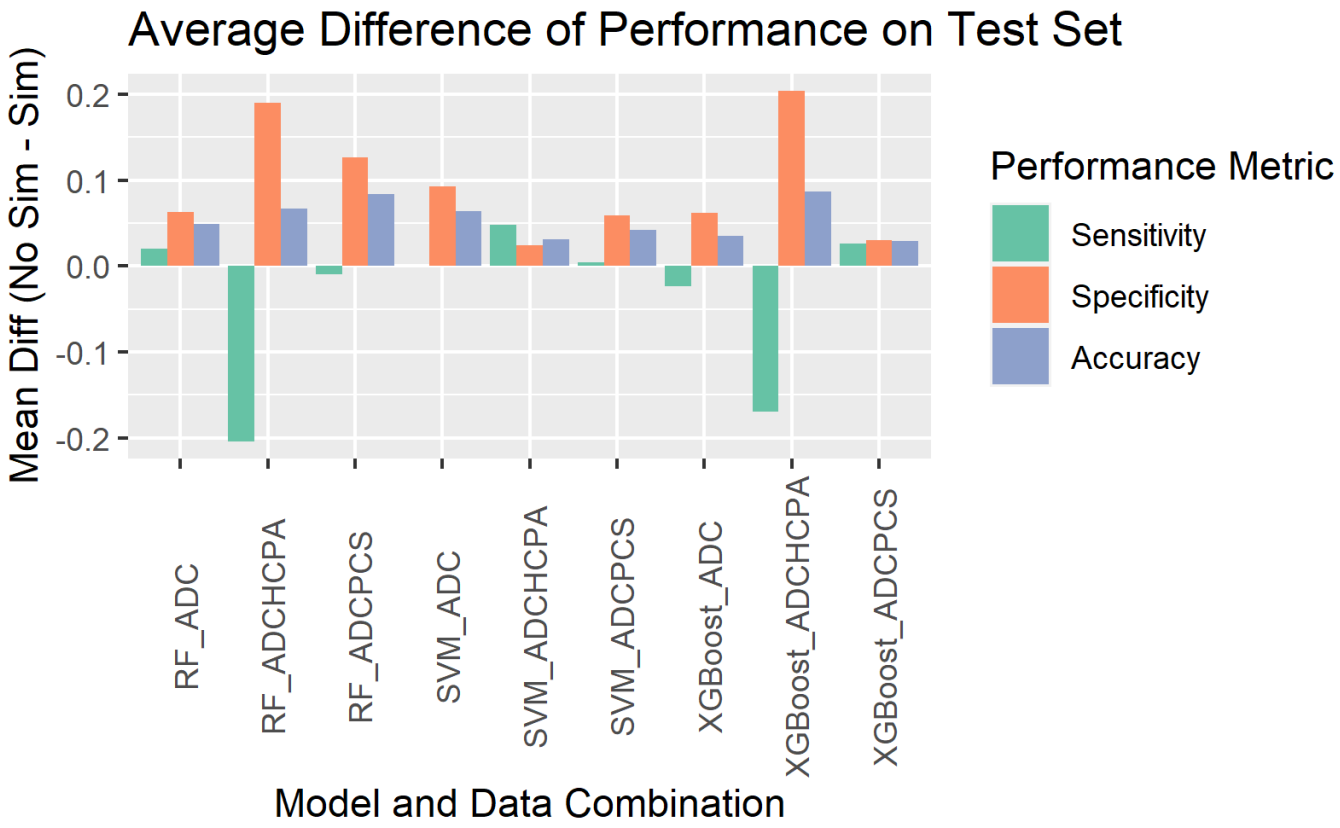

Figure 4.23: Average difference (result from model that did use simulation - result from model that did not use simulation) of sensitivity, specificity, and accuracy for each model type and data combination. AD = AD and control data, ADPCS = AD and control data reflected as principal components, ADCHCPA = AD and control with additional controls from HCPA in training and simulation

Chapter 5

DISCUSSION

The purpose of this research was to determine a hill-climbing approach the performs best in forming a Bayesian network of MRS and MRI derived parameters, and to determine if simulated data increased performance of classification models in uncovering AD status.

