Diagnosis of Alzheimer's Disease Based on a Parsimonious Serum Autoantibody Biomarker Derived from Multivariate Feature Selection

A. James Viscio

A Thesis

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Master of Science in Data Mining

Department of Mathematical Sciences

Central Connecticut State University

New Britain, Connecticut

December 2017

Thesis Advisor:

Dr. Darius M. Dziuda

Department of Mathematical Sciences

Diagnosis of Alzheimer's Disease Based on a Parsimonious Serum Autoantibody Biomarker Derived from Multivariate Feature Selection

A. James Viscio

An Abstract of a Thesis

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Master of Science in Data Mining

Department of Mathematical Sciences

Central Connecticut State University

New Britain, Connecticut

December 2017

Thesis Advisor:

Dr. Darius M. Dziuda

Department of Mathematical Sciences

Keywords: Alzheimer's Disease, blood-based diagnostic biomarker, autoantibodies, multivariate feature selection, recursive feature elimination, random forests, support vector machines.

Abstract

Alzheimer's Disease [AD] is a terminal neurodegenerative disorder comprising most dementia cases that is thus far without a known disease-modifying treatment. Major strides toward a potential cure have been made as research advances further elucidate the pathological intricacies of the disease including the identification of cerebral-spinal fluid [CSF] diagnostic biomarkers and neuroimaging prognostic biomarkers. However, these biomarkers are either too invasive or expensive for both widespread diagnostic screening of at-risk patients or enabling the large research sample sizes needed for an effective search for a disease modifying treatment. In addition to recent technological advances such as microarrays, the need for an inexpensive and non-invasive biomarker has prompted a surge of research focused on identifying and validating potential blood-based biomarkers of Alzheimer's Disease.

This thesis advances the research of a promising blood-based biomarker study (Nagele, E. et al., 2011), which identified a diagnostic biomarker of AD composed of 10 serum autoantibodies with 93.33% accuracy on both out-of-bag [OOB] and test set data. While several other contributions were made, the main contribution of this thesis was to replace the univariate feature selection filter, which arbitrarily selected the 10-autoantibody biomarker based on m-statistical analysis, with two multivariate feature selection wrappers that each identified a parsimonious biomarker determined heuristically to maximize the expectation of the value of the area under the ROC curve [AUC] on unseen samples: random forests – recursive feature elimination [RF-RFE] and support vector machines – recursive feature elimination [SVM-RFE]. Bootstrap aggregation [bagging] was also integrated into recursive feature elimination to lower the variance of the feature selection heuristic. The six-autoantibody random forests [RF] biomarker resulting from RF-RFE and the nine-autoantibody support vector machines [SVM]

biomarker resulting from SVM-RFE had mean OOB accuracy values of 0.9637 and 0.9929, respectively. Given that these results were validated successfully on the test set, both represent an improvement upon the 10-autoantibody biomarker from E. Nagele et al. (2011), which scored 0.9333 for both mean OOB and test set accuracies. The two improved biomarkers were also compared and validated based on sensitivity, specificity, and area under the ROC curve [AUC], and were interpreted using model-based variable importance rankings and two-way hierarchical clustering.

Acknowledgements

I would like to thank Professor Darius Dziuda for being an excellent thesis advisor and inspiring my interest in bioinformatics. His dedicated guidance and support made a big impact. I would also like to extend my gratitude to Professors Daniel Larose and Roger Bilisoly for serving on my thesis committee and for their feedback and good advice.

I would also like to thank my girlfriend of five years Christine, my parents, and my brother Peter for their unfailing support and encouragement throughout the thesis process and during my years of study. This accomplishment would not have been possible without them. Thank you.

Table of Contents

ABSTRACT **3**

TABLE OF CONTENTS **6**

LIST OF FIGURES **8**

LIST OF TABLES **10**

LIST OF ABBREVIATIONS **12**

1 INTRODUCTION **14**

1.1 ABOUT ALZHEIMER'S DISEASE 14

1.1.1 A Public Health and Economic Crisis 14

1.1.2 Alzheimer's Disease Pathology 16

1.1.3 Biomarkers Facilitate Research Advances 19

1.1.4 The Search for Multivariate Blood-based Biomarkers 23

1.2 LITERATURE REVIEW 25

1.2.1 Genomics Biomarker Studies 25

1.2.2 Proteomics Biomarker Studies 27

1.2.3 The Adaptive Immune System 33

1.3 DATASET 36

1.4 STATEMENT OF PURPOSE 37

2 METHODS **39**

2.1 ROBUST LINEAR MODEL NORMALIZATION 39

2.2 BOOTSTRAP AGGREGATING 43

2.3 RANDOM FORESTS 44

2.4 SUPPORT VECTOR MACHINES 47

2.5 RECURSIVE FEATURE ELIMINATION 51

2.6 GRID SEARCH 52

2.7 AREA UNDER THE ROC CURVE 53

2.8 HIERARCHICAL CLUSTERING WITH EUCLIDEAN DISTANCE AND WARD LINKAGE 56

3 EXPERIMENTS AND RESULTS 58

3.1 DATA PREPARATION 58

3.2 EXPLORATORY DATA ANALYSIS 59

3.3 PRE-PROCESSING 62

3.3.1 Background Correction 62

3.3.2 Normalization 65

3.4 BIOMARKER DISCOVERY 67

3.4.1 Feature Selection 69

3.4.2 Hyperparameter Optimization 78

3.4.3 Biomarker Selection 79

3.5 BIOMARKER INTERPRETATION 82

3.6 BIOMARKER VALIDATION 87

4 DISCUSSION 91

5 LIMITATIONS 94

REFERENCES 96

APPENDIX 108

BIOGRAPHICAL STATEMENT 109

List of Figures

Figure 1. Percent change in age-adjusted death rates in selected causes of death in the US 2000 vs 2010. 15

Figure 2. Bagged ensemble classifier algorithm. 43

Figure 3. Bagged model performance algorithm. 44

Figure 4. Random forest algorithm. 46

Figure 5. Recursive feature elimination with bagging. 52

Figure 6. Grid search algorithm. 53

Figure 7. Confusion matrix structure. 53

Figure 8. Test set ROC curve analysis. 56

Figure 9. Custom Data Preparation Function 59

Figure 10. Array boxplots of log2 transformed foreground autoantibody expression after data preparation. 60

Figure 11. Array boxplots of log2 transformed foreground control protein expression after data preparation. 62

Figure 12. Array boxplots of log2 transformed autoantibody expression after background correction. 63

Figure 13. Array boxplots of log2 transformed control protein expression after background correction. 65

Figure 14. Boxplots of the autoantibody feature intensities in each array after RLM normalization. 66

Figure 15. Random forest variable ranking algorithm. 71

Figure 16. Mean OOB performance per subset size during RF-RFE. 72

Figure 17. Strip plot of AUC performance per subset size during RF-RFE. 73

Figure 18. Mean OOB performance per subset size during SVM-RFE. 75

Figure 19. Strip plot of AUC performance per subset size during SVM-RFE. 76

Figure 20. 95% confidence intervals of the mean OOB performance measures for the RF and SVM biomarkers corresponding to their final model parameters. 80

Figure 21. Random forest biomarker autoantibodies ranked using the mean decrease in accuracy criterion. 83

Figure 22. Heatmap of two-way hierarchical clustering applied to the RF biomarker using Euclidean distance and Ward linkage. 84

Figure 23. SVM biomarker autoantibodies ranked by the square of their weight. 85

Figure 24. Heatmap of two-way hierarchical clustering applied to the SVM biomarker using Euclidean distance and Ward linkage. 86

Figure 25. Test set confusion matrix – RF. 88

Figure 26. Test set confusion matrix – SVM. 89

Figure 27. RF biomarker's test set ROC curve and threshold performance plot. 90

Figure 28. SVM biomarker's test set ROC curve and threshold performance plot. 90

List of Tables

Table 1. Average sample stats per class of autoantibody foreground expression after data preparation in $\log_2$ scale. 61

Table 2. Average sample stats per class of autoantibody expression after background correction in $\log_2$ scale. 63

Table 3. Average sample stats per class of autoantibody expression after RLM normalization. 66

Table 4. Distribution of Fisher's SNR before and after RLM Normalization. 67

Table 5. 92 training set samples. 68

Table 6. 31 test set samples. 68

Table 7. RFE initializations independent of learning algorithm. 70

Table 8. RF-specific RFE initializations. 70

Table 9. SVM-specific RFE initializations. 71

Table 10. The six-autoantibody random forest biomarker. 74

Table 11. The nine-autoantibody support vector machine biomarker. 77

Table 12. Parameter tuning grids of final classifiers. 78

Table 13. SVM grid search results for classifier trained on SVM biomarker. 78

Table 14. RF grid search results for classifier training on RF biomarker. 79

Table 15. Mean OOB performance of RF and SVM biomarkers. 79

Table 16. Paired t-test information and 95% confidence intervals of the mean difference in each performance metric between the final two classifiers. RF - SVM. 81

Table 17. The four autoantibody clusters of the RF biomarker. 85

Table 18. The five clusters of features in the SVM biomarker. 87

Table 19. Test set classification probabilities per final classifier. 88

Table 20. Test set performance summary of the two final classifiers. 89

Table 21. Mean OOB performance of the biomarker identified by E. Nagele et al. (2011) and the two identified in this project. AUC not measured by Nagele et al. (2011).......................... 92

Table 22. Test set performance of the biomarker identified by E. Nagele et al. (2011) and the two identified in this project. AUC not measured by Nagele et al. (2011).......................... 93

List of Abbreviations

Abbreviation	Description
Biomarker	Biological marker
AD	Alzheimer's Disease
PD	Parkinson’s Disease
BC	Breast Cancer
NDC	Non-demented control
MCI	Mild Cognitive Impairment
AD/MCI	Alzheimer’s Disease or Mild Cognitive Impairment
DAT	Dementia of Alzheimer’s type
ADNI	Alzheimer’s Disease Neuroimaging Initiative
Aβ	Amyloid-beta
$A\beta_{42}$	Amyloid-beta with 42 amino acids
$A\beta_{40}$	Amyloid-beta with 40 amino acids
CSF	Cerebral-spinal fluid
CSF t-tau	Amount of total tau in cerebral-spinal fluid
CSF p-tau	Prevalence of phosphorylation among tau in cerebral-spinal fluid
ELISA	Enzyme-linked immunosorbent assays
PET	Positron emission tomography
MRI	Magnetic resonance imaging
DNA	Deoxyribonucleic acid
RNA	Ribonucleic acid
mRNA	Messenger ribonucleic acid
APoE4	A mutation in the ε4 allele of apoliprotein E
SNP	Single-nucleotide polymorphism
BNP	Brain natriuretic peptide
CNP	C-reactive protein
IL-3	Interleukin-3
BBB	Blood-brain barrier
GWAS	Genome-wide association analysis
RLM	Robust linear model

Abbreviation	**Description**
SNR	Signal-to-noise ratio
QC	Quality control
CoV	Coefficient of Variation
PAM	Prediction Analysis of Microarrays
CV	Cross-validation
OOF	Out-of-fold
OOB	Out-of-bag
Bagging	Bootstrap aggregating
AUC	Area under the ROC curve
PLS	Partial-least-squares
CART	Classification and Regression Tree
GEO	Gene Expression Omnibus
GPR	Genepix Results
RFE	Recursive feature elimination
RF	Random forests
RF-RFE	Random forests – recursive feature elimination
SVM	Support vector machines
SVM-RFE	Support vector machines – recursive feature elimination

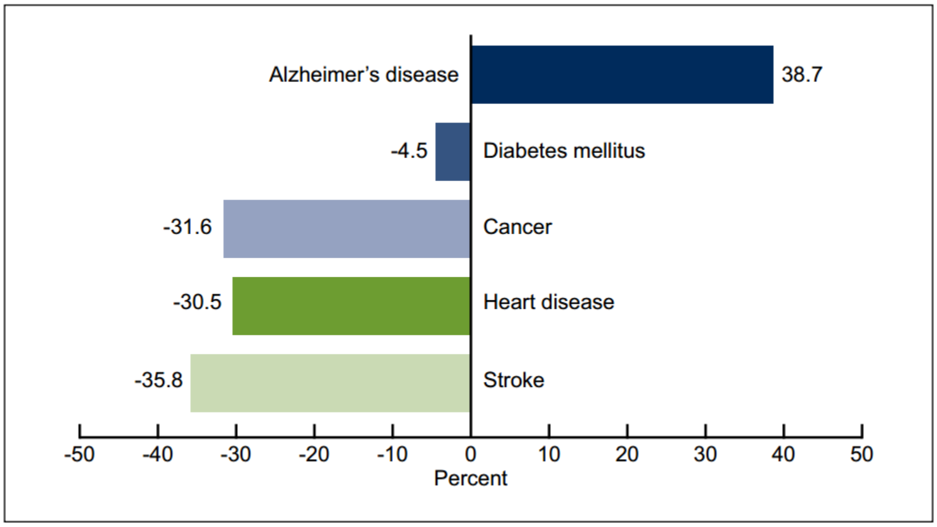
1 Introduction

1.1 About Alzheimer’s Disease

1.1.1 A Public Health and Economic Crisis

Alzheimer’s Disease is a terminal neurodegenerative disorder characterized by a progressive decline in memory and cognition. It is the most common type of dementia, accounting for 60-80% of cases (Alzheimer's Association, 2015). The other types are Vascular Dementia, Dementia with Lewy Bodies, Frontotemporal Dementia, and Creutzfeldt-Jakob Disease. The latest statistics provided by the Centers for Disease Control and Prevention (2017) report Alzheimer’s Disease as the 5th leading cause of death for Americans age 65 and older behind heart disease, cancer, chronic lower respiratory diseases, and stroke, in that order. However, this statistic severely underrepresents the dire and escalating public health threat of Alzheimer’s Disease, which is arguably already greater than even heart disease and cancer.

For one, while promising medical advancements in heart disease and cancer research have buffered the mortality rate for those disease, Alzheimer’s disease is the only condition among the top ten causes of death that has no cure, no means of prevention, and no disease-modifying treatments (UsAgainstAlzheimer's, 2012). With age as its primary risk factor and life expectancy increasing, Alzheimer’s Disease is on track to soon become the top cause of death in America. This trend can be visualized in the following graph taken from the National Center for Health Statistics Data Brief No. 116 (Tejada-Vera, 2013), which displays the percent change in top causes of death between 2000 and 2010, adjusted for age:



SOURCE: National Vital Statistics System, Mortality.

Figure 1. Percent change in age-adjusted death rates in selected causes of death in the US 2000 vs 2010.

It should be noted that part of the reason for this shift can be attributed to the severe discrepancy in research funding between the major diseases. The U.S. government spends over $6 billion a year on cancer research, over $4 billion a year on heart disease research, over $3 billion a year on HIV research, and only $480 million a year on Alzheimer's research (Fagan, 2014).

The causes of death statistics also belie the comparatively higher emotional and financial toll the disease takes on patients and their families. Few would disagree that the progressive, slow, unstoppable deterioration of the patient's mental faculties, memories, and ability to take care of themselves is a more extreme form of suffering than diseases primarily restricted to the physical body. The patient's family members also must watch and care for their loved one as they go through this horrible condition, only to be eventually forgotten by their family member with the disease, often for many years before their loved one passes away.

A good example of the comparatively higher financial costs is the study conducted by Kelly, McGarry, Gorges, & Skinner (2015), which measured the healthcare costs in the last 5 years of life of 1,702 Medicare fee-for-service beneficiaries, aged 70 years or older, who died between 2005 and 2010. The patients were stratified into four groups according to cause of death: dementia, cancer, heart disease, or other causes. The average cost per decedent with dementia ($287,038) was much greater than the average cost of those who died of heart disease ($175,136), cancer ($173,383), or other causes ($197,286). Many of these expenses were not covered by the patients' insurance plans. For Alzheimer's Disease, the average out-of-pocket spending was 32% of household wealth. Further, the disease lasts an average of 8-10 years after the initial diagnosis and can last up to 20 years or more (Mayo Clinic, 2015), much longer than the 5 years studied.

The true financial burden of the disease is also much greater than the official costs indicated by Kelly et al. (2005). The Alzheimer's Association estimated that 17.9 billion hours of unpaid care was given to Alzheimer's patients in 2014 by 15 million family members and other caregivers, which is a contribution valued at over $217 billion. In 2015, total payments for health care, long-term care, and hospice services for dementia patients age 65 and over were expected to be $226 billion (Alzheimer's Association, 2015), which is more than 400 times the amount spent on trying to find a cure (UsAgainstAlzheimer's, 2012).

1.1.2 Alzheimer's Disease Pathology

There are two defining hallmarks of dementia of Alzheimer's type: the presence in the brain parenchyma of both extracellular amyloid plaques composed of aggregated amyloid-β [Aβ] peptides and intracellular neurofibrillary tangles composed of hyperphosphorylated tau (Fagan & Holtzman, 2010). Aβ are secreted peptides composed of 37-43 amino acids that are formed when the large amyloid-precursor protein is cleaved into peptide fragments by secretases (Clark & Kodadek, 2013). The most abundant species of Aβ has 40 amino acids [$A\beta_{40}$], but it is version with 42 amino acids [$A\beta_{42}$] that most contributes to intraneuronal deposition of Aβ due to its increased tendency to form plaques (Clark & Kodadek, 2013). While the exact nature of the underlying processes behind intraneuronal Aβ accumulation is not known, research has revealed that the majority of Aβ is produced in the brain and, under healthy conditions, is prevented from aggregating by various degradation and clearance mechanisms (Ritter & Cummings, 2015). As $A\beta_{42}$ accumulates in the brain, the mechanisms which normally degrade and clear excess $A\beta_{42}$ from the brain into the interstitial space begin to malfunction under too much stress. The result is between 100 and 1,000-fold increases in the quantity of Aβ in the brain (Ritter & Cummings, 2015), although the precise relative contributions to intraneuronal accumulation of $A\beta_{42}$ from increased amyloid production vs. reduced amyloid clearance have yet to be determined.

The burden of too much intraneuronal Aβ has two profound consequences. For one, $A\beta_{42}$ fragments increasingly aggregate into oligomers of various sizes and shapes. Per Ritter and Cummings (2015), these oligomers exert effects typical of the pathophysiology of Alzheimer's such as neurotoxicity, memory inhibition, and inhibition of synaptic function despite not being considered among the established hallmarks of the disease. In fact, after exposing Aβ oligomers to hippocampal neurons, Esparza et al. (2013) demonstrated that Aβ oligomers were more neurotoxic than monomeric or fibrillar forms of amyloid.

The second result of rising amyloid concentrations in the brain is that Aβ fragments increasingly fibrillize into cross-β-sheets, creating amyloid plaques, the first hallmark of the disease. The normal physiological function of Aβ plaques is theorized to be reservoirs of soluble amyloid, which serve as a buffer against sudden changes in the amount of circulating amyloid (Ritter & Cummings, 2015). This is useful because there is 100-fold more insoluble fibrillar amyloid in the brain than soluble Aβ (Ritter & Cummings, 2015). Despite their theorized utility, Aβ plaques are also locally neurotoxic (Ritter & Cummings, 2015). As plaque-load increases with disease progression, the local neurotoxicity gradually compounds and becomes an integral factor driving further pathological progression.

Tau is an intracellular protein, whose various isoforms are ubiquitous in neurons and present, but much less abundant, in astrocytes and oligodendrocytes (Avila, 2004). Tau provides the vital function of stabilizing microtubule proteins, which is achieved through isoforms and various post-translational modifications such as phosphorylation (Hampel et al., 2010). Through its role in stabilizing microtubule proteins, tau indirectly helps maintain neuronal integrity, cellular signaling, and axonal transport (Ritter & Cummings, 2015). In Alzheimer's, tau becomes excessively hyperphosphorylated, which causes it to no longer bind to microtubule proteins (Hampel et al., 2010). The absence of bound tau destabilizes the microtubule proteins, which reduces the mobility of nutrients and other important substances in the cell, eventually culminating in the collapse of the microtubule protein (Hampel et al., 2010).

Compounding the situation, unbound hyperphosphorylated tau has a tendency to aggregate into insoluble paired helical filaments (Hampel et al., 2010), which comprise the second hallmark of the disease - neurofibrillary tangles. The buildup of neurofibrillary tangles in neuronal axons is toxic to nerve cells (Hampel et al., 2010). Thus, tau-related neuronal dysfunction in Alzheimer's pathology is thought to result from two factors: the loss of cellular integrity stemming from the collapse of microtubule proteins and the toxicity from the accumulation of neurofibrillary tangles in neuronal axons.

Once the hallmark amyloid-β and tau protein constituents were identified in the 1980s, researchers began to propose etiological theories of the disease. While not the first theory proposed, the most widely accepted version has for many years been the Amyloid Cascade Hypothesis (Thambisetty & Lovestone, 2010). First proposed by John Hardy (1992), this theory posits that the initiating event that leads to Alzheimer's disease pathogenesis is the abnormal and excessive accumulation of amyloid-β in the brain parenchyma. This initiating event leads to a cascade of pathological changes including tau hyperphosphorylation, which leads to the hallmark neurofibrillary tangles, as well as a range of non-disease-specific pathological changes including rampant inflammation, glial activation, catastrophic dendrite and synaptic loss, and progressive neurodegeneration (Clark & Kodadek, 2013).

1.1.3 Biomarkers Facilitate Research Advances

A biological marker [biomarker] is any substance, structure, or process that can be objectively measured in the body or its products and evaluated as an indicator of an interaction between a biological system and a potential hazard, which may be chemical, physical, or biological (World Health Organization, 1993). Examples of biomarkers include common characteristics such as heart rate and blood pressure, neuronal damage to specific brain regions, and various types of molecules in body fluids such as genes, proteins, and metabolites. The type of biomarker depends on what it indicates. Diagnostic biomarkers, which are the type used in this project, indicate the presence or absence of a disease. Other biomarker types include prognostic biomarkers, which forecast the likely course of a disease if no treatment intervention is made, and predictive biomarkers, which assess the most probable outcome of a treatment intervention.

So far, a definitive diagnosis of Alzheimer's disease can only be ascertained through post-mortem examination of the functional anatomy of the brain to detect the presence of the hallmark amyloid-β plaques and neurofibrillary tangles (Petrella, 2013). For an ante-mortem diagnosis, which is also used to label samples (disease vs. control) for clinical trials of potential Alzheimer's therapeutics, a patient is diagnosed with probable Dementia of Alzheimer's Type

[DAT] based on a combination of behavioral history assessments, neuropsychological tests, and blood work to rule out other diseases (McKhann et al., 2011). This type of evaluation is primarily dependent on subjective analysis of dementia symptoms, for which Alzheimer's Disease is not the only possible causal factor. Other neurodegenerative diseases that lead to dementia are Lewy body disease, frontotemporal lobar degeneration, and strokes. A study examining the efficacy of clinical diagnosis at predicting 919 autopsy-confirmed cases of Alzheimer's Disease that comprise the National Alzheimer's Coordinating Center database found the clinical evaluation method to be a modest 71-88% sensitive, but an abysmal 44-71% specific (Beach, Monsell, Phillips, & Kukull, 2012). The poor accuracy of a clinical diagnosis has been further supported by a study using amyloid PET scans (Rinne et al., 2010). The inaccuracy of the clinical diagnosis has been a strong impediment to the development of effective disease-modifying treatments for Alzheimer's Disease.

Insights acquired from longitudinal analysis of Aβ and tau cerebrospinal fluid and neuroimaging biomarkers (often in multi-national, multi-center biomarker studies such as the Alzheimer's Disease Neuroimaging Initiative, founded in 2003) were instrumental in revealing the limitations of the clinical diagnosis and in re-defining the disease altogether. For example, there is now a consensus among researchers that the initiation of Alzheimer's Disease pathology, such as the formation of the hallmark plaques and tangles, occurs 5-10 years prior to the emergence of symptoms (Sperling et al., 2011). This revelation, along with increased understanding of other neurodegenerative disorders, led to a change in how Alzheimer's Disease is defined: from a binary disease considered solely in terms of clinical expression of dementia symptoms to a multi-stage disease process that begins with a long asymptomatic period, during which $A\beta_{42}$ deposition gradually increases; leading to an intermediate prodromal stage termed mild cognitive impairment [MCI], where neuronal damage first becomes noticeable; eventually escalating into full-blown dementia of Alzheimer's type [DAT] with the hallmark amyloid plaques and neurofibrillary tangles (Fagan, 2014).

Along with the extreme difficulty distinguishing between Alzheimer's and other diseases with similar phenotypes, the implication of these insights is that many of the samples used in clinical trials of potential Alzheimer's therapeutics are often incorrectly labeled in one of two following ways:

- As AD when the sample really has a different disease with similar symptoms.
- As NDC when the sample really has pre-symptomatic AD.

Further, many correctly labeled AD samples were obtained from patients with late-stage AD, where the disease has already caused too much brain damage for a viable treatment to be reasonably expected to have efficacy. These three sources of error are commonly thought among researchers to be a major factor behind the lack of success in validating potential Alzheimer's drugs in clinical trials.

Fortunately, thanks to the past 25 years of biomarker research, six biomarkers thus far have been validated widely enough to be included into the latest research criteria (Fagan, 2014). Three are cerebral-spinal fluid [CSF] biomarkers, measured using enzyme-linked immunosorbent assays [ELISAs]. CSF is a clear, colorless, fluid found in the brain and spine that serves as a protective buffer and filtration system for the brain.   Due to its connection to the interstitial space, it is also the most direct fluid source to measure biochemical changes in the central nervous system. The first two of these biomarkers are the amounts of the protein constituents of the hallmark plaques and tangles in CSF: CSF $A\beta_{42}$ and CSF total tau [CSF t-tau].

The third is CSF phospho-tau-181 [CSF p-tau], which is the prevalence of phosphorylation, a type of post-translational modification, present in tau in the CSF.

Among the three CSF biomarkers, only CSF Aβ_{42} has been found to be effective at detecting pre-symptomatic Alzheimer's disease, which provides further support to the Amyloid Cascade Hypothesis (Fagan, 2014). CSF t-tau and CSF p-tau have some utility as progression markers, as they are indicative of non-specific synaptic loss and neuronal injury (Humpel, 2011). For instance, in Creutzfeldt-Jacob disease, CSF t-tau levels are dramatically enhanced to over 3,000 pg/ml (Humpel, 2011). The most effective diagnostic biomarker for Alzheimer's thus far has been shown to be the combination of the three univariate CSF biomarkers. A recently published meta-analysis of CSF and blood-based biomarkers estimated the combination biomarker to diagnose AD in the dementia phase with between 85-90% sensitivity and specificity and in the prodromal phase with 83-95% sensitivity and 71-90% specificity (Ritter & Cummings, 2015). It was also shown by Humpel (2011) to be highly predictive of Alzheimer's pathology at autopsy. However, Vanderstichele & Kodadek (2014) argue that in a typical clinical setting, the accuracy can be much lower due to various analytical and methodological flaws in many of these studies.

