Cardiff University

Doctoral Thesis

Machine Learning for the Genetic Prediction of Alzheimer's Disease

---

Author:
Thomas Rowe

Supervisors:
Professor Valentina Escott-Price
Professor Peter Holmans
Dr Dobril Ivanov

A thesis submitted in fulfilment of the requirements.
for the degree of Doctor of Philosophy

November 2023

Declaration of Authorship

Statements

This thesis is being submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy.

This work has not been submitted in substance for any other degree or award at this or any other university or place of learning, nor is it being submitted concurrently for any other degree or award (outside of any formal collaboration agreement between the University and a partner organisation).

I hereby give consent for my thesis, if accepted, to be available in the University's Open Access repository (or, where approved, to be available in the University's library and for inter-library loan), and for the title and summary to be made available to outside organisations, subject to the expiry of a University-approved bar on access if applicable.

Declaration

This thesis is the result of my own independent work, except where otherwise stated, and the views expressed are my own. Other sources are acknowledged by explicit references. The thesis has not been edited by a third party beyond what is permitted by Cardiff University's Use of Third Party Editors by Research Degree Students Procedure.

Signed: T W ROWE

Date: 12.11.2023

Abstract

Alzheimer's disease (AD) is the most common form of dementia in humans, with disease course involving initial memory loss, a subsequent debilitative state and eventually death. It is a polygenic disorder, meaning its genetic component comprises many known and unknown mutations. This complexity alongside further influences from a range of lifestyle factors, have made the prediction of disease risk a challenging pursuit.

The initial attempts to predict AD risk from genetic data arose due to the identification of risk loci in genome wide association studies (GWAS). Resulting variants are used to assess risk of disease onset through polygenic risk scoring (PRS). This score is generated through the summation of risk alleles multiplied by their respective effect sizes derived from GWAS. Publication results demonstrate PRS to be a useful method for assessing lifetime risk, however it has also been proven that PRS can only cover a fraction of genetic liability for AD. A possible explanation for this inadequacy is the inability for PRS to assess non-linear relationships between loci due to the use of linear modelling. Given AD is a complex polygenic disorder, it is likely that onset is the result of interactions between loci. A format which holds the capability to analyse non-linear patterns is machine learning (ML). Interest in these algorithms has increased in recent decades due to their predictive power, ability to analyse large datasets, and capabilities in disease prediction.

Initial results demonstrated a superior performance for PRS compared to ML when using datasets comprising smalls amount of AD associated single nucleotide polymorphisms (SNPs). However, in some instances ML achieved accuracies close to that of PRS. This occurred when using the algorithm support vector machine with various kernels. However, it was acknowledged these algorithms would result in excessive training times when using larger datasets in subsequent chapters. Therefore, only decision tree-based algorithms were employed moving forwards. It was also deduced that techniques such as balancing by age and sex had made no discernible difference on model performance.

Further investigation involved the use of variants sourced on a genome wide scale, as it was reasoned that using a greater number of SNPs might improve upon results from the previous

chapter. However, increasing the number of variants resulted in issues relating to high dimensionality. Despite efforts to alleviate this through the use of feature selection techniques, prediction performance for ML models was still inferior to PRS. Further avenues were also explored such as using a more lenient threshold of r^2 when clumping and removing this step completely for SNP selection, but this again failed to improve upon ML prediction accuracy. PRS continued to achieve better performance when using an imputed version of the dataset used in previous analyses, this was still evident when again exploring method such as feature selection. However, the observed difference between ML and PRS was reduced in the final investigations conducted in this thesis. Analysis on datasets comprising SNPs derived from biologically associated AD pathways resulted in improved ML performance. This result identified the possibility of focusing on the underpinning biological mechanisms of AD when selecting datasets.

Acknowledgements

I would like to acknowledge by main supervisor Valentina for her excellent support and guidance over the previous 3-4 years. I am also very grateful for the help provided by my other supervisors Peter and Dobril, who spent many hours reading chapters etc. I must also praise the assistance I received from both Matthew Bracher-Smith and Ganna Leonenko, whose expertise was invaluable throughout. I would also like to thank the Dementia Research Institute for their support and funding over the past 3-4 years, the continuous work they conduct in dementia research has immeasurable impacts on people's lives. I also cannot overstate the support that both my parents and brother have given me throughout my academic and wider journey. Finally, I would like to dedicate this thesis to my four grandparents. One who departed many years ago, another who passed on midway through my research and two that still give me strong support in my endeavours.

Contents

Table of Contents

List of Figures

probability of being a case. With the y-axis denoting observed class frequencies. Classifier probabilities are recalibrated using either the sigmoidal or isotonic approach (Page No 241).

List of Tables

List of Abbreviations

Abbreviation	Full Title
Aβ	Amyloid Beta
ACC	Accuracy
AD	Alzheimer’s Disease
ADNI	Alzheimer’s Disease Neuroimaging Initiative
AI	Artificial Intelligence
APOE	Apolipoprotein E Protein
APP	Amyloid-B Precursor Protein
AUC	Area Under the Curve
BDR	Brains for Dementia Research
CART	Classification and Regression Trees
CHARMS	Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies
CIs	Cholinesterase Inhibitors
CLU	Clusterin
CNS	Central Nervous System
CSF	Cerebrospinal Fluid
CV	Cross-Validation
DPUK	Dementias Platform UK
EADB	European Alzheimer & Dementia Biobank
EADI	European Alzheimer's Disease Initiative
EOAD	Early Onset Alzheimer’s Disease
EPV	Events Per Variable
FN	False Negatives
FP	False Positives
FPR	False Positive Rate
GERAD	Genetic and Environmental Risk in Alzheimer’s Disease
GO	Gene Ontology
GP	Gaussian Process
GWAS	Genome-wide Association Studies
GWS	Genome Wide Significant
HMM	Hidden Markov Model
HPC	High Performance Computing
HWE	Hardy-Weinberg Equilibrium
IBD	Identical-by-descent
KEGG	Kyoto Encyclopaedia of Gene
LASSO	Least Absolute Shrinkage and Selection Operator
LD	Linkage Disequilibrium
LOAD	Late Onset Alzheimer’s Disease
LR	Logistic Regression
MAF	Minor Allele Frequency
MAR	Missing at Random
MCAR	Missing Completely at Random
MCI	Mild Cognitive Impairment
ML	Machine Learning

MMSE	Mini-Mental State Examination
MR	Magnetic Resonance
mRMR	Minimum Redundancy Relevance
NB	Naïve Bayes
NFTs	Neurofibrillary Tangles
NIA-LOAD	National Institute on Aging-Late-Onset Alzheimer's Disease Family Study
NNs	Neural Networks
OLS	Ordinary Least Squares
PCA	Principal Component Analysis
PCs	Principal Components
PET	Positron Emitting Topographers
PICALM	Phosphatidylinositol Binding Clathrin Assembly Protein
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PROBAST	Prediction Model Risk of Bias Assessment Tool
PRS	Polygenic Risk Score
PSEN1	Presenilin 1
PSEN2	Presenilin 2
QC	Quality Control
RBF	Radial Basis Function
RF	Random Forest
ROB	Risk of Bias
ROC	Receiver Operating Characteristics
SMOTE	Synthetic Minority Sampling Technique
SNPs	Single Nucleotide Polymorphisms
SVM	Support Vector Machine
TN	True Negatives
TOPMED	Trans-Omics for Precision Medicine
TP	True Positives
TPR	True Positive Rate
WGCNA	Weighted Correlation Network Analysis

1 Introduction to Alzheimer’s disease

1.1 Introduction

Mankind has been aware of dementia for millennia, with knowledge of cognitive decline with age recorded in ancient Egypt (Yang et al., 2016). A common cause of cognitive decline is Alzheimer's Disease (AD), a form of dementia found primarily in individuals aged 65 and over (Yang et al., 2016b). Recognition of AD came in the early 1900’s due to the work of the German scientist Alois Alzheimer, in which he treated a patient at Frankfurt psychiatric hospital named Auguste Deter. Aged 51, Deter displayed different symptoms to other patients such as amnesia and disorientation. Following her death, Alzheimer conducted a biopsy of Deter’s brain in which he discovered substantial thinning in the hippocampus, a region of the brain associated with controlling memory, language and thinking. Further investigation revealed the presence of senile plaques in neurons and neurofibrillary tangles within nerve fibres. It had long been considered that plaques were present in only those over 70 years old, whilst neurofibrillary tangles were a novel discovery (Yang et al., 2016b).

As understanding of AD developed, it became clear that two forms of the disease existed in humans. These are classified by age of onset, with those before the age of 65 diagnosed with early onset Alzheimer’s disease (EOAD), whilst persons above the age of 65 are classified as having late onset Alzheimer’s disease (LOAD). EOAD can be broken down into two further categories, the autosomal dominant type, also known as ‘familial’ EOAD, and the sporadic disease. Familial EOAD is caused by single genetic mutations and comprises 10-15% of all EOAD cases. However, the genetic component of the sporadic form of EOAD is polygenic in nature, due to the involvement of multiple loci (Awada, 2015). This is also the case for LOAD. However, a large proportion of the heritability for both forms of AD remains uncovered (Lynn M Bekris et al., 2010). Despite one hundred years of subsequent research and growing prevalence, an effective treatment to reverse the progression of AD has yet to be found (Yiannopoulou and Papageorgiou, 2020).

1.2 Overview

Alzheimer’s Disease is the most common form of dementia, with estimates suggesting that it accounts for over 50% of all cases (Lynn M Bekris *et al.*, 2010). As of 2019, it was estimated that 27 million people were affected by AD globally (Silva *et al.*, 2019). Additionally, it is estimated that five million new cases of AD are diagnosed each year. AD is also ranked among the top ten leading causes of death in the United States, with 121,499 deaths in 2019 (‘Alzheimer’s disease facts and figures’, 2021). Alongside the traumatic effect of AD on a patient, the disease also inflicts economic burdens at both personal and national levels. As of 2005, AD was surpassed by only cancer and heart disease when considering the costliest disorders in the United States with the global estimate of total financial cost at around 315 million dollars (Castro *et al.*, 2010). The high costs associated with AD originate from the need for progressive levels of care for patients as the disease develops. As an individual enters the latter stages of the disease, they will lose all cognitive and physical abilities, resulting in the need for 24-hour intensive care. These can have significant financial implications for the individual, their next of kin and health systems (Castro *et al.*, 2010). Concerns are growing regarding the capability of health services to cope with these costs, given that prevalence within the global over 65's population is estimated to rise from 6.8% to 16.2% percent by 2040 (Castro *et al.*, 2010).

1.3 Epidemiology

EOAD is diagnosed in individuals who are below the age of 65, with development possible as early as the 5th decade of life (Mendez, 2017). When considering the overall incidence of AD, EOAD comprises around 5% of cases. Despite this, the early onset form of AD is still the most common form of dementia in the 45-64 age category, with an incidence rate of about 6.3/100,000 in the United States. There is often a delay in diagnosis of EOAD compared to LOAD (Mendez, 2017). This can be attributed to clinicians assigning symptoms to other conditions, whilst also possibly overlooking their severity. Such oversight originates from the belief that AD is an ‘elderly persons’ disease only (Mendez, 2019).

LOAD is diagnosed in those individuals who are aged 65 and above. The incidence of LOAD increases with age, with rates increasing from 2/1000 from ages 65 to 74, to 30/1000 at age 85 (Qiu, Kivipelto and von Strauss, 2009). Women are more likely to develop LOAD than males, with two-thirds of patients being female. This is partially explained by women having longer life expectancies than men, however analyses corrected for age have identified further factors. These include greater risks compared to men for comorbidities such as depression, myocardial infarction and coronary heart disease (Nebel *et al.*, 2018). In terms of incidence amongst different ethnicities, black populations are more at risk of developing LOAD than Caucasians. This is also the case for Hispanic individuals; however, it has been shown that Asian populations are at a similar risk to Caucasians (Barnes, 2022). Survival periods post diagnosis vary across individuals, with average life expectancy between 4-5 years (Rait *et al.*, 2010). This period can be affected by several factors including age of onset, sex, and medical comorbidities (Rountree *et al.*, 2012).