Using data in which each AD observation is doubled, thus combating class imbalance, hill climbing scored via BIC was the best performing hill climbing algorithm in terms of posterior predictive correlation of ascorbate, signal to noise, and percent white matter from the resulting network after fitting. Values for posterior predictive correlation did not widely differ between AIC and BIC scored networks when utilizing data in which each AD observation was over-sampled. Arcs for Bayesian networks created in this fashion show conditional dependencies of cohort (one's AD status) for each variable used to compare posterior predictive correlation. While these variables may therefore be important in classification models to distinguish AD status, it may be wise to revisit this analysis and compare performance with more variables including ones assumed to appear deeper within network structure. One of the benefits of simulating via Bayesian networks is the ability to use domain knowledge to manually require or disallow arcs between variables. Networks may benefit from establishing such known or assumed dependencies in Bayesian networks with assistance from domain experts. Additionally, structure learning throughout the whole project was determined by network performance on one data set ,AD and control data, when the process could be repeated for principal component data and the AD and control data with additional controls from the Human

Connectome Project in Aging used in training and simulation to ensure the best method of structure learning is being utilized in each case. Lastly, in future projects, highly correlated variables should be removed prior to fitting Bayesian networks,as the inclusion of highly correlated variables can increase false positive associations [81].

It was widely demonstrated that the inclusion of additional training data simulated from Bayesian networks created in the fashion outlined above did not lead to improved performance of classification models across random forest, support vector machines, and XGBoost models. It was also demonstrated that specificity was more widely and negatively affected than sensitivity when using simulated data in training. Of 18 total comparisons (3 models x 3 datasets evaluated in 2 ways) average specificity was significantly and negatively affected in 13 comparisons. Sensitivity only significantly differed in 4 comparisons, all of which favored the use of simulated data. This may suggest that Bayesian networks in this research are better able to characterize conditional dependencies of AD status for AD observations over controls. A direction of future work may then be to only simulate AD observations and see the resulting effect on specificity.

Random Forest and XGBoost models created using additional control data from The Human Connectome Project in Aging performed with better sensitivity when simulated data was utilized, though with lower average specificity. When simulating control observations, conditional dependencies likely more closely reflected data from the Human Connectome project in Aging, a study in which ages range from 36-100+ years, as the number of age matched controls were outnumbered by this group. Models trained on this simulated data could have then relied more on age-related changes in neurochemistry, gray matter, white matter, and cerebrospinal fluid when classifying control participants as such. This could have theoretically lead to the age matched controls being misclassified at a higher rate. This theory is partially supported by variable importance from models trained with AD and control data with additional controls. Phosphorylethanolamine (PE) notably received higher variable importance in these models compared to models trained only using AD and control data, a neurochemical previously demonstrated to exhibit age-related changes in the post-cingulate cortex region [67]. Percent cerebrospinal fluid, percent white matter, percent gray

matter,and linewidth were also given higher variable importance comparatively, all of which have demonstrated age-related changes [82, 83]. SVM models, not being a tree based method, were less impacted by these age related changes. A overarching limitation of incorporating this data includes fitting used for MRS, which is age dependant. Spectra fit with young macromolecules in HCPA data may not exhibit the same relationships found only in AD and control data. This should be investigated further.

Variable importance from random forest and XGBoost models that did and did not utilize additional simulated data for AD and control data suggest a strong use of ascorbate (Asc) and myo-inositol (Ins), consistent with previous findings [2]. When simulated data was utilized, there was a large uptick in variable importance for percent white matter in both random forest and XGboost models. This suggests that percent white matter is being utilized more in these models when training. Posterior predictive correlation for white matter using Bayesian Networks suggest a high reliability when simulating, which may explain why it was being used more in models that used simulated data. Splits in these tree based algorithms regarding percent white matter may not be applicable to test data, causing a decrease in specificity. Variable importance resulting from models that did not utilize simulated data can provide variables that should be most prioritized when evaluating fit of Bayesian Networks to be used for simulation. Additionally, variable importance could be utilized in future projects as a means of feature selection to reduce computation time, increase interpretability, and reduce noise. In particular, feature selection could be very influential for SVM performance, and the task should be revisited with this step included.