The other three validated biomarkers are neuroimaging biomarkers, meaning they are structural or functional characteristics of the brain that are measured using neuroimaging techniques such as magnetic resonance imaging [MRI] or positron emission tomography [PET] scans (Ritter & Cummings, 2015):

- Hippocampal atrophy, which is detected using a structural MRI
- Intraneuronal amyloid plaque load, which is detected by the level of retention of either the Florbetapir F18 or Pittsburgh Compound B amyloid tracers from a PET scan.

- The level of glucose metabolism in characteristic brain regions, which is detected by the amount of uptake of 18F-2fluoro-2-deoxy-D-glucose from a PET scan.

Specific cut points for the biomarker concentrations that differentiate disease vs. control are dependent on the specific lab doing the testing, as different laboratories use different methods to establish cut-points and there is even substantial variability in biomarker concentrations between laboratories and assays (Olsson et al., 2016). Neuroimaging biomarkers are mainly used for disease prognosis, since their sensitivity coincides with the onset of symptoms (Clifford et al., 2013), which is later than for the CSF biomarkers.

1.1.4 The Search for Multivariate Blood-based Biomarkers

Per Humpel (2011), the most important factors for the utility of a diagnostic biomarker are sensitivity, specificity, and ease-of-use. The estimated 85-90% sensitivity and specificity of the combined CSF biomarker is quite good for a complex neurodegenerative disease often seen in patients with other comorbidities, yet is not ideal due to the serious consequences of misclassifications in a diagnostic for a terminal illness. The prognostic detection provided by the imaging biomarkers also fills a distinct need. However, while their diagnostic and prognostic advantages over the phenotypic examinations that constitute the traditional diagnostic methodology are clear, they have thus far proven too impractical for widespread implementation and have been largely limited to research settings. Collecting CSF requires an invasive lumbar puncture procedure many patients are unwilling to go through, especially for a disease that is lacking a viable disease-modifying treatment. The neuroimaging biomarkers, on the other hand, are cost prohibitive as they require expensive specialized machinery to detect. There is thus a great unmet need to discover and validate an accurate diagnostic biomarker of Alzheimer's Disease that is non (or minimally)-invasive and has limited barriers to implementation.

Recognition of this need has sparked a surge of research in recent years into potential blood-based biomarkers, although none have yet been validated successfully in any large multi-center biomarker study.

Blood is highly complex biological information resource with four components (American Red Cross, 2017): plasma (~55%), red blood cells (~45%), and white blood cells and platelets (< 1%). A wide variety of molecular information is contained in blood including DNA markers such as single nucleotide polymorphisms [SNPs], copy number variants, and other static variation and epigenetic changes; RNA markers such as messenger RNA [mRNA] and microRNA; protein markers such as protein concentrations in serum or plasma, isoforms, immunoglobulins, and post-translational modifications; as well as lipids and other metabolic products (Thambisetty & Lovestone, 2010). Other than containing a wide variety of biological information, a blood diagnostic has several advantages in comparison to those from CSF and neuroimaging. For one, only a simple blood draw would be needed for diagnosis, which is fast, cheap, and does not require expensive machinery or sophisticated techniques. Second, a blood-test is much less invasive than a lumbar puncture, allowing for repeated measurements in frail elderly people. Thanks to the advent of DNA microarrays and other high throughput multiplex assays, which can detect the simultaneous expression of thousands of blood-based analytes from a blood sample, the complex well of biological data in blood is now in the initial stages of being mined for useful biomarkers. Most blood-based diagnostic biomarker research thus far has involved either genomics or serum/plasma proteomics. The following section first summarizes the major genomic and proteomic studies. Then, another study is explored with a novel approach towards the discovery of a viable blood-based diagnostic biomarker of AD: mining the adaptive immune system.

1.2 Literature Review

1.2.1 Genomics Biomarker Studies

Study 1: *Variant of TREM2 Associated with the Risk of Alzheimer's Disease* (Jonsson et al., 2012)

In 1993, a sequence variant (mutation) in the ε4 allele of apoliprotein E [APoE4] was identified as a risk factor for late onset Alzheimer's disease. Today it remains the most well validated sequence variant risk factor for late onset Alzheimer's disease due to its high prevalence and effect size. For carriers of APoE4, a meta-analysis (Bertram, McQueen, Mullin, Blacker, & Tanzi, 2007) determined the odds ratio of developing Alzheimer's to be between 3 and 4.

Jonsson et al.'s goal was to identify other useful sequence variants that may predispose an individual to Alzheimer's disease. To accomplish this, whole-genome sequencing was performed on samples taken from 2,261 Icelanders. Genome-wide association analysis [GWAS] was then performed on sequence variants in those genomes that were considered likely to affect protein function. Resulting from the analysis, 191,777 nonsynonymous SNPs, frameshift variants, splicing variants, and stop gain/loss variants were identified. Approximately 34 million markers (SNPs and insertion-deletion polymorphisms), including the 191,777 functional variants, were then imputed in a dataset of Alzheimer's and control samples using haplotype phasing and chip-genotype information and tested for an association with Alzheimer's disease.

A rare missense mutation labeled rs75932628-T in the TREM2 gene, which encodes the triggering receptor on myeloid cells 2 and is predicted to induce a R47H substitution, was found to be a strong risk factor of Alzheimer's Disease, at least among the study population of Icelanders. The odds ratio of this effect was 2.92 with a 95% confidence interval of (2.09, 4.09), and p-value of 3.42×10^{-10}. In brain cells, the *TREM2* gene is primarily expressed on microglia. Activation of microglia can lead to phagocytosis of cellular debris and amyloid. It is theorized that the R47H substitution induced by the rs75932628-T mutation in the *TREM2* gene results in reduced *TREM2* expression and consequently increased buildup of amyloid plaques, due to reduced phagocytosis of toxic amyloid products in the brain.

Study 2: *A gene expression pattern in blood for the early detection of Alzheimer's disease* (Booij et al., 2011).

The goal of this study was to identify a gene expression diagnostic biomarker of Alzheimer's disease in peripheral blood. The gene expression dataset was collected using AB1700 Whole Genome Survey Microarrays and consisted of 251 samples organized into 28 batches due to experimental steps. These batches were randomly partitioned such that 21 batches comprise the training set and 7 batches comprise the test set. The resulting training and test sets contained 188 samples (94 AD, 94 NDC) and 63 samples (31 AD, 32 NDC), respectively. Each sample contained expression values for 32,878 oligonucleotide probes. After quality control filtering, 11,013 gene probes remained. The training and test datasets were log2 transformed, mean normalized, and adjusted for batch effects using a probe-wise ANOVA correction.

Partial Least Squares [PLS] classification with Jack-knife feature selection was used for modeling with a double cross-validation [CV] procedure to minimize the risk of overfitting and to assess the efficacy of the variable selection method. A modified Hotelling's T^2 test was used in the Jack-knife method to test for variable significance based on a 0.05 threshold for false discovery rate-corrected p-values. Any gene not found significant in any fold of leave-one-out cross-validation was eliminated. Jack-knife variable selection reduced the number of gene probes to 1,239. A grid search was then applied with a PLS model on the training set with the reduced variable set and the number of components that had the highest mean out-of-fold accuracy was six. A final PLS model with six components was then fit on the reduced training set, which was subsequently validated on the test set samples. The final PLS model correctly predicted 55/63 samples in the test set (accuracy = 87.3 %) with a sensitivity of 84%, a specificity of 91%, and an AUC of 0.94.

1.2.2 Proteomics Biomarker Studies

Study 3: *Classification and prediction of preclinical Alzheimer's disease based on plasma signaling proteins* (Ray et al., 2007).

Ray et al. complied a dataset of 259 archived blood plasma samples from seven different clinical centers, each containing relative concentrations of 120 plasma signaling proteins as measured by filter-based sandwich ELISA arrays. Their goal was to identify an effective plasma protein diagnostic biomarker of Alzheimer's Disease. The class breakdown of the 259 samples was 85 AD, 79 NDC, 47 MCI, 11 other-dementia [OD], 21 other neurological disease [OND], and 16 Rheumatoid arthritis [RA]. Prior to modeling, the dataset was randomly split into training (43 AD, 40 NDC) and two test sets. The first test set was composed of 42 AD, 39 NDC, and the 11 OD samples, while the second was composed of the 47 MCI samples. The OND and RA samples were not used.

For biomarker discovery and classification, a shrunken-centroids algorithm implemented in Prediction Analysis of Microarrays [PAM] was applied on the 120 normalized plasma signaling protein measurements for samples in the training set and evaluated using 10-fold cross-validation. PAM identified an 18-protein signature that was 89% accurate during cross-validation with 95% sensitivity and 83% specificity with the clinical diagnosis. When the 18-

protein signature was applied to the independent test set, 89% accuracy was again achieved, with 90% sensitivity and 88% specificity with the clinical diagnosis. Of note, the 10/11 OD samples were correctly classified as not having Alzheimer's.

The 18-protein biomarker signature was then used to classify the 47 MCI subjects. This time, a longitudinal follow-up diagnosis on the MCI subjects was used as the "true measure" instead of the original diagnosis. The biomarker correctly classified 81% of these samples, with 91% sensitivity and 72% specificity with the follow-up diagnosis. The authors concluded that this 18-protein biomarker composed of plasma signaling proteins can accurately diagnose Alzheimer's disease, even in its prodromal stage, and is involved in two independent regulatory networks: one centered on tumor necrosis factor – α and monocyte-colony simulating factor, the other on epidermal growth factor.

Study 4: *A serum protein-based algorithm for the detection of Alzheimer's disease* (O'Bryant, 2010).

O'Bryant analyzed the serum expression of 30 proteins in a dataset of 400 subjects (197 AD, 203 NDC) enrolled in the Texas Alzheimer's Research Consortium with the following study goals:

- Assess the predictive utility of a serum-based biomarker for the diagnosis of AD.
- Find out if including demographic variables into the algorithm would improve its predictive power.
- Determine if inflammatory proteins are influential in detecting AD.

The 30 serum proteins in each assay were quantified through multiplex fluorescent immunoassay utilizing colored microspheres with protein-specific antibodies. The 400 subjects were randomly assigned to training and test sets. Random Forest models, one for serum data only and one with both serum and demographic data, were trained on the training set using the default settings of the Random Forest package in R. When the serum data only model was applied to the test set, the AUC, sensitivity, and specificity were 0.91, 0.80, and 0.91, respectively. For the model incorporating serum data with demographic data, the AUC, sensitivity, and specificity on the test set were 0.94, 0.84, and 0.95, respectively.

The authors concluded that proteins quantified in serum from blood samples could accurately predict diagnosis of Alzheimer's disease and that demographic data can boost the predictive utility of such a test. Of the 30 proteins with the highest variable importance in the final Random Forest biomarker signature, 10 were inflammatory in nature. This led the authors to conclude that there is evidence of an inflammatory-related endophenotype of Alzheimer's that could identify a subset of Alzheimer's patients to receive targeted inflammatory-related therapeutics. However, the authors also note two potential flaws in the conclusion. One, it is possible the biomarker signature is not AD-specific and suggest further testing of the algorithm's effectiveness at detecting AD when mixed-in with non-AD dementia samples. Second, it is possible that the inflammatory component of the signature is not specific to AD, but rather is indicative of other co-morbid factors such as cardiovascular disease.

Study 5: *Evaluation of a previously selected plasma biomarker panel to identify Alzheimer's disease* (Björkqvist, Ohlsson, Minthon, & Hansson, 2012).

Bjorkqvist et al. sought to validate the 18-marker signature of plasma signaling proteins identified by Ray et al. (2007) in a dataset of 433 plasma samples (142 AD, 174 NDC, 29 depression, and 88 non-AD dementia) obtained at Skåne University Hospital, Sweden. Three different classification algorithms were implemented using the 18 proteins for binary classification of samples as either AD or NDC: multiple logistic regression, artificial neural network (bagging ensemble of 30 multilayer perceptrons, each with 1 hidden layer and 2 hidden nodes), and nearest shrunken centroid. 100 repeats of 10-fold cross-validation were used to estimate the performance of the classification models, with mean out-of-fold [OOF] area under the ROC curve [AUC] as the metric of classification performance. The resulting AUC for the multiple logistic regression, artificial neural network, and nearest shrunken centroid classifiers were 0.60, 0.63, and 0.63, respectively.

Only three of the 18 proteins were found to be differentially expressed using Kruskal-Wallis tests followed by Mann-Whitney tests for continuous variables and Pearson's X^2 tests for dichotomous variables. Using only these three proteins, the authors repeated the classification methodology above, leading to a mean AUC of 0.66 for all three learning algorithms.

The authors then applied the same classification methodology to the dataset of 142 AD and 88 OD samples to assess the ability of the 18 proteins to discriminate between AD and other types of dementia. The resulting mean AUC was below 0.50 for all three models, leading to the conclusion that these proteins lack the ability to discriminate between AD and other dementias.

Study 6: *Identification of a blood-based biomarker panel for classification of Alzheimer's disease* (Laske et al., 2011).

Laske et al. sought out to identify a diagnostic biomarker based on ELISA measurements of 24 blood markers that are involved in various biological pathways that are involved in the pathogenesis of Alzheimer's disease using a dataset of 155 age-matched samples (85 AD and 70 NDC). The AD samples fulfilled the National Institute of Neurological Disorders and Stroke – Alzheimer's Disease and Related Disorders Association criteria for probable AD and underwent further verification through either cranial computed tomography or MRI tests. The dataset was randomly split 3:1 into training and test sets. Using the training set, feature selection was implemented using logistic regression combined with methods from Weka's machine learning toolkit such as CfsSubsetEval, ConsistencySubsetEval, and InfoGainAttributeEval. Feature selection decided on the following 3-variable set: vWF, Cortisol, and OLAB. A SVM classifier with a radial basis function kernel was then trained on these three protein markers. After parameter tuning, the final SVM classifier predicted the training set with 81.7% accuracy (86.2% sensitivity and 76% specificity) and predicted the test set with 87.1% accuracy (93.8% sensitivity and 80% specificity).

Study 7: *Plasma multianalyte profiling in mild cognitive impairment and Alzheimer's disease* (Hu et al., 2012).

Hu et al. analyzed the $\log_2$ expression of 190 plasma proteins and peptides in two independent discovery datasets with the goal of identifying plasma analyte biomarkers that can effectively discriminate between two classes of samples: Alzheimer's Disease or Mild Cognitive Impairment [AD/MCI] and NDC [Non-demented Control]. The 600 discovery set samples were obtained from University of Pennsylvania [UPenn] and Washington University [WU]. A third independent dataset of 566 samples obtained from the Alzheimer's Disease Neuroimaging Initiative [ADNI] was used for biomarker validation. First, a logistic regression model adjusting for age and gender was fit to both discovery datasets. Significant variables in each logistic regression model were identified using a modified Intersection Union Test, an alternative to the likelihood ratio hypothesis test, with a 0.10 significance threshold. 41/190 and 51/190 plasma analytes were found significant in the UPenn and WU datasets, respectively. 23 analytes were found significant in both datasets. Of these 23 analytes, 17 had a common direction of change between the two datasets.

The ADNI dataset was then used to validate the results. A logistic regression model was fit using the 14 of the 17 plasma analytes available in this dataset. With a strict Bonferroni correction threshold of 0.0036 (0.05/14), the six analytes in the model significantly associated with AD/MCI were APoE4, brain natriuretic peptide [BNP], cortisol, c-reactive protein [CRP], interleukin-3 [IL-3], and pancreatic polypeptide. Out of the 566 ADNI samples, 352 came with CSF biomarker data, which provided the class variables of the samples with a higher confidence diagnosis than those labeled solely by the clinical diagnostic methodology. Two logistic regression models were fit using the 14 plasma analytes on these 352 samples, one with CSF $A\beta_{42}$ as the dependent variable, the other with CSF $\frac{\text{t−tau}}{A\beta_{42}}$. Using the same Bonferroni corrected threshold of 0.0036, the four of the previous six analytes that were found to be significantly associated with both CSF biomarkers were APoE4, BNP, CRP, and pancreatic polypeptide. These four plasma analytes were then considered the final biomarkers.

The correlation between the four plasma biomarkers and the CSF biomarkers was then evaluated using multivariate linear regression with two CSF biomarkers as dependent variables: CSF $A\beta_{42}$ and CSF $\frac{\text{t−tau}}{A\beta_{42}}$. Age and gender were entered as independent variables in the first stage. The four plasma analytes were subsequently entered the model in a step-wise fashion. Three plasma analytes were significantly correlated with CSF $A\beta_{42}$ ($p < 0.001$): APoE4, BNP, and pancreatic polypeptide. APoE4 and BNP were negative correlated with $A\beta_{42}$ ($p < .001$), while pancreatic polypeptide was also negative correlated, but with $p = .007$. While further validation is needed, these results demonstrate the potential for a viable plasma biomarker of AD. Further, due to their correlation with CSF $A\beta_{42}$, which is known to become sensitive before the onset of symptoms, the results further suggested that plasma proteins may also be useful for detecting the disease in its early stages.

1.2.3 The Adaptive Immune System

For a long time, the role of the immune system was thought to be exclusively host defense against external threats such as bacterial and viral pathogens. However, there is increasing awareness of another vital role of the immune system: reconnaissance and defense against internal aberrations in the body (Abel et al., 2014). Autoantibodies, which are antibodies directed against endogenous proteins that the immune system decides are a threat, help maintain homeostasis through initiating protective cellular processes such as adaptive clearance of apoptotic cellular debris (Acharya et al., 2012). Levin et al. (2010) demonstrated that brain-reactive autoantibodies are ubiquitous in human sera, with high individual variation in specific autoantibody expression, but very low variation in total prevalence regardless of age, gender, or disease status.

In situations of blood-brain barrier [BBB] breakdown, which has been shown to occur in brains with Alzheimer's Disease, or those that have suffered some other form of neuronal injury, the brain-reactive autoantibodies in the blood gain access to the brain parenchyma and begin to selectively bind to abundant reactive antigen targets on the exposed surfaces of neurons (Nagele R. et al., 2011). As the brain is normally an immunoprivileged region of the body, there is no buildup of tolerance, which causes the immune response to be more intense than otherwise (Levin et al., 2010). Incessant binding of brain-reactive autoantibodies to membrane-associated antigens stimulates chronic endocytosis, which exposes even more immunogenic proteins and protein fragments to the autoimmune system, causing a self-reinforcing chain-reaction (Nagele R. et al., 2011). This process exacerbates intraneuronal deposition of Aβ peptides, particularly $A\beta_{42}$, in the cells targeted by these autoantibodies in brains with Alzheimer's pathology (Nagele R. et al., 2011). In the long run, neuronal surface receptors are downregulated and eventually completely stripped, significantly impairing neuronal function (Nagele R. et al., 2011). These factors provide a cogent explanation of the role of the autoimmune system in the Amyloid Cascade Hypothesis (Nagele E. et al., 2011).

There is thus growing evidence that brain-reactive autoantibodies that infiltrate the brain parenchyma during instances of BBB breakdown are a major factor in either the initiation of Alzheimer's Disease, its escalation, or both. Given that the abundance of autoantibodies in an individual is independent of age, gender, or the presence of disease, it is thus hypothesized that an individual's autoantibody profile, in concert with the level of intactness of their BBB, is predictive of Alzheimer's pathology (Levin et al., 2010).

Study 8: *Diagnosis of Alzheimer's disease based on disease-specific autoantibody profiles in human sera* (Nagele E. et al., 2011).

Seeking to validate this hypothesis, Nagele E. et al. (2011) used ProtoArray v5.0 Human Protein Microarrays to detect the expression of 9,486 autoantibodies in 90 human sera samples (40 AD, 50 NDC) with the goal of identifying an autoantibody signature that can accurately diagnose Alzheimer's Disease from a blood serum sample. Most of the analysis was implemented in the Prospector software package as per the recommended biomarker discovery workflow suggested by the manufacturer of the microarrays. The 90 samples were randomly partitioned into training and test sets such that each partition contained 25 AD and 20 NDC samples each, with equal proportions of both early-stage to late-stage AD samples and old to young controls.

Two-group discrimination analysis was implemented on the training set using the minimum M-Statistic, which has the advantageous property of being sub-group-specific (Sboner et al., 2009). Out of the total 9,486 autoantibodies, 451 had a significantly higher prevalence in the AD samples than in the NDC samples with $p < 0.01$. The 10 of these 451 autoantibodies with the largest difference in prevalence between the two classes were selected as the final biomarker. For additional verification, the PAM method was used to analyze the level of differential expression among the 9,486 autoantibodies. PAM verified that the 10 autoantibodies previously selected were among the features best able to discriminate between the classes. The analysis then switched over to the R environment. Random forests [RF], a multivariate ensemble classification algorithm (Breiman, 2001) provided by the *randomForest* package (Liaw & Wiener, 2015), was utilized to evaluate the ability of the 10-autoantibody biomarker to classify sera samples as AD or NDC. The RF algorithm classified the out-of-bag [OOB] samples with 93.33% accuracy, 96% sensitivity, and 90% specificity, while the 45 test set samples were classified with 93.33% accuracy, 88% sensitivity, and 100% specificity.

Nagele E. et al. (2011) then sought to assess the ability of the RF biomarker to discriminate between AD and two other major diseases: Parkinson's Disease [PD] and Breast Cancer [BC]. 59 additional protein microarray serum samples were acquired for this, 30 from patients with BC and 29 from patients with PD. A RF classifier fit on the 50 AD and 30 BC samples with the previous 10-autoantibody biomarker was able to discriminate between AD and BC with 92.5% OOB accuracy. For discriminating between AD and PD, which unlike BC is a closely-related neurological disorder, the entire analysis pipeline was re-implemented. This time, five autoantibodies were selected using M-statistical analysis, subsequently verified by PAM, and then were used to fit a RF model on the 50 AD and 29 PD samples. The RF model was able to discriminate between AD and PD with 86% OOB accuracy.

The results thus indicated that the expression levels of a small subset of autoantibodies in human sera can accurately distinguish both between AD and controls and between AD and other diseases, even similar neurological ones. However, given the small sample sizes used in the study and other potential sources of variability, much further validation is required to confirm its classification efficacy. The utility of this finding also depends on whether the harmful autoimmune reaction is an etiological factor leading to the initiation of the disease or a secondary factor that plays a role in its escalation. If the former, a validated autoantibody biomarker could yield promising therapeutic benefits with a strategy for removing or suppressing the problematic immunoglobulins. If the latter, a validated biomarker would still be useful for aiding diagnosis, and potentially monitoring progression, but its utility would be confined to Alzheimer's patients with a disease-stage passed the point where potential disease-modifying therapeutics would be reasonably expected to have efficacy. Regardless, the results of this study represented a promising breakthrough.

1.3 Dataset

123 human sera samples were extracted for this project from two publicly available datasets on the Gene Expression Omnibus [GEO] website: GSE29676 and GSE39087. The samples were uploaded, respectively, by the authors of the Nagele E. et al. (2011) and Nagele E. et al. (2013) autoantibody studies from Section 1.2.3. In both studies, ProtoArray v5.0 Human Protein Microarrays were used to detect serum expression measurements of 9,480 autoantibodies from the samples with no differences in data collection methodology. The class breakdown of the 123 samples was 50 AD and 73 NDC. The 50 AD samples and 40 of the NDC samples were from GSE29676, while the other 33 NDC samples were from GSE39087.

Each protein microarray contained 9,480 unique full-length human proteins that are known and addressable autoantigens, 384 human-IgG and anti-human IgG control proteins (8 unique human-IgG and anti-human IgG per each of the 48 subarrays), and various other housekeeping controls that are not used in this thesis. Each protein was printed in-duplicate on 130 µm spots across 48 subarrays (blocks). After blocking and incubation with a human blood serum sample, each array was probed with a secondary fluorescently-labeled antibody to detect the expression intensity of the 9,480 autoantibodies that target the 9,480 printed autoantigens.

Each array was then scanned with a GenePix 4000B Fluorescence Scanner, which used image processing to measure the reactivity of each analyte with the human sera sample. The scanner first partitions each spot on an array into foreground and background regions. The foreground partition is the center section of each spot containing the analyte and the background partition is the remaining area. In contrast with DNA microarrays where analytes are generally contained in a well-defined circle within a spot, the background regions of spots on protein microarrays can be easily skewed by artifacts such as small speckles. For this reason, the arrays were scanned using the irregular segmentation setting, where no assumption was made about the specific shape of the foreground region. The output of the scanning software for each sample was a Genepix Results [GPR] file containing foreground and background expression intensity measurements of the 9,480 duplicated autoantibodies and 384 duplicated controls; quality control indicators for each analyte such as % Saturation, and various meta data for each analyte and sample.

1.4 Statement of Purpose

Among the recently proposed multivariate blood-based biomarkers of Alzheimer's Disease, none have yet been validated successfully enough for widespread acceptance. While there are many potential confounding forces at play, a major roadblock to the development of a successful blood-based biomarker of Alzheimer's Disease is the analytical insufficiencies often present in the feature selection methodologies used to identify the putative biomarkers.

A common example is when the feature selection method is not multivariate, which is bad for two reasons. One, important discriminatory information is otherwise often removed, since a feature that is useless by itself may be useful when considered jointly with other features (Guyon & Elisseeff, 2003). Two, there is otherwise often redundancy in the final feature-set, since correlations and interactions between features were not taken into account (Dziuda, 2010, p. 122-123).

A second example is when a resampling method such as bootstrap aggregating [bagging] is not used to help stabilize the feature selection process. Biomarker discovery datasets obtained using high throughput microarrays often contain a much higher proportion of variables than samples. This low degrees-of-freedom situation, termed the *curse of dimensionality*, greatly increases the risk of analysis overfitting the training data, leading to solutions with overconfident performance estimates that fail to generalize to new data. Bagging the feature selection process can help lower the variance and lead to a more stable and generalizable biomarker.

The contribution of this thesis was to improve upon the feature selection process used in Study 8 from Section 1.2.3, the promising autoantibody biomarker study *Disease Associated Autoantibodies for Diagnosis of Alzheimer's Disease in Human Sera* (Nagele E. et al., 2011), in the following two ways:

- By replacing shrunken centroid feature selection, a single-variable-centered multivariate filter with a univariate bias, with recursive feature elimination, a multivariate feature selection wrapper.
- By aggregating results of multiple feature selection experiments performed on bootstrap samples of the training data. This helps stabilize the optimum subset size selection, and consequently, the selected features.

While these were the primary contributions, the biomarker discovery learning process was also improved in the following ways:

- By comparing the performance of two non-similar non-parametric classification algorithms (random forest and linear support vector machine) commonly applied to biomarker discovery analysis.
- By using bagging with a grid search to optimize the model parameters of the classifiers trained on the final biomarkers.
- By increasing the sample size of the dataset from 90 to 123 samples.
- In the original study, there is no mention of what pre-processing methods were applied to the dataset prior to implementing the shrunken-centroid filter. In this project, the following pre-processing methods were applied as recommended by the literature: background correction and robust linear model normalization.
- By replacing accuracy with area under the ROC curve [AUC] as the primary performance metric used for biomarker discovery. AUC is the sole metric used to evaluate subset sizes during feature selection and model parameters during hyperparameter optimization, and is greatly involved in biomarker selection. Accuracy, sensitivity, and specificity were also examined to get a more complete characterization of performance.