1.4 Neuropathology of AD

Despite extensive research into methods to diagnose AD, post-mortems remain the only definitive test for diagnosis (Weller and Budson, 2018). However, upon first inspection with the human eye, the brain of an AD patient may not present differently to that of a non-diseased individual. Visible diagnostic factors such as brain lesions or other alterations are not typical (Perl, 2010). Also, it has been shown that clinical features such as reduced brain weight and cerebral cortical thickness cannot be used as markers for AD. This is due to age matched cognitively healthy brains displaying similar reductions. The only change in brain anatomy which might be an indicator of AD is significant atrophy of the hippocampus, with associated dilation of the adjacent temporal horn of the lateral ventricle. However, this alone cannot be used as definitive evidence that an individual had AD (Perl, 2010). Due to this, the only method to reach a conclusive diagnosis is through examination of tissue using microscopes. Two pathological features must be present to confirm AD, these are known as senile plaques and neurofibrillary tangles (NFTs) (Perl, 2010).

1.4.1 Neurofibrillary Tangles

The protein tau has the primary role of maintaining the stability of microtubules in axons. However, alterations in the function of tau leads to the development of neurofibrillary tangles. As this protein undergoes abnormal changes, its ability to support microtubules reduces, leading to a reduction in nutrient transport between neurons. The development of NFTs typically follows three levels of maturity. Initially, abnormal tau deposits accumulate outside of the nucleus of the cell. Eventually, this accumulation of tau overwhelms the entire neuron, as fibres are formed. This leads to the development of the mature tangle, in which the neuron becomes shrunken. The final stage of the process is the death of the neuron, with the NFT remaining in a tomb like presence (Moloney, Lowe and Murray, 2021).

The presence of NFTs in a diseased brain typically presents in the entorhinal cortex, hippocampus, amygdala, and the neocortex. The degree to which NFTs have spread in these regions has been positively associated with the severity of disease and duration. This suggests that NFTs have a direct impact on the impact of AD. However, the presence of NFTs doesn't exclusively suggest that AD is present, as they are also linked to other neurological disorders (Perl, 2010).

1.4.2 Senile Plaques

The second neuropathological sign of AD is the presence of senile plaques in the hippocampus and surrounding cortical regions. Typically, these are spherical deposits comprising of the 38-42 amino acid long peptide called amyloid beta (Aβ). This peptide is derived from the protein amyloid-B precursor protein (*APP*). The *APP* protein is cleaved to form various species of Aβ. The most common of these is the form Aβ 40, mostly produced by astrocytes and neurons (Brothers, Gosztyla and Robinson, 2018). Aβ has been shown to have several beneficial roles within the human brain. These include the suppression of tumour growth, promoting recovery from brain injuries and regulating synaptic function. Levels of Aβ in the brain are regulated by cerebrospinal fluid (CSF) and microglia (Brothers, Gosztyla and Robinson, 2018).

Mutations in the *APP* protein have been linked to the development of AD. In a healthy functioning brain, the peptide Aβ42 is less prevalent than Aβ 40. However, mutations of *APP* have been linked to increased levels of Aβ42. This peptide of Aβ has a regular function and has also been associated with the early development of senile plaques. This association has been reinforced in those individuals with Down's syndrome, in which an additional copy of chromosome 21 contributes to increased expression of *APP* and in turn greater levels of Aβ. Patients with Down's syndrome develop senile plaques and eventually succumb to AD, suggesting a relationship between *APP* mutations and plaque development (Findeis, 2007).

The role of Aβ in AD development was termed the amyloid cascade hypothesis. This postulated the deposition of Aβ into senile plaques was the initial trigger of AD progression. This trigger leads to further degeneration such as neuritic injury, NFTs and cell death. Experimental approaches have produced results which reinforced this hypothesis. Approaches aimed at reducing the levels of Aβ within mice achieved reductions in synaptic loss and reprisal of some memory functions (Ricciarelli and Fedele, 2017). However, in recent decades questions have been asked regarding the suitability of the cascade hypothesis, due to the failure of Aβ targeted drug therapies. Furthermore, some investigations suggest that Aβ deposition may not have a definitive correlation with neuronal loss. Individuals have been clinically assessed as having significant levels of simile plaques and NFTs, whilst presenting as cognitively healthy. Given this, it is clear that both tau and Aβ have significant roles in AD development, however the aetiology of AD is more complicated than initially thought (Ricciarelli and Fedele, 2017).

Figure 1.1: Microscopic image of both senile plaques and NFTs relating to AD (Perl, 2010).

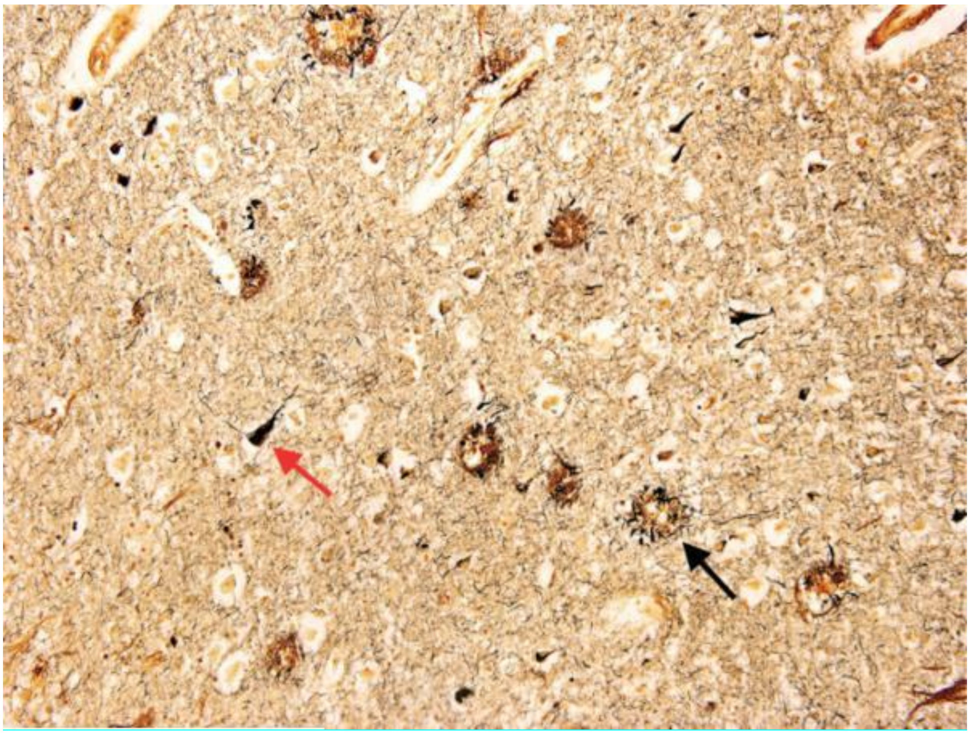


Figure 1.1: A Photomicrograph image of the temporal cortex of a patient with Alzheimer's disease. A senile plaque is identified with a black arrow, whilst the red arrow points towards an NFT. This image has not been altered from the original source.

1.5 Symptoms

The symptoms of AD can be categorised into three main areas, these are cognitive impairments, physiological illness, and physical disabilities. The combination of these three areas makes AD a severe disorder.

1.5.1 Cognitive symptoms

1.5.1.1 Early onset AD

Several studies have suggested that the cognitive symptoms of EOAD are partially different to those of LOAD, e.g., (Koedam *et al.*, 2010). Patients with EOAD tend to have better memory recognition than those at a similar stage of LOAD. However, they exhibit worse attention skills, whilst also having more deteriorated visuospatial skills. EOAD is also linked with a greater decline in executive functions, such as planning and organisation (Toyota *et al.*, 2007). Several studies have also indicated that the disease progression of EOAD is more aggressive and rapid than LOAD, e.g., (Toyota *et al.*, 2007).

1.5.1.2 Late onset AD

The most well-known symptom of LOAD is increasing memory loss, observed in all cases of the disease (Ricciarelli and Fedele, 2017). The deterioration in memory is often a gradual process, with onset being decades before clinical diagnosis. Difficulties exist in distinguishing this symptom from the natural process of ageing, in which cognitive abilities also decline. However, symptoms in AD patients develop well beyond those of normal ageing (Weller and Budson, 2018). The initial loss of memory function may manifest itself in signs such as difficulty with word finding, naming of individuals or locations and locating objects. Suspicion of AD may arise when symptoms begin to impact an individual's social and work activities. These may include failure to remember important information such as home addresses and existence of close relationships.

Progression of the disease will lead to an inability to conduct everyday tasks. Examples of these include driving, cooking and employment. As the disorder develops into the final stages, the decline in cognitive ability becomes absolute. Individuals will no longer recognise even those with closest relationships (Weller and Budson, 2018).

1.5.1.3 Psychological symptoms

Alongside cognitive issues, AD has been linked with a range of mental health issues. Depression is a common comorbidity of AD, with estimates of incidence ranging from 25% to 75% (Li *et al.*, 2014). Despite its high prevalence, it is still unclear whether depression leads to an increased risk of AD, or whether it is a result of developing AD. Studies which have examined the link between AD and depression have shown association between illness and the level of Aβ 42 in in the brain. This suggests depression maybe linked to AD development. However, other schools of thought have asked whether depression results from AD, or whether it is a response to the diagnosis. Patients will be initially aware of their prognosis, with the likelihood of a reduced life span. Such knowledge could potentially lead to a reduction in general mood (Li *et al.*, 2014).

Apathy is also a common psychological symptom of both EOAD and LOAD (Toyota *et al.*, 2007). This is categorised by a reduction in responses such as concern and interest. For example, this might be observed through a lack of interest in loved ones or a reduced motivation in previously favoured tasks such as hobbies. Incidence of apathy in AD patients is estimated to be around 40% of all cases (Li *et al.*, 2014). Similarly, to other psychological symptoms, an increased level of apathy has been linked with faster disease progression (Li *et al.*, 2014).

Two further symptoms of AD which can cause issues for both patients and caregivers are aggressiveness and psychosis. Aggressive behaviour can manifest in both physical actions and verbal attack. These issues have been more prevalent in male patients, whilst their presence has been associated with more rapid cognitive decline (Li *et al.*, 2014). Psychosis is classed as a severe mental health disorder, with patients experiencing a significant alteration in the way they perceive their surroundings. Examples of symptoms include both hallucinations and delusions, such as fear of being in danger. Such symptoms tend to be linked to the later stages of AD progression. The presence of psychosis has also been more associated with the LOAD form of the disease (Toyota *et al.*, 2007). Clearly, psychosis can be associated with increased levels of stress for both care givers and the patient. At a certain level, the presence of psychosis may result in the involvement of mental health services and police intervention (Frederick, O'Connor and Koziarski, 2018).

1.5.1.4 Physical symptoms

Greater focus has often been placed on the cognitive symptoms of AD. However, as the function of the brain deteriorates, a patient will also experience a range of physical symptoms. These will typically become more apparent in the mid-later stages of disease. Unlike memory loss, which is a universal symptom, some patients experience different physical symptoms to others. Some of the more common issues relate to the weakening of important muscle groups. This can cause symptoms such as difficulty walking, loss of coordination and general fatigue.

Despite leading to a reduced lifespan in most cases, AD is usually not the direct cause of death in individuals, rather patients succumb to complications caused by the disease (Manabe *et al.*, 2019). One of the leading reasons for death in AD patients is the reduced ability to swallow. This will impact the quality of diet for an individual, which can lead to issues such as malnourishment and dehydration. Alongside this, a reduction in swallowing ability leads to an increased risk of choking ('2020 Alzheimer's disease facts and figures', 2020). With a significant chance of food becoming trapped in the trachea. The presence of trapped food can also increase the risk of infection, which the patient's immune system may struggle to treat. Other types of infection are also common in AD patients, such as pneumonia and sepsis ('2020 Alzheimer's disease facts and figures', 2020).