Though not a goal of this research, results can also be used to provide insight as to which models and data allowed for the best performance, depending on the goal of the algorithm's application. The model and data combination that resulted in the highest average sensitivity, 73.75%, was XGBoost applied to the AD and Control data with additional control from the Human Connectome project on aging and simulated training data. A possible cause for this increased sensitivity is outlined above. SVM models applied to AD and Control data with additional control from the Human Connectome project on aging without simulated training data gave rise to the highest average specificity, 91.81%.

Highest overall accuracy, 80.20%, was achieved via random forest applied to the principal component data. These findings should be further supported with statistical comparisons and effect size analysis.

While improving classification was largely not achieved, Bayesian networks may have other applications in MRS research. Biologically defended associations can be evaluated for strength via coefficients of fit of Gaussian network variables, for example. New insights may be able to be had if there is a strong association uncovered that has yet to be well documented. Additionally, it could be used to give rise to understanding how methods of fitting spectra influence relationships in the data.

In conclusion, hill-climbing score via BIC was deemed the best method of a structure learning by posterior predictive correlation of ascorbate, SN, and percent white matter, though more variables should likely be used in this decision process. Additional simulated training data did not lead to widespread significant increases in performance of random forest, support vector machines, and XGBoost classification models. Variable importance of random forest and XGBoost models may provide insights that can be used in defining features to be prioritized in Bayesian networks and feature selection for models, which may influence the effectiveness of additional simulated training data.

BIBLIOGRAPHY

[1] M. Scutari, M. M. Scutari, and H.-P. Mmpc, "Package 'bnlearn'," *Bayesian network structure learning, parameter learning and inference, R package version*, vol. 4, no. 1, 2019.

[2] M. Marjańska, J. R. McCarten, J. S. Hodges, L. S. Hemmy, and M. Terpstra, "Distinctive neurochemistry in alzheimer's disease via 7 T in vivo magnetic resonance spectroscopy," *J. Alzheimers. Dis.*, vol. 68, no. 2, pp. 559–569, 2019.

[3] S. Y. Bookheimer, D. H. Salat, M. Terpstra, B. M. Ances, D. M. Barch, R. L. Buckner, G. C. Burgess, S. W. Curtiss, M. Diaz-Santos, J. S. Elam, B. Fischl, D. N. Greve, H. A. Hagy, M. P. Harms, O. M. Hatch, T. Hedden, C. Hodge, K. C. Japardi, T. P. Kuhn, T. K. Ly, S. M. Smith, L. H. Somerville, K. Uğurbil, A. van der Kouwe, D. Van Essen, R. P. Woods, and E. Yacoub, "The lifespan human connectome project in aging: An overview," *Neuroimage*, vol. 185, pp. 335–348, Jan. 2019.

[4] P. Scheltens, B. De Strooper, M. Kivipelto, H. Holstege, G. Chételat, C. E. Teunissen, J. Cummings, and W. M. van der Flier, "Alzheimer's disease," *Lancet*, vol. 397, no. 10284, pp. 1577–1590, Apr. 2021.

[5] S. E. Counts, M. D. Ikonomovic, N. Mercado, I. E. Vega, and E. J. Mufson, "Biomarkers for the early detection and progression of alzheimer's disease," *Neurotherapeutics*, vol. 14, no. 1, pp. 35–53, Jan. 2017.

[6] J. M. Tognarelli, M. Dawood, M. I. F. Shariff, V. P. B. Grover, M. M. E. Crossey, I. J. Cox, S. D. Taylor-Robinson, and M. J. W. McPhail, "Magnetic resonance spectroscopy: Principles and techniques: Lessons for clinicians," *J. Clin. Exp. Hepatol.*, vol. 5, no. 4, pp. 320–328, Dec. 2015.