2 Methods

2.1 Robust Linear Model Normalization

The goal of normalization is to reduce the non-biological variation both between and within the arrays (Dziuda, 2010, p. 26). Sources of non-biological variance in protein microarrays include spatial artifacts introduced by the printing process; systematic noise from the scanning process, which can happen if the scanner is not properly aligned; differences in the total quantity of either serum or secondary antibody; and the heterogeneity of the array surface (Sboner et al., 2009). This non-biological variability induced by the printing, probing, and scanning processes of protein microarrays can distort and even completely mask the true biological variability, leading to inaccurate modeling results driven more by noise rather than signal (Sboner et al., 2009).

Tools and techniques for the analysis and pre-processing of protein microarray data have been largely adapted from those already existing for DNA microarrays, which have a much longer and more extensive history. As both DNA and protein microarrays involve high throughput detection of the simultaneous expression of analytes that are arranged by sample on a grid, the data pre-processing and analysis tools used are largely interchangeable, but not entirely so. One area where separate techniques are needed is normalization. With DNA microarrays, an assumption is made that the amount of mRNA is approximately equal across samples, which is often valid due to large numbers of "neutral" genes with intensity expression values that cover the entire dynamic range of the microarray (Sboner et al., 2009). This then permits the assumptions made by DNA microarray normalization methods such as global and quantile normalization, which assume that arrays have the same median expression value and that arrays have the same distribution of expression values, respectively. These assumptions are invalid for protein microarrays, since the quantity and intensity of reactive proteins can vary greatly across samples, leading to different median intensities and signal distributions for each sample (Sboner et al., 2009).

Sboner and colleagues (2009) compared global and quantile normalization with Robust Linear Model [RLM] normalization, a method specifically developed for protein microarrays that utilizes the control proteins printed on each subarray for intra-array and inter-array normalization. For the comparison, each normalization method was applied to a dataset consisting of three replicated samples of the disease group and 3 replicated samples of the control group. The normalization methods were evaluated via a before/after comparison of intra-array variability, inter-array variability, and class separability. Intra-array variability was measured by examining the distribution of the coefficient of variation [CoV] of the control proteins replicated in each subarray (block) within each array. Inter-array variability was measured by examining the distribution of the CoV of the protein variables spotted at the same location in each array. Class separation was measured by Fisher's signal-to-noise ratio [SNR] applied to the $\log_2$ transformed proteins, which is calculated as follows:

$$SNR = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2}, \tag{2}$$

where μ_1 and μ_2 are the mean values of a protein expression variable for the two classes, and σ_1 and σ_2 are the corresponding standard deviations.

An ideal normalization method should both lower the non-biological variability, leading to lower coefficients of variation; and increase sample separability between classes, leading to higher SNRs (Sboner et al., 2009). Global and quantile normalization showed no significant effect on the CoV distribution for the control proteins within each subarray, which was expected since these methods are designed to only address inter-array variability. RLM normalization did reduce intra-array variability, lowering the median CoV of the control proteins from 0.24 to 0.19-0.20. All three normalization methods significantly reduced inter-array variability, with quantile normalization the most effective by a slight margin. RLM normalization, however, also significantly increased class separation, whereas global and quantile normalization reduced class separation, distorted the signal, and lowered the signal range.

The results of the study thus confirmed that RLM normalization is a more effective than global or quantile normalization for protein microarray data, since it significantly reduces non-biological intra and inter-array variability without distorting the biological signal. Further, it can exploit the control proteins on the arrays to capture a variety of error sources, such as variations local to subarrays and overall differences in brightness between arrays, and utilizes a robust model-fitting algorithm, which is resistant to outliers (Sboner et al., 2009). Given these advantages, RLM normalization was the method used for this thesis.

The RLM normalization method used in this project was adapted from the article by Sboner et al. (2009) as follows:

1. Both autoantibody and control protein data are first $\log_2$ transformed to stabilize the variation across the signal intensities, which turned multiplicative errors into additive errors. Note that the control protein data only includes the 384 human and anti-human IgG controls and not the other housekeeping variables.
2. The following robust linear model is then trained on the control protein dataset using the *MASS* package in R with default settings:

$$y_{ijk} = \alpha_i + \beta_j + \tau_k + \varepsilon_{ijk}, \tag{3}$$

   where y_{ijk} is the $\log_2$ expression of control protein k located in subarray j on array *i*; α_i is the sample effect, which captures inter-array non-biological variations; β_j is the block effect, which captures spatial and printing pin variations; τ_k is the feature effect, which captures variations in the amount of spotted protein and its binding affinity; and ε_{ijk} is the random error, which is assumed to have a normal distribution centered at zero with constant variance.

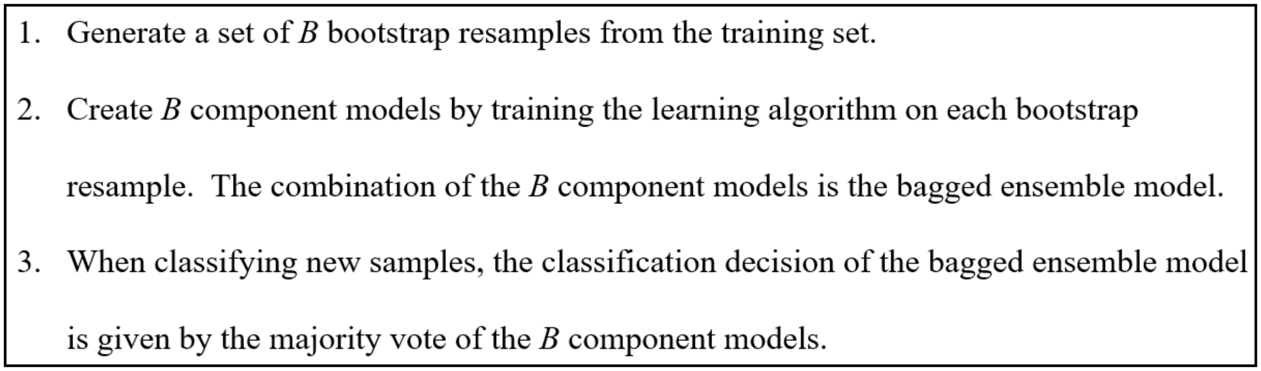
3. The sample and block coefficients from the RLM model (α_i and β_j) are then used to normalize the background corrected and $\log_2$ transformed autoantibody dataset as follows:

$$y'_{ijk} = y_{ijk} - \alpha_i - \beta_j, \tag{4}$$

   where y'_{ijk} and y_{ijk} are the normalized and pre-normalized $\log_2$ expression, respectively, of autoantibody k on block j of sample i.

2.2 Bootstrap Aggregating

Bootstrap aggregating [bagging] is an ensemble method introduced by Leo Breiman (1996) to improve the stability of machine learning algorithms used in statistical classification or regression. When a learning algorithm is bagged, the resulting ensemble model makes lower-variance classifications than a single model trained using the same learning algorithm on the full training set. However, bias is largely unaffected by bagging. For this reason, bagging works best with unstable learning algorithms, which are learners that produce model outputs that tend to be significantly affected by small perturbations in the training data. For a given classification learning algorithm, bagging works as follows:

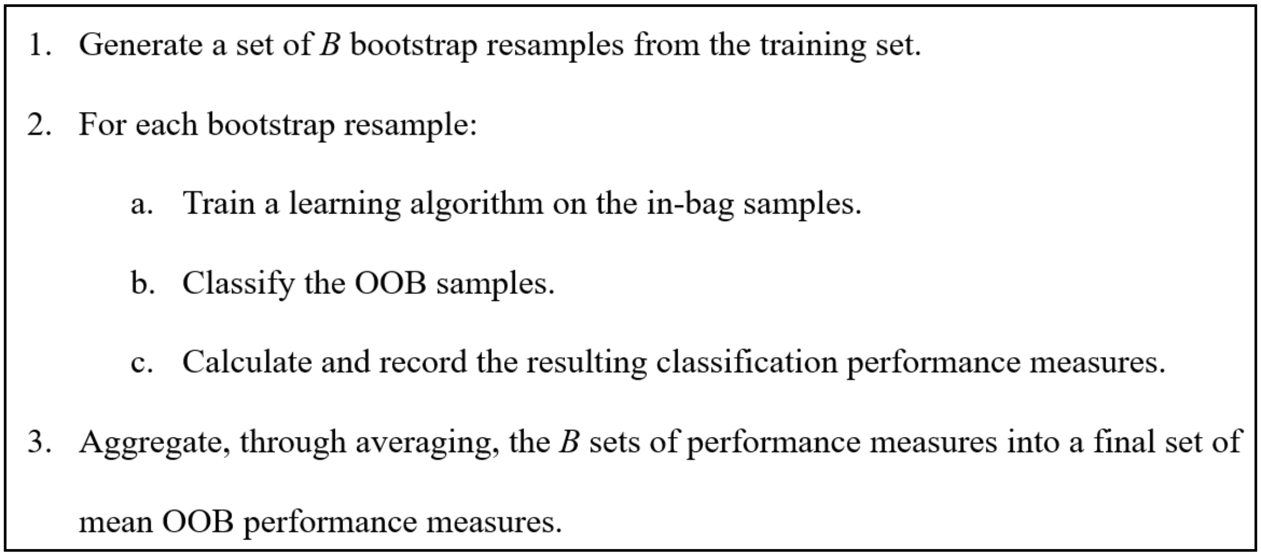


1. Generate a set of B bootstrap resamples from the training set.
2. Create B component models by training the learning algorithm on each bootstrap resample. The combination of the B component models is the bagged ensemble model.
3. When classifying new samples, the classification decision of the bagged ensemble model is given by the majority vote of the B component models.

Figure 2. Bagged ensemble classifier algorithm.

While there are multiple ways bootstrap resampling can be implemented, the version used in this project is the standard nonparametric bootstrap (Efron, 1979), which is also the type used in the random forest algorithm (Breiman, 2001) described in Section 2.3 of this project. In this context, each bootstrap resample of a training set *D* that has *n* total samples is a set of *n* samples randomly selected with replacement from *D*. On average, each bootstrap resample contains 63.2% of the *n* samples from the original training data (Dziuda, 2010). The samples not selected, which are on average 36.8% of the original *n* samples, are called out-of-bag [OOB] samples.

Thus far, bagging has been introduced as an ensemble modeling technique used to stabilize classification decisions or regression predictions. However, bagging can also be used to stabilize the assessment of model performance. The algorithm for using bagging in this way is as follows:



```
1. Generate a set of $B$ bootstrap resamples from the training set.
2. For each bootstrap resample:
     a. Train a learning algorithm on the in-bag samples.
     b. Classify the OOB samples.
     c. Calculate and record the resulting classification performance measures.
3. Aggregate, through averaging, the $B$ sets of performance measures into a final set of
   mean OOB performance measures.
```

Figure 3. Bagged model performance algorithm.

2.3 Random Forests

The random forests learning algorithm was first proposed by Leo Breiman (2001) as an improved version of bagged decision trees. Decision trees recursively partition the data into smaller groups that are more homogenous with respect to the response using binary splits in the form of nested if-then statements. The classification and regression tree [CART] algorithm

(Breiman et al., 1984) is the most prevalent decision tree algorithm. The CART algorithm starts by searching every distinct value of every variable in the entire training dataset *S* and identifies the combination of variable and split value that partitions the root node (entire dataset) into two child nodes (non-overlapping subgroups) S_1 and S_2 such that the class impurity of the child nodes is minimized (Kuhn & Johnson, 2013). The measure of class impurity of the child nodes can be calculated in multiple ways, with the most common method being the Gini impurity index. The Gini impurity index of a node *S* of a decision tree used for binary classification is calculated as follows:

$$Gini(S) = p_1(S)p_2(S), \tag{5}$$

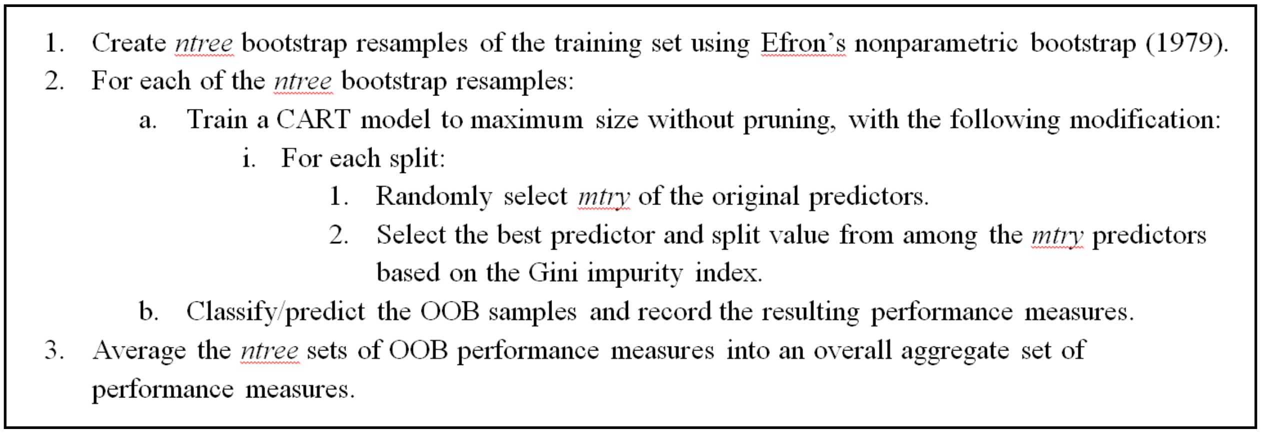
where $p_1(S)$ and $p_2(S)$ represent the proportion of data points in node *S* that belong to class 1 and class 2, respectively. Now let *S* represent any parent node and S_1 and S_2 represent the two child nodes resulting from a binary split of *S*. The decrease in the Gini impurity index of the split is then:

$$\Delta Gini(S) = Gini(S) - Gini(S_1)n_{S_1} - Gini(S_2)n_{S_2}, \tag{6}$$

where n_{S_1} and n_{S_2} represent the fraction of observations in node S that wind up in child nodes S_1 and S_2, respectively.

The optimization problem of finding the variable and split value that maximizes the decrease in Gini impurity index resulting from the split occurs recursively for each node of the tree until a stopping criterion is met. This stopping criterion can be customized, but the rule used in this thesis is the default option of the *rpart* package, which halts growing when the depth of the tree is 30 nodes. The final nodes of each branch of the tree are the nodes that do not split into child nodes. These nodes are called leaves. New observations are predicted by starting at the root node of the tree, and progressing down the branches until a leaf node is reached. In the classification case, the predicted class of the new observation is then the majority class of the training observations that fall into that leaf node.

The hierarchical structure of decision trees is intrinsically unstable, which leads it to synergize well with the variance-reducing properties of bagging, the ensemble learning method described in Section 2.2. However, the trees in a bagged-tree model are not completely independent of each other since, despite the trees being constructed from different bootstrap resamples, all variables are considered at every split of every tree. The random forest algorithm was developed by Leo Breiman (2001) as a modified version of bagged trees that resolves the tree correlation issue. Instead of evaluating all predictors at every split in every tree, a randomly selected subset of predictors of size *mtry* are evaluated, where *mtry* is a tuning parameter. For classification, Breiman (2001) recommends initially setting *mtry* to the square root of the total number of variables. Having the algorithm only considering a randomly selected subset of variables when determining optimal splits introduces additional randomness that helps de-correlate the trees. Breiman (2001) also showed that this additional randomness protects random forest from overfitting as the number of trees in the forest increases. The pseudocode of the random forest algorithm is as follows:



```
1. Create ntree bootstrap resamples of the training set using Efron's nonparametric bootstrap (1979).
2. For each of the ntree bootstrap resamples:
     a. Train a CART model to maximum size without pruning, with the following modification:
          i. For each split:
               1. Randomly select mtry of the original predictors.
               2. Select the best predictor and split value from among the mtry predictors
                  based on the Gini impurity index.
     b. Classify/predict the OOB samples and record the resulting performance measures.
3. Average the ntree sets of OOB performance measures into an overall aggregate set of
   performance measures.
```

Figure 4. Random forest algorithm.

By averaging the *ntree* sets of performance measures from step 2b. in the RF algorithm above, a final aggregate set of performance measures is produced that serves as a stabilized estimate of the ability of the RF model to classify unseen data, so long as the training set is a representative sample of the population. In fact, Breiman (1996) showed that these aggregate measures can be considered approximations of classification performance on a test set that is of the same size as the training set (Breiman, 1996). New observations are classified by the random forest ensemble learner via a majority vote of the classification decisions returned by the *ntree* component CART models.

2.4 Support Vector Machines

Support vector machines [SVMs] are a class of learning algorithms that identify the optimal hyperplane separating the two classes, often in high-dimensional space. Assume we have a binary classification dataset with *n* samples and *p* variables where the data are linearly separable, meaning there exists a (p – 1)-dimensional hyperplane that can perfectly partition the data in *p*-dimensional space by the class of the target variable. In this situation, there are an infinite number of (p – 1)-dimensional hyperplanes that can linearly separate the two classes of this dataset. Since each of these infinite qualifying hyperplanes would produce classifications with perfect accuracy, an alternative metric called the *geometric margin* is used to discriminate between them, which is the distance between the hyperplane and the closest p-dimensional sample. The optimal separating hyperplane is the one with the maximum geometric margin. For this reason, the optimal separating hyperplane is termed the *maximum margin classifier*. The two hyperplanes parallel to the maximum margin classifier that pass through the data points that lie on the boundaries of the margin are called *support hyperplanes*.

A separating hyperplane (**w**, b) is represented by the following equation (Dziuda, 2010):

$$\boldsymbol{w}'\boldsymbol{x} + b = 0, \tag{7}$$

where:

- **w** is a p-dimensional vector of feature-weights that is orthogonal to the hyperplane.
- **x** is a matrix of training data with dimensions *n* x *p*, where *n* is the total number of samples and *p* is the total number of variables.
- *b* is a bias term that defines how offset the hyperplane is from the origin.
- $\boldsymbol{w}'\boldsymbol{x}$ denotes the cross product between **w** and **x**

The maximum margin classifier (**w**, b) can be identified by the solution to the following minimization problem (Dziuda, 2010):

$$\underset{\boldsymbol{w}, b}{minimize} \quad ||\boldsymbol{w}||^2, \tag{8}$$

subject to the following condition:

- $y_i(\boldsymbol{w}'\boldsymbol{x}_i + b) \geq 1$ for all training points $\boldsymbol{x}_i, i = 1, \dots, n,$

where $||\boldsymbol{w}||$ denotes the Euclidian norm, also called the L_2 norm, of **w**.

However, it is often more useful to calculate the equivalent dual optimization problem, which optimizes a vector of weights corresponding to samples, instead of variables (Dziuda, 2010):

$$\text{maximize} \quad \mathrm{W}(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i' \boldsymbol{x}_j, \tag{9}$$

subject to the following two conditions:

- $\sum_{i=1}^{n} y_i \alpha_i = 0$
- $\alpha_i \geq 0$, for $i = 1, \dots, n,$

where:

- $\boldsymbol{\alpha}$ is a vector of Lagrange multipliers, one for each of the *n* samples.
- y_i is the value of the target class for observation *i*. The first class is encoded as -1 and the second class is encoded as 1.
- $\mathbf{x}_i$ is the vector of variable values for sample *i*.
- $\boldsymbol{x}_i'\boldsymbol{x}_j$ denotes the dot product between vectors $\mathbf{x}_i$ and $\mathbf{x}_j$.

The weight vector **w** can then be derived from the $\boldsymbol{\alpha}$ coefficients (Dziuda, 2010):

$$\boldsymbol{w} = \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x}_i \tag{10}$$

In the case where the separating hyperplane can linearly separate the two classes in p-dimensional space, it can be shown that, for each sample *i*, its Lagrange multiplier α_i is only non-zero if the sample lies on one of the two support hyperplanes (Dziuda, 2010). These samples with non-zero Lagrange multipliers are called *support vectors* and are the only samples that are used to make classification decisions, which is where the name *support vector machines* comes from. Support vector machines classification is thus driven by the samples with the least amount of classification certainty (closest to the maximum margin hyperplane).

Once the weight vector **w** has been calculated, the offset term *b* can be calculated by plugging any of the support vectors into

$$b = y_i - \boldsymbol{w}'\boldsymbol{x}_i \tag{11}$$

and solving for *b* (Dziuda, 2010). The decision function for classifying a new sample **u** is then (Dziuda, 2010):

$$D(\boldsymbol{u}) = \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x}_i'\boldsymbol{u} + b \tag{12}$$

If D($\boldsymbol{u}$) > 0, then the new sample is classified as the positive class ($Y_u = 1$). Else, it is classified as negative ($Y_u = -1$).

Often in many real-world analysis applications, including this thesis, a linear separation cannot perfectly partition the classes in *p*-dimensional space. In this situation, a hard-margin SVM has no solution and a soft-margin SVM must be used. In the soft-margin SVM algorithm, candidate separating hyperplanes are permitted to misclassify samples, but with a penalty proportional to the extent of the violation of the margin. Each training set sample *i* is assigned a non-negative slack variable ε_i that represents the amount that the sample violates the margin. Value ranges for samples that are classified correctly, incorrectly, or unclassifiable due to being located on the optimal separating hyperplane are $0 \leq \varepsilon_i < 1$, $\varepsilon_i > 1$, and $\varepsilon_i = 1$, respectively (Dziuda, 2010).

The vector of feature weights **w** corresponding to the maximum margin hyperplane in the soft-margin SVM algorithm is identified by the following minimization problem (Dziuda, 2010):

$$\underset{\boldsymbol{w},\ b,\boldsymbol{\varepsilon}}{minimize} \quad \|\boldsymbol{w}\|^2 + C\sum_{i=1}^{n} \varepsilon_i\,, \tag{13}$$

subject to:

- $y_i(\boldsymbol{w}'\boldsymbol{x}_i + b) \geq 1 - \varepsilon_i$ for all training points $\boldsymbol{x}_i, i = 1, \dots, n$

The scalar *C* controls the degree of penalization. As with hard margin SVMs, it is often more useful to solve for the corresponding dual optimization problem (Dziuda, 2010):

$$maximize \quad W(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i\alpha_j y_i y_j \boldsymbol{x}_i'\boldsymbol{x}_j, \tag{14}$$

subject to:

- $\sum_{i=1}^{\mathrm{n}} y_i\alpha_i = 0$

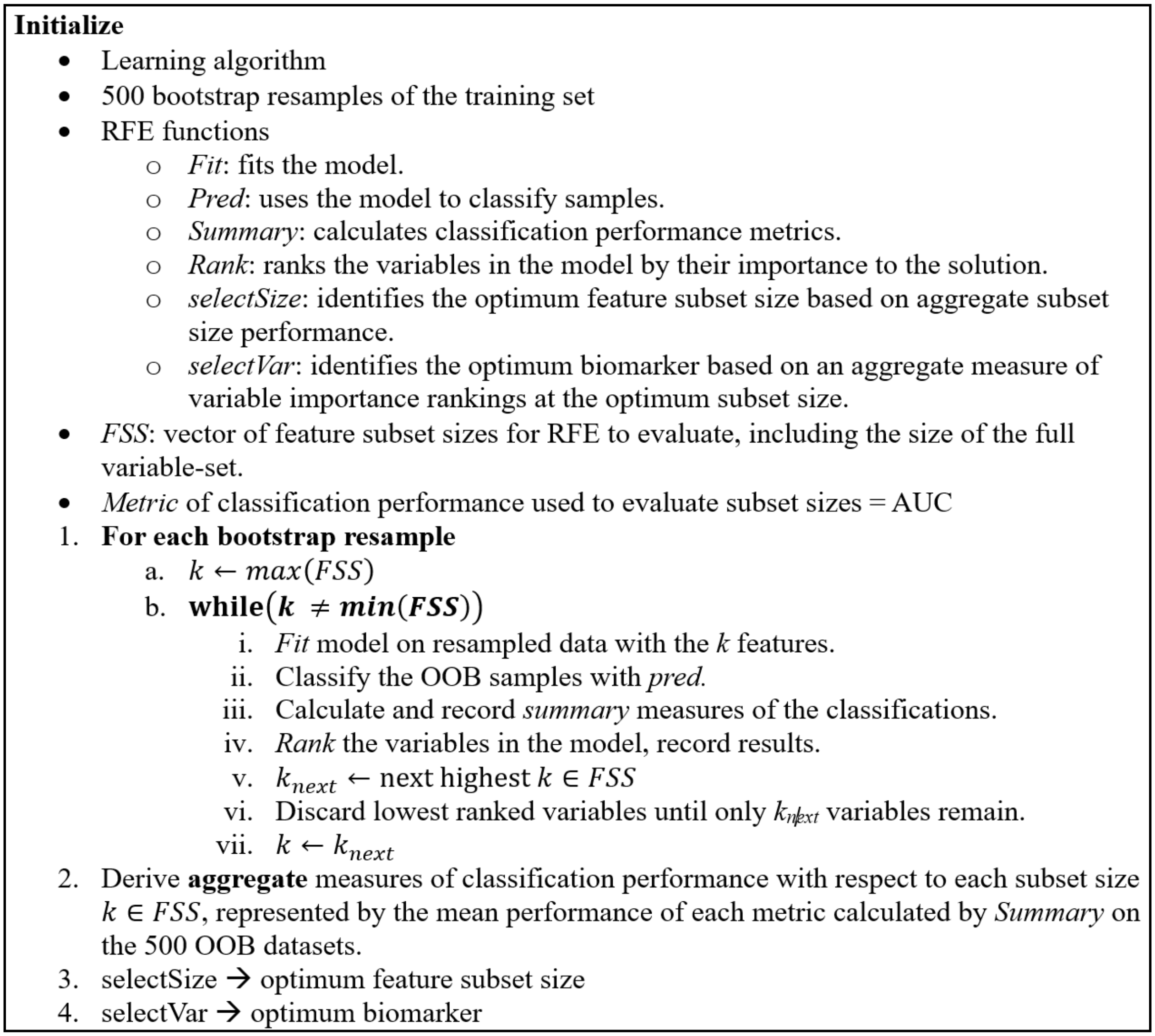
- $0 \leq \alpha_i \leq C$, for $i = 1, \ldots, n$

The only difference between the dual optimization problems for hard-margin and soft-margin SVMs is the maximum-bound for the α coefficients. The vector of feature weights can be calculated using the α coefficients in the same way as with hard-margin SVMs as seen in Equation 10. The calculations for the offset b and the prediction function for new samples are also the same as with hard-margin SVMs (Equation 11).