1.6 Diagnosis

The only definitive method to diagnose AD remains a postmortem biopsy (DeTure and Dickson, 2019). However, the requirement to attempt to diagnose AD during a patient's lifetime remains important. There are two main reasons for this. Treatments for the different forms of dementia are diverse, with varying forms of drug treatments. The use of incorrect drug treatment has the potential to cause harm to an individual. Therefore, it is important the correct form of dementia is diagnosed, and the corresponding medication is prescribed (Iddi *et al.*, 2019). Also, research has shown that medication for AD works most effectively when administered in the early stages of disease development. Therefore, it is important to diagnose AD early when attempting to sustain an individual's quality of life (Rasmussen and Langerman, 2019).

1.6.1 Issues with diagnosing EOAD

EOAD manifests in patients typically between 40-60 years of age. This presents problems from both the patient and clinician point of view when diagnosing the disease. An individual in their 40's may be reluctant to seek medical advice following atypical symptoms, due to a perceived belief of good health (Mendez, 2017). Similarly, a doctor may not consider the possibility of AD due to the individual's age. A clinician may also have limited experience of EOAD, as well as low confidence in diagnosing it. These issues arise from the perception that AD is disease prevalent only in those aged over 65. This hesitancy can lead to a delay in diagnosis for patients, with a large percentage waiting a year of more (Mendez, 2019).

1.6.2 Issues with diagnosing LOAD

The difficulty of diagnosing the correct form of dementia lies in the similarity that LOAD shares with other forms of illness, such as frontotemporal dementia, vascular dementia and dementia with Lewy bodies (Gaugler *et al.*, 2013). Studies have shown that clinical diagnosis differs from examination for a certain percentage of individuals, with estimates ranging from 20-30% (Klatka *et al.*, 1996).

1.6.3 Cognitive tests

Advances in technology have led to the use of biomarkers in partnership with cognitive tests for AD diagnosis. However, access to biomarkers may be limited in some areas due to the costly nature of obtaining them (Palmqvist *et al.*, 2012). In this instance, the more traditional technique of cognitive tests is used. The most used psychometric test used in AD diagnosis is the mini-mental state examination (MMSE) (Arevalo-Rodriguez *et al.*, 2015). This assesses a range of cognitive functions including memory, attention, and language. Advantages of the MMSE lie in its ease in implementation, consistent scoring system and implementation in many countries. The title of 'mini' relates to the short nature of the test, with 11 questions used. These focus on topics such as dates, location, and self-awareness. Despite its widespread usage, concerns have arisen regarding the MMSE's ability to distinguish between MCI and AD, with some published studies suggesting MMSE has a predictive ability of zero (Palmqvist *et al.*, 2012).

1.6.4 Biomarkers

Despite their widespread usage, sole employment of cognitive tests is often not considered sufficient for diagnosis. Another set of tools used for decision making are biomarkers. Biomarkers are elements of the human physiology which can be used to indicate phenotypes. The increase in the use of biomarkers in AD diagnosis has occurred due to the development of computer technology (Mayeux, 2004). Neuroimaging has had initial success in the diagnosis of tumours; however, its use is also becoming more prevalent in AD (Ferreira and Busatto, 2011). One form of this is the use of positron emitting topographers (PET) scanners, which are used to assess cerebral metabolic rates of glucose. As the human brain ages, the average metabolic rate of neurons reduces (de la Monte and Tong, 2014). However, these rates decline at a greater rate at an earlier stage in AD patients. Early studies suggest a promising ability for PET scanning to differentiate between cognitively healthy patients and AD. This has also extended to separating those with MCI and AD (Grueso and Viejo-Sobera, 2021). Published studies have also shown that PET scanning may also be effective in distinguishing between the different types of dementia and are important when considering treatment options (Marcus, Mena and Subramaniam, 2014).

In the last decade, several radioactive imaging agents have been developed. These highlight possible AD affected areas of the brain when scanned. Cerebrospinal Fluid (CSF) is a clear fluid which surrounds the brain and spinal cord. A form of this fluid named CSF Aβ42 has been identified as a biomarker for AD. Studies have shown that decreased levels of this fluid have been linked to an increased likelihood of AD (Tarawneh, 2020). Further to this, another CSF variation which has been linked to AD is CSF p-tau181. Measurements of this biomarker have been proven to be strong predictors for AD progression. However, extracting CSF from individuals is both a costly and invasive process. Therefore, efforts have been concentrated on finding simpler methods, including assessment of blood samples (Tarawneh, 2020).

1.7 Treatment methods

1.7.1 Current treatment

Current treatment strategies for AD are aimed at reducing the severity of symptoms. Unfortunately, treatments to reverse or prevent the diseases are still yet to be developed. The most common form of treatment is Cholinesterase Inhibitors (CIs). The cholinergic hypothesis of AD states that systems in the basal forebrain are damaged early in the disease process. This damage is thought to result in the common symptom of memory loss, as well as other neuropsychiatric symptoms. The use of cholinesterase inhibitors has been shown to slow the progression of this decline (Yiannopoulou and Papageorgiou, 2013). Currently three CIs have been approved for general use, donepezil, rivastigmine and galantamine. However, these drugs are only effective when treating the initial stages of the disease. This again emphasises the importance of early diagnosis for AD (Yiannopoulou and Papageorgiou, 2013).

A drug used in the more advanced stages of AD is memantine. This treatment is believed to protect neurons from excitotoxicity, which is a phenomenon involving the toxic actions of certain neurotransmitters. Studies have shown improvements in cognition in individuals with an advanced stage of the disease (Yiannopoulou and Papageorgiou, 2013). Alongside the common cognitive symptoms of AD, patients also often experience certain behaviour and psychological symptoms, including psychosis, affective symptoms, hyperactivity, and apathy. In the early stages of these symptoms, CIs and memantine can reduce severity (Yiannopoulou and Papageorgiou, 2013). However, as the disease progresses and symptoms worsen, further drug treatment may be required.

Serotonin reuptake inhibitors are often used to treat AD related depression. These can also be used to reduce symptoms of psychosis. However, this is more often treated with antipsychotics such as olanzapine, risperidone, and ziprasidone. The use of these drugs has been met with some criticism, as patients have been linked higher mortality risk after use (Yiannopoulou and Papageorgiou, 2013). To treat anxiety related to AD, benzodiazepines are commonly used. Similarly, to antipsychotic drugs, caution has been exercised over the use of these. This is due the association with more rapid cognitive and functional decline

(Yiannopoulou and Papageorgiou, 2013). Despite the partial success of some treatment methods such as CIs, an effective treatment for AD has yet to be derived. A possible avenue for research is the identification of the biological mechanisms for disease development. These can be determined by assessing which genes are significant for the onset of disease, with subsequent analysis of their role within the body. Treatments can then be tailored to reversing the effect of genes in AD development (Calabrò *et al.*, 2021).

1.8 Genetics in Alzheimer's disease

1.8.1 Twin studies

Twin studies are used to establish the heritability of a particular trait, with the aim of removing the effect of an individual's shared environment. Heritability is the measure of the variation of a phenotype in a population which can be explained by the genetic variation. Twin studies are derived by analysing twins raised within the same family. Reasoning behind the use of twin studies is the monozygotic nature of identical twins who are expected to have the same genetic material (Sahu and Prasuna, 2016). Dizygotic (fraternal) twins on the other hand share around 50 percent of all genetic material. This is a similar percentage to nontwin siblings. Analysis of both types of twins leads to the calculation of a concordance rate, which can be defined as the probability that two individuals with the same genetic makeup will develop a certain trait.

Several twin studies for AD have been conducted over the past 80 years. One study used a cohort of Finnish twins to assess concordance for a range of dementia types. It was found that monozygotic twins had significantly higher concordance rates of AD than dizygotic twins. This relationship was not prevalent in both vascular and mixed dementia (Raiha *et al.*, 1996). Breitner et al., 1993, identified a significantly higher rate of concordance rate for monozygotic twins (78%), than dizygotic twins (39%). Similarly, to the previous study, this relationship was not observed for vascular dementia. Alongside this, the study estimated that the total heritability for AD was around 60%.

One of the largest twin studies compiled to date used a cohort comprised of Swedish twins (392 pairs). Results found in the two previously mentioned studies were reinforced. However, unlike earlier studies, findings were adjusted for factors such as age and genetic differences between genders (Gatz *et al.*, 2006). From this, the study estimated the heritability of AD to be 58-79%. Additionally, results demonstrated no significant difference with respect to gender for the prevalence or heritability of AD. Therefore, when assessing results across all twin studies, it can be estimated that the heritability of AD is around 60-80% (Ertekin-Taner, 2007). However, this estimate varies with age of individuals (Baker et al., 2022)

1.8.2 Early onset AD

The 'familial' version of EOAD is a hereditary illness, with estimates of heritability ranging between 92-100%. Risk of familial EOAD is extensive in families, with nearly half of all patients having at least one first-degree relative with the disease (Cacace, Sleegers and Van Broeckhoven, 2016). Genetic studies have revealed three causative genes for EOAD. These are the *amyloid precursor protein* (*APP*), *presenilin 1* (*PSEN1*) and *presenilin* 2 (*PSEN2*). The genetic component of the more common sporadic form of EOAD is similar to LOAD, this will be covered in greater detail in the next section.

1.8.3 Late onset Alzheimer's disease

Research has continually identified the presence of genetic variants associated with LOAD (Lynn M Bekris et al., 2010). The genetic aspect of AD has been shown to be polygenic in nature, with several genes being associated with disease risk but not causative, e.g., (Wightman et al., 2021). The gene which has shown to be greatest risk factor for AD is *apolipoprotein E protein* (*APOE*) (Husain, Laurent and Plourde, 2021). The *APOE* gene is a major lipid transporter, which plays a role in the development, maintenance and repair of the central nervous system (CNS). Three different forms of the *APOE* status exist within the human population, these are defined by ε2, ε3 and ε4 alleles. These isoforms are the result of the haplotype combinations of two *APOE* SNPs. These are rs429358 (C > T) and rs7412 (C > T) which lead to an amino acid change at position 112 and 158 within the *APOE* protein. The ε3 and ε4 variants are the most common alleles, with 70% and 25% of all alleles respectively. The remaining 5% of the *APOE* alleles are comprised of the ε2 form (Husain, Laurent and Plourde, 2021).

The presence of the ε4 allele has been associated with an increased risk of developing AD. With heterozygote carriers having a 3-4-fold increased risk of disease compared to non-carriers, this increases to 9-15 times for homozygote individuals (Husain, Laurent and Plourde, 2021). The presence of ε4 allele a has been linked to increases in Aβ accumulation and deposition in the brain, as well as exacerbating the construction of senile plaques. Microglia are cells of the CNS which have various roles within the brain. The removal of unwanted Aβ is one of these, however the presence of the ε4 allele has also been associated with a reduced ability to clear extracellular Aβ (Fernandez *et al.*, 2019).

1.8.4 Genome wide association studies

Genome-wide Association Studies (GWAS) examine the relationship between a range of genetic variants (single nucleotide polymorphisms (SNPs)) and a phenotype. SNPs are defined as an alteration in the DNA base sequence of an individual, where one of the four bases has been replaced by another. A variation in the base sequence is defined as a SNP if it is present in at least one percent of the population (Johnson, 2009). The purpose of a GWAS is to compare the allele frequencies between individuals with the disease (cases) and unaffected individuals (controls), which is typically done via a logistic regression. The significance of a SNP within a GWAS is defined by its effect size and p-value. A SNP's effect size is the log-odds ratio for disease risk associated with one copy of the minor allele. A p-value is a result of the test for association, representing the likelihood of a result being due to random chance. The smaller the value, the less likely the result is spurious (Andrade, 2019).