[7] E. Novotny, S. Ashwal, and M. Shevell, "Proton magnetic resonance spectroscopy: an emerging technology in pediatric neurology research," *Pediatr. Res.*, vol. 44, no. 1, pp. 1–10, Jul. 1998.

[8] M. A. Brown and R. C. Semelka, *MRI: Basic Principles and Applications*. John Wiley & Sons, Jan. 2011.

[9] A. D. Harris, M. G. Saleh, and R. A. E. Edden, "Edited 1 H magnetic resonance spectroscopy in vivo: Methods and metabolites," *Magn. Reson. Med.*, vol. 77, no. 4, pp. 1377–1389, Apr. 2017.

[10] F. Gao and P. B. Barker, "Various MRS application tools for alzheimer disease and mild cognitive impairment," *AJNR Am. J. Neuroradiol.*, vol. 35, no. 6 Suppl, pp. S4–11, Jun. 2014.

[11] R. Mullins, D. Reiter, and D. Kapogiannis, "Magnetic resonance spectroscopy reveals abnormalities of glucose metabolism in the alzheimer's brain," *Ann Clin Transl Neurol*, vol. 5, no. 3, pp. 262–272, Mar. 2018.

[12] K. S. Button, J. P. A. Ioannidis, C. Mokrysz, B. A. Nosek, J. Flint, E. S. J. Robinson, and M. R. Munafò, "Power failure: why small sample size undermines the reliability of neuroscience," *Nat. Rev. Neurosci.*, vol. 14, no. 5, pp. 365–376, May 2013.

[13] M. Khalilia, S. Chakraborty, and M. Popescu, "Predicting disease risks from highly imbalanced data using random forest," *BMC Med. Inform. Decis. Mak.*, vol. 11, p. 51, Jul. 2011.

[14] M. Scutari, "Learning bayesian networks with the bnlearn R package," Aug. 2009.

[15] T.-T. Wong, "Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation," *Pattern Recognit.*, vol. 48, no. 9, pp. 2839–2846, Sep. 2015.

[16] D. Geiger, T. Verma, and J. Pearl, "Identifying independence in bayesian networks," *Networks*, vol. 20, no. 5, pp. 507–534, Aug. 1990.

[17] R. Li, J. Yu, S. Zhang, F. Bao, P. Wang, X. Huang, and J. Li, "Bayesian network analysis reveals alterations to default mode network connectivity in individuals at risk for alzheimer's disease," *PLoS One*, vol. 8, no. 12, p. e82104, Dec. 2013.

[18] C. Zhang, C. Liu, X. Zhang, and G. Almpanidis, "An up-to-date comparison of state-of-the-art classification algorithms," *Expert Syst. Appl.*, vol. 82, pp. 128–150, Oct. 2017.

[19] R. E. Schapire, "The boosting approach to machine learning: An overview," in *Nonlinear Estimation and Classification*, D. D. Denison, M. H. Hansen, C. C. Holmes, B. Mallick, and B. Yu, Eds. New York, NY: Springer New York, 2003, pp. 149–171.

[20] Y. Qi, "Random forest for bioinformatics," in *Ensemble Machine Learning: Methods and Applications*, C. Zhang and Y. Ma, Eds. Boston, MA: Springer US, 2012, pp. 307–323.

[21] A. Sarica, A. Cerasa, and A. Quattrone, "Random forest algorithm for the classification of neuroimaging data in alzheimer's disease: A systematic review," *Front. Aging Neurosci.*, vol. 9, p. 329, Oct. 2017.

[22] A. Sarica, A. Cerasa, P. Valentino, J. Yeatman, M. Trotta, S. Barone, A. Granata, R. Nisticò, P. Perrotta, F. Pucci, and A. Quattrone, "The corticospinal tract profile in amyotrophic lateral sclerosis," *Hum. Brain Mapp.*, vol. 38, no. 2, pp. 727–739, Feb. 2017.