SVMs can be extended to the nonlinear case with the use of the kernel trick, but that is not necessary for this thesis. The dataset used in this thesis suffers greatly from the *curse of dimensionality* and protecting against overfitting was thus a primary concern. The use of a nonlinear kernel such as the radial basis function would have been explored if there was evidence that the linear SVM model in Section 3.4.3 underfit the data and a more complex model was required.

2.5 Recursive Feature Elimination

Recursive feature elimination [RFE] is a multivariate feature selection wrapper algorithm that begins with the full set of variables and recursively eliminates variables based on their multivariate importance in a fitted model. In this way, the search process is "wrapped" around the learning algorithm. The version of RFE used in this project is a modified form of the recursive feature elimination procedure that was introduced by Guyon, Weston, Barnhill, & Vapnik (2002) in concert with support vector machines. The modified RFE is implemented using the framework provided by the *rfe* function in the *caret* package (Kuhn, 2016) and incorporates bagging into the algorithm to help lower the variance of the feature selection process. Pseudocode of the modified recursive feature elimination algorithm is as follows:

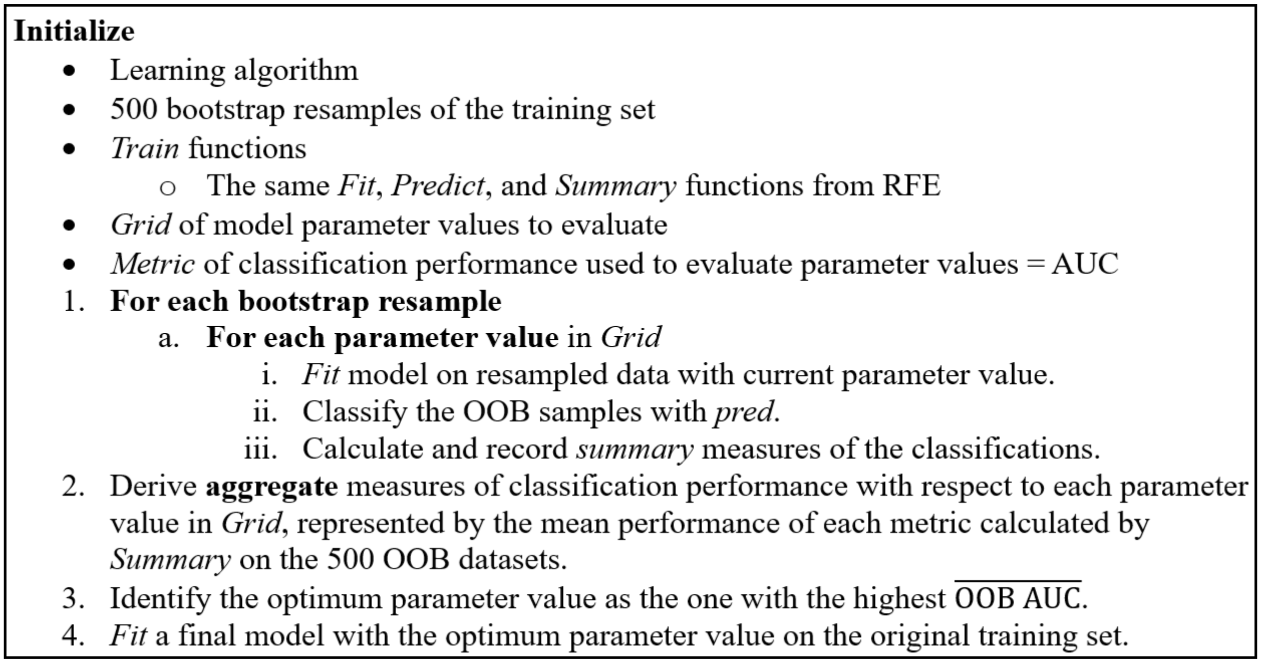


```
Initialize
  • Learning algorithm
  • 500 bootstrap resamples of the training set
  • RFE functions
      o Fit: fits the model.
      o Pred: uses the model to classify samples.
      o Summary: calculates classification performance metrics.
      o Rank: ranks the variables in the model by their importance to the solution.
      o selectSize: identifies the optimum feature subset size based on aggregate subset
        size performance.
      o selectVar: identifies the optimum biomarker based on an aggregate measure of
        variable importance rankings at the optimum subset size.
  • FSS: vector of feature subset sizes for RFE to evaluate, including the size of the full
    variable-set.
  • Metric of classification performance used to evaluate subset sizes = AUC
  1. For each bootstrap resample
      a. $k \leftarrow max(FSS)$
      b. $\mathbf{while}(\boldsymbol{k} \neq \boldsymbol{min}(\boldsymbol{FSS}))$
           i. Fit model on resampled data with the k features.
          ii. Classify the OOB samples with pred.
         iii. Calculate and record summary measures of the classifications.
          iv. Rank the variables in the model, record results.
           v. $k_{next} \leftarrow$ next highest $k \in FSS$
          vi. Discard lowest ranked variables until only $k_{next}$ variables remain.
         vii. $k \leftarrow k_{next}$
  2. Derive aggregate measures of classification performance with respect to each subset size
     $k \in FSS$, represented by the mean performance of each metric calculated by Summary on
     the 500 OOB datasets.
  3. selectSize → optimum feature subset size
  4. selectVar → optimum biomarker
```

Figure 5. Recursive feature elimination with bagging.

2.6 Grid Search

A grid search uses resampling, in this case bagging, to optimize the performance of a classifier by stabilizing the selection of model parameters. This method is analogous to how the modified RFE algorithm incorporates bagging to stabilize the selection of the optimal feature subset size. For this project, a grid search was used to stabilize the selection model parameters for the final random forest and support vector machine classifiers that were fit on their respective optimum biomarkers determined during the feature selection stage. The grid search algorithm was implemented using the framework provided by the *train* function in the caret package (Kuhn, 2016) and had the following pseudocode:

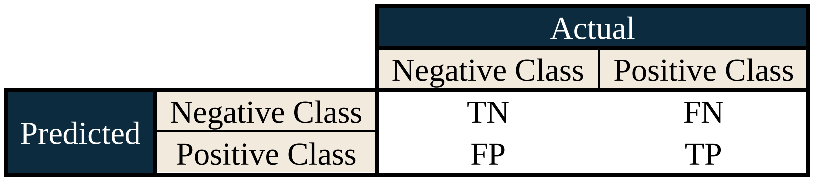


```
Initialize
    • Learning algorithm
    • 500 bootstrap resamples of the training set
    • Train functions
        o The same Fit, Predict, and Summary functions from RFE
    • Grid of model parameter values to evaluate
    • Metric of classification performance used to evaluate parameter values = AUC
    1. For each bootstrap resample
        a. For each parameter value in Grid
            i. Fit model on resampled data with current parameter value.
            ii. Classify the OOB samples with pred.
            iii. Calculate and record summary measures of the classifications.
    2. Derive aggregate measures of classification performance with respect to each parameter
       value in Grid, represented by the mean performance of each metric calculated by
       Summary on the 500 OOB datasets.
    3. Identify the optimum parameter value as the one with the highest $\overline{\text{OOB AUC}}$.
    4. Fit a final model with the optimum parameter value on the original training set.
```

Figure 6. Grid search algorithm.

2.7 Area under the ROC Curve

For binary classification problems, there are four possible outcomes for each classification decision: true positive, true negative, false positive, and false negative. A confusion matrix is used to visualize the relative quantities of these four outcomes among a set of classification decisions and consists of the following structure:



		Actual	
		Negative Class	Positive Class
Predicted	Negative Class	TN	FN
	Positive Class	FP	TP

Figure 7. Confusion matrix structure.

Three standard metrics of classification performance can be calculated directly from a confusion matrix based on the relative proportions of the four possible classification outcomes: accuracy, sensitivity, and specificity. Accuracy is the proportion of samples that were classified correctly, represented by the following equation:

$$Accuracy = \frac{TP + TN}{n}, \tag{15}$$

where *n* is the total number of classified samples. Sensitivity and specificity are class-specific accuracy measures. Sensitivity measures the proportion of positive class samples that were correctly classified as being positive:

$$Sensitivity = \frac{TP}{TP + FN}, \tag{16}$$

Analogously, specificity measures the proportion of negative class samples that are correctly classified as negative:

$$Specificity = \frac{TN}{TN + FP}. \tag{17}$$

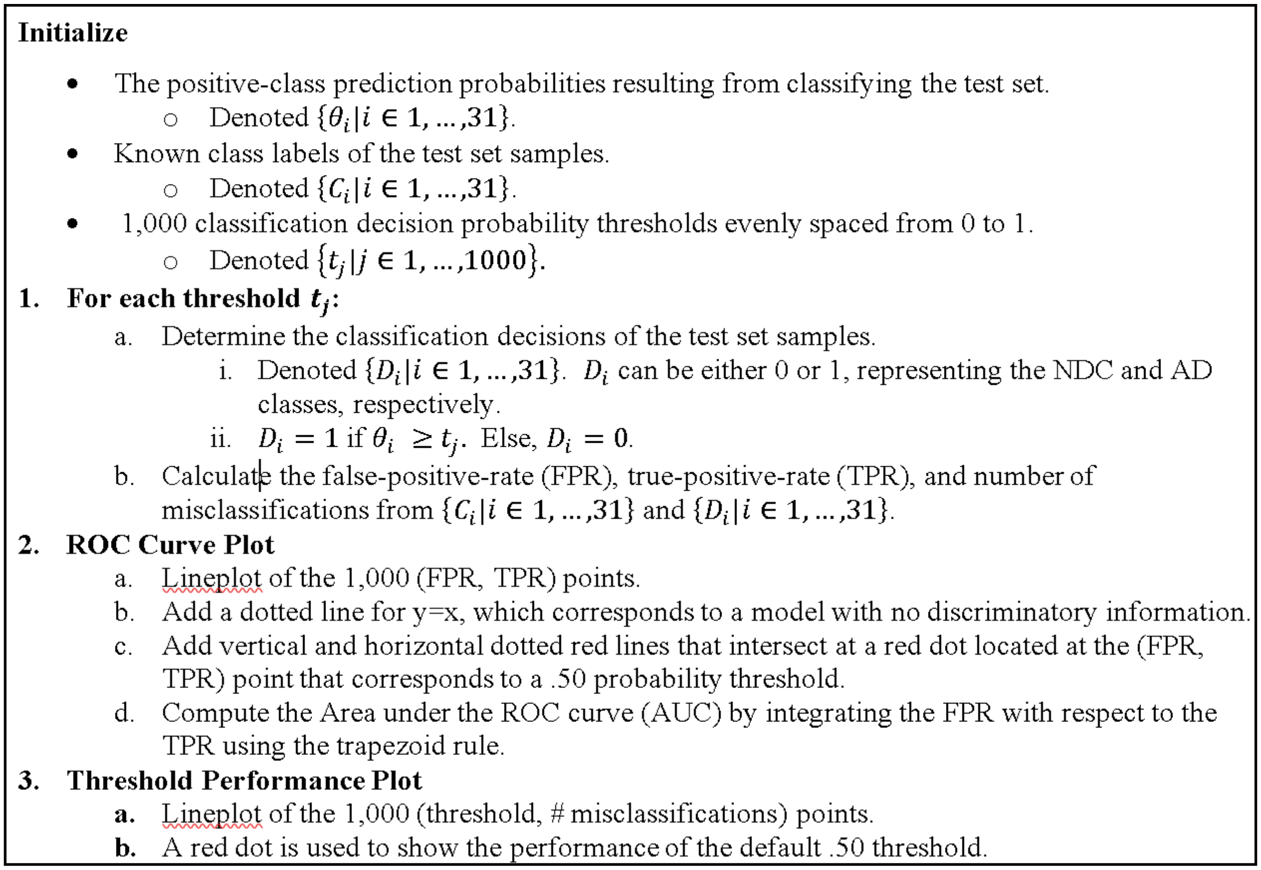
When classifying a sample, most models first estimate θ, which is the probability that the sample belongs to the positive class. A classification decision is then made for that sample based on θ and a probability threshold *t*. Samples where $\theta \geq t$ are classified positive, and negative otherwise. By default, $t = 0.5$, which assumes that false positives and false negatives are equally costly. Increases in the threshold sacrifice sensitivity for higher specificity, while lowering of the threshold produces the opposite effect.

An ROC curve is a two-dimensional graph used to evaluate the tradeoff a classifier makes between sensitivity and specificity as *t* varies between 0 and 1 (Hanley & McNeil, 1982). The x-axis is the false-positive rate, or 1 – specificity, and the y-axis is the true-positive rate, another name for sensitivity. The area under the ROC Curve [AUC] is a summary statistic of classification performance that represents the expected probability that, out of a randomly chosen pair of positive and negative samples, the positive sample will be determined to have a higher probability of belonging to the positive class than the negative sample (Hanley & McNeil, 1982).

AUC has several advantages as a classification performance metric in comparison to accuracy, the standard and more intuitive metric: it factors in the uncertainty of the predictions, is insensitive to unbalanced class proportions, and is independent of the final probability threshold used to determine classification decisions (Hanley & McNeil, 1982). Given these advantages, as well as the following additional reasons, AUC was the sole metric used in this project for the selection of subset sizes in Section 3.3.1, model parameters in Section 3.3.2, and the final model in Section 3.3.3, as well as the primary metric used for model validation in Section 3.4:

- The optimal classification decision threshold was unknown.
- There was some class imbalance in the training set and potentially more serious class imbalances in the bootstrap resamples used during feature selection and model parameter tuning.
- Accuracy is not a very meaningful statistic without also considering sensitivity and specificity. For instance, the null model would have 90% accuracy in a dataset with 90% class A and 10% class B. This model would seem decent if only accuracy were considered. Yet it would also have 0% specificity, making it useless for disease diagnosis. Since the feature selection and grid search methods are driven by a single summary metric of performance, it made more sense to use a stand-alone metric such as AUC.

Additionally, ROC curve analysis was performed for the test set performance of the final classifier trained on the optimum biomarker. While the *pROC* package was used to calculate AUC during feature selection and model parameter tuning, the ROC curve in this case was created manually from the test set classifications, the actual class labels, and 1,000 random decision thresholds evenly spaced from 0 to 1. The AUC statistic was then derived from the ROC curve using the trapezoidal rule. Specifically, the ROC curve analysis was implemented as follows:



```
Initialize
    • The positive-class prediction probabilities resulting from classifying the test set.
        o Denoted $\{\theta_i | i \in 1, \dots, 31\}$.
    • Known class labels of the test set samples.
        o Denoted $\{C_i | i \in 1, \dots, 31\}$.
    • 1,000 classification decision probability thresholds evenly spaced from 0 to 1.
        o Denoted $\{t_j | j \in 1, \dots, 1000\}$.
1. For each threshold $t_j$:
    a. Determine the classification decisions of the test set samples.
        i. Denoted $\{D_i | i \in 1, \dots, 31\}$. $D_i$ can be either 0 or 1, representing the NDC and AD
           classes, respectively.
        ii. $D_i = 1$ if $\theta_i \geq t_j$. Else, $D_i = 0$.
    b. Calculate the false-positive-rate (FPR), true-positive-rate (TPR), and number of
       misclassifications from $\{C_i | i \in 1, \dots, 31\}$ and $\{D_i | i \in 1, \dots, 31\}$.
2. ROC Curve Plot
    a. Lineplot of the 1,000 (FPR, TPR) points.
    b. Add a dotted line for y=x, which corresponds to a model with no discriminatory information.
    c. Add vertical and horizontal dotted red lines that intersect at a red dot located at the (FPR,
       TPR) point that corresponds to a .50 probability threshold.
    d. Compute the Area under the ROC curve (AUC) by integrating the FPR with respect to the
       TPR using the trapezoid rule.
3. Threshold Performance Plot
    a. Lineplot of the 1,000 (threshold, # misclassifications) points.
    b. A red dot is used to show the performance of the default .50 threshold.
```

Figure 8. Test set ROC curve analysis.

2.8 Hierarchical Clustering with Euclidean Distance and Ward Linkage

Clustering is an unsupervised learning method that seeks to partition objects into groups (clusters) such that objects in the same group are more similar than objects in separate groups.

The ClustVis web tool (Metsalu & Vilo, 2015) was used to cluster both the autoantibodies comprising the optimal biomarker determined in Section 3.3.3 and the samples in the training set, the results of which were visualized in a heatmap. Both the autoantibodies and samples were clustered using hierarchical clustering with the Euclidean distance metric and the ward linkage method. With Euclidean distance, the distance $d(\boldsymbol{x}, \boldsymbol{y})$ between two n-dimensional vectors **x** = $(x_1, x_2, \dots, x_n)$ and **y** = $(y_1, y_2, \dots, y_n)$ is:

$$d(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}\,, \tag{18}$$

Per Ward's minimum variance criterion method, the clustering algorithm used the Euclidean distance metric as follows:

1. Assign each object (autoantibody or sample) to a single cluster.
2. Set the initial distances between clusters to be the Euclidean distance between the objects.
3. Until there is only 1 cluster:
    a. Merge the two most similar clusters, which are defined to be the ones that lead to the minimum increase in total within-cluster variance after merging.
    b. For the newly merged cluster, re-calculate the distances to the other clusters using the Lance-Williams dissimilarity update formula, which for newly merged clusters C_i and C_j and another cluster C_k is as follows:

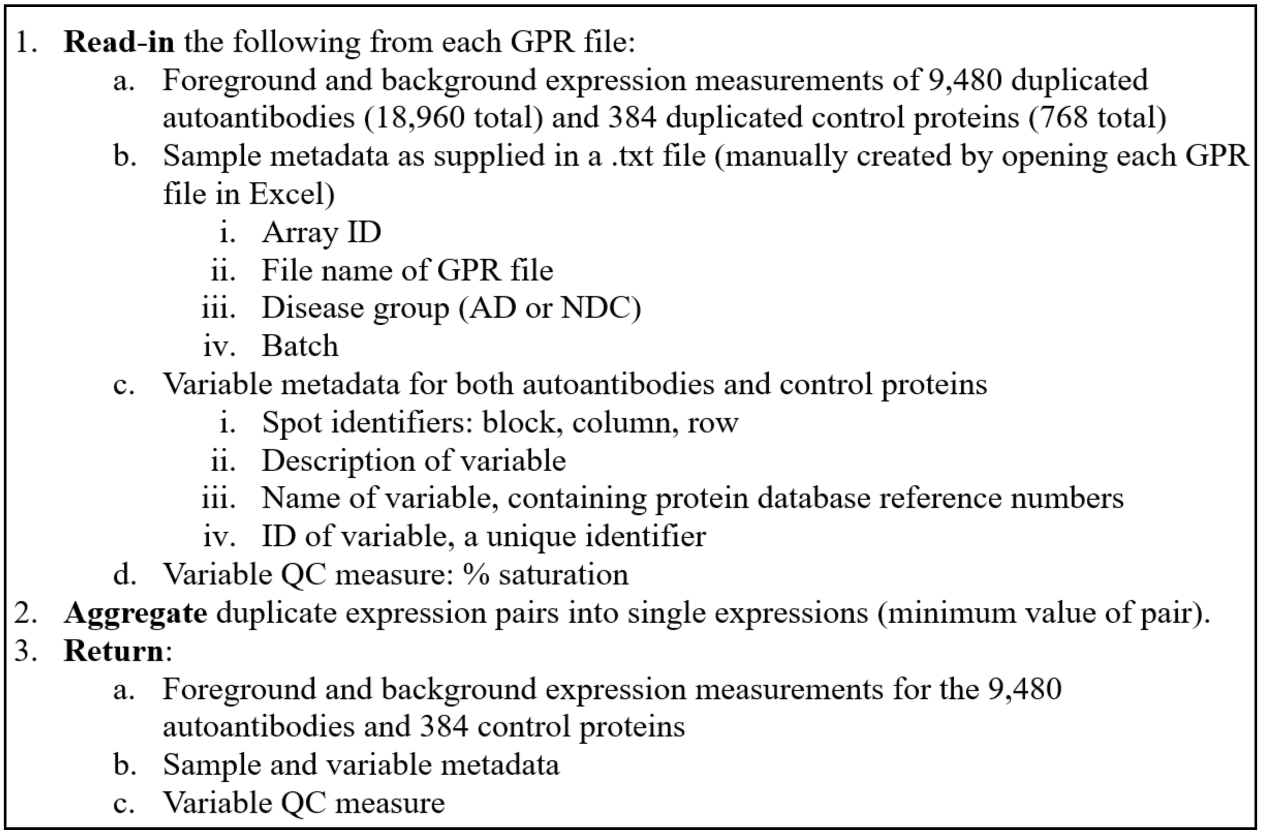
$$d(C_i \cup C_j, C_k) = \frac{n_i + n_k}{n_i + n_j + n_k} d(C_i, C_k) + \frac{n_j + n_k}{n_i + n_j + n_k} d(C_j, C_k) - \frac{n_k}{n_i + n_j + n_k} d(C_i, C_j), \tag{19}$$

where n_i, n_j, and n_k are the cluster sizes of clusters C_i, C_j, and C_k, respectively, and *d()* is the Euclidean distance function.

3 Experiments and Results

3.1 Data Preparation

The data in the downloaded GPR files were in a very raw format and required several data preparation and pre-processing steps to be converted into the final protein expression matrix used for supervised learning, where rows and columns correspond to samples and autoantibodies, respectively. The data preparation tasks, which are summarized in *Figure 9* below, began with importing the following for each of the 123 samples' GPR files: foreground and background expression of duplicated autoantibodies and control proteins, protein and sample metadata, and a protein quality control [QC] measure. The next step was to aggregate the pairs of expression measurements corresponding to each duplicated protein into single measures by retaining the minimum of each duplicate pair. These two initial data preparation steps were implemented via a custom R function that was a slightly modified version of the *loadGPR* function from the PAA package (Turewicz, 2016), which itself is a modified wrapper to the *read.maimages* function from the Limma package (Ritchie et al., 2015). The main addition to what was already provided by the *loadGPR* function was the inclusion of the % saturation quality control variable. The structure of the custom function was as follows:



1. **Read-in** the following from each GPR file:
    a. Foreground and background expression measurements of 9,480 duplicated autoantibodies (18,960 total) and 384 duplicated control proteins (768 total)
    b. Sample metadata as supplied in a .txt file (manually created by opening each GPR file in Excel)
        i. Array ID
        ii. File name of GPR file
        iii. Disease group (AD or NDC)
        iv. Batch
    c. Variable metadata for both autoantibodies and control proteins
        i. Spot identifiers: block, column, row
        ii. Description of variable
        iii. Name of variable, containing protein database reference numbers
        iv. ID of variable, a unique identifier
    d. Variable QC measure: % saturation
2. **Aggregate** duplicate expression pairs into single expressions (minimum value of pair).
3. **Return**:
    a. Foreground and background expression measurements for the 9,480 autoantibodies and 384 control proteins
    b. Sample and variable metadata
    c. Variable QC measure

Figure 9. Custom Data Preparation Function

Once the data was loaded into R, a QC filter was applied to eliminate autoantibodies whose spot expression intensities were, on average, too saturated to be trustworthy. Specifically, autoantibodies with $\overline{\%\,Saturation} \geq 2$ were discarded. The QC filter eliminated 31 autoantibodies, leaving 9,449 remaining in the dataset.

3.2 Exploratory Data Analysis

Before any pre-processing was done to the data, a quick exploratory data analysis [EDA] stage was implemented to examine the properties of the foreground expression of the autoantibodies and control proteins. Since the data was $\log_2$ transformed prior to normalization (see Section 3.3.2), the summary statistics and exploratory plots used in this section and Section 3.3.1 were presented using $\log_2$ transformed data to facilitate easy comparison with the normalized data.

Autoantibody foreground expression values in the dataset ranged from 4.858 to 16 (29 to 65,535 in original scale) with a mean of 9.2640 and standard deviation of 1.4279. The sample boxplots of foreground autoantibody expression shown in *Figure 10* provided a visual overview of the sample and class distributions:

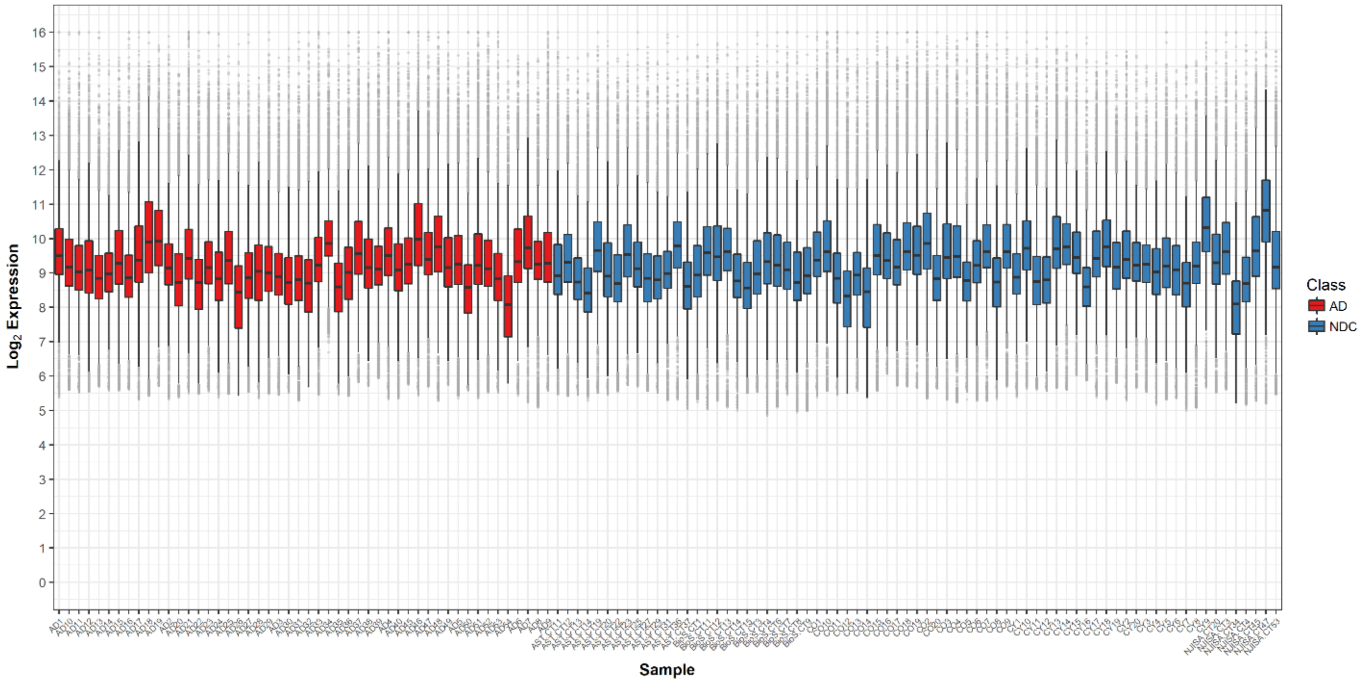


Figure 10. Array boxplots of log2 transformed foreground autoantibody expression after data preparation.