Statistical power in GWAS refers to the ability to correctly reject the null hypothesis and thereby infer a SNP has significant effect on disease status. Therefore, considerable importance is placed upon using an adequate number of samples (Sham and Purcell, 2014). A GWAS typically comprises 100,000-1,000,000 genotyped SNPs, however the estimated number of common variants (minor allele frequency ≥ 0.05) is greater than ten million (Li et al., 2009a). Therefore, modern GWAS can study only a proportion of possible disease heritability.

1.8.5 Imputation of missing genotypes

A method developed to increase the number of SNPs within GWAS is genotype imputation, in which missing genotypes are estimated from collections of previously sequenced individuals known as reference panels. Imputation techniques allow geneticists to include genetic markers not directly genotyped when conducting association analyses, whilst also allow the combining of separate GWAS and thereby increasing the number of available samples. This boosts the statistical power of GWAS, increasing the likelihood of deriving significant markers (Li et al., 2009b), due to rare variants (minor allele frequency (MAF) < 1%) being more likely to occur within larger cohorts (Korte and Farlow, 2013). It has been estimated that 24% of the phenotypic variance of AD can be attributed to directly genotyped common variants. This estimation reaches 33% following the inclusion of imputed SNPs (Ridge et al., 2013). Whilst this is still below the estimated level of 58-79% of heritability from twin studies, it is clear that increasing the number of SNPs can account for some of the missing heritability in AD (Wang *et al.*, 2021).

A common cause for limitations in the number of SNPs for GWAS is the use of genotyping arrays, as these can only genotype a certain number of SNPs and regions of the genome. The limited nature of genotyping arrays has resulted in the development of imputation, this process is achieved using a ‘reference’ dataset, which contains many genotyped SNPs. Reference markers are then used to impute missing SNPs in a separate sample, with the underlying assumption that individuals from the reference set were sampled from a genetically similar population as the target samples, as this increases the likelihood of accurate imputation (Halperin and Stephan, 2009). Furthermore, the genome coverage of genotyping arrays can also be restricted due to the choice of reference panels used (Tam *et al.*, 2019). For example, early genome-wide SNP arrays were mostly designed from reference panels based on European populations only. Due to different population structures across ethnic groups, these arrays provide poor coverage for non-European populations (Tam *et al.*, 2019).

Several reference datasets have been compiled, these include the 1000 Genomes project, the Human Genome Diversity Project, the Haplotype Reference Consortium (Schurz *et al.*, 2019) and recently the trans-omics for precision medicine (TOPMED) (Taliun *et al.*, 2021). As the number of samples used in reference panels continue to increase, the ability to impute rare variants will also rise. For example, the use of TOPMED has resulted in the imputation of SNPs with minor allele frequency (MAF) of 0.01%. However, some rare variants may still only be detectable using whole genome sequencing (Prokopenko *et al.*, 2021). In addition, imputation can also be used to aid meta-analyses of GWAS. It is often the case that different genotyping platforms are used across different studies. These platforms can in practice have little overlap in SNPs. Therefore, the most common practice is to impute individual cohorts separately, to maximise the number of SNPs common across studies (Verma *et al.*, 2014). As well as the reference dataset, the type of imputation software used is also important (Hancock *et al.*, 2012). Several 'free to use' packages have been developed to achieve high imputation accuracy. The method IMPUTE2 is a popular two-step algorithm used in imputation (Roshyara *et al.*, 2016). Haplotype information is inferred using a Markov Chain Monte Carlo approach. Following this, a hidden Markov model (HMM) is used to impute missing genotypes. SHAPEIT2 uses a graph based HMM to impute genotypes (Delaneau, Zagury and Marchini, 2013), whilst Minimac2 employs a state space reduction HMM to reduce computational burden (Das *et al.*, 2016).

Non-imputed genotypes are often coded as zero, one or two, according to the number of copies of the minor allele (Strandén and Christensen, 2011). The result of imputation is often three probability values, which represent the likelihood of the possible genotypes for a sample. For example, for two alleles 'A' and 'B', possible genotypes would be AA, AB and BB (Shin *et al.*, 2020). These are known as genotypic probabilities and always sum to one, for instance, a possible set of genotypic dosages could be 0.6, 0.3, 0.1 for the AA, AB and BB genotypes respectively. For further analysis such as the calculation of polygenic risk score (PRS), the three imputed genotype probabilities must be converted to a single value. This can be done in two ways, either through the usage of allelic dosages or by best-guess genotypes (Collister, Liu and Clifton, 2022a). Allelic dosages are calculated from the three genotype probabilities derived from imputation. This depends on which allele is determined as the effect allele. For instance, when using the example in the previous paragraph and delegating 'A' as the effect allele, the allelic dosage would be calculated as:

2*(0.6) +1*(0.3) = 1.5

This is the format with which the statistical software *PLINK* handles dosage values (Purcell, Neale, Todd-Brown, Thomas, Manuel A R Ferreira, *et al.*, 2007). Once in this format, values can then be used for analysis such as PRS. The second option known as best-guess genotypes involves the conversion of allelic dosages to 'best guess' genotypes, with resulting values in the same format as directly genotyped SNPs. This process is carried out using a set of thresholds for the conversion of allelic dosages, an example of which is given below:

$$hardcall\ dosage = \begin{cases} 0\ if\ allelic\ dosage\ \in [0.0,\ 0.1] \\ 1\ if\ allelic\ dosage\ \in [0.9,\ 1.1] \\ 2\ if\ allelic\ dosage\ \in [1.9,\ 2.0] \\ \quad Missing\ otherwise \end{cases}$$

Whilst the conversion of dosages to best-guess genotypes allows imputed values to be stored in the same format as genotyped SNPs, the process results in some loss of information (Collister, Liu and Clifton, 2022b). This is due to the possibility of some dosage values not falling within thresholds and subsequently being marked as missing (Collister, Liu and Clifton, 2022a). For example, the value of 1.5 calculated in the above example does not fall within the given thresholds, therefore the best-guess genotype would be recorded as missing for this sample. This is perhaps not surprising given that the highest genotype probability was 0.6, suggesting that the imputation of this genotype was not made with high confidence when compared to the true genotype. If this occurs often enough for any given SNP, the variant might be removed due to quality control measures.

The process of imputing genotypes is susceptible to errors as the quality of imputation relies on the reference dataset and imputation software used (Verma *et al.*, 2014). To reduce the possibility of spurious results from subsequent analysis, the quality of imputation should be assessed. A metric commonly used to examine imputation quality is the r^2 method. This is calculated by comparing the variance of the result of imputation (dosages) to genotypes if known by certainty (Chanda *et al.*, 2012). The formula for this is given by:

$$r^2 = \frac{var(g)}{2p(1-p)}$$

where $\underline{g}$ is the set of dosage values for a sample, with var(g) its variance. The frequency of the effect allele (p) is calculated as the sum of the dosage values across all samples divided by 2N. A score between zero and one is returned, with imputation quality increasing as r^2 increases (Schurz *et al.*, 2019).

1.8.6 Multiple testing correction methods

An element related to the statistical power of GWAS is multiple-testing correction. A typical GWAS may analyse several hundreds of thousands of markers, resulting in many association tests (Moskvina and Schmidt, 2008). The large number of tests performed in GWAS is therefore likely to result in some spurious results. A method to correct for this issue is multiple-testing correction, whereby the significance level α is adjusted to a value which reduces the possibility of spurious associations. A commonly used method known to achieve this is Bonferroni correction. In this process, a new value of α is derived by dividing the p-value 0.05 by the number of variants to be tested 'm'. This reduces the possibility of spurious associations due to a lower significance threshold. The Bonferroni correction assumes that tests are independent, which is not the case for a GWAS due to linkage disequilibrium between SNPs and will thus be overly conservative. The genome-wide testing burden was estimated to be the equivalent of approximately one million independent tests, leading to the commonly used p-value criterion of 5x10-8 for genome-wide significance (Pe'er *et al.*, 2008). Given this stringent significance level, sample sizes of several thousand individuals are typically required to achieve power in GWAS of complex traits (Wang and Xu, 2019).

Most historical studies used cohorts formed of mostly Caucasian individuals. One limitation of GWAS is the lack of diversity of study populations (Haga, 2010). As the geographic range of humans expanded past the African continent, sub populations have arisen due to isolation, interbreeding and adaptation to the local environment. To rectify this, a limited number of subsequent studies have been conducted in non-Caucasian cohorts. Findings have shown that results often derived in Caucasian cohorts are not similar to those in non-Caucasian populations (Haga, 2010). In some instances, variants were deemed associated with a disease in both cohorts, however the level of their significance was different, whilst in other instances, SNPs associated with a phenotype in one set of samples were not significant in a different population. These variations in results are of concern when considering using prediction algorithms developed using purely Caucasian cohorts.

Given the concerns surrounding differences in the genetic structure of populations, methods have been developed to correct for this issue if a sample of mixed population was used for a GWAS. The most popular method was developed by Price et al., 2006 and implemented in a program called EIGENSTRAT. Here, principal components (PCs) are estimated using SNPs within a GWAS. Prior to the calculation of PCs, it is usually necessary to" prune" the entire SNP set by removing variants in high linkage disequilibrium (LD) with other variants, thus preventing the PCs being dominated by genomic regions of high LD, such as the major histocompatibility complex (MHC) region. Such regions might introduce 'nuisance' clusters which can be falsely interpreted as population structure (Zhao *et al.*, 2018). The results of EIGENSTRAT PCs then used as covariates in a logistic regression when computing the association between each SNP and disease status. Each included PC can be thought of as a representative of separate population structures within the overall dataset. Therefore, the inclusion of these components adjusts each SNP for all possible variations of ancestry (Zhao *et al.*, 2018). The selection of the appropriate number of PCs can involve plotting PCs against one another in a scatter plot. If underlying populations exist within the set of SNPs, these will be identified by clusters of samples. The maximum number of PCs plotted before clusters disappear is equal to the number used for adjustment (Zhao *et al.*, 2018).

1.8.6.1 Quality control measures

A key step for GWAS is quality control (QC) for both samples and SNPs. The avoidance of such steps may lead to systematic biases in the outcome and increased risk of false positive associations (Coleman *et al.*, 2016). The likelihood of false positive results in a GWAS is increased due to the large amounts of SNPs involved, which can be susceptible to random errors in genotyping. Therefore, stringent QC steps are required prior to analysis (Coleman *et al.*, 2016).

Quality control measures focus on both SNPs and samples. In terms of variants, the majority of GWAS begin by filtering SNPs based upon their minor allele frequency (MAF). Low MAF has been linked with issues related to spurious results, including an increased risk of genotyping error and incorrect size of the association statistic. The threshold for removing SNPs is typically a MAF presence of less than 1% in the sample. This removes SNPs termed as rare variants. On certain occasions in which a sample size is deemed small, a threshold of 5% might be employed. This is due to SNPs with low MAF having greater potential effect on results within small sample sizes (Coleman *et al.*, 2016).

A method used to detect genotyping errors is assessing deviation from Hardy-Weinberg equilibrium (HWE). In this process, observed genotypes are compared to expected values. These expected values are derived from the HWE assumption. This law states that the genotype frequencies AA, AB and BB will occur in proportions of p^2, 2pq and q^2. In which p is the allele frequency of A, with q = 1-p being the respective value for B (Graffelman and Weir, 2016). A P-value of P < 1 x 10-5 is typically used, to remove SNPs showing large deviations from HWE, while retaining SNPs with small deviations that may have arisen by chance. This process is often only carried out in controls only, since, under the assumption that penetrance of risk genotypes is low, deviations from HWE due to associations of genotypes with disease risk will be small (Coleman *et al.*, 2016). Individuals who are closely related will typically share more of their genome than randomly chosen members of the population. The presence of related individuals in a GWAS may skew analysis. The identical-by-descent (IBD) is a metric derived by assessing the overlap in alleles between samples, with a typical score used to remove individuals being a pi-hat of > 0.1875 (Coleman *et al.*, 2016). This value removes closely related samples whilst avoiding reducing the cohort size drastically (Coleman *et al.*, 2016).