[23] M. Tanveer, B. Richhariya, R. U. Khan, A. H. Rashid, P. Khanna, M. Prasad, and C. T. Lin, "Machine learning techniques for the diagnosis of alzheimer's disease," *ACM Trans. Multimed. Comput. Commun. Appl.*, vol. 16, no. 1s, pp. 1–35, Jan. 2020.

[24] K. Thurnhofer-Hemsi, E. López-Rubio, M. A. Molina-Cabello, and K. Najarian, "Radial basis function kernel optimization for support vector machine classifiers," Jul. 2020.

[25] T. M. Deist, A. Patti, Z. Wang, D. Krane, T. Sorenson, and D. Craft, "Simulation-assisted machine learning," *Bioinformatics*, vol. 35, no. 20, pp. 4072–4080, Oct. 2019.

[26] B. G. Marcot and T. D. Penman, "Advances in bayesian network modelling: Integration of modelling technologies," *Environmental Modelling & Software*, vol. 111, pp. 386–393, Jan. 2019.

[27] M. Scutari, P. Auconi, G. Caldarelli, and L. Franchi, "Bayesian networks analysis of malocclusion data," *Sci. Rep.*, vol. 7, no. 1, pp. 1–11, Nov. 2017.

[28] D. Kaur, M. Sobiesk, S. Patil, J. Liu, P. Bhagat, A. Gupta, and N. Markuzon, "Application of bayesian networks to generate synthetic health data," *J. Am. Med. Inform. Assoc.*, vol. 28, no. 4, pp. 801–811, Mar. 2021.

[29] J. Young, P. Graham, and R. Penny, "Using bayesian networks to create synthetic data," http://www.scb.se/contentassets/ff271eeeca694f47ae99b942de61df83/using-bayesian-networks-to-create-synthetic-data.pdf, 2009, accessed: 2023-4-5.

[30] S. L. Ang, H. C. Ong, and H. C. Low, "Classification using the general bayesian network," *Pertanika J. Sci. Technol.*, vol. 24, no. 1, 2016.

[31] C. Bielza and P. Larrañaga, "Bayesian networks in neuroscience: a survey," *Front. Comput. Neurosci.*, vol. 8, p. 131, Oct. 2014.

[32] R. P. Adhitama and D. R. S. Saputro, "Hill climbing algorithm for bayesian network structure," *AIP Conf. Proc.*, vol. 2479, no. 1, p. 020035, Jul. 2022.

[33] J. A. Gámez, J. L. Mateo, and J. M. Puerta, "Learning bayesian networks by hill climbing: efficient methods based on progressive restriction of the neighborhood," *Data Min. Knowl. Discov.*, vol. 22, no. 1-2, pp. 106–148, Jan. 2011.

[34] M. Scanagatta, A. Salmerón, and F. Stella, "A survey on bayesian network structure learning from data," *Progress in Artificial Intelligence*, vol. 8, no. 4, pp. 425–439, Dec. 2019.

[35] P. Kaewprag, C. Newton, B. Vermillion, S. Hyun, K. Huang, and R. Machiraju, "Predictive models for pressure ulcers from intensive care unit electronic health records using bayesian networks," *BMC Med. Inform. Decis. Mak.*, vol. 17, no. Suppl 2, p. 65, Jul. 2017.

[36] Z. Liu, B. Malone, and C. Yuan, "Empirical evaluation of scoring functions for bayesian network model selection," *BMC Bioinformatics*, vol. 13 Suppl 15, no. Suppl 15, p. S14, Sep. 2012.

[37] M. Scutari, "Using custom scores in structure learning," https://www.bnlearn.com/examples/custom-score/, Nov. 2022.

[38] S. Beretta, M. Castelli, I. Gonçalves, R. Henriques, and D. Ramazzotti, "Learning the structure of bayesian networks: A quantitative assessment of the effect of different algorithmic schemes," *Complexity*, vol. 2018, Sep. 2018.