There was considerable inter-array variation in the distributions of foreground autoantibody expression. However, this was not cause for concern, since as mentioned in Section 2.1, the assumption of equal distribution of expression between arrays is not valid for protein microarray data. Per *Table 1* below, the differences in average sample statistics of foreground autoantibody expression between the AD and NDC classes were small. This was expected, since only a small subset of a person’s autoantibodies would be expected to be involved in any disease pathology.

Table 1. *Average sample stats per class of autoantibody foreground expression after data preparation in $\log_2$ scale.*

Class	Min	Max	Mean	SD	Q1	Median	Q3	IQR
AD	5.5315	15.7372	9.2411	1.3733	8.5171	9.1395	9.9371	1.4201
NDC	5.4269	15.6795	9.2797	1.3336	8.5797	9.1850	9.9640	1.3845

Some additional information about the samples in the dataset:

- While high-outliers for each sample extended further away from their medians than low-outliers in Figure 10, there were a similar number of high and low-outliers for each sample. On average, the AD samples had 305 high-outliers and 315 low outliers, while the NDC samples had an average of 277 high-outliers and 319 low-outliers.
- AD54 had the lowest median (8.09) and Q1 (7.14).
- NJISA.CT34 had the lowest Q3 (8.76) and IQR (1.02)
- NJISA.CT47 had the highest median (10.83), Q1 (9.89), Q3 (11.70), and fewest high-outliers (36).
- BioS.CT4 had the most low-outliers (525).
- AD26, AD54, CO14, NJISA.CT34, CO12, and AD32 all had zero low-outliers.

A major benefit of the protein microarrays used to collect the data is that they contain internal control features that can be utilized to assess the data quality and amount of non-biological variability in each array. Sample boxplots of foreground control protein expression are shown below:

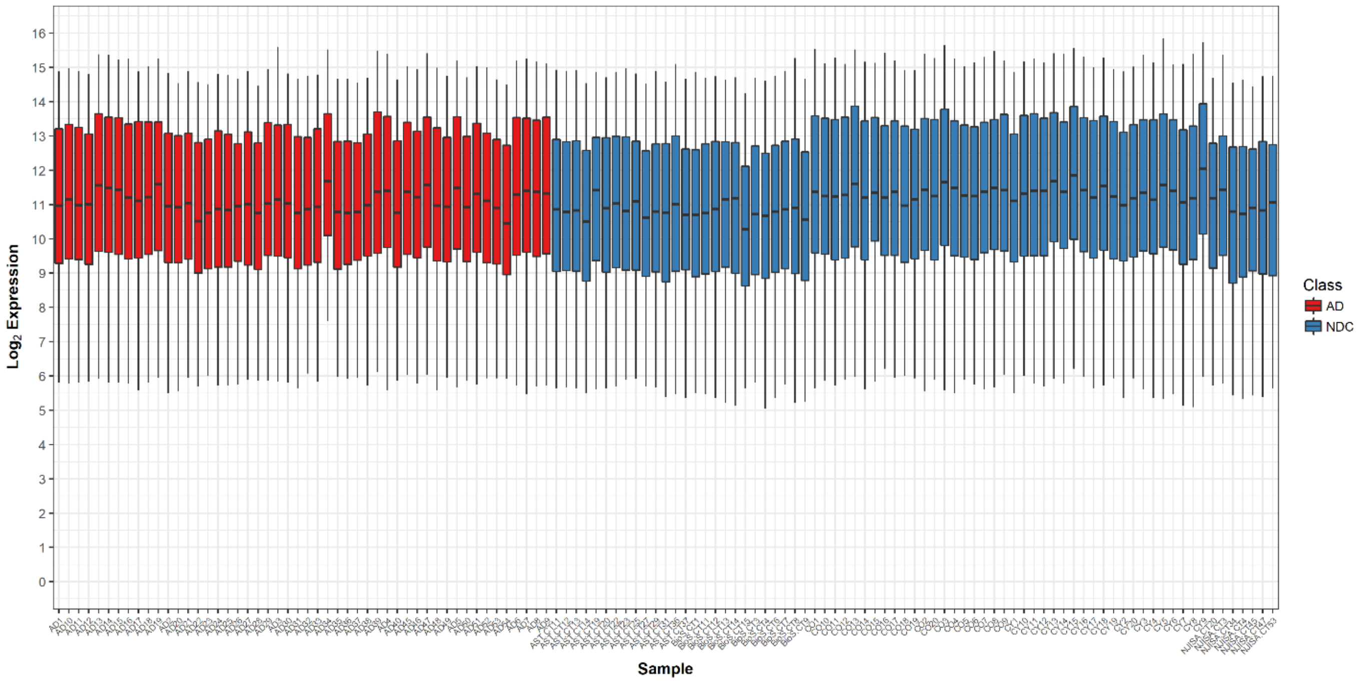


Figure 11. Array boxplots of log2 transformed foreground control protein expression after data preparation.

There was distinct inter-array distributional variation in control protein foreground expression, but much less than for the distributions of foreground autoantibody expression. However, unlike for autoantibodies, this inter-array distributional variation was evidence for non-biological variability, which was corrected for in section 3.3.

3.3 Pre-processing

3.3.1 Background Correction

The first pre-processing step was to correct the foreground signal for each autoantibody and control protein by eliminating the background noise and non-specific binding present on their spots on the array. The standard method of background subtraction was used, where background-corrected intensities of each autoantibody and control feature were calculated by subtracting the background intensity from the foreground intensity. Any corrected intensity that was below 1 was then set to 1 so that they would be 0 in $\log_2$ form, indicating a lack of signal.

Five autoantibodies were discarded due to having a corrected intensity of one for all 123 samples, leaving 9,444 remaining in the dataset.

In the background-corrected dataset, autoantibody expression in $\log_2$ form ranged from 0 to 15.9990 with a 4.86% decreased mean of 8.8141 and a 55.61% increased standard deviation of 2.2219. Sample boxplots and average sample statistics per class of background-corrected autoantibody expression are shown in the figure and table below:

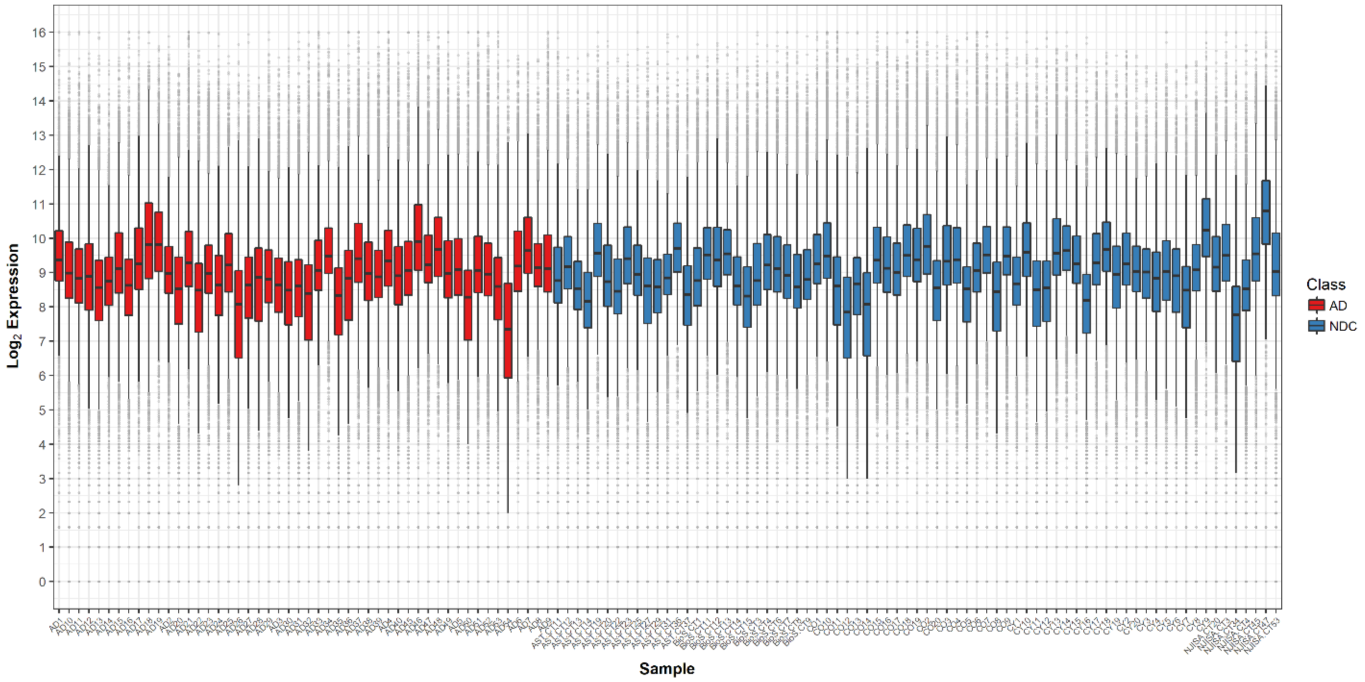


Figure 12. Array boxplots of log2 transformed autoantibody expression after background correction.

***Table 2**. Average sample stats per class of autoantibody expression after background correction in $\log_2$ scale.*

Class	Min	Max	Mean	SD	Q1	Median	Q3	IQR
AD	0	15.7348	8.7507	2.2083	8.0986	8.9395	9.8318	1.7333
NDC	0	15.6774	8.8575	2.1048	8.2393	9.0097	9.8699	1.6304

Per *Figure 12*, the inter-array distributional variation remained and had many similar characteristics as the dataset prior to background correction. However, there was a distinct downward shift in the expression distributions of each sample as background noise and non-specific binding were adjusted for. The downward shift was most noticeable in the lower-end of expression as seen in the sample box plots as well as the lowering of the average minimum and Q1 of the samples in each class in *Table 2*. The average number of low outliers per sample also increased from 318 to 445. The less-noticeable, small downward shift in the upper-end of the sample distributions was evidenced by the decrease in average sample Q3 for AD and NDC samples by 1.1% and 1%, respectively, and the decrease in the average number of high outliers per sample from 288 to 208. Additional information about samples in the background-corrected dataset:

- **Q1**: lowest was AD54 (5.93), highest was NJISA.CT47 (9.83).
- **Median**: lowest was AD54 (7.34), highest was NJISA.CT47 (10.79).
- **Q3**: lowest was NJISA.CT34 (8.59), highest was NJISA.CT47 (11.68).
- **IQR**: lowest was AST.CT31 (1.19), highest was AD54 (2.75).
- **Low-outliers**: lowest was CO12 (335), highest was BioS.CT4 (650).
- **High-outliers**: lowest was NJISA.CT47 (31), highest was AD23 (497).

Sample boxplots of background-corrected control protein expression shown below in *Figure 13* revealed that there was still a lot of non-biological variability remaining in the background-corrected data. This was expected, since background-correction only addressed the level of background noise and non-specific binding per spot on each array individually. It was not designed to correct for differences in confounding background effects between arrays, nor for the multitude of other possible sources of non-biological variability. These were corrected for in the next section.

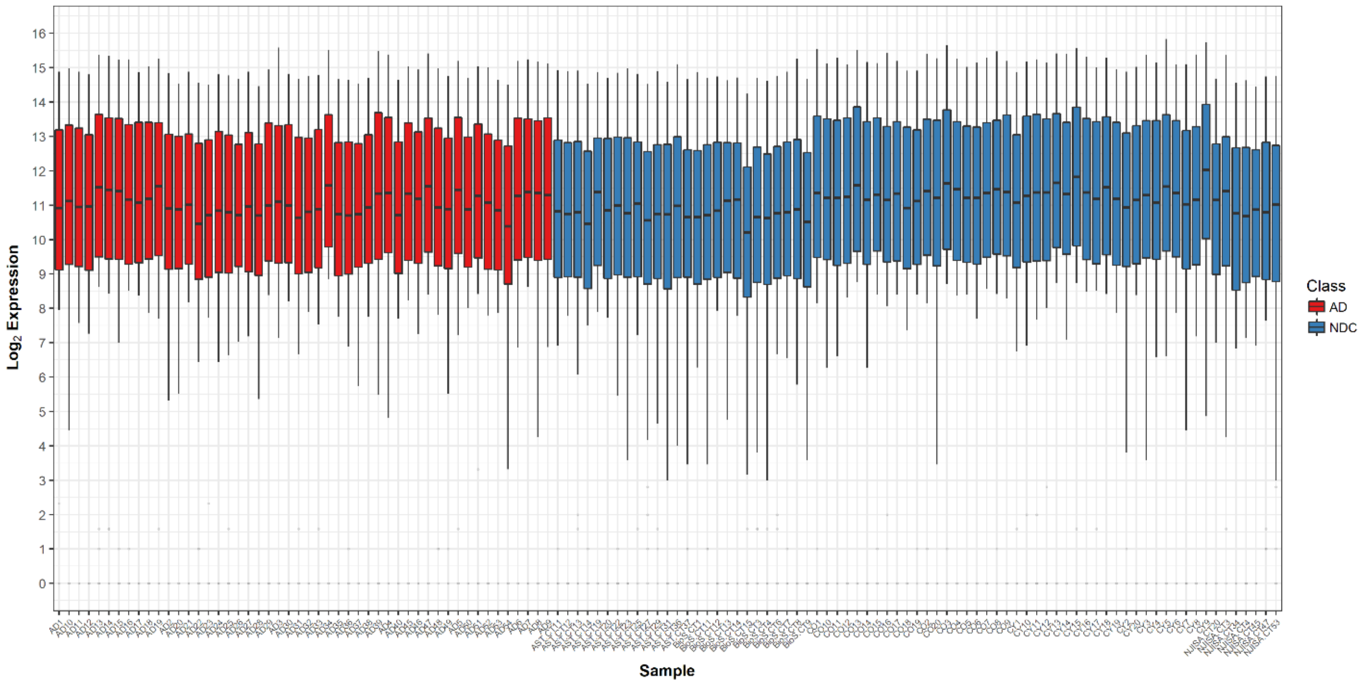


Figure 13. Array boxplots of log2 transformed control protein expression after background correction.

3.3.2 Normalization

RLM normalization, preceded by the log$_2$ transformation of protein expression, was then applied to the background-corrected dataset using the method described in Section 2.1 to eliminate intra-array and inter-array nonbiological variance from the data. As with background correction, any autoantibody log$_2$ expression value below zero after RLM normalization was subsequently set to zero. Autoantibody expression in the RLM-normalized dataset ranged from 0 to 16.6813 with a 0.61% decreased mean of 8.7605 and a 0.26% increased standard deviation of 2.2276. Sample boxplots and average sample statistics per class are shown below:

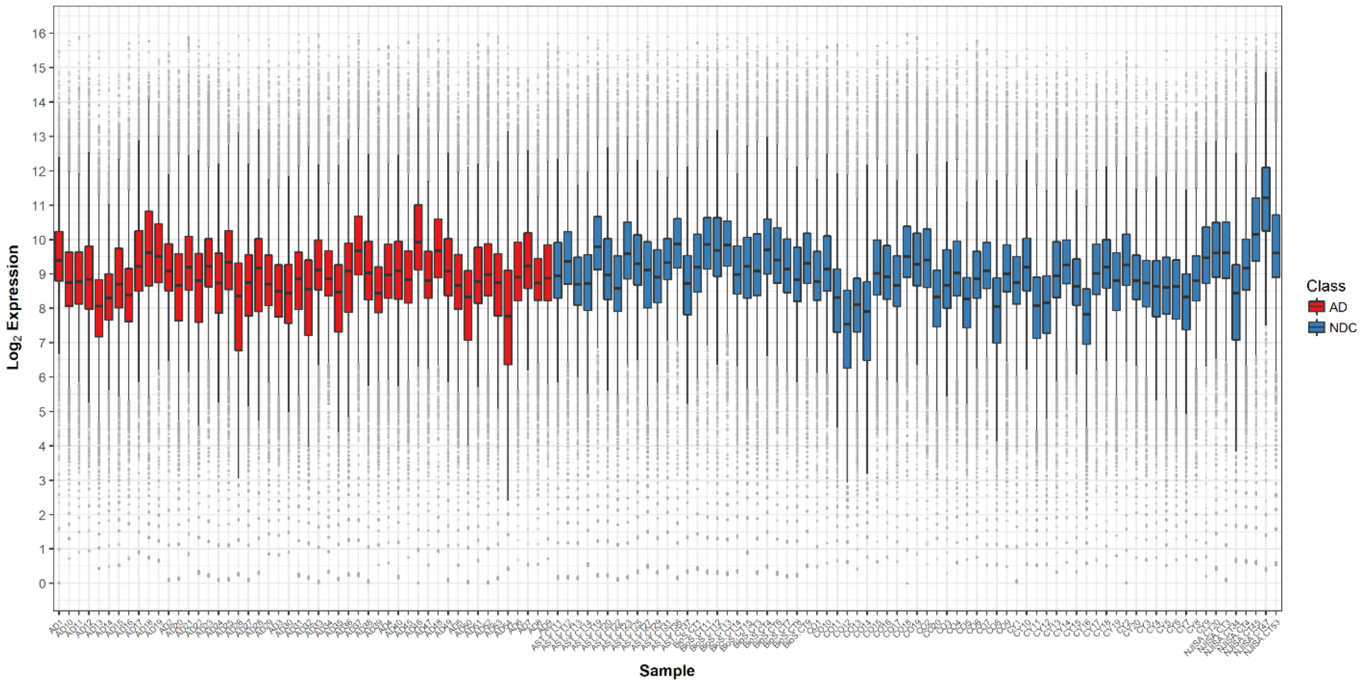


Figure 14. Boxplots of the autoantibody feature intensities in each array after RLM normalization.

***Table 3**. Average sample stats per class of autoantibody expression after RLM normalization.*

Class	Min	Max	Mean	SD	Q1	Median	Q3	IQR
AD	0	16.3082	8.6671	2.2583	8.0420	8.8644	9.8010	1.7595
NDC	0	16.6813	8.8244	2.2040	8.1960	9.9885	9.9410	1.7448

The sample distributions in *Figure 14* were very like the those in the background corrected dataset with very minor adjustments in the Q1, median, Q3, and location of outliers for various samples. Per *Table 3*, one exception is that the average of the median expression for NDC samples increased 10.86% from 9.0097 to 9.9885. This effect was not seen in the AD samples, as the average of their median expressions decreased 0.84% from 8.9395 to 8.8644. Other than the increased NDC sample medians, there were no drastic changes in the sample distributions after RLM normalization. From this, it was concluded that the level of non-biological variance in the data was not excessive and that there were no major data quality issues. Perhaps the most noticeable difference in the sample boxplots before and after normalization was that there was large uptick in variance among the extreme low end of expression for the samples. The background-corrected sample boxplots showed horizontal rows of dots at zero and one, whereas expression values ≤ 2 in *Figure 14* were much more randomly distributed. Additional information about samples in the RLM-normalized dataset:

- **Q1**: lowest was CO12 (6.14), highest was NJISA.CT47 (10.24).
- **Median**: lowest was CO12 (7.47), highest was NJISA.CT47 (11.21).
- **Q3**: lowest was CO12 (8.49), highest was NJISA.CT47 (12.10).
- **IQR**: lowest was AST.CT31 (1.19), highest was AD54 (2.74).
- **Low-outliers**: lowest was CO12 (339), highest was BioS.CT4 (650). Average amount per sample stayed the same (445).
- **High-outliers**: lowest was NJISA.CT47 (32), highest was AD23 (499). Average amount per sample stayed the same (208).

As described by Sboner et al. (2009), RLM normalization not only reduces intra-array and inter-array variability, but also improves class separation. This effect can be seen in the table below, which compares the distribution of Fisher's SNR for each autoantibody before and after the RLM normalization step:

***Table 4**. Distribution of Fisher's SNR before and after RLM Normalization.*

Dataset	Min	Q1	Median	Q3	Max
Background Corrected	0	.0048	.0213	.0650	1.7850
RLM Normalized	0	.0061	.0262	.0747	2.2400

RLM normalization increased the Q1 27.08%, the median 23.58%, the Q3 14.92%, and the maximum 25.49%.

3.4 Biomarker Discovery

Before implementing any biomarker discovery analysis, 80% of the AD samples and 70% of the NDC samples in the final pre-processed dataset were randomly selected into the training set, while the remaining samples served as an independent test set. The partitioning was done in this way, as opposed to an even 70% split, so that the training set would be a little larger and more balanced, without compromising the size of the test set too much. All analysis steps were conducted on the training set samples, while the test set was used solely for validating the final biomarkers. The specific 40 AD and 52 NDC samples that comprised the training set were as follows:

Table 5. *92 training set samples.*

AD				NDC			
AD2	AD17	AD36	AD50	CO1	CO14	CY12	AST.CT36
AD3	AD18	AD39		CO2	CO15	CY13	BioS.CT1
AD4	AD19	AD40		CO3	CO16	CY16	BioS.CT6
AD5	AD20	AD46		CO4	CO17	CY18	BioS.CT7
AD6	AD21	AD47		CO5	CO18	CY19	BioS.CT11
AD7	AD22	AD49		CO6	CO19	AST.CT13	BioS.CT12
AD8	AD24	AD51		CO7	CY2	AST.CT14	BioS.CT13
AD9	AD25	AD52		CO8	CY4	AST.CT19	BioS.CT14
AD10	AD27	AD53		CO9	CY5	AST.CT20	BioS.CT15
AD11	AD30	AD23		CO10	CY6	AST.CT22	NJISA.CT3
AD13	AD31	AD26		CO11	CY8	AST.CT23	NJISA.CT4
AD14	AD33	AD35		CO12	CY9	AST.CT25	NJISA.CT20
AD15	AD34	AD37		CO13	CY11	AST.CT29	NJISA.CT34

The 10 AD and 21 NDC samples that comprised the test set:

Table 6. *31 test set samples.*

AD	NDC		
AD1	CO20	AST.CT12	NJISA.CT53
AD12	CY1	AST.CT27	
AD16	CY3	AST.CT31	
AD28	CY7	AST.CT37	
AD29	CY10	BioS.CT3	
AD32	CY14	BioS.CT4	
AD38	CY15	BioS.CT8	
AD45	CY17	BioS.CT9	
AD48	CY20	NJISA.CT45	
AD54	AST.CT11	NJISA.CT47	

3.4.1 Feature Selection

To identify a parsimonious subset of autoantibodies with the optimal expectation of classification efficacy on unseen samples, multivariate feature selection was implemented on the training dataset using the modified recursive feature elimination method described in Section 2.5 with two non-parametric learning algorithms popular in bioinformatics applications where there are many more features than samples. These two learning algorithms were random forests and support vector machines, described in Sections 2.3 and 2.4, respectively. The two feature selection processes are termed *random forest – recursive feature elimination* [RF-RFE] and *support vector machine – recursive feature elimination* [SVM-RFE].

Many of the initialization parameters were the same for both RF-RFE and SVM-RFE, while others depended on the learning algorithm. The initializations which remained constant regardless of the learning algorithm were as follows:

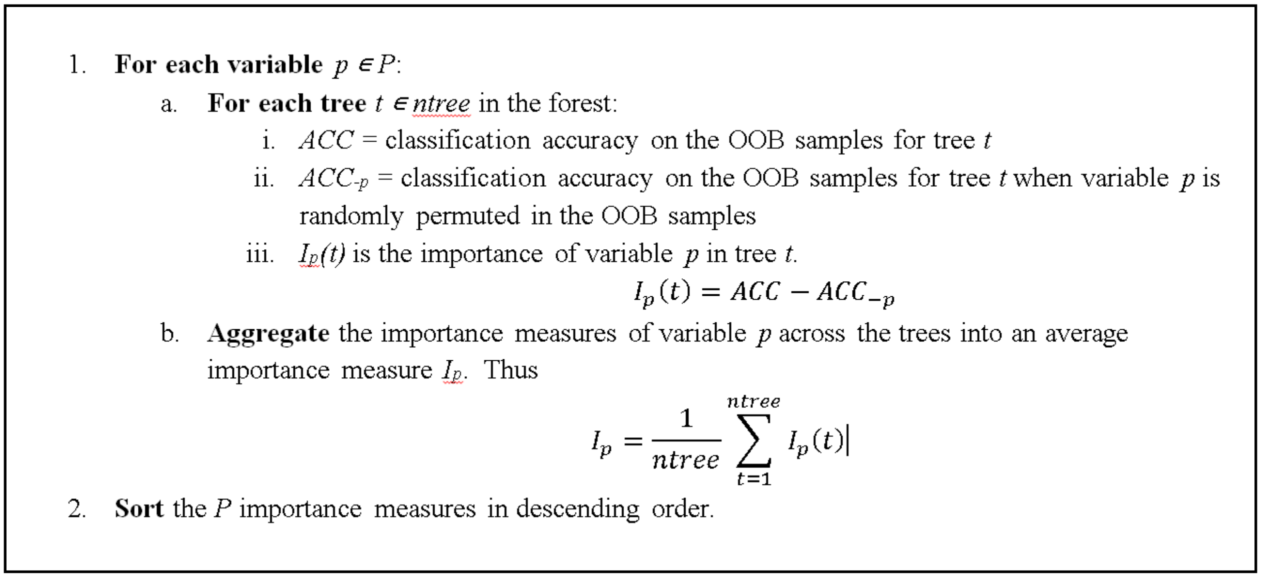
Table 7. RFE initializations independent of learning algorithm.

Initialization	Description
FSS	9444, 4722, 2361, 1181, 591, 296, 148, 74, 37, 30, 29, 28, …, 2
Pred	Default prediction function.
Summary	Calculates four summary measures of classification performance: AUC, accuracy, sensitivity, and specificity.
selectSize	Identifies the optimum subset size using the 1% tolerance rule, a common alternative to the one-standard-error rule recommended by Breiman, Friedman, Stone, & Olshen (1984) to protect against overfitting. The 1% tolerance rule is applied as follows: 1. Let the subset size S_{max} be the subset size ∈ FSS with the maximum $\overline{\text{OOB AUC}}$. If $S_{max} > 30$, then modify S_{max} to be the subset size ≤ 30 with the highest $\overline{\text{OOB AUC}}$ to protect against overfitting. 2. Let S_{tol} be the smallest subset size ∈ FSS with $\overline{\text{OOB AUC}}$ that is within 1% of S_{max}'s $\overline{\text{OOB AUC}}$. 3. Return S_{tol} as the optimum subset size.
selectVar	Identifies the optimum biomarker based on aggregate variable importance rankings at the optimum subset size S_{tol} using the "most frequent" rule, which works as follows: 1. Let {P} represent the set of variables that appeared at least once in the 500 models fit at subset size S_{tol}. 2. Calculate the number of times out of the 500 resamples that each variable in {P} was retained at subset size S_{tol}. Sort the variables by frequency in descending order. 3. Return the S_{tol} variables most frequently retained at subset size S_{tol}.