A QC step for imputed genotypes is to assess the accuracy of imputation, as the quality of derived alleles can affect subsequent analysis. Several factors can influence the quality of imputation, these include the type of software used, as well as the reference panel chosen for genetic information. Ideally, samples in the GWAS to be imputed should be from the same ethnic population of the reference panel (Stahl, Gola and König, 2021). The accuracy of imputation is calculated by comparing predicted genotypes to genotype values directly genotyped in separate cohorts. Several different methods have been developed to calculate the quality of imputation. The concordance rate is defined as the proportion of the correctly imputed genotypes with respect to all imputations (Krithika *et al.*, 2012). The r^2 statistic is also used as a common method to assess accuracy and is measured as the correlation between imputed and true genotypes (Iwata and Jannink, 2010). Values lie between zero and one, with imputation quality increasing as values tend towards one (Zheng *et al.*, 2015).

1.8.6.2 GWAS in AD

The first GWAS in AD research used a cohort of 1808 LOAD cases and 2062 control from the United Kingdom and United States of America. 17,343 SNPs were tested for association in a case-control sample. Three AD significant markers present on Chromosome 19 and in high LD with the *APOE* gene were discovered (Grupe *et al.*, 2007). Following this, several subsequent GWAS emerged which used greater numbers of SNPs and samples. As techniques for collecting samples and computing greater amounts of data developed, sample sizes used in GWAS continued to increase. The GWAS of Harold et al., 2009 analysed a cohort of 3,941 AD cases and 7,848 controls, using 529,218 SNPs. In agreement with previous GWAS, the significance of the *APOE* region was again demonstrated. However, two novel loci were also associated with AD at genome-wide significance (p-value threshold of 9.4x10-8), near to the *clusterin* (*CLU) and phosphatidylinositol blinding clathrin assembly protein* (*PICALM)* genes. These were further replicated in a secondary dataset comprising 2,023 cases and 2,340 controls, strengthening evidence for association. A further 13 SNPs were shown to be associated with AD but not at genome wide significance. The association of the *CLU* gene (rs11136000 SNP) with AD was also identified at genome-wide significance in a separate study (Lambert, Heath, Even, Campion, Sleegers, Amouyel, et al., 2009a). This study conducted a GWAS using a French cohort comprising of 2,032 cases and 5,328 controls.

Lambert et al., 2013 continued research with a GWAS comprising of 74,046 total samples. Association was tested using a two-stage meta-analysis. In contrast to both Harold et al., 2009 and Lambert et al., 2009, this study used imputed variants. The use of the 1000 genomes reference panel resulted in a dataset of 7,055,881 SNPs (17,008 cases and 37,154 controls), with samples collated from four previously published GWAS. These were the Alzheimer's Disease Genetic Consortium (ADGC) (Naj *et al.*, 2011), the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium (Frisoni et al., 2008), the European Alzheimer's Disease Initiative (EADI) (Lambert, Heath, Even, Campion, Sleegers, Hiltunen, et al., 2009) and the Genetic and Environmental Risk in Alzheimer's Disease (GERAD) (Harold *et al.*, 2009) Consortium.

For replication purposes, 11,632 SNPs reaching a significance level of P < 1 x 10-3 in the original dataset were tested for association in an independent sample (8,572 cases and 11,312 controls). In addition to the *APOE* region, a further 19 SNPs were identified as genome-wide significant in both stages. Of these 19 variants, 11 were newly associated with AD. More recently, further GWAS have been undertaken using larger sample sizes. The study of Kunkle et al., 2019 used 94,737 individuals (35,274 cases and 59,163 controls). Twenty previous risk variants were confirmed for GWAS significance, whilst a further five novel loci were also identified. Pathway analysis also confirmed the influence of genes related to *APP* and amyloid beta. This implies that these processes are not only associated with EOAD, but also with the development of LOAD (Kunkle *et al.*, 2019).

Figure 1.2: Manhattan plot of Genes containing SNPs significant at a genome wide level from Kunkle *et al.*, 2019.

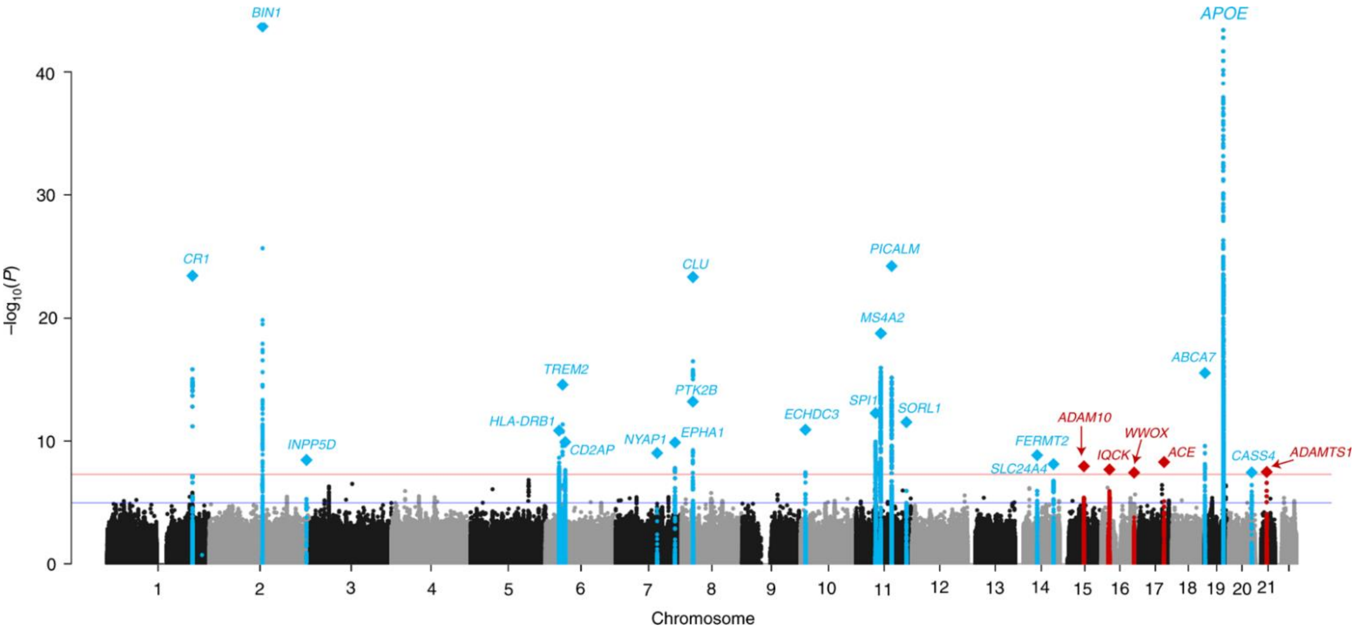


Figure 1.2: This Manhattan plot demonstrates the results from (Kunkle *et al.,* 2019). The red horizontal line signifies the genome wide significant level of P < 5x10-8. Genes marked in blue resemble previously associated loci, whilst those newly associated are shown in red. This image has not been altered from the original source.

One limitation of GWAS is the difficulty of obtaining sufficiently large numbers of AD cases (Marioni *et al.*, 2018). Marioni et al., 2018, used a novel GWAS approach. This study used a proxy-AD status for analyses, with cases defined as those with a familial history of AD (maternal and/or paternal). This enabled the use of the UK Biobank cohort, whose members are typically too young to develop LOAD. A dataset comprising of 314,1278 samples was used, with 27,696 maternal cases and 14,338 paternal cases, meta-analysed with the Lambert et al., 2013 sample. The proxy AD status increased the sample size; however, the disease measure was less accurate due to an increased risk of including non-AD dementia cases. Analyses resulted in the discovery of three novel AD associated loci (Marioni *et al.*, 2018). A similar approach was also investigated in (Jansen *et al.*, 2019), in which both clinically diagnosed cases and AD by proxy were used (71,880 cases, 383,378 controls). Analysis identified 29 risk loci and implicated a further 215 potential related genes (Jansen *et al.*, 2019).

The GWAS of Wightman et al., 2021, comprised 13 cohorts totalling 1,126,563 individuals (90,338 cases and 1,036,225 controls). Meta-analysis identified 3,915 SNPs across 38 independent loci as being genome-wide significant. Of these, five loci were associated with AD for the first time. Further analysis also implicated both immune cells and microglia as cells of interest (Wightman et al., 2021). The most recent GWAS for AD Bellenguez et al., 2022 performed a two-stage GWAS, in which 75 risk loci were identified, 42 were unknown at the time of analysis (Bellenguez, Küçükali, Iris E Jansen, et al., 2022). The first stage of analysis was carried out in 39,016 cases and 46,828 controls, with results replicated in 25,392 cases and 276,086 controls. Pathway analysis also implicated both amyloid and tau in AD development, as well as microglia involvement.

When reviewing all GWAS in AD research, the number of genome-wide significant loci discovered has increased as sample sizes have become larger. This is a similar outcome to other disorders considered polygenic such as schizophrenia and major depression. Therefore, it can be hypothesised that results of GWAS support the existence of a polygenic component of AD. However, this is still subject to debate amongst various researchers. Zhang *et al.*, 2020 investigated the genetic loading of AD using a genetic risk score. They concluded the genetic component of LOAD was the result of several hundred common variants, equivalent to 0.01% of the SNPs used with MAF > 1%. This percentage suggests an oligogenic model of AD, in which the genetic component of the disease is the result of a few genes and not many (polygenic). This estimate was also lower than other disorders considered polygenic such as schizophrenia 17.5%, major depression 3.2% and Parkinson's disease 16.4% (Zhang *et al.*, 2020).

1.9 Biological factors

1.9.1 Microglia

Microglia are stable, long living cells with low renewal rates, with the primary role of supporting the brain's neurons from issues such as infection, trauma, or neurodegeneration (Bachiller *et al.*, 2018). Despite their role of preserving homeostasis in the brain, microglia have been associated with AD development under the amyloid cascade-inflammation hypothesis, which suggests microglial activation links the development of Aβ plaques and the creation of NFTs.

1.9.2 The synapse

Synapses are defined as cellular junctions, which pass information from a presynaptic neuron to a postsynaptic cell (Burns and Augustine, 1995). These synaptic junctions are diverse, with differences in types of neurotransmitters, synapse composition and roles. As discussed previously, the two main hallmarks of AD are the presence of both Aβ and tau protein. Neurofibrillary tangles have become strongly associated with the development of AD. However, the presence of tangles has not been associated with the reduction of synaptic density in neurons. Rather, the presence of soluble forms of tau might be more related to the reduction of synapse function (Robbins, Clayton and Kaminski Schierle, 2021). Brains with AD have been shown to have higher levels of hyperphosphorylated species of tau, with these forms of tau spreading between synapses, encouraging the spread of disease (Robbins, Clayton and Kaminski Schierle, 2021). The presence of such forms of tau have been related to impaired axonal transport in post-mortem AD brains. This reduces the transport of organelles such as mitochondria, which are important in maintaining synapses (Robbins, Clayton and Kaminski Schierle, 2021).

1.9.3 Biological pathways enriched in AD.

Pathway analysis, also known as gene-set enrichment analysis, is a strategy in which prior biological information is used to assess the relationship between sets of variants and a phenotype. A common method for defining gene sets is known as functional annotation, in which genes are selected based upon their involvement in molecular and biological functions. Information regarding a gene's function is typically extracted from a knowledge base. A widely used database is Gene Ontology (GO), which breaks gene function down into molecular function, cellular component and biological process (Ashburner *et al.*, 2000). However, a more descriptive database known as the Kyoto Encyclopaedia of Genes and Genomes (KEGG) is also used. This details how groups of genes result in biomolecular activities through interactions and reactions (Silberstein *et al.*, 2021). However, the use of databases to define gene sets relies on prior knowledge of the biological function of genes. This is not always available for genes which have not been studied at length. A further technique to derive sets of related genes is the use of omics data, which can be defined as sets of biological data used to assess biological functions. An example of this is the weighted correlation network analysis (WGCNA), which produces groups of related genes by computing correlations of gene expression data (Langfelder and Horvath, 2008).