[39] S. Beretta, M. Castelli, I. Goncalves, I. Merelli, and D. Ramazzotti, "Combining bayesian approaches and evolutionary techniques for the inference of breast cancer networks," Mar. 2017.

[40] I. Tsamardinos, L. E. Brown, and C. F. Aliferis, "The max-min hill-climbing bayesian network structure learning algorithm," *Mach. Learn.*, vol. 65, no. 1, pp. 31–78, Oct. 2006.

[41] Y. Wang, "Analysis of the max-min hill-climbing algorithm," in *Proceedings of the 2018 International Conference on Transportation & Logistics, Information & Communication, Smart City (TLICSC 2018)*. Paris, France: Atlantis Press, Dec. 2018, pp. 509–511.

[42] W. Fu, Q. Kan, B. Li, and X. Zhang, "Prognosis model of advanced Non-Small-Cell lung cancer based on Max-Min Hill-Climbing algorithm," *Comput. Math. Methods Med.*, vol. 2022, p. 9173913, Mar. 2022.

[43] M. Gasse, A. Aussem, and H. Elghazel, "An experimental comparison of hybrid algorithms for bayesian network structure learning," in *Machine Learning and Knowledge Discovery in Databases*. Springer Berlin Heidelberg, 2012, pp. 58–73.

[44] M. Scutari, "Simulate random samples from a given bayesian network," https://www.bnlearn.com/documentation/man/rbn.html, 2023.

[45] A. Crawford, "Posterior predictive model checking in bayesian networks," Ph.D. dissertation, Arizona State University, Ann Arbor, United States, 2014.

[46] G. Biau and E. Scornet, "A random forest guided tour," *Test*, vol. 25, no. 2, pp. 197–227, Jun. 2016.

[47] H. Ali, M. N. M. Salleh, R. Saedudin, K. Hussain, and M. F. Mushtaq, "Imbalance class problems in data mining: A review," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 14, no. 3, pp. 1560–1571, 2019.

[48] S. RColorBrewer and M. A. Liaw, "Package 'randomforest'," *University of California, Berkeley: Berkeley, CA, USA*, 2018.

[49] R. Geetha, S. Sivasubramanian, M. Kaliappan, S. Vimal, and S. Annamalai, "Cervical cancer identification with synthetic minority oversampling technique and PCA analysis using random forest classifier," *J. Med. Syst.*, vol. 43, no. 9, p. 286, Jul. 2019.

[50] S. Kabiraj, M. Raihan, N. Alvi, M. Afrin, L. Akter, S. A. Sohagi, and E. Podder, "Breast cancer risk prediction using XGBoost and random forest algorithm," in *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, Jul. 2020, pp. 1–4.

[51] G. N. Dimitrakopoulos, A. G. Vrahatis, V. Plagianakos, and K. Sgarbas, "Pathway analysis using XGBoost classification in biomedical data," in *Proceedings of the 10th Hellenic Conference on Artificial Intelligence*, ser. SETN '18, no. Article 46. New York, NY, USA: Association for Computing Machinery, Jul. 2018, pp. 1–6.

[52] L. Akter and Ferdib-Al-Islam, "Dementia identification for diagnosing alzheimer's disease using XGBoost algorithm," in *2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD)*. ieeexplore.ieee.org, Feb. 2021, pp. 205–209.

[53] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. New York, NY, USA: Association for Computing Machinery, Aug. 2016, pp. 785–794.

[54] I. Ghosal and G. Hooker, "Boosting random forests to reduce bias; One-Step boosted forest and its variance estimate," *J. Comput. Graph. Stat.*, vol. 30, no. 2, pp. 493–502, Apr. 2021.

[55] T.-T. Dai and Y.-S. Dong, "Introduction of SVM related theory and its application research," in *2020 3rd International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE)*. ieeexplore.ieee.org, Apr. 2020, pp. 230–233.