The remaining two initializations which varied depending on the learning algorithm were the *Fit* function and the *Rank* function. The RF-specific initializations are described below:

Table 8. RF-specific RFE initializations.

Initialization	Description
Fit	Trains a RF classifier with 2,000 trees using the randomForest package (Liaw & Wiener, 2002). The *mtry* parameter is always set to the square root of current number of variables in the model.
Rank	Ranks the variables in the model by the mean decrease in accuracy that occurs after each variable is randomly permuted in the forest. Pseudocode for this ranking function is shown in Figure 15 below, where *P* is the number of variables in the model, *ntree* is the total number of trees in the forest.



```
1. For each variable $p \in P$:
    a. For each tree $t \in ntree$ in the forest:
        i. $ACC$ = classification accuracy on the OOB samples for tree $t$
        ii. $ACC_{-p}$ = classification accuracy on the OOB samples for tree $t$ when variable $p$ is
            randomly permuted in the OOB samples
        iii. $I_p(t)$ is the importance of variable $p$ in tree $t$.
                 $I_p(t) = ACC - ACC_{-p}$
    b. Aggregate the importance measures of variable $p$ across the trees into an average
       importance measure $I_p$. Thus
                 $I_p = \frac{1}{ntree} \sum_{t=1}^{ntree} I_p(t)$
2. Sort the $P$ importance measures in descending order.
```

Figure 15. Random forest variable ranking algorithm.

The SVM-specific RFE initializations are described below:

***Table 9**. SVM-specific RFE initializations.*

Initialization	Description
Fit	Trains a soft-margin linear SVM classifier using the *kernlab* package (Karatzoglou, Smola, Hornik, & Zeileis, 2004). As is default, the cost parameter C is kept constant at 1.
Rank	Ranks each variable p in the model by the square of its weight w_p as is recommended by Rakotomamonjy (2003). Higher squared weights indicate greater multivariate importance in the SVM model.

The results of the two feature selection procedures are reviewed below, first for RF-RFE, then for SVM-RFE:

RF-RFE:

Displayed below are two plots summarizing the results of RF-RFE. One shows the mean OOB values of each performance metric per subset size and the other is a strip plot that shows the distribution of AUC performance for each subset size:

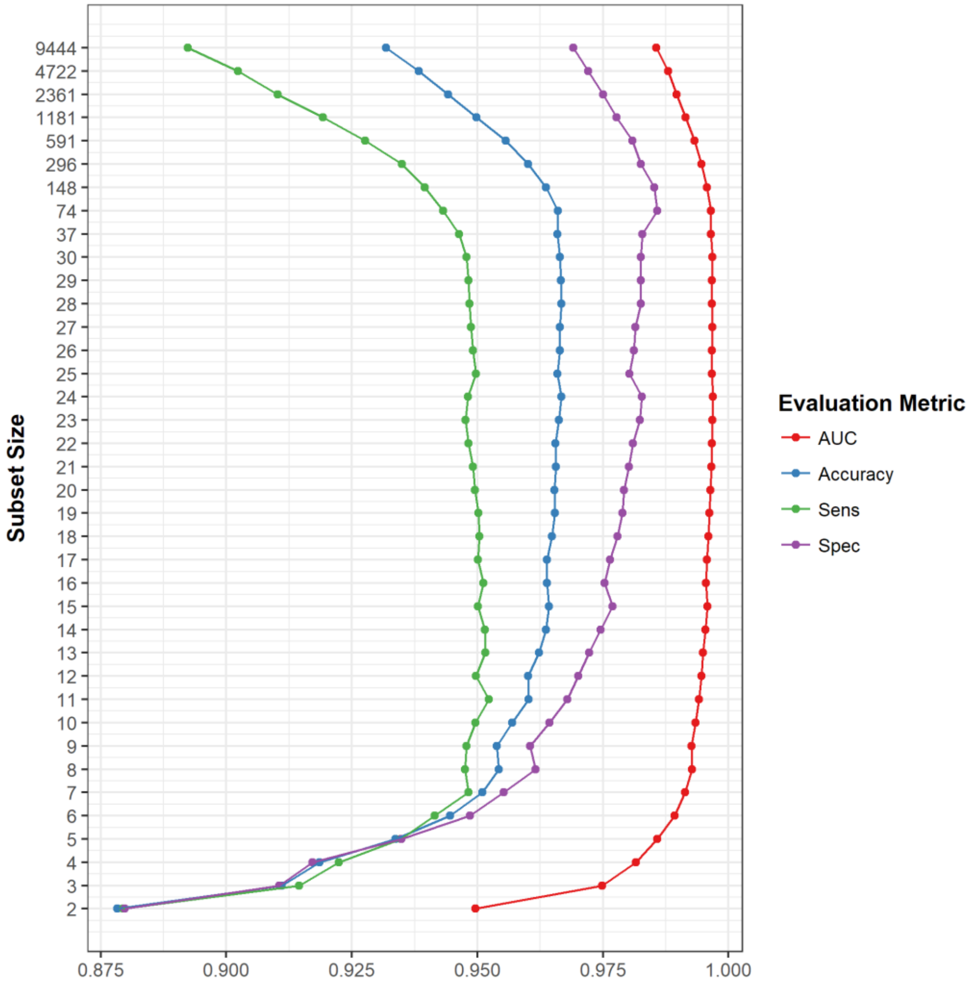


Figure 16. Mean OOB performance per subset size during RF-RFE.

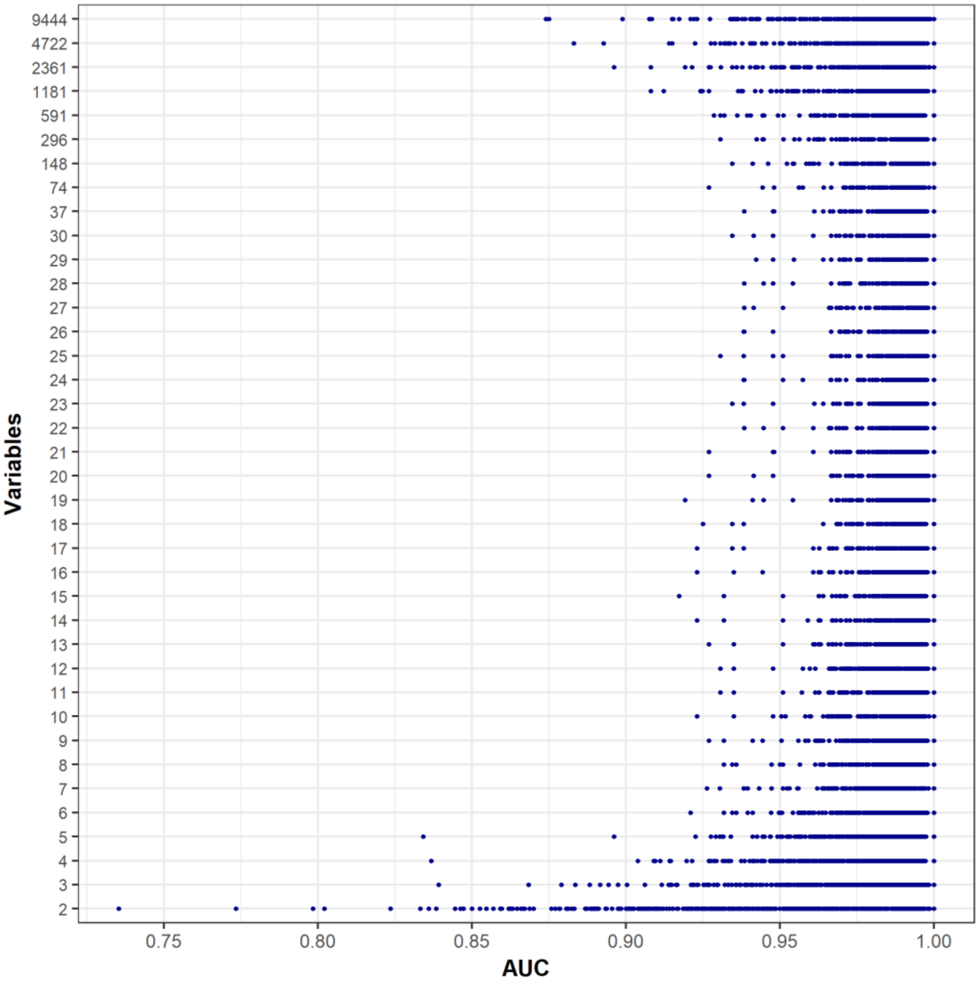


Figure 17. Strip plot of AUC performance per subset size during RF-RFE.

Per *Figure 16*, the subset size S_{max} with the maximum $\overline{\text{OOB AUC}}$ was 24. Applying the 1% tolerance rule, the subset size S_{tol} with the maximum $\overline{\text{OOB AUC}}$ that is still within 1% of S_{max}'s $\overline{\text{OOB AUC}}$ was 6. The final size of the RF biomarker was thus 6. Additional observations about the results of RF-RFE from *Figures 16-17* (all performance values refer to mean OOB performance unless otherwise specified):

- AUC performance had distinct improvements in each iteration as it went from 0.9856 at subset size 9444 to 0.9965 at size 74. After 74, AUC improved slightly until reaching its maximum of 0.9969 at size 24. After 24, AUC performance slowly declined until 0.9927 at size 8, followed by a faster decline until reaching its minimum of 0.9497 at size 2.

- As expected, the worst performing subset size was 2 with AUC = 0.9497, accuracy = 0.8783, sensitivity = 0.8795, and specificity = 0.8797.
- Size 24 had the highest accuracy (0.9667) as well as highest AUC (0.9969).
- Size 11 had the highest sensitivity (0.9523).
- Size 74 had the highest specificity (0.9858).
- Size 29 had the lowest variance in OOB AUC ($4.9987e^{-5}$). Size two had the highest ($1.6412e^{-3}$).
- The increase in performance as the subset size dropped from 9444 to 74 was driven by increases in both sensitivity and specificity, but more so by increases in specificity.
- Specificity was greater than sensitivity for all subset sizes greater than 5. At five, they were equivalent. At sizes 3 and 4, sensitivity was greater than specificity. At size 2, specificity was again slightly larger.

Aggregate variable importance rankings at subset size six were then used to determine the six autoantibodies most frequently selected into 500 six-feature random forest classifiers during RF-RFE. These six autoantibodies were the final RF biomarker and are displayed below:

***Table 10**. The six-autoantibody random forest biomarker.*

Feature	BRC ID	Description
FRMD8	B17R09C11	FERM domain containing 8
MGC24125	B36R03C09	Hypothetical protein MGC24125
IL20	B26R16C03	interleukin 20
PCBD2	B24R17C03	pterin-4 alpha-carbinolamine dehydratase/dimerization cofactor of hepatocyte nuclear factor 1 alpha (TCF1) 2
DNAJC8	B18R07C13	DnaJ homolog subfamily C member 8
PRELID2	B07R02C17	PRELI domain-containing protein 2

SVM-RFE:

The mean OOB performance per subset size returned by SVM-RFE:

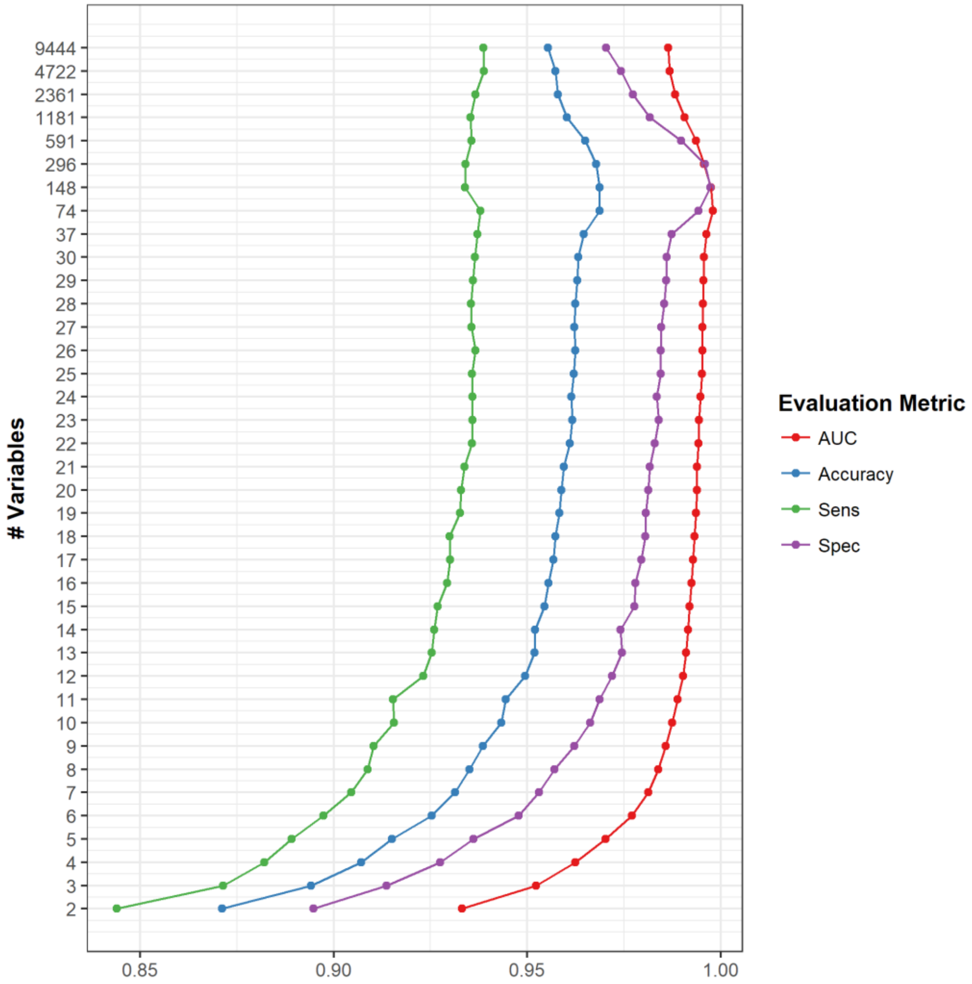


Figure 18. Mean OOB performance per subset size during SVM-RFE.

Strip plot of the distribution of OOB AUC performance per subset size:

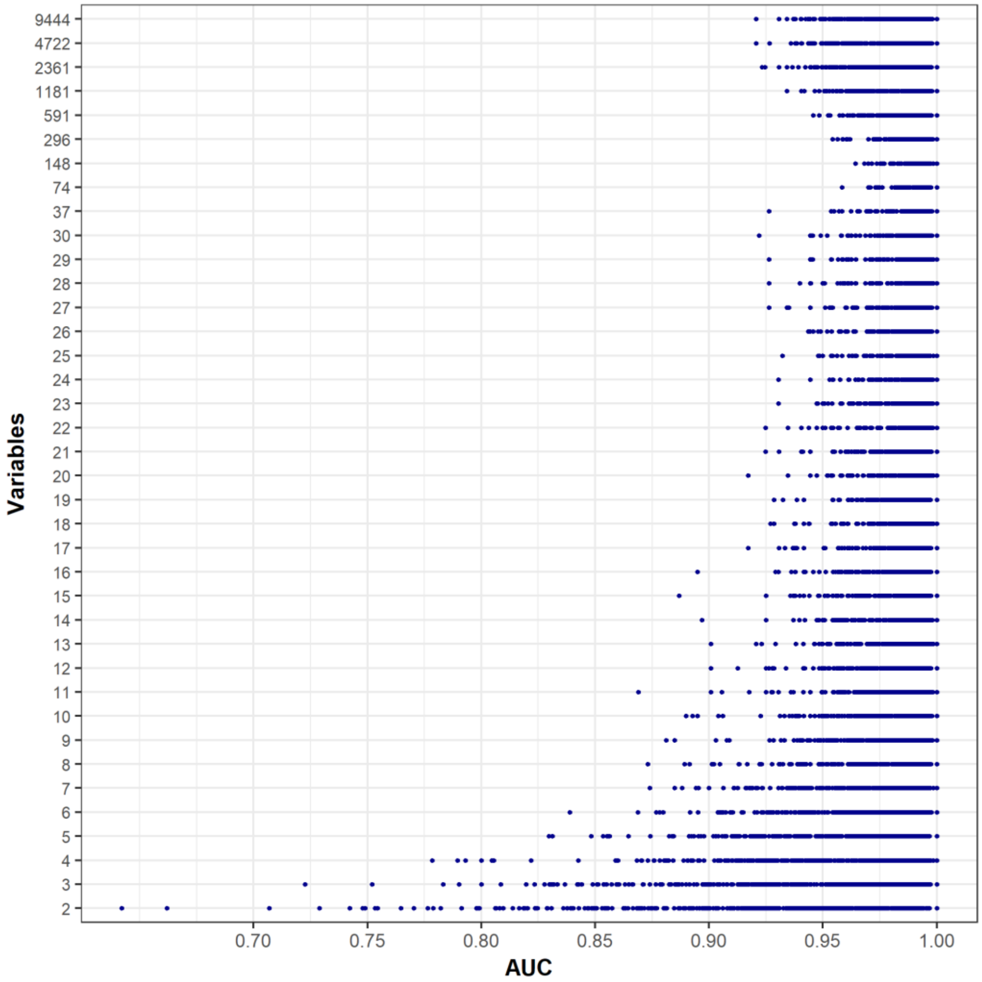


Figure 19. Strip plot of AUC performance per subset size during SVM-RFE.

S_{max}, the subset size ≤ 30 with the highest $\overline{\text{OOB AUC}}$, was 30 with AUC = 0.9957. Applying the 1% tolerance rule, the subset size S_{tol} with the maximum $\overline{\text{OOB AUC}}$ that is still within 1% of S_{max}'s $\overline{\text{OOB AUC}}$ was 9 with AUC = .9858. The final size of the SVM biomarker was thus 9. Additional observations about the results of SVM-RFE from *Figures 18 and 19* (all performance values refer to mean OOB performance unless otherwise specified):

- As the subset size dropped from 9444 to 74, AUC improved gradually from 0.9865 to 0.9980. From 74 to 17, AUC declined very slowly to 0.9929, after which a more accelerated decline began until reaching size 2 with AUC = 0.9331. Size 2 also had the worst accuracy (0.8711), sensitivity (0.8439), and specificity (0.8948).

- The performance gains from subset size 9444 to 148 were driven by increases in specificity, while sensitivity fell slightly in that subset size interval. From 148 to 74, performance improved slightly due to an increase in sensitivity and decrease in specificity. From 74 to 2, sensitivity and specificity declined mostly in sync.
- While size 30 was the subset size ≤ 30 with the highest AUC, size 74 had the highest AUC (0.9980) out of all evaluated subset sizes. Size 74 also had the highest accuracy (0.9687).
- Size 4722 had the highest sensitivity (0.9388).
- Size 148 had the highest specificity (0.9973).
- Size 74 had the lowest variance in OOB AUC ($2.5e^{-5}$). Size two had the highest ($3.2e^{-3}$).
- At all subset sizes, specificity was higher than sensitivity.

Aggregate variable importance rankings at subset size nine were then used to determine the nine autoantibodies most frequently selected into 500 nine-feature support vector machine classifiers during SVM-RFE. These nine autoantibodies were the final SVM biomarker and are displayed below:

***Table 11**. The nine-autoantibody support vector machine biomarker.*

Feature	BRC ID	Description
GTF2I	B02R19C01	General transcription factor II-I
PHF15	B21R12C15	PHD finger protein 15
LGALS1	B35R11C21	lectin, galactoside-binding, soluble, 1
FRMD8	B17R09C11	FERM domain containing 8
PCBD2	B24R17C03	pterin-4 alpha-carbinolamine dehydratase/dimerization cofactor of hepatocyte nuclear factor 1 alpha 2
IL20	B26R16C03	interleukin 20
PTCD2	B25R11C21	pentatricopeptide repeat domain 2
IL4	B28R16C01	interleukin 4, transcript variant 1
ANKHD1	B15R12C09	ankyrin repeat and KH domain containing 1, transcript variant 3

3.4.2 Hyperparameter Optimization

After the RF and SVM biomarkers were identified, the parameters of the RF and SVM classifiers were then optimized using the grid search method described in Section 2.6. The tuning grid evaluated for each classifier is shown below:

Table 12. Parameter tuning grids of final classifiers.

Classifier	Model Parameter	Tuning Grid
Random Forest	*mtry*	{2, 3, 4, 5, 6}
Support Vector Machine	*C*	{.25, .50, 1, 2, 4}

The mean OOB performances of each evaluation metric returned by the SVM grid search were as follows for each evaluated cost parameter:

Table 13. SVM grid search results for classifier trained on SVM biomarker.

C	AUC	Sensitivity	Specificity	Accuracy
0.25	**1.0000**	0.9892	**0.9962**	**0.9929**
0.50	0.9999	**0.9896**	0.9921	0.9906
1	0.9996	0.9881	0.9885	0.9878
2	0.9994	0.9894	0.9854	0.9867
4	0.9994	0.9883	0.9865	0.9868

Since *C* = 0.25 had the highest mean AUC when predicting the OOB samples, a perfect 1 in fact, 0.25 was considered the optimal value of the cost parameter. It also scored the highest mean OOB specificity (0.9962) and accuracy (0.9929), but its mean OOB sensitivity (0.9892) was slightly lower than for *C* = 0.50 (0.9896). A final SVM classifier was then fit using the nine-autoantibody SVM biomarker on the full 92-sample training set with *C* = 0.25. The resulting classifier had 10 support vectors, meaning only 10 training set samples contributed to the classification function.

The mean OOB performances of each evaluation metric returned by the RF grid search were as follows for each evaluated *mtry* parameter value:

***Table 14**. RF grid search results for classifier training on RF biomarker.*

mtry	AUC	Sensitivity	Specificity	Accuracy
2	**0.9950**	**0.9677**	**0.9615**	**0.9637**
3	0.9927	0.9564	0.9478	0.9509
4	0.9907	0.9459	0.9375	0.9407
5	0.9885	0.9366	0.9297	0.9323
6	0.9868	0.9305	0.9260	0.9276

Since *mtry* = 2 had the highest mean AUC (0.9950) when predicting the OOB samples, the number of randomly selected variables (out of the six in the biomarker) estimated to be optimal to use for each decision tree in the RF classifier was two. In addition to AUC, *mtry* = 2 also scored the highest mean OOB sensitivity (0.9677), specificity (0.9615), and accuracy (0.9637). A final RF classifier was then fit using the six-autoantibody RF biomarker on the full 92-sample training set with mtry = 2.

3.4.3 Biomarker Selection

In this section, the OOB classification results for the RF and SVM biomarkers were compared with the goal of selecting the one expected to classify unseen samples with higher efficacy. While AUC was the primary evaluation metric, sensitivity, specificity, and accuracy were also used for the biomarker comparison. *Table 15* below shows the mean OOB performance measures of the SVM and RF biomarkers (taken from the grid search in the previous section) corresponding to their final model parameters, while *Figure 20* shows the corresponding 95% confidence intervals:

***Table 15**. Mean OOB performance of RF and SVM biomarkers.*

	Biomarker Size	Model Parameter	AUC	Sensitivity	Specificity	Accuracy
SVM	9	*C* = 0.25	1.0000	0.9892	0.9962	0.9929
RF	6	*mtry* = 2	0.9950	0.9677	0.9615	0.9637

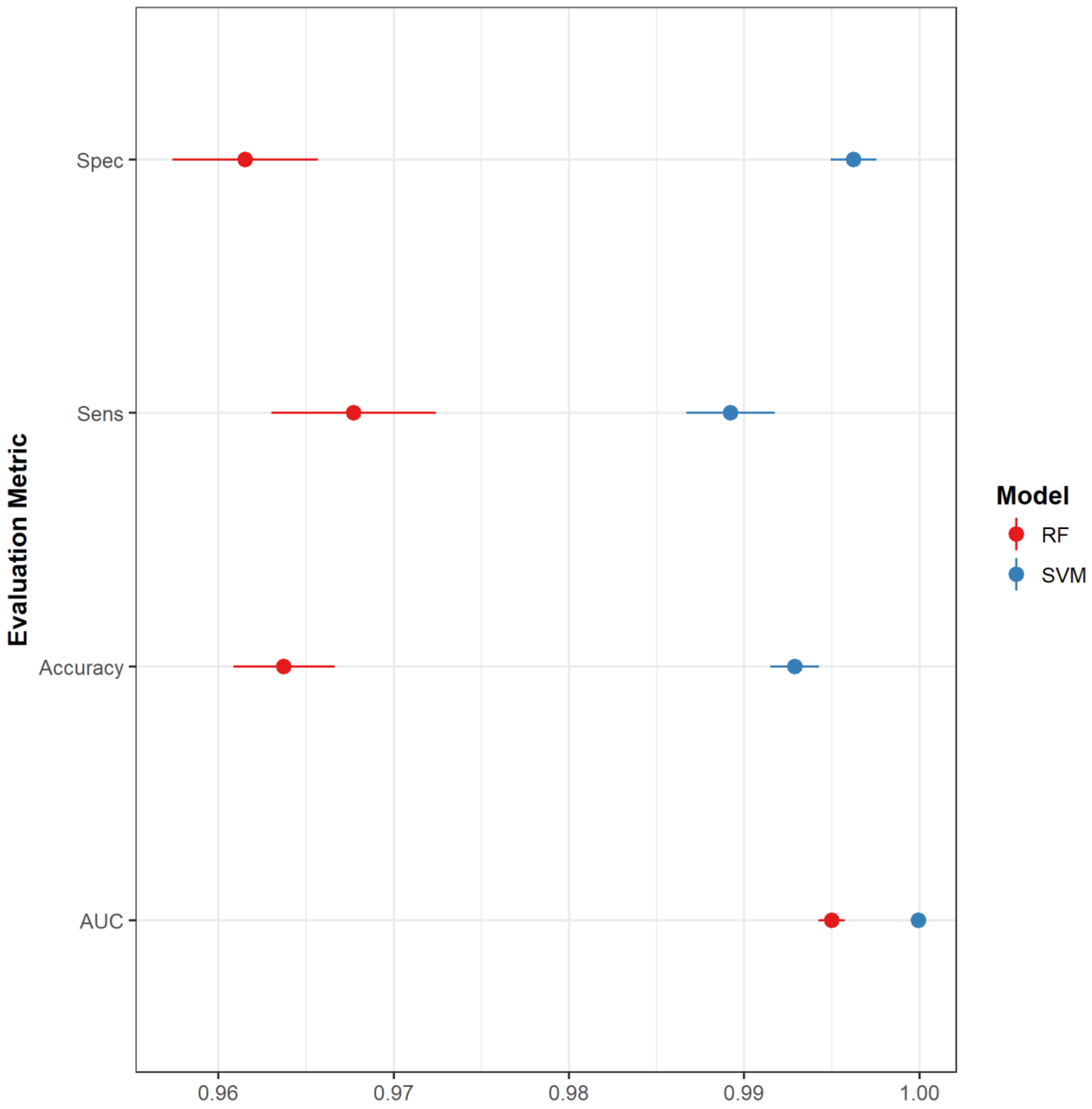


Figure 20. 95% confidence intervals of the mean OOB performance measures for the RF and SVM biomarkers corresponding to their final model parameters.