Numerous studies have focused on identifying further pathways related to AD. The first systematic pathway analysis of late-onset AD GWAS was performed by Jones et al., 2010. A combination of two GWAS GERAD (Harold *et al.*, 2009) and EADI (Lambert, Heath, Even, Campion, Sleegers, Hiltunen, et al., 2009) comprising of 19,000 participants were used. Processes related to both cholesterol metabolism and innate immune response were established as significant for disease development. These biological aspects have been previously associated with disease onset (Jones et al., 2010). However, uncertainty still exits on whether these are causal effects, or the result of the disease process. Analysis identified that both processes were aetiologically relevant in this instance (Jones et al., 2010). Cui et al., 2018, also conducted analysis into biological pathways linked with AD. Results demonstrated that pathways linked with both Toll-like receptors and natural killer cell mediated cytotoxicity were associated with AD. Toll-like receptors are a class of protein which form part of the immune system. They recognise certain types of microbes and trigger appropriate responses. It is accepted that Toll-like receptors play a role in microglia induced Aβ settlement. Natural killer cells are lymphocytes which form part of the innate immune system and function by destroying virus-infected cells. However, research has shown that the presence of these cells might increase neuroinflammation in certain circumstances (Cui *et al.*, 2018).

Identification of AD associated pathways was carried out by Silver et al., 2012. Single nucleotide polymorphisms (SNPs) from the Alzheimer' Disease Neuroimaging Initiative (ADNI) dataset were mapped to genes previously implicated in biological pathways. In total 66,162 SNPs were allocated to 4425 genes, which in turn were mapped to 185 known pathways, derived from the KEGG database. A form of sparse reduced-rank regression was used to assess the association between pathways and AD. Phenotypes were derived from magnetic resonance (MR) images for 99 cases and 164 controls. The top 30 most associated pathways were ranked, with results replicating previous findings (Silver *et al.*, 2012). This included pathway implicated in processes such as insulin production, cardiac motion, melanogenesis and Huntington's disease. Type 2 diabetes through the disruption of insulin production has been linked with increased risk of AD development. Research has identified insulin involvement in the maintaining of synapses, with reductions linked to learning and memory loss. Insulin has also been associated with the metabolism of both beta-amyloid and tau, the two main hallmarks of AD (Biessels, Kappelle and Utrecht Diabetic Encephalopathy Study Group, 2005).

Pathway analysis was conducted by Kunkle et al., 2019, a large GWAS which identified five new genome-wide loci. Analysis implicated pathways related to immunity, lipid metabolism, tau binding proteins and *APP*. This showed that variants associated with *APP* and Aß are not only linked to EOAD but LOAD as well. These pathways are used for analyses conducted in Chapter 7 of this thesis. The most recent pathway analysis for AD was conducted by Bellenguez, Küçükali, Iris E. Jansen, et al., 2022), in which a two-stage GWAS comprising 111,326 diagnosed and proxy cases with 677,663 controls identified 75 risk loci (42 novel). Further pathway enrichment analysis identified the involvement of amyloid/tau pathways, as well as links to microglia implication. These pathways were not used in this thesis (Chapter 7) as this research was not published at time of analysis.

1.10 Risk prediction in Alzheimer's disease

1.10.1 Polygenic risk score

Since the development of the GWAS, it has become apparent that many disorders are polygenic in nature, suggesting the genetic components of these diseases are the result of multiple variants (Lvovs, Favorova and Favorov, 2012). Each SNP will most likely have a minimal effect on disease development and cannot be used on an individual basis for assessing disease risk. A common method to calculate risk is to combine the effects of SNPs in an individual's genome, commonly achieved through polygenic risk score (PRS). This is calculated through a weighted sum of an individual's risk variants and their effect sizes calculated in a GWAS (Lewis and Vassos, 2020a).

An important aspect of PRS calculation is the use of two independent cohorts. The first of these is the training set, in which p-values and effect sizes for each SNP are generated. The second cohort termed the test set, is used for the generation of PRS values for each individual. SNP effect sizes generated in the training set are multiplied to their allele counts in the test set. It is important that samples in the test are independent to those in the training cohort, as this separation of individuals will avoid spurious results when estimating PRS association (Choi, Mak and Paul F O'Reilly, 2020). Following the derivation of PRSs for each individual in a cohort of cases and controls, these values are typically used to predict the disease risk using logistic regression (LR). The model's ability to discriminate between cases and controls is then assessed.

SNP association effect sizes generated for each variant can be uncertain, and not all SNPs within a GWAS will influence the trait (Choi, Mak and Paul F O'Reilly, 2020). To remove variants unrelated to disease when deriving the PRS, SNPs can be filtered by their respective p-values from a GWAS, however, the optimal threshold is not always known prior to analysis. Therefore, PRS is typically calculated across a range of thresholds. The best performing threshold is then chosen from results. This process can be seen as a form of feature selection in machine learning field, as a forward selection is used to choose a set of optimal variants. One further issue when selecting the relevant risk variants for a trait is Linkage Disequilibrium (LD). Variants in the genome often share strong correlation, making it difficult to extract the desired signal between risk alleles and a trait (Choi, Mak and Paul F O'Reilly, 2020). Two methods are used to reduce this correlation, known as pruning. For pruning, a window of SNPs is selected, within this window SNPs are then sorted by their genomic position. Correlations are calculated between the first ordered SNP and all others within the window. If a correlation is greater than a predefined value, one of the correlated pair is removed. Once either the index SNP or the correlated SNP has been removed, the algorithm moves on to the second SNP. The result of this process is a set of (almost) uncorrelated (or LD-pruned) variants (Calus and Vandenplas, 2018).

Clumping also removes correlated SNPs but achieves this is in a different way. Similarly, to pruning, a predefined genomic window is selected. Within this window, the most statistically significant SNP is chosen to be the index SNP. SNPs in the same window whose correlation with the index is greater than a predefined value are removed. Following this, the next remaining SNP is chosen as the index variant and the process repeats. The result of this is a set of significant SNPs which are largely independent of each other (Choi, Mak and O'Reilly, 2020). Clumping is most often preferred to the pruning process due to preferentially selecting the most associated SNPs with the trait, as well as retaining further significant variants in the same genomic region. These variants run the risk of being removed during pruning due to the almost random nature of how variants are discarded (Choi, Mak and O'Reilly, 2020). The method of clumping is often combined with the p-value thresholding process, which produces multiple sets of variants.

Despite the clumping and thresholding method for PRS successfully producing results in AD research, studies have shown that discarded SNPs through the clumping process might be limiting prediction accuracy (Ge *et al., 2019*). Bayesian methods such as LDpred have allowed for the incorporation of genome-wide markers, as the effect of LD is allowed for when deriving statistics. However, Bayesian priors used in this instance often introduce significant computational challenges. A Bayesian method which has reduced this burden is PRS-CS, in which effect sizes are inferred using a high-dimensional Bayesian regression framework. This method has been known to outperform traditional PRS methods in simulation trials (Ge *et al.*, 2019).

Several published studies have used PRS for AD prediction purposes. Leonenko *et al.*, 2021 used samples from a range of different cohorts to calculate PRS (p-value thresholds ≤ 5e-8.1e-5,0.1 and 0.5). Performance using these scores ranged from 55.7 – 73.7% AUC. Differences in prediction accuracy were due to alterations in methodology for modelling. These included removing SNPs within the *APOE* region, reintroducing these removed variants and using scores for the *APOE* (ε2,ε4) alleles. A further study calculated PRS for LOAD patients from the brains for dementia research (BDR) cohort (Hayes, Hudspith and Francis, 2012). A multivariate regression was used for AD prediction with PRS, SNPs within the *APOE* region, age and gender used as predictors. These features were used to discriminate between controls and cases, with a prediction accuracy 82.5% AUC. Discrimination was then assessed in a separate cohort, in which prediction accuracy between mild cognitive impairment (MCI) and cases was calculated. A prediction score of 61.0% AUC was achieved (Chaudhury *et al.*, 2019). Further research assessed the prediction performance of PRS in the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset. Similarly, to previous studies, age, gender and both the *APOE* alleles were included as covariates to the model, alongside PRS generated from chosen SNPs. When discriminating between cases and controls, an accuracy of 80% AUC was obtained. Further analysis was then conducted between MCI and cases, with an accuracy of 73.5% AUC (Daunt *et al.*, 2021).

1.10.2 The use of machine learning for prediction of complex genetic disorders

Following the widespread usage of PRS for prediction in AD, questions have been raised regarding the possibility of using machine learning (ML). ML can be defined as a set of algorithms which derive patterns within datasets, these insights can then be used to make informed predictions on independent data. Despite achieving AUC of 80% in certain studies, PRS is limited by its use of aggregating the SNP's effects with linear methods. These limitations include an assumption of independent predictor effects, normally distributed data and uncorrelated features (Ho et al., 2019a). Neurodegenerative conditions such as AD are known to be polygenic in nature, in which a complexity of genetic factors is thought to contribute to disease development (McCarroll and Hyman, 2013). The assumptions of linear modelling and the restriction to additive effects only reduce the possibility of capturing complex interactions between genes (Ho et al., 2019). In comparison, ML algorithms use non-parametric methods which can derive insights from large datasets with the ability to assess non-linear interactions between variables. It is these factors which have increased the interest in ML approaches within bioinformatics.

1.11 Outline of thesis

Interest for ML in disease prediction is increasing on a yearly basis. This thesis assesses whether the ability of ML algorithms to identify non-linear patterns in the data can enable greater prediction accuracy than PRS for AD prediction. As discussed throughout the introductory chapter, AD is the most common form of dementia (Duong, Patel and Chang, 2017) and it was for this reason it was chosen for a subject matter. Common issues such as population structure, imbalances between samples and features and computional burdens are discussed.

The central aim of this thesis is to compare the performance of both ML and PRS for AD prediction using genetic data. It is theorised that ML algorithms could have future potential to be used in clinical settings. The ability to predict disease status from genetic data could improve outcomes for patients through early detection and subsequent correct treatments. However, models must achieve high levels of accuracy for this possibility, as the impact of incorrect predictions could be severe on patients (Kappen *et al.*, 2018). Initially, the prediction performance of a range of ML techniques shall be assessed on a small set of AD associated SNPs. The best achieving algorithms will then be used on much larger sets of variants. This includes the use of both imputed and non-imputed genotypes, as well as SNPs related to specific AD related biological pathways.

Chapter 2 provides an overview of the field of ML. This includes a brief introduction to the history of ML. Then aspects of supervised learning are discussed at length. Aspects of the learning process such as loss functions and optimisation are then covered. Focus then shifts to common challenges when developing ML models, including overfitting, high dimensionality issues and missing data. The remainder of the chapter covers the ML methods used in the thesis. This includes in depth descriptions of decision tree-based methods, support vector machines, gradient boosting and naïve bayes approaches.

A systematic review of the literature for ML in AD is detailed in Chapter 3. In this review articles were included only if the main source of data used was SNPs. However, exceptions were made for studies who used imaging data alongside genetic variants. The included studies were assessed for risk of bias using the prediction model risk of bias assessment tool (PROBAST). Inferences drawn from the review included common usage of the ADNI dataset, alongside the consistent underreporting of metrics such as AUC and calibration. Sample size was also an area of focus, as most studies used datasets with imbalances between samples and features.

Analyses in Chapter 4 introduce the comparison between ML and PRS for AD prediction. ML algorithms were selected to make predictions based upon a small set of SNPs deemed AD significant in Kunkle et al., 2019. These predictions accuracies were then compared to PRS. Focus is also given to the impact on prediction from adjusting predictors using principal components (PCs), in order to correct for possible population stratification. Further techniques to avoid confounding due to age and sex were also investigated.