[56] H. Talabani and E. Avci, "Impact of various kernels on support vector machine classification performance for treating wart disease," in *2018 International Conference on Artificial Intelligence and Data Processing (IDAP)*. ieeexplore.ieee.org, Sep. 2018, pp. 1–6.

[57] A. Patle and D. S. Chouhan, "SVM kernel functions for classification," in *2013 International Conference on Advances in Technology and Engineering (ICATE)*. ieeexplore.ieee.org, Jan. 2013, pp. 1–9.

[58] A. Murugan, S. A. H. Nair, and K. P. S. Kumar, "Detection of skin cancer using SVM, random forest and kNN classifiers," *J. Med. Syst.*, vol. 43, no. 8, p. 269, Jul. 2019.

[59] X. Yang, Q. Song, and Y. Wang, "A WEIGHTED SUPPORT VECTOR MACHINE FOR DATA CLASSIFICATION," *Int. J. Pattern Recognit Artif Intell.*, vol. 21, no. 05, pp. 961–976, Aug. 2007.

[60] I. Dimou, I. Tsougos, E. Tsolaki, E. Kousi, E. Kapsalaki, K. Theodorou, M. Kounelakis, and M. Zervakis, "Brain lesion classification using 3T MRS spectra and paired SVM kernels," *Biomed. Signal Process. Control*, vol. 6, no. 3, pp. 314–320, Jul. 2011.

[61] D. Berrar, "[no title]," https://www.researchgate.net/profile/Daniel-Berrar/publication/324701535_Cross-Validation/links/5cb4209c92851c8d22ec4349/Cross-Validation.pdf, 2019, accessed: 2023-4-5.

[62] Q. Lin, M. D. Rosenberg, K. Yoo, T. W. Hsu, T. P. O'Connell, and M. M. Chun, "Resting-State functional connectivity predicts cognitive impairment related to alzheimer's disease," *Front. Aging Neurosci.*, vol. 10, p. 94, Apr. 2018.

[63] Y. Zhang, S. Liu, and X. Yu, "Longitudinal structural MRI analysis and classification in alzheimer's disease and mild cognitive impairment," *Int. J. Imaging Syst. Technol.*, vol. 30, no. 2, pp. 421–433, Jun. 2020.

[64] A. Saribudak, A. A. Subick, and M. Ü. Uyar, "Computation of pharmacologic therapy effects on cognitive abilities of alzheimer's disease patients," in *2016 IEEE 16th International Conference on Bioinformatics and Bioengineering (BIBE)*. ieeexplore.ieee.org, 2016, pp. 129–136.

[65] "RAVLT," in *Encyclopedia of Clinical Neuropsychology*, J. S. Kreutzer, J. DeLuca, and B. Caplan, Eds. New York, NY: Springer New York, 2011, pp. 2115–2115.

[66] R. M. Reitan, *Trail Making Test: TMT*. Testzentrale, 1979.

[67] M. Marjańska, J. R. McCarten, J. Hodges, L. S. Hemmy, A. Grant, D. K. Deelchand, and M. Terpstra, "Region-specific aging of the human brain as evidenced by neurochemical profiles measured noninvasively in the posterior cingulate cortex and the occipital lobe using 1H magnetic resonance spectroscopy at 7 T," *Neuroscience*, vol. 354, pp. 168–177, Jun. 2017.

[68] R. Gruetter, "Automatic, localized in vivo adjustment of all first- and second-order shim coils," *Magn. Reson. Med.*, vol. 29, no. 6, pp. 804–811, Jun. 1993.