Per *Table 15*, the SVM classifier had superior OOB classification performance for each evaluation metric. The mean OOB sensitivity, specificity, and accuracy for the SVM biomarker were 2.22%, 3.61%, and 3.03% higher, respectively, than for the RF biomarker. However, mean OOB AUC for the SVM biomarker was only 0.5% higher. Additionally, for each evaluation metric in *Figure 20*, the SVM biomarker's lower bound of the 95% confidence interval of mean

OOB performance was higher than the upper bound of the RF biomarker's confidence interval. The SVM biomarker's confidence intervals were also distinctly narrower than those for the RF biomarker, which indicated greater confidence in the performance estimations.

A paired t-test was also applied for each evaluation metric, each of which tested the hypothesis of zero difference in mean OOB performance (on the same 500 bootstrap resamples) between the final RF and SVM classifiers. The table below displays the results of the four paired t-tests (RF – SVM), as well as 95% confidence intervals of the difference in mean OOB performance for each evaluation metric:

***Table 16**. Paired t-test information and 95% confidence intervals of the mean difference in each performance metric between the final two classifiers. RF - SVM.*

					95% Confidence Interval	
	Mean Difference	**df**	**t**	**p-value**	**LB**	**UB**
AUC	-0.0050	499	-12.910	< 2.2e^{-16}	-0.0057	-0.0042
Accuracy	-0.0292	499	-18.125	< 2.2e^{-16}	-0.0323	-0.0260
Sensitivity	-0.0215	499	-7.998	8.875e^{-15}	-0.0268	-0.0162
Specificity	-0.0347	499	-16.554	< 2.2e^{-16}	-0.0388	-0.0306

With α = 0.05, all four null hypotheses of equivalent performance were rejected, since each p-value was less than 0.05. Each 95% confidence intervals' upper-bound was also less than 0, further validating the rejections of the hypotheses. Note that the p-values for each hypothesis test were extremely low and each hypothesis was rejected with much greater confidence than 95%. Since the performance differences were calculated by subtracting the SVM performance from the RF performance, the results also further validated the conclusion of the SVM biomarker having superior expectations of classification performance on new data.

Given the OOB results, the SVM biomarker would seem the clear choice for the final biomarker. However, there were two other factors to consider. One, the RF biomarker had three fewer autoantibodies and still had almost as good performance as the SVM biomarker. Having as small a biomarker as possible is a vital concern when conducting biomarker discovery on a dataset suffering from the curse of dimensionality (many more variables than samples), where overfitting is a major threat to the identification of a generalizable biomarker. Two, if only AUC is considered and not the other three performance metrics, which are based on a specific decision threshold (0.50), then the performance of the two biomarkers were much closer to equivalent. Given these two factors, both biomarkers were considered the final biomarkers and were subsequently interpreted and validated.

3.5 Biomarker Interpretation

The RF and SVM biomarkers were interpreted using two methods. The first was variable importance rankings, which used the same model-based variable importance measures that were used to drive the elimination of variables during feature selection. The second method was the two-way hierarchical clustering method described in Section 2.8, which used Euclidean distance and Ward linkage. The clustering results were visualized using heatmaps created with the ClustVis web tool (Metsalu & Vilo, 2015), which revealed for each biomarker clusters of both autoantibodies and samples in the training set with similar expression patterns. The dendrograms on the top of each heatmap show the clustering structure of the autoantibodies, while the dendrograms on the left show the structure of the sample clusters. Both autoantibody and sample dendrograms were ordered by median expression value in descending order.

Random Forest:

The following variable importance rankings were extracted from the RF biomarker with its final model parameter (*mtry = 2*) using the mean decrease in accuracy criterion (Figure 21):

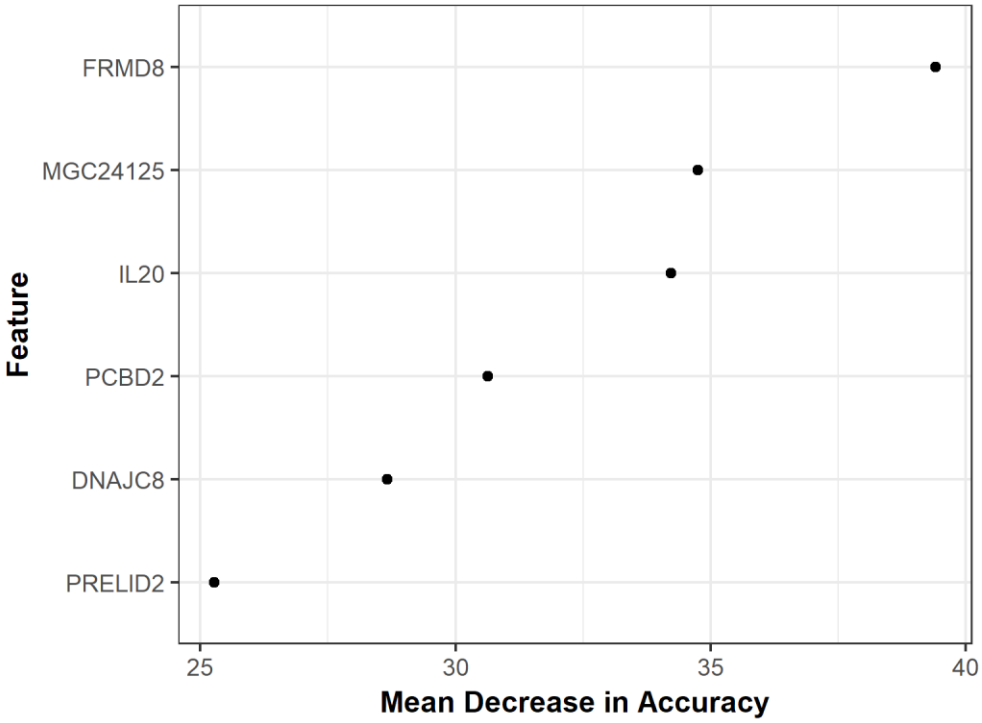


Figure 21. Random forest biomarker autoantibodies ranked using the mean decrease in accuracy criterion.

The most important feature in the RF model was FRMD8. When this autoantibody's expression values were randomly perturbed in the trees of the forest, the average decrease in accuracy was 39.87%. The next important features in descending order were: MGC24125, IL20, PCBD2, DNAJC8, and PRELID2. The least important of the six variables in the RF biomarker still caused an average decrease in mean accuracy of over 25% when its values were randomly perturbed in the trees.

The heatmap resulting from two-way hierarchical clustering applied to the RF biomarker in the 92-sample training set using Euclidean distance and Ward linkage (trees in row and column dendrograms are ordered by median expression in descending order):

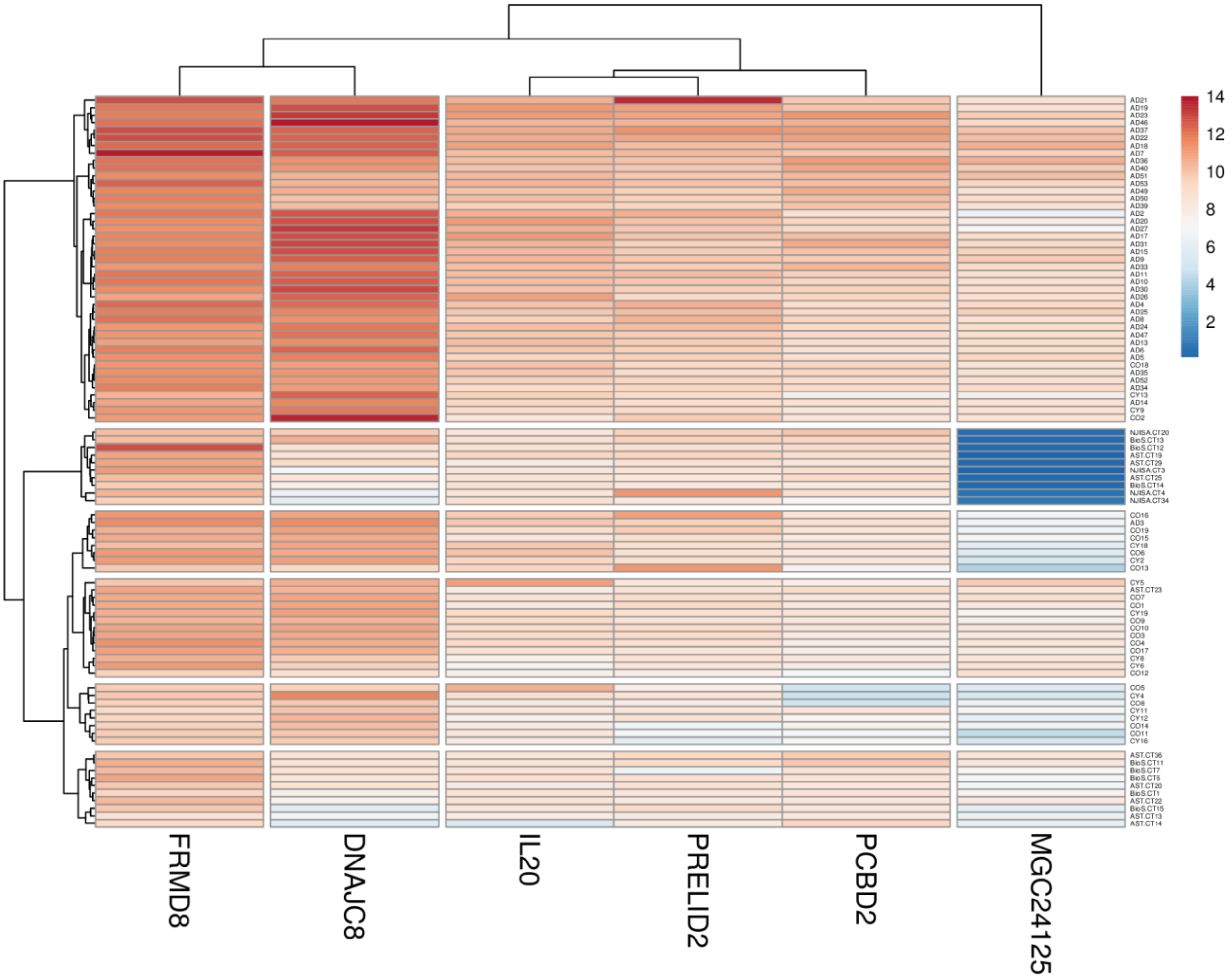


Figure 22. Heatmap of two-way hierarchical clustering applied to the RF biomarker using Euclidean distance and Ward linkage.

Four distinct clusters comprise the six autoantibodies and six distinct clusters comprise the 92 training samples. The sample clusters were mostly grouped by the class variable: the top cluster was 91% AD samples, the next cluster was 87.5% NDC samples, and the bottom four clusters were 100% NDC samples. The autoantibody-clusters along with the descriptions of the autoantibodies within them were as follows:

Table 17. The four autoantibody clusters of the RF biomarker.

Cluster	Autoantibody	Description
1	FRMD8	FERM domain containing 8
2	DNAJC8	DnaJ homolog subfamily C member 8
3	IL20	interleukin 20
	PRELID2	PRELI domain-containing protein 2
	PCBD2	pterin-4 alpha-carbinolamine dehydratase/dimerization cofactor of hepatocyte nuclear factor 1 alpha 2
4	MGC24125	hypothetical protein MGC24125

SVM:

The result of ranking the nine autoantibodies that comprise the SVM biomarker by the square of their contribution to the SVM weight vector was as follows:

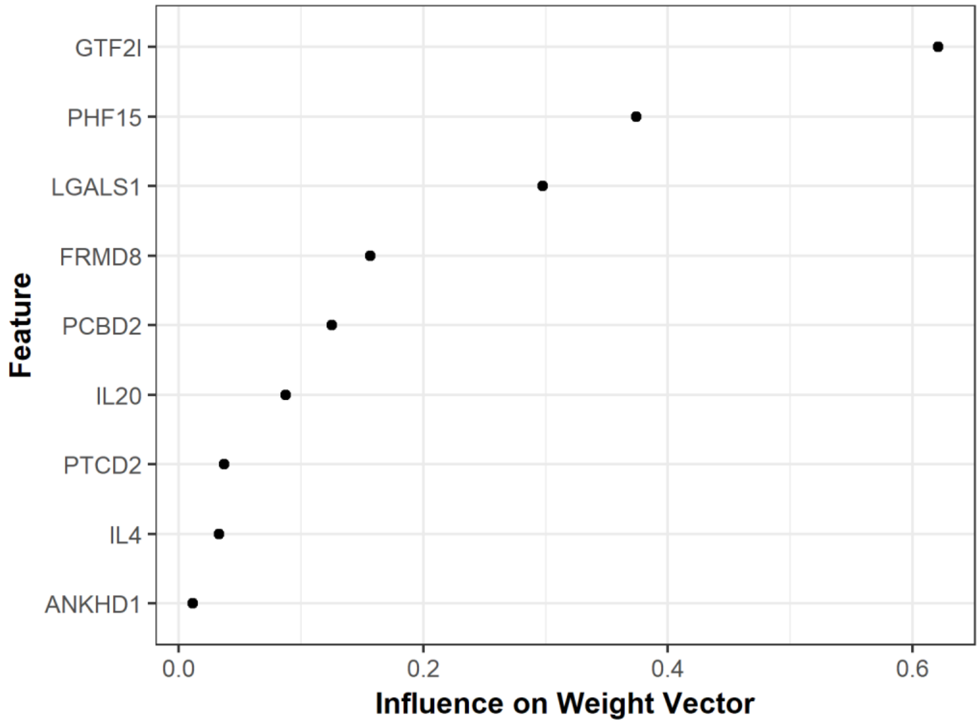


Figure 23. SVM biomarker autoantibodies ranked by the square of their weight.

By far the most important feature was GTF2I. The next most important features in descending order were PHF15, LGALS1, FRMD8, PCBD2, IL20, PTCD2, IL4, and ANKHD1. There was a greater distance between the most important and second-most important autoantibodies in the SVM biomarker than in the RF biomarker.

The heatmap resulting from two-way hierarchical clustering applied to the SVM biomarker in the 92-sample training set using Euclidean distance and Ward linkage (trees in row and column dendrograms are ordered by median expression in descending order):

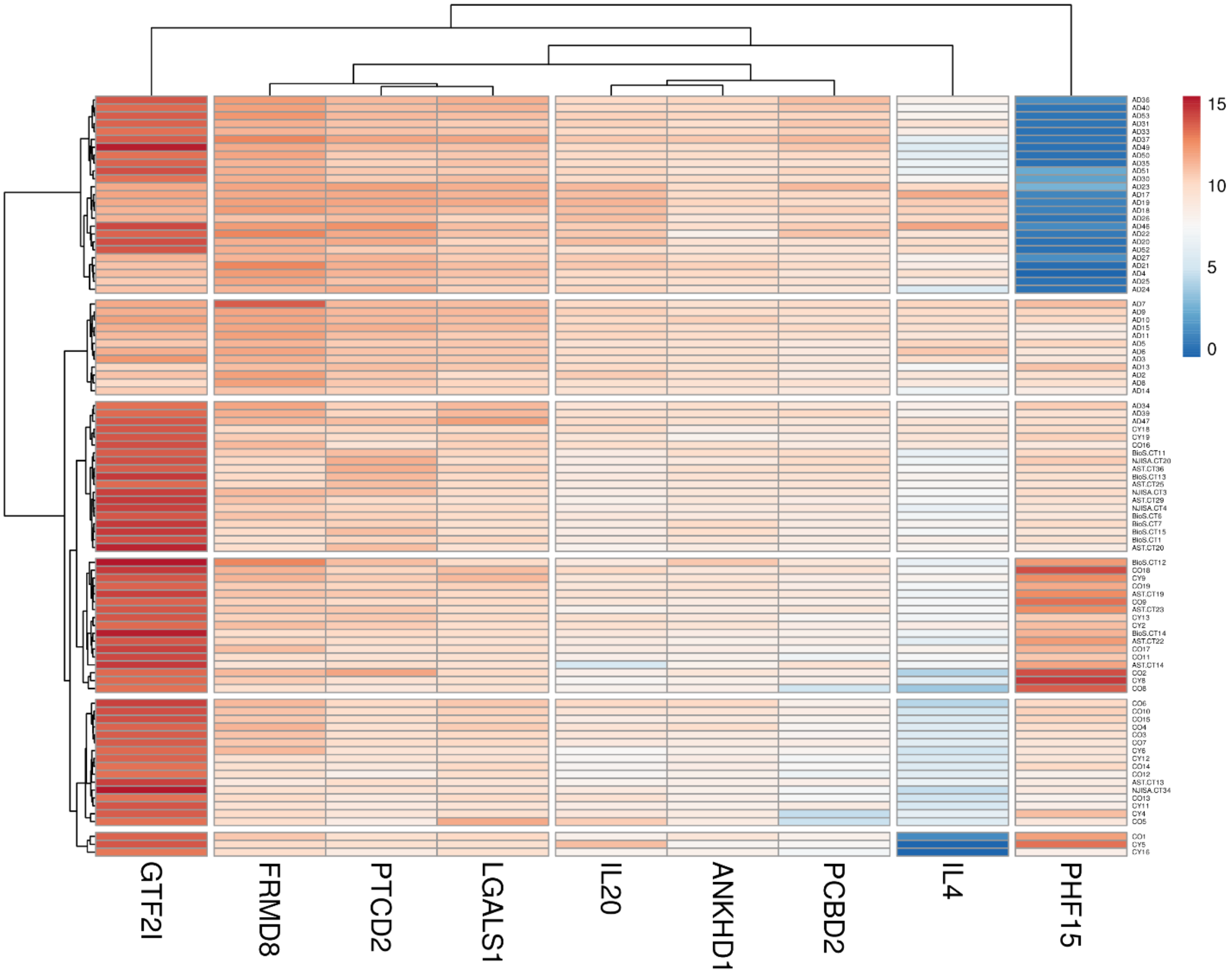


Figure 24. Heatmap of two-way hierarchical clustering applied to the SVM biomarker using Euclidean distance and Ward linkage.

The SVM biomarker had five distinct autoantibody-clusters and six distinct sample-clusters. The samples were again grouped mostly by the class variable. If we consider the clustered to be ordered 1-6 from top to bottom, then: clusters 1-3 were 100% AD samples, clusters 4 and 6 were 100% NDC samples, and cluster 5 was 90% NDC samples. The autoantibody-clusters along with the descriptions of the features within them:

Table 18. The five clusters of features in the SVM biomarker.

Cluster	Feature	Description
1	GTF2I	General transcription factor II-I
2	FRMD8	FERM domain containing 8
	LGALS1	lectin, galactoside-binding, soluble, 1
	PTCD2	pentatricopeptide repeat domain 2
3	PCBD2	pterin-4 alpha-carbinolamine dehydratase/dimerization cofactor of hepatocyte nuclear factor 1 alpha 2
	ANKHD1	ankyrin repeat and KH domain containing 1, transcript variant 3
	IL20	interleukin 20
4	IL4	interleukin 4, transcript variant 1
5	PHF15	PHD finger protein 15

3.6 Biomarker Validation

The 31 test set samples set aside prior to biomarker discovery were used to validate the classification performance of the final RF and SVM biomarkers. Upon classifying the test set, each biomarker produced the following classification probabilities for the test set samples, which represent the estimated probability of that sample having Alzheimer's Disease:

Table 19. Test set classification probabilities per final classifier.

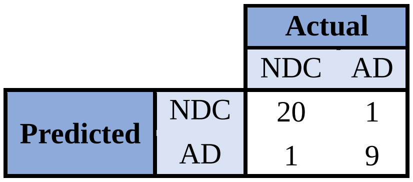
AD Samples	RF	SVM
AD1	0.9305	0.9981
AD12	0.9990	0.9955
AD16	0.9285	0.2799
AD28	0.9990	1.0000
AD29	0.7195	0.7684
AD32	0.9935	0.9590
AD38	0.9995	0.9917
AD45	0.8465	0.8675
AD48	1.0000	0.9982
AD54	0.7630	0.9927

NDC Samples	RF	SVM
AST.CT11	0.1250	0.0003
AST.CT12	0.0000	0.0048
AST.CT27	0.0000	0.0018
AST.CT31	0.0020	0.0044
AST.CT37	0.0000	0.0037
BioS.CT3	0.0020	0.0192
BioS.CT4	0.0160	0.0058
BioS.CT8	0.0000	0.0051
BioS.CT9	0.0255	0.0250
CO20	0.2195	0.0240
CY1	0.0930	0.0896
CY10	0.3625	0.0669
CY14	0.2690	0.0124
CY15	0.0160	0.0090
CY17	0.2100	0.0202
CY20	0.0495	0.1349
CY3	0.0000	0.0015
CY7	0.0000	0.0006
NJISA.CT45	0.5880	0.4297
NJISA.CT47	0.4515	0.5008
NJISA.CT53	0.4295	0.0758

The classification probabilities above were then used in conjunction with the default 0.50 decision threshold to classify the test set samples. The biomarkers' test set classification performances are summarized in the following confusion matrices:

		Actual	
		NDC	AD
Predicted	NDC	20	0
	AD	1	10

Figure 25. Test set confusion matrix – RF.



		Actual	
		NDC	AD
Predicted	NDC	20	1
	AD	1	9

Figure 26. Test set confusion matrix – SVM.

The RF biomarker only misclassified 1/31 test set samples, which was a false positive. The SVM classifier misclassified 2/31 test set samples with 1 false positive and 1 false negative. The test set classification performance of the two biomarkers is summarized in *Table 20*:

Table 20*. Test set performance summary of the two final classifiers.*

Biomarker	Autoantibodies	AUC	Accuracy	Sensitivity	Specificity
SVM	9	0.9905	0.9355	0.9	0.9524
RF	6	1	0.9677	1	0.9524

The test set performance measures of the RF biomarker were very close to its corresponding OOB performance. The SVM biomarker's test set performance measures were less comparable to its OOB performance, with a few percent lower values of accuracy and specificity. However, this performance difference is insubstantial given the small size of the test set and the consequent high relative impact of single sample misclassification. Therefore, both biomarkers were considered successfully validated based on the test set.

The biomarkers were also validated by implementing the ROC curve analysis introduced in Section 2.7, which produced the following test set ROC curves and threshold performance plots for the two biomarkers:

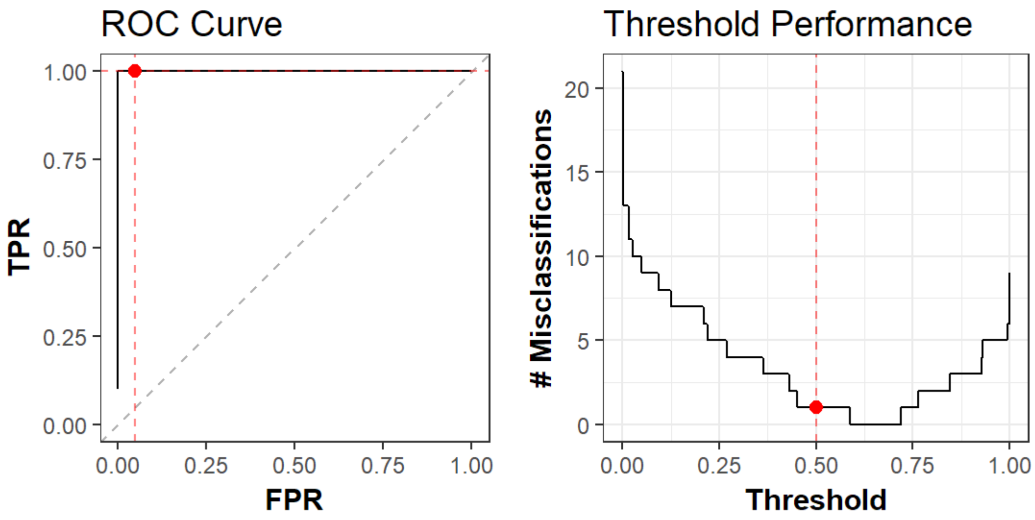


Figure 27. RF biomarker's test set ROC curve and threshold performance plot.

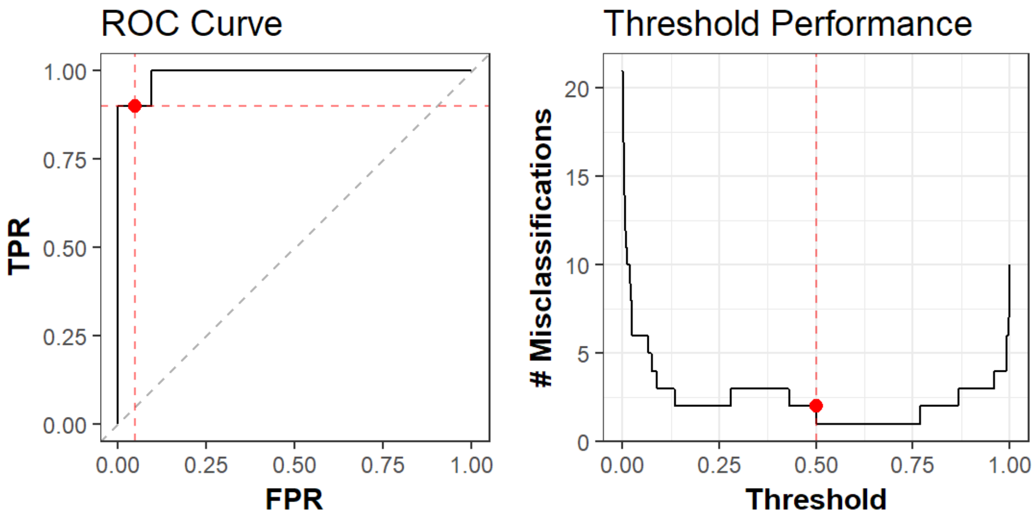


Figure 28. SVM biomarker's test set ROC curve and threshold performance plot.

Using the trapezoidal rule, the AUC for the RF and SVM biomarkers were 1 and 0.9905, respectively. The reason the AUC of the RF biomarker was a perfect 1 despite misclassifying a test set sample with the .50 decision threshold is because AUC can be interpreted as the probability that, out of a randomly chosen pair of positive and negative samples, the positive sample will be ranked higher (as more likely to be positive) than the negative sample. As seen in Table 19, the lowest probability among the AD samples was 0.7195 (AD11) and the highest probability among the NDC samples was 0.5880 (NJISA.CT45). This means that for any pair of positive and negative test set samples, the RF biomarker will always rank the positive sample higher, which produces an AUC of 1. The SVM biomarker falls short of a perfect AUC because it has one positive sample (AD16) with a lower probability (0.2799) than the probability of two of the NDC samples (NJISA.CT45 and NJISA.CT47).