Following the establishment of the most accurate ML algorithms in Chapter 4, a greater number of SNPs are used for AD prediction in Chapter 5. The Genetic and Environmental Risk in Alzheimer's Disease (GERAD) (Harold *et al.*, 2009) dataset was clumped at differing p-value thresholds to obtain sets of variants, with the largest SNP set containing over one hundred thousand predictors. To reduce the burden of high dimensionality, several feature selection techniques are tested. This includes both traditional methods such as decision tree algorithms, embedded techniques such as regularisation and biological importance. The performance of ML algorithms following the use of feature selection was then compared to not using dimensionality reduction. This is alongside comparing the predictive capability of both PRS and ML.

Chapter 6 focuses on using imputed variants. The progression from non-imputed to imputed variants increases the number of SNPs for analysis from 400,000 to over 6,000,000. Therefore, the most efficient feature selection techniques from Chapter 5 are used, alongside the same ML techniques. To compare the ML performance with the PRS, SNPs were chosen using the clumping method.

The selection of SNPs for prediction purposes in Chapter 7 altered from the methods used in Chapters 5&6. Instead of selecting variants across the whole genome, SNPs associated with the genetic pathways deemed AD significant in Kunkle et al., 2019 are used. The performance of ML is compared against to both PRS using thresholding and polygenic risk score continuous shrinkage (PRS-CS) for each pathway. This is then expanded upon, within a multivariate model. The 9 sets of genotypes are used in one model, for both ML and PRS methods. This analysis is then replicated using PRS generated from the 9 pathways as inputs to both ML and PRS methods. A univariate model is then assessed, in which SNPs from all 9 pathways are combined into one set, following the removal of duplicate SNPs, PRSs generated from these SNPs are used as an input to ML, as well as a LR.

Finally, Chapter 8 provides a summary of results across all chapters. This is followed by interpretation of what these results suggest for the prediction of AD using ML. Discussion is also had regarding possible further steps for the development of ML in AD research, as well as limitations for analyses in this thesis.

2 Machine learning methods

2.1 Background of machine learning

Definitions of the term 'learning' vary depending on the subject matter at hand. In general terms, it can be defined as a change in behaviour caused by an experience. The ability to learn has enabled species to adapt and climatise to their surroundings. Homo-sapiens have the greatest ability to do this, due to their superior intelligence in the animal kingdom. Therefore, the ability to learn and reason in the biological sense has been present for billions of years (Roth, Krochmal and Németh, 2015). Despite this, the question of whether a machine could learn has only come to the forefront in recent centuries.

The increased usage of computers in recent decades has resulted in greater amounts of data being generated and stored. Sources of this data include smartphones, social media, healthcare and businesses (Elgendy and Elragal, 2014). The presence of this data has resulted in increased demand to understand patterns and nuances which can be beneficial to users. A paradigm fitting this requirement known as machine learning (ML) has increased in popularity in recent decades (Sarker, 2021). ML can be defined as the study of algorithms which learn from patterns within datasets with the intention of then making informed decisions. The term was introduced in 1959 by computer scientist Arthur Samuel (Awad and Khanna, 2015). The increase in popularity for ML algorithms is the result of their ability to analyse large datasets and assess complex non-linear relationships between features. This is an advantage over linear methods which only assess linear relationships (Ryo and Rillig, 2017).

2.2 Types of machine learning

Machine Learning models can be separated into four broad categories. These are supervised, semi-supervised, unsupervised and reinforcement algorithms. The following section provides background knowledge of the supervised paradigm, as all algorithms used in this thesis fall into this category.

2.2.1 Supervised learning

Supervised learning is a ML technique used for data sets which contain labelled examples, in which each data point comprises a set of features related to a label (output). Features can be defined as individual measurements of properties related to the label and can be both continuous and categorical in nature. The core aim of supervised learning is the approximation of a function which maps features to outputs. The prediction to a discrete output is known as classification, whereby inputs are mapped to a set of class labels (categorical output). Whilst mapping inputs to a continuous output is known as regression. A formal depiction of the supervised learning process is shown in Figure 2.1:

Figure 2.1: A Block diagram that outlines the supervised learning process within machine learning (Liu and Wu, 2012).

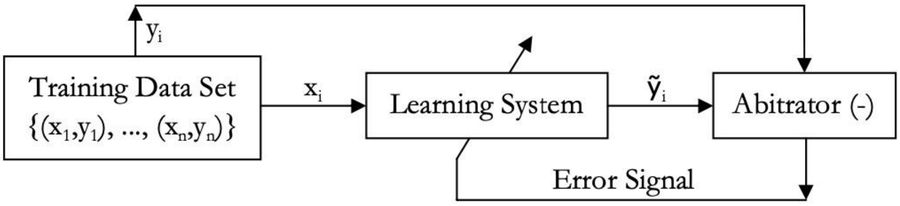


Figure 2.1: (Xi, Yi) represents a supervised training sample, with ‘x’ representing an input and ‘y’ representing the corresponding label. The data inputs Xi are provided to the learning system, which generates an output of ỹi. These outputs are then compared to the ground truth labels Yi. This image has not been altered from the original source.

The difference between the predicted output (ỹi) of the learning system and truth labels (yi) is termed the error signal. This value is propagated back to the learning system in order to update model parameters, with the central aim of minimizing the difference between model outputs and truth labels (Liu and Wu, 2012). This difference is termed ‘generalisation error’, which can be defined as a measure of how accurately an algorithm can predict on unseen data.

Broadly there are three types of classification. Binary classification refers to observations that are to be predicted into one of two classes only. These classes are often termed ‘positive’ and ‘negative’, in which the positive class represents the target, such as having a disease, whilst the negative represents an individual without the disease (Yousef, 2019). Multi-class classification is a scenario in which observations can be predicted into one of three or more classes. In this instance, none of the classes are defined as positive or negative. An example of this might be a face recognition system, in which an image is classified as belonging to a particular person. Therefore, the number of potential classes can be very large depending on the problem. Multi-label classification refers to when a sample can be predicted to have 3 or more labels. This scenario may occur in the field of photo classification, in which an image might contain multiple objects.

Supervised learning for regression purposes has certain parallels with classification. The data in question contains a set of features, however these are linked to a continuous target variable. Linear regression uses the mathematical equation, i.e., y =a+ bx + e, which describes the line of best fit between a dependent variable (y) and independent variable (x). The regression coefficient b represents the amount of variability between y and x, which is also known as the 'slope' in linear modelling. Whilst e represents the error term, and a is the 'intercept'. This can also be extended to multivariate regression, in which more than one input is used to influence a dependent variable (Kumari and Yadav, 2018).

2.3 Algorithmic learning

All analyses conducted in this thesis predicted to a binary class containing either 'control' or 'case', resulting in the requirement of supervised techniques only. Therefore, the rest of this chapter is going to focus on methodology concerning classification algorithms.

2.3.1 Optimisation

Optimisation is the process of finding a set of inputs which results in the minimum or maximum of an objective function. The field of supervised learning can be defined in terms of a function proximation, whereby the unknown function which maps inputs to outputs can be approximated by a learning algorithm. To determine this approximate function, an optimisation algorithm is used in most ML architectures. The function of which is to calculate optimal parameters learned from the given data (Sun et al., 2019). Popular optimisation methods can be divided into three general categories, these are first-order, higher-order and heuristic derivative-free optimisation methods.

When considering optimisation problems in supervised learning, the goal is to find a function f(x) which minimises the error in training samples. This is defined below:

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} L\left(y^i, f(x^i, \theta)\right), \qquad (2.1)$$

, where N is defined as the number of training samples, θ is the parameter of the mapping function. The variables xi and yi are a feature vector and label respectively, with as the loss function.

When considering first order methods, one of the most used in ML is the gradient descent procedure (Ruder, 2016). The function of this process is to update the weights of a model in an iterative fashion, with the aim of minimising the error of classification. This is achieved by moving in the negative direction of the gradient for the objective function. This process is controlled by a learning rate η, which controls the size of adjustment for weights with respect to the loss gradient. A formal definition of gradient descent for regression purposes is as follows. F(x) is a function to be learned, L(θ) is the objective function and θ is a parameter to be optimised. The function to be optimised is:

$$L(\theta) = \frac{1}{2N} \sum_{i=1}^{N} \left(y^i - f_\theta(x^i)\right)^2, \qquad (2.2)$$

$$f_\theta(x) = \sum_{j=1}^{D} \theta_j x_j, \qquad (2.3)$$

, where N is the number of training samples, D is the number of features, whilst x^i is an independent feature with yi as the corresponding target variable. The gradient descent process alternates between the next two steps:

Compute L(θ) for θj to arrive at the gradient for each θj:

$$\frac{dL(\theta)}{d\theta_J} = -\frac{1}{N} \sum_{i=1}^{N} \left(y^i - f_\theta(x^i)\right) x_j^i \qquad (2.4)$$

Update each θj in the opposite direction to minimise the function:

$$\theta_j' = \theta_j + \eta . \frac{1}{N} \sum_{i=1}^{N} (y^i - f\theta(x^i))x_j^i \qquad (2.5)$$

Despite its popularity, the gradient descent algorithm has a computational drawback. All training samples are used during the optimisation process, this can incur long training times when learning within large data sets (Ruder, 2016).

2.3.2 Loss functions

Loss functions are a key component of ML as they aid the process of learning. They compare model predictions to actual observations, with the aim of minimising the difference between these two sets of values. The selection of an appropriate cost function is related to the effectiveness and development of a model (Wang *et al.*, 2022). The type of loss function used is often decided by the learning problem, with different options for supervised and unsupervised learning.

For classification purposes, loss functions range in complexity. The perceptron loss function is a piecewise function. For a predicted class membership equal to the real label, the loss value is zero. Otherwise, the value is considered as the absolute value of the predicted value. This loss function is easy to optimise as it continuous and therefore differentiable everywhere. However, the function has poor generalisability and is not robust in the presence of noisy data (Wang *et al.*, 2022). An alternative method is the logarithmic loss function that operates by calculating the conditional probability of a sample being predicted as its label, with a greater conditional probability leading to a smaller loss value (Wang *et al.*, 2022). Another example of a loss function for classification is the sigmoid cross entropy. The prediction probability of a sample is projected using the sigmoid activation function. The loss of each is sample is then calculated by taking the difference between projected score and observed values. This process is most often termed cross-entropy loss in ML (Wang *et al.*, 2022). A combination of logarithmic loss and sigmoid cross entropy was used for algorithms in this thesis.

2.4 Model training and validation

Training and validation are important steps in producing accurate ML algorithms (Tan *et al.*, 2021). The central terms involved are ‘training’, ‘test’ and ‘validation’. These represent sections of the dataset used for differing purposes. The training phase allows the algorithm to learn the underlying patterns between features and class targets, with model parameters such as feature weights and biases estimated. This also extends to hyperparameters, which are external model parameters set prior to the training phase and control the learning process. Optimal values can be determined through trials during training. To maximise the possibility of accurate representation, a sufficiently large proportion of the dataset is required for training. The percentage of samples used is often determined by the user, however the recommended amount is 70% (Maleki et al., 2020). Following the training phase, the next step assesses how successfully the algorithm learned the nuances of the dataset. This is achieved by testing the trained model on a separate section of the dataset. Sections of a dataset containing independent samples are known as both test and validation sets, derived from the remaining 30% of the dataset. Confusion has arisen due to these terms being used interchangeably, with their definitions altering depending on the method of development.

The process known as holdout validation described in the previous paragraph performs poorly when estimating hyperparameters. This is due to the same split of data being used to both select hyperparameter values and assess model performance, which has been shown to result in over optimistic predictions (Maleki et al., 2020). Another method of algorithm development used is the nested approach. This uses a training, validation and test set to derive a prediction model. The nested method overcomes this issue by deriving hyperparameter values in the validation set, a separate entity to both the training and test set. This separation of hyperparameter selection and model evaluation reduces the possibility of over-optimistic model performance. A typical split of a dataset might be 70% training data, 15% validation and 15% test data (Maleki et al., 2020).