[69] M. Wilson, O. Andronesi, P. B. Barker, R. Bartha, A. Bizzi, P. J. Bolan, K. M. Brindle, I.-Y. Choi, C. Cudalbu, U. Dydak, U. E. Emir, R. G. Gonzalez, S. Gruber, R. Gruetter, R. K. Gupta, A. Heerschap, A. Henning, H. P. Hetherington, P. S. Huppi, R. E. Hurd, K. Kantarci, R. A. Kauppinen, D. W. J. Klomp, R. Kreis, M. J. Kruiskamp, M. O. Leach, A. P. Lin, P. R. Luijten, M. Marjańska, A. A. Maudsley, D. J. Meyerhoff, C. E. Mountford, P. G. Mullins, J. B. Murdoch, S. J. Nelson, R. Noeske, G. Öz, J. W. Pan, A. C. Peet, H. Poptani, S. Posse, E.-M. Ratai, N. Salibi, T. W. J. Scheenen, I. C. P. Smith, B. J. Soher, I. Tkáč, D. B. Vigneron, and F. A. Howe, "Methodological consensus on clinical proton MRS of the brain: Review and recommendations," *Magn. Reson. Med.*, vol. 82, no. 2, pp. 527–550, Aug. 2019.

[70] A. Gussew, M. Erdtel, P. Hiepe, R. Rzanny, and J. R. Reichenbach, "Absolute quantitation of brain metabolites with respect to heterogeneous tissue compositions in 1H-MR spectroscopic volumes," pp. 321–333, 2012.

[71] R. Ihaka and R. Gentleman, "R: A language for data analysis and graphics," *J. Comput. Graph. Stat.*, vol. 5, no. 3, pp. 299–314, Sep. 1996.

[72] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *Selected Papers of Hirotugu Akaike*, E. Parzen, K. Tanabe, and G. Kitagawa, Eds. New York, NY: Springer New York, 1998, pp. 199–213.

[73] G. Schwarz, "Estimating the dimension of a model," *Ann. Stat.*, vol. 6, no. 2, pp. 461–464, 1978.

[74] "[no title]," https://www.researchgate.net/profile/Russ-Greiner/publication/221345098_Model_Selection_Criteria_for_Learning_Belief_Nets_An_Empirical_Comparison/links/02e7e53a369b3a3b0c000000/Model-Selection-Criteria-for-Learning-Belief-Nets-An-Empirical-Comparison.pdf, accessed: 2023-3-19.

[75] R. D. Shachter and M. Peat, "Simulation approaches to probabilistic inference for general probabilistic inference on belief networks," *Fifth Workshop on Uncertainty in Artificial Intelligence*.

[76] A. Liaw and M. Wiener, "Classification and regression by randomforest," https://cogns.northwestern.edu/cbmg/LiawAndWiener2002.pdf, accessed: 2023-3-16.

[77] E. Dimitriadou, K. Hornik, F. Leisch, D. Meyer, and A. Weingessel, "The e1071 package," https://www.cs.upc.edu/~belanche/Docencia/mineria/Practiques/R/e1071.pdf, accessed: 2023-3-16.

[78] V. Apostolidis-Afentoulis and K.-I. Lioufi, "SVM classification with linear and RBF kernels," *July): 0-7. http://www. academia. edu/13811676/SVM Classification with Linear and RBF kernels. [21]*, 2015.

[79] J. Wainer and P. Fonseca, "How to tune the RBF SVM hyperparameters? an empirical evaluation of 18 search algorithms," *Artificial Intelligence Review*, vol. 54, no. 6, pp. 4771–4797, Aug. 2021.

[80] M. Kuhn, "The caret package," https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=e96acd28910c3fa7f5312d4b24915b7dd01d4044, accessed: 2023-3-16.

[81] H. Bae, S. Monti, M. Montano, M. H. Steinberg, T. T. Perls, and P. Sebastiani, "Learning bayesian networks from correlated data," *Sci. Rep.*, vol. 6, p. 25156, May 2016.

[82] C. R. Guttmann, F. A. Jolesz, R. Kikinis, R. J. Killiany, M. B. Moss, T. Sandor, and M. S. Albert, "White matter changes with normal aging," *Neurology*, vol. 50, no. 4, pp. 972–978, Apr. 1998.

[83] A. A. Maudsley, V. Govind, and K. L. Arheart, "Associations of age, gender and body mass with 1H MR-observed brain metabolites and tissue distributions," *NMR Biomed.*, vol. 25, no. 4, pp. 580–593, Apr. 2012.