The threshold performance plots showed that the default 0.50 decision threshold does not maximize test set accuracy for either biomarker. The maximum test set classification accuracy for the RF biomarker of 1.0 was achieved with thresholds between 0.5886 and 0.7187 and the maximum accuracy for the SVM biomarker of 0.9677 was achieved with thresholds between 0.5015 and 0.7677 . Note that the purpose of the threshold performance analysis was solely exploratory. Given the higher proportion of NDC samples in the test set compared to AD samples, and the lack of apriori information regarding the prevalence of the classes in the targeted population, the expectation was that false positives would be more common than false negatives. Consequently, the optimal decision threshold, where test set classification accuracy is maximized, was expected to be higher than 0.50. A larger and more balanced test dataset would be required to draw meaningful conclusions regarding the likelihood of type 1 and type 2 errors and an understanding of the true proportions of disease/healthy among recipients of the potential diagnostic test would be needed to optimize the decision threshold for practical use. Thus, no recommendation for modifying the decision threshold was made.

4 Discussion

This project sought to identify an autoantibody biomarker with a greater expectation of classification efficacy on unseen samples than the 10-autoantibody biomarker identified by Nagele, E. et al. (2011), which was identified by arbitrarily selecting the top 10 ranked autoantibodies from a m-statistical analysis filter, a univariate method. Several analytical improvements to the biomarker discovery methodology used by Nagele, E. et al. (2011) were made with this goal in mind. Most notably, univariate feature selection (M-statistical analysis filter) was replaced with multivariate feature selection (RF-RFE and SVM-RFE) and bagging was integrated into RFE to lower the variance of the feature selection results. Several additional improvements were made such as expanding the sample size of the dataset from 90 to 123, pre-processing the data with background correction and RLM normalization, using bagging to tune the model parameters of the final biomarkers, making the primary evaluation measure area under the ROC curve instead of accuracy, and interpreting the biomarkers using model-based variable importance rankings and two-way hierarchical clustering.

Ultimately, the six-autoantibody RF biomarker and the nine-autoantibody SVM biomarker were both improvements upon the 10-autoantibody biomarker identified by Nagele E. et al. (2011) and are recommended for further study. The mean OOB performance of the 10-autoantibody biomarker resulting from random forest classification reported by Nagele E. et al. (2011) was the following in comparison to the OOB performance of the RF and SVM biomarkers identified in this project:

Table 21. *Mean OOB performance of the biomarker identified by E. Nagele et al. (2011) and the two identified in this project. AUC not measured by Nagele et al. (2011).*

Biomarker	Size	AUC	Sensitivity	Specificity	Accuracy
E. Nagele et al. (2011)	10	-	0.9600	0.9000	0.9333
SVM	9	1.0000	0.9892	0.9962	0.9929
RF	6	0.9950	0.9677	0.9615	0.9637

In terms of mean OOB performance, the RF biomarker improved on the 10-autoantibody biomarker with 0.80%, 6.83%, and 3.26% higher values of sensitivity, specificity, and accuracy, respectively. The SVM biomarker was an even larger improvement since its mean OOB values of sensitivity, specificity, and accuracy were 3.04%, 10.69%, and 6.39% higher, respectively.

As seen in the table below, comparing test set classification performance between the three biomarkers told a slightly different story:

Table 22. Test set performance of the biomarker identified by E. Nagele et al. (2011) and the two identified in this project. AUC not measured by Nagele et al. (2011).

Biomarker	Size	AUC	Sensitivity	Specificity	Accuracy
E. Nagele et al. (2011)	10	-	0.8800	1.0000	0.9333
SVM	9	1.0000	0.9000	0.9524	0.9355
RF	6	0.9905	1.0000	0.9524	0.9677

In comparison to the test set performance of the E. Nagele et al. (2011) biomarker, the SVM biomarker achieved 2.27% higher sensitivity and 0.24% higher accuracy, but 4.76% lower specificity. Despite being the smallest biomarker, the RF significantly improved upon the sensitivity (13.64% higher) and accuracy (3.69% higher) of the E. Nagele et al. (2011) biomarker, while having the same drop in specificity as the SVM biomarker (4.76% lower). According to the test set results, the RF and SVM biomarkers were both improvements in terms of sensitivity and accuracy, but were a setback in terms of specificity.

However, there are two reasons why the comparison of OOB performance made for a more trustworthy method of comparing biomarkers than test set performance. One, the test sets used in E. Nagele et al. (2011) and in this project contained 45 and 31 samples, respectively. These sample sizes are both too small to avoid excessive variance in any performance estimations drawn from them. Second, while the test set used by E. Nagele et al. (2011) was nearly balanced (25 AD, 20 NDC), the test set used in this project was certainly not (10 AD, 21 NDC). Larger and more balanced test sets would be needed to expand their role beyond simple validation and enable meaningful inter-biomarker performance comparisons.

5 Limitations

There were several limitations to this project. One was that, even after increasing the sample size to 123, the 9,444 variables in the dataset meant that the curse of dimensionality was a major threat of overfitting the training data. While steps were taken to mitigate the risk such as using two learning algorithms less affected by the curse of dimensionality than others, incorporating bagging into feature selection, and using an independent test set for validation, these efforts might not fully alleviate the threat.

A second limitation was that, as described by M. Turewicz et al. (2013), batch effects are a serious limitation to biomarker studies using samples from more than one production lot (batch). Batch effects are systematic sources of error that can arise due to differences between microarrays produced in separate batches such as protein spot concentrations and other spotting conditions. One way to adjust for this is to apply a batch adjustment procedure such as described by C. Chen et al. (2011). However, batch adjustment was considered unviable for the dataset used in this thesis since three of the four total batches contained samples belonging to only one of the two classes. The batch adjustment would thus be obfuscated by the effect of the class variable and would introduce more bias than it resolved.

Third, as mentioned in Nagele E. et al. (2011), it remains uncertain whether the harmful autoimmune response represents an etiological factor leading to Alzheimer's Disease or a secondary factor that exacerbates disease pathology. The former would have much greater potential clinical utility than the latter, given the urgent need for an early diagnostic of Alzheimer's Disease that can be implemented with the standard benefits of a blood test: inexpensive, noninvasive, and widely accessible due to not requiring specialized machinery. An early diagnostic biomarker of Alzheimer's Disease with these benefits would provide a major boost to Alzheimer's Disease research and greatly increase the likelihood of finally discovering an effective disease-modifying treatment. If the autoimmune response is a consequent of Alzheimer's Disease pathology, the utility of the diagnostic biomarkers would be limited to the potential monitoring of disease progression and helping to verify the class labels of samples in research studies. However, this would still be quite useful as the current CSF and imaging biomarkers used for these tasks are invasive and expensive.

Finally, even if the estimated performance of the RF and SVM biomarkers revealed in this study holds up in multiple independent studies with much larger sample sizes, it may still not meet the steep sensitivity and specificity requirements for a widely implementable diagnostic biomarker for such a serious illness. Even 95% specificity may be too low, as that would mean that 5% of the millions of people who would be taking the diagnostic each year would receive a false terminal diagnosis. However, for research purposes, the sensitivity and specificity demands are not as severe, since a main roadblock to discovering a disease modifying treatment for Alzheimer's Disease is not the lack of an effective early diagnostic, but an effective one that is also inexpensive and noninvasive.

References

Abel, L., Kutschki, S., Turewicz, M., Eisenacher, M., Stoutjesdijk, J., Meyer, H. E., … May, C. (2014). Autoimmune profiling with protein microarrays in clinical applications. *Biomarkers: A Proteomic Challenge*, *1844*(5), 977–987. https://doi.org/10.1016/j.bbapap.2014.02.023

Acharya, N. K., Nagele, E. P., Han, M., Coretti, N. J., DeMarshall, C., Kosciuk, M. C., … Nagele, R. G. (2012). Neuronal PAD4 expression and protein citrullination: possible role in production of autoantibodies associated with neurodegenerative disease. *Journal of Autoimmunity*, *38*(4), 369–380.

Alzheimer's Association. (2015). 2015 Alzheimer's disease facts and figures. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, *11*(3), 332–384. https://doi.org/10.1016/j.jalz.2015.02.003

Ambroise, C., & McLachlan, G. J. (2002). Selection bias in gene extraction on the basis of microarray gene-expression data. *Proceedings of the National Academy of Sciences*, *99*(10), 6562–6566.

American Red Cross (2017). *Blood Components.* Retrieved from http://www.redcrossblood.org/learn-about-blood/blood-components.

Avila, J. (2004). Role of Tau Protein in Both Physiological and Pathological Conditions. *Physiological Reviews*, *84*(2), 361–384. https://doi.org/10.1152/physrev.00024.2003

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., … Holko, M. (2013). NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Research*, *41*(D1), D991–D995.

Beach, T. G., Monsell, S. E., Phillips, L. E., & Kukull, W. (2012). Accuracy of the clinical diagnosis of Alzheimer disease at National Institute on Aging Alzheimer Disease Centers, 2005–2010. *Journal of Neuropathology & Experimental Neurology*, *71*(4), 266–273.

Bertram, L., McQueen, M. B., Mullin, K., Blacker, D., & Tanzi, R. E. (2007). Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database. *Nature Genetics*, *39*(1), 17–23. https://doi.org/10.1038/ng1934

Biomarkers Definitions Working Group (2001). Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clinical Pharmacology and Therapeutics*, *69*(3), 89–95. https://doi.org/10.1067/mcp.2001.113989

Björkqvist, M., Ohlsson, M., Minthon, L., & Hansson, O. (2012). Evaluation of a previously suggested plasma biomarker panel to identify Alzheimer's disease. *PloS One*, *7*(1), e29868.

Booij, B. B., Lindahl, T., Wetterberg, P., Skaane, N. V., S\a ebø, S., Feten, G., … Lönneborga, A. (2011). A gene expression pattern in blood for the early detection of Alzheimer's disease. *Journal of Alzheimer's Disease*, *23*(1), 109–119. https://doi.org/10.3233/JAD-2010-101518

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers (pp. 144–152). Presented at the Proceedings of the fifth annual workshop on Computational learning theory, ACM.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Bron, E. E., Smits, M., Niessen, W. J., & Klein, S. (2015). Feature Selection Based on the SVM Weight Vector for Classification of Dementia. *IEEE Journal of Biomedical and Health Informatics, 19*(5), 1617–1626.

Buerger, K., Ewers, M., Pirttilä, T., Zinkowski, R., Alafuzoff, I., Teipel, S. J., … Hampel, H. (2006). CSF phosphorylated tau protein correlates with neocortical neurofibrillary pathology in Alzheimer's disease. *Brain*, *129*(11), 3035–3041. https://doi.org/10.1093/brain/awl269

Centers for Disease Control and Prevention. (2017, March 17). National Center for Health Statistics. *Health, Unites States, 2016, table 20.* Retrieved from https://www.cdc.gov/nchs/fastats/leading-causes-of-death.htm

Chen, C., Grennan, K., Badner, J., Zhang, D., Gershon, E., Jin, L., & Liu, C. (2011). Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. *PloS One, 6*(2), e17238.

Chong, M. S., Lim, W. S., Chan, M., Tay, L., Chen, G., Feng, L., … Lee, T. S. (2013). Gene expression profiling of peripheral blood leukocytes shows consistent longitudinal downregulation of TOMM40 and upregulation of KIR2DL5A, PLOD1, and SLC2A8 among fast progressors in early Alzheimer's disease. *Journal of Alzheimer's Disease, 34*(2), 399–405. https://doi.org/10.3233/JAD-121621

Clark, L. F., & Kodadek, T. (2013). Advances in blood-based protein biomarkers for Alzheimer's disease. *Alzheimer's Research & Therapy*, *5*(3), 18. https://doi.org/10.1186/alzrt172

Clifford, R. J. J., Knopman, D. S., Jagust, W. J., Petersen, R. C., Weiner, M. W., Aisen, P. S., … Trojanowski, J. Q. (2013). Update on hypothetical model of Alzheimer's disease biomarkers. *Lancet Neurology*, *12*(2), 207–216. https://doi.org/10.1016/S1474-4422(12)70291-0.Update

Dziuda, D. M. (2010). *Data mining for genomics and proteomics: analysis of gene and protein expression data* (Vol. 1). John Wiley & Sons.

Efron, B. (1979). Computers and the theory of statistics: thinking the unthinkable. *SIAM Review, 21*(4), 460–480.

Esparza, T. J., Zhao, H., Cirrito, J. R., Cairns, N. J., Bateman, R. J., Holtzman, D. M., & Brody, D. L. (2013). Amyloid-beta oligomerization in Alzheimer dementia versus high-pathology controls. *Annals of Neurology*, *73*(1), 104–119.

Fagan, A. M. (2014). CSF biomarkers of Alzheimer's disease: impact on disease concept, diagnosis, and clinical trial design. *Advances in Geriatrics, 2014*.

Fagan, A. M. & Holtzman, D. M. (2010). Cerebrospinal fluid biomarkers of Alzheimer's disease. *Biomarkers in Medicine*, *4*(1), 51–63. https://doi.org/10.2217/BMM.09.83

Guerreiro, R., Wojtas, A., Bras, J., Carrasquillo, M., Rogaeva, E., Majounie, E., … Hardy, J. (2012). TREM2 Variants in Alzheimer's Disease. *New England Journal of Medicine, 368*(2), 117–127. https://doi.org/10.1056/NEJMoa1211851

Guyon, I. & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, *3*(Mar), 1157–1182.

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, *46*(1–3), 389–422.

Hampel, H., Blennow, K., Shaw, L. M., Hoessler, Y. C., Zetterberg, H., & Trojanowski, J. Q. (2010). Total and Phosphorylated tau protein as biological markers of Alzheimer's disease. *Exp Gerontol*, *45*(1), 30. https://doi.org/10.1016/j.exger.2009.10.010.Total

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*(1), 29–36.

Hardy, J., & Higgins, G. (1992). Alzheimer's disease: the amyloid cascade hypothesis. *Science*, *256*(5054), 184. https://doi.org/10.1126/science.1566067

Henriksen, K., O'Bryant, S. E., Hampel, H., Trojanowski, J. Q., Montine, T. J., Jeromin, A., … Soares, H. (2014). The future of blood-based biomarkers for Alzheimer's disease. *Alzheimer's & Dementia*, *10*(1), 115–131.

Hu, W. T., Holtzman, D. M., Fagan, A. M., Shaw, L. M., Perrin, R., Arnold, S. E., … Soares, H. (2012). Plasma multianalyte profiling in mild cognitive impairment and Alzheimer disease. *Neurology*, *79*(9), 897–905. https://doi.org/10.1212/WNL.0b013e318266fa70

Humpel, C. (2011). Identifying and validating biomarkers for Alzheimer's disease. *Trends in Biotechnology*, *29*(1), 26–32.

Johnstone, D., Milward, E. A., Berretta, R., Moscato, P., & Alzheimer's Disease Neuroimaging Initiative. (2012). Multivariate protein signatures of pre-clinical Alzheimer's disease in the Alzheimer's disease neuroimaging initiative (ADNI) plasma proteome dataset. *PLoS One*, *7*(4), e34341.

Jonsson, T., Stefansson, H., Steinberg, S., Jonsdottir, I., Jonsson, P. V., Snaedal, J., … Stefansson, K. (2012). Variant of TREM2 Associated with the Risk of Alzheimer's Disease. *New England Journal of Medicine*, *368*(2), 107–116. https://doi.org/10.1056/NEJMoa1211103

Karatzoglou, A., Smola, A., Hornik, K., & Zeileis, A. (2004). kernlab - An S4 Package for Kernel Methods in R. *Journal of Statistical Software*, *11*(9), 1–20.

Kelley, A. S., McGarry, K., Gorges, R., & Skinner, J. S. (2015). The Burden of Health Care Costs in the Last 5 Years of Life. *Annals of Internal Medicine*, *163*(10), 729–736. https://doi.org/10.7326/M15-0381

Kuhn, M. (2016). caret: Classification and Regression Training (Version 6.0-70). Retrieved from https://CRAN.R-project.org/package=caret

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer.

Kursa, M. B. (2014). Robustness of Random Forest-based gene selection methods. *BMC Bioinformatics*, *15*(1), 1.

Laske, C., Leyhe, T., Stransky, E., Hoffmann, N., Fallgatter, A. J., & Dietzsch, J. (2011). Identification of a blood-based biomarker panel for classification of Alzheimer’s disease. *The International Journal of Neuropsychopharmacology / Official Scientific Journal of the Collegium Internationale Neuropsychopharmacologicum (CINP)*, *14*(9), 1147–1155. https://doi.org/10.1017/S1461145711000459

Lebedev, A., Westman, E., Van Westen, G., Kramberger, M., Lundervold, A., Aarsland, D., … Tsolaki, M. (2014). Random Forest ensembles for detection and prediction of Alzheimer’s disease with a good between-cohort robustness. *NeuroImage: Clinical, 6,* 115–125.

Levin, E. C., Acharya, N. K., Han, M., Zavareh, S. B., Sedeyn, J. C., Venkataraman, V., & Nagele, R. G. (2010). Brain-reactive autoantibodies are nearly ubiquitous in human sera and may be linked to pathology in the context of blood–brain barrier breakdown. *Brain Research, 1345,* 221–232.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, *2*(3), 18–22.

Liaw, A., & Wiener, M. (2015). Breiman and Cutler's Random Forests for Classification and Regression (Version 4.6-12). Retrieved from https://www.stat.berkeley.edu/~breiman/RandomForests/

Liu, Q., Chen, C., Zhang, Y., & Hu, Z. (2011). Feature selection for support vector machines with RBF kernel. *Artificial Intelligence Review*, *36*(2), 99–115. https://doi.org/10.1007/s10462-011-9205-2

Lunnon, K., Sattlecker, M., Furney, S. J., Coppola, G., Simmons, A., Proitsi, P., … Hodges, A. (2013). A blood gene expression marker of early Alzheimer's disease. *Journal of Alzheimer's Disease*, *33*(3), 737–753. https://doi.org/10.3233/JAD-2012-121363

Mayo Clinic (2015, Nov. 24). *Alzheimer's stages: how the disease progresses*. Retrieved from http://www.mayoclinic.org/diseases-conditions/alzheimers-disease/in-depth/alzheimers-stages/art-20048448?pg=2

McKhann, G. M., Knopman, D. S., Chertkow, H., Hyman, B. T., Jack, C. R., Kawas, C. H., … Mayeux, R. (2011). The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia*, *7*(3), 263–269.

Metsalu, T., & Vilo, J. (2015). ClustVis: a web tool for visualizing clustering of multivariate data using Principal Component Analysis and heatmap. *Nucleic Acids Research*, *43*(W1), W566–W570.

Nagele, E., Han, M., DeMarshall, C., Belinka, B., & Nagele, R. (2011). Diagnosis of Alzheimer’s disease based on disease-specific autoantibody profiles in human sera. *PLoS One*, *6*(8), e23112.

Nagele, E., Han, M., Acharya, N., DeMarshall, C., Kosciuk, M., & Nagele, R. (2013). Natural IgG Autoantibodies Are Abundant and Ubiquitous in Human Sera, and Their Number Is Influenced By Age, Gender, and Disease. *PLoS ONE, 8*(4), e60726. https://doi.org/10.1371/journal.pone.0060726

Nagele, R., Clifford, P., Siu, G., Levin, E., Acharya, N., Han, M., … Zarrabi, S. (2011). Brain-reactive autoantibodies prevalent in human sera increase intraneuronal amyloid-β1-42 deposition. *Journal of Alzheimer’s Disease*, *25*(4), 605–622.

O’Bryant, S. (2010). A Serum Protein–Based Algorithm for the Detection of Alzheimer Disease. *Archives of Neurology*, *67*(9), 1077. https://doi.org/10.1001/archneurol.2010.215

Olsson, B., Lautner, R., Andreasson, U., Öhrfelt, A., Portelius, E., Bjerke, M., … Zetterberg, H. (2016). CSF and blood biomarkers for the diagnosis of Alzheimer’s disease: a systematic review and meta-analysis. *The Lancet Neurology*, *15*(7), 673–684. https://doi.org/10.1016/S1474-4422(16)00070-3

Petrella, J. R. (2013). Neuroimaging and the search for a cure for Alzheimer disease. *Radiology, 269*(3), 671–691.

Platt, J. C. (1999). 12 fast training of support vector machines using sequential minimal optimization. *Advances in Kernel Methods*, 185–208.

Provost, F. J., Fawcett, T., & Kohavi, R. (1998). The case against accuracy estimation for comparing induction algorithms. (Vol. 98, pp. 445–453). Presented at the ICML.

R Core Team. (2016). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rakotomamonjy, A. (2003). Variable selection using SVM-based criteria. *Journal of Machine Learning Research*, *3*(Mar), 1357–1370.

Ray, S., Britschgi, M., Herbert, C., Takeda-Uchimura, Y., Boxer, A., Blennow, K., … Wyss-Coray, T. (2007). Classification and prediction of clinical Alzheimer’s diagnosis based on plasma signaling proteins. *Nature Medicine*, *13*(11), 1359–1362. https://doi.org/10.1038/nm1653

Rinne, J. O., Brooks, D. J., Rossor, M. N., Fox, N. C., Bullock, R., Klunk, W. E., … Okello, A. A. (2010). 11 C-PiB PET assessment of change in fibrillar amyloid-β load in patients with Alzheimer’s disease treated with bapineuzumab: a phase 2, double-blind, placebo-controlled, ascending-dose study. *The Lancet Neurology*, *9*(4), 363–372.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, *43*(7). https://doi.org/10.1093/nar/gkv007

Ritter, A., & Cummings, J. (2015). Fluid Biomarkers in Clinical Trials of Alzheimer’s Disease Therapeutics. *Frontiers in Neurology*, *6*, 186. https://doi.org/10.3389/fneur.2015.00186

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, *12*, 77. https://doi.org/10.1186/1471-2105-12-77

Rye, P. D., Booij, B. B., Grave, G., Lindahl, T., Kristiansen, L., Andersen, H. M., … Lönneborg, A. (2011). A novel blood test for the early detection of Alzheimer's disease. *Journal of Alzheimer's Disease*, *23*(1), 121–129. https://doi.org/10.3233/JAD-2010-101521

Sboner, A., Karpikov, A., Chen, G., Smith, M., Dawn, M., Freeman-Cook, L., … Gerstein, M. B. (2009). Robust-Linear-Model Normalization To Reduce Technical Variability in Functional Protein Microarrays. *Journal of Proteome Research*, *8*(12), 5451–5464. https://doi.org/10.1021/pr900412k

Snyder, H. M., Carrillo, M. C., Grodstein, F., Henriksen, K., Jeromin, A., Lovestone, S., … Sjøgren, M. (2014). Developing novel blood-based biomarkers for Alzheimer's disease. *Alzheimer's & Dementia*, *10*(1), 109–114.

Sperling, R. A., Aisen, P. S., Beckett, L. A., Bennett, D. A., Craft, S., Fagan, A. M., … Phelps, C. H. (2011). Toward defining the preclinical stages of Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia*, *7*(3), 280–292. https://doi.org/10.1016/j.jalz.2011.03.003

Tejada-Vera, B. (2013). Mortality From Alzheimer's Disease in the United States: Data for 2000 and 2010. NCHS data brief, no. 116. *National Center for Health Statistics: Hyattsville, MD*.

Thambisetty, M., & Lovestone, S. (2010). Blood-based biomarkers of Alzheimer's disease: challenging but feasible. *Biomarkers in Medicine*, *4*(1), 65–79.

UsAgainstAlzheimer's (2012, October 4). *The Crisis*. Retrieved October 5, 2016, from http://www.usagainstalzheimers.org/crisis

Turewicz, M. (2016). ProteinArrayAnalyzer (PAA): A Novel R/Bioconductor Package for Autoimmune Biomarker Discovery with Protein Microarrays (Version 1.4.1). Retrieved from http://www.medizinisches-proteom-center.de/PAA

Turewicz, M., May, C., Ahrens, M., Woitalla, D., Gold, R., Casjens, S., … Eisenacher, M. (2013). Improving the default data analysis workflow for large autoimmune biomarker discovery studies with ProtoArrays. *PROTEOMICS*, *13*(14), 2083–2087. https://doi.org/10.1002/pmic.201200518

Vanderstichele, H., & Kodadek, T. (2014). Roadblocks for integration of novel biomarker concepts into clinical routine: the peptoid approach. *Alzheimer's Research & Therapy*, *6*(2), 1.

Vapnik, V. (1963). Pattern recognition using generalized portrait method. *Automation and Remote Control*, *24*, 774–780.

Venables, W. N., Ripley, B. D., & Venables, W. N. (2002). *Modern Applied Statistics with S* (Fourth). New York: Springer. Retrieved from http://www.stats.ox.ac.uk/pub/MASS4

Wickham, H. (2007). Reshaping Data with the reshape Package. *Journal of Statistical Software*, *21*(12), 1–20.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. Retrieved from http://ggplot2.org

Wickham, H., & Francois, R. (2016). dplyr: A Grammar of Data Manipulation (Version 0.5.0). Retrieved from https://CRAN.R-project.org/package=dplyr

World Health Organization (1993). *Biomarkers and risk assessment: Concepts and principles.* Retrieved from http://www.inchem.org/documents/ehc/ehc/ehc155.htm

Yang, H., Lyutvinskiy, Y., Herukka, S.-K., Soininen, H., Rutishauser, D., & Zubarev, R. a. (2014). Prognostic Polypeptide Blood Plasma Biomarkers of Alzheimer's Disease Progression. *Journal of Alzheimer's Disease : JAD*, *40*, 1–8. https://doi.org/10.3233/JAD-132102

Appendix

Code

All Rmarkdown code and required files used in this thesis are publicly available on the author's github at: https://github.com/viscioalj/Thesis.

Software (all open-source)

The analysis for this project was implemented in The R Project for Statistical Computing (up to version 3.3.3): https://www.r-project.org/.

- Bioconductor for R: bioconductor.org/
- R packages: caret, dplyr, reshape2, ggplot2, tidyr, PAA, limma, MASS, doParallel, randomForest, kernlab, knitr, e1071, Hmisc, gridExtra, DT, miscset, and caTools.
- R code was run using Rmarkdown documents via the RStudio IDE: https://www.rstudio.com/

Heatmaps displaying the results of two-way hierarchical clustering were created using the ClustVis web tool: http://biit.cs.ut.ee/clustvis/

Data

The original data for the 123 samples used in this thesis were contained in .tar files downloaded from two datasets available publically at the Gene Expression Omnibus website:

- GSE29676: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE29676
- GSE39087: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE39087

Biographical Statement

For the past two years, A. James Viscio has worked directly with the Founder of Infrasonic Monitoring Inc as a Data Science Intern developing predictive models in R for a wearable device for athletes that noninvasively monitors cardiac output. Prior to completing his M.S. in Data Mining at Central Connecticut State University, he earned a B.A. in Mathematics from Franklin and Marshall College with a minor in Computer Science. He enjoys adventure hiking, having climbed Mt. Kilimanjaro and the Inca Trail, and lives with his girlfriend of five years in East Haddam, CT.