An issue which affects all methods described in this section is termed the ‘easy test set’. This refers to the possibility that model performance might be associated with how the data was split. If an algorithm is developed through one dataset split only, this might result in bias when reporting performance, as the composition of samples in the test could differ significantly on each occasion (Crisci, Ghattas and Perera, 2012). A method to overcome this is to report average performance across multiple splits, this reduces the effect of test set selection bias (Crisci, Ghattas and Perera, 2012).

2.4.1 Cross-validation

One method developed to deal with the shortcomings of the train/test split process is cross-validation (CV). This is a resampling approach, which aims to provide an unbiased estimate of model performance. This is relevant to the issue of small datasets, in which resampling can reduce the issue of single split bias on model performance (Maleki et al., 2020). This process also uses a train-test split; however, it is used multiple times. Instead of developing one model, several are built and validated using different sections of the data. Prediction accuracy is then aggregated across all models.

A commonly used approach is K-fold CV. In this approach, the dataset is firstly split into K equal portions. A commonly used number for K is 10 (Maleki et al., 2020). Subsequently, 10 rounds of CV are processed, with one section used as the validation set and K-1 sections used as training. This process is repeated K times, with a different portion of the data used as the test set on each occasion. Following K rounds of training, prediction performance is averaged across all rounds of CV. The advantage of CV over train-test split is the use of all data for validation. This reduces the effect of random chance in prediction performance, as variances in test sets are averaged across multiple splits (Maleki et al., 2020). However, one drawback of K-fold CV is the increased computational resources required in comparison with hold out validation. This is due to the requirement of building K number of models (Maleki et al., 2020). The process of K-fold CV is detailed in Figure 2.2 below:

Figure 2.2: The process of K-fold CV for algorithm development (Cristianini and Shawe-Taylor, 2000).

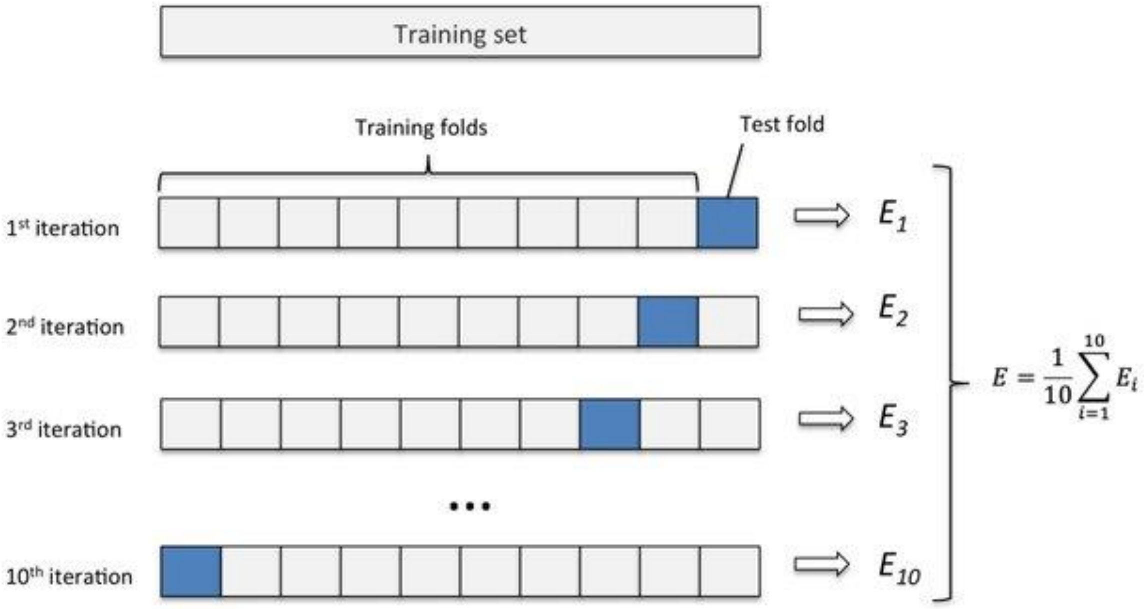


Figure 2.2: The dataset has been split into ten folds (denoted by boxes), with the training and test sets altering in each iteration. Model error denoted by E is averaged across all runs of training and validation. This image has not been altered from the original source.

There are other types of CV used in ML development. A slight variation of the K-fold CV is stratified K-fold CV. In this version, each fold is stratified to contain approximately the same proportion of class labels as the original dataset. This is important addition due to the common issue of class imbalance in ML, in which there is a substantial difference between the number of samples in the classes. If such a scenario occurs when using ordinary K-fold CV, the presence of very few samples in any fold might lead to large differences in prediction performance (Maleki et al., 2020). Despite reducing the random variation in test data when compared to the holdout method, K-fold CV uses only K-1 partitions for learning. A process which uses further information for training and validation is leave one out cross-validation (LOOCV). This uses a single sample as the validation set for each round of CV. In this instance, the remaining samples are used for training. Therefore, the number of CV rounds is equal to the number of samples in the dataset. However, this method can cause significant computational cost due to using all samples as validation sets (Berrar, 2019).

Most ML algorithms rely on several hyperparameters to achieve optimisation. These are parameters which are derived outside of the initial training phase. Optimum values for these parameters are rarely known prior to model building, therefore they are estimated experimentally (Kassraian-Fard *et al.*, 2016). As detailed in Section 2.4, research has shown that if these values are estimated in the same validation set as model testing, model accuracy can be artificially inflated (Kassraian-Fard *et al.*, 2016). Therefore, hyperparameter optimisation is often carried out in separate test set. This has led to the development of nested CV. In this process, two rounds of CV are used during development. An outer loop of CV is used for model validation, whilst an inner loop optimises model hyperparameters. The process of nested CV is shown in Figure 2.3.

Figure 2.3: The process of nested CV for algorithm development (Zhong, Chalise and He, 2020).

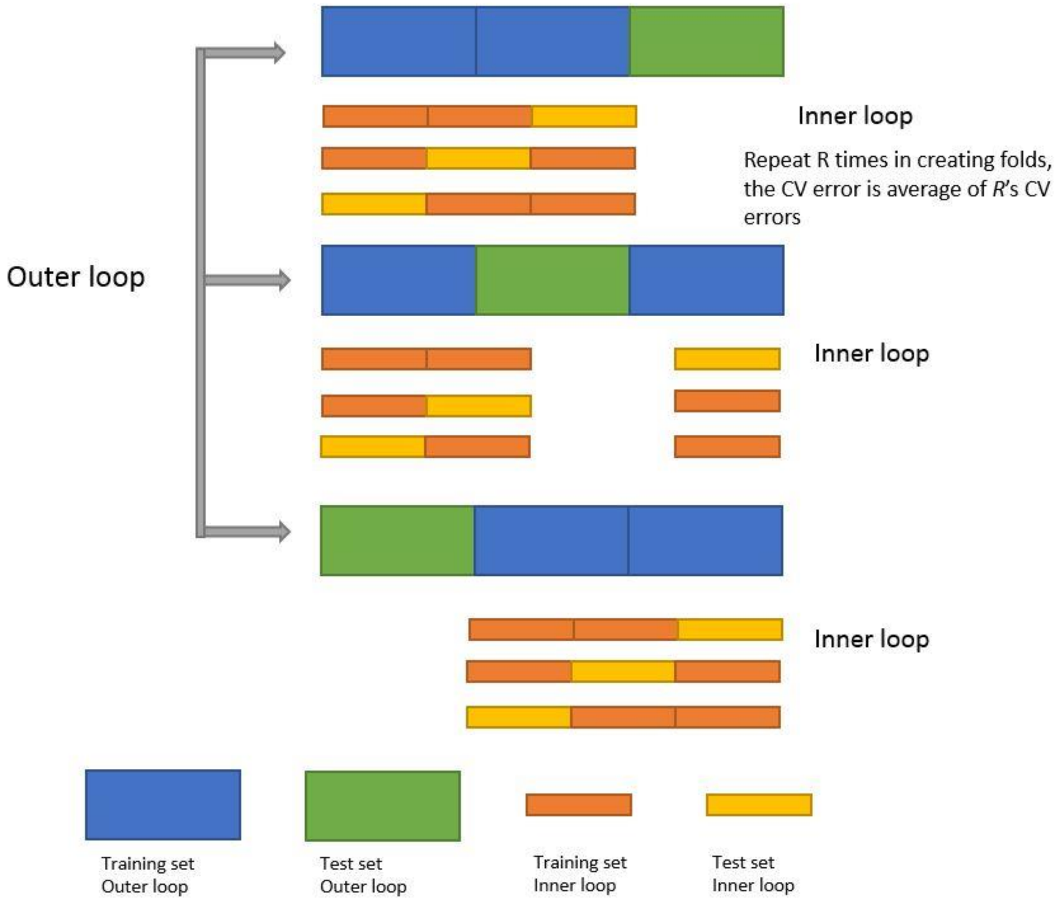
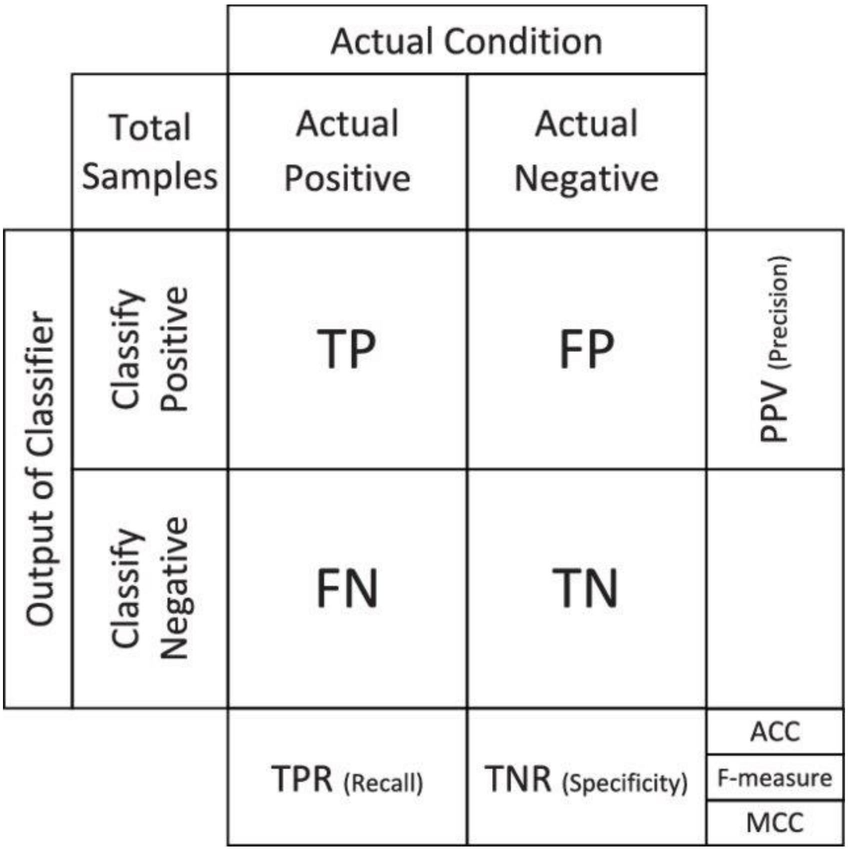


Figure 2.3: The dataset is split into an outer loop of CV which is used to assess model performance, whilst an inner split of data is used for the training of hyperparameters. This image has not been altered from the original source.

2.4.2 Evaluation metrics

Following training and validation, algorithm performance can be evaluated using a range of metrics. Most metrics can be based upon the confusion matrix for binary classification. During classification, true positives (TP) and true negatives (TN) denote the number of positive and negative instances correctly classified. Conversely, false positives (FP) and false negatives (FN) represent the number of misclassified positive and negative instances (M and M.N, 2015). This matrix is shown in Figure 2.4.

Figure 2.4: A confusion matrix for a binary classification problem (Xia et al., 2015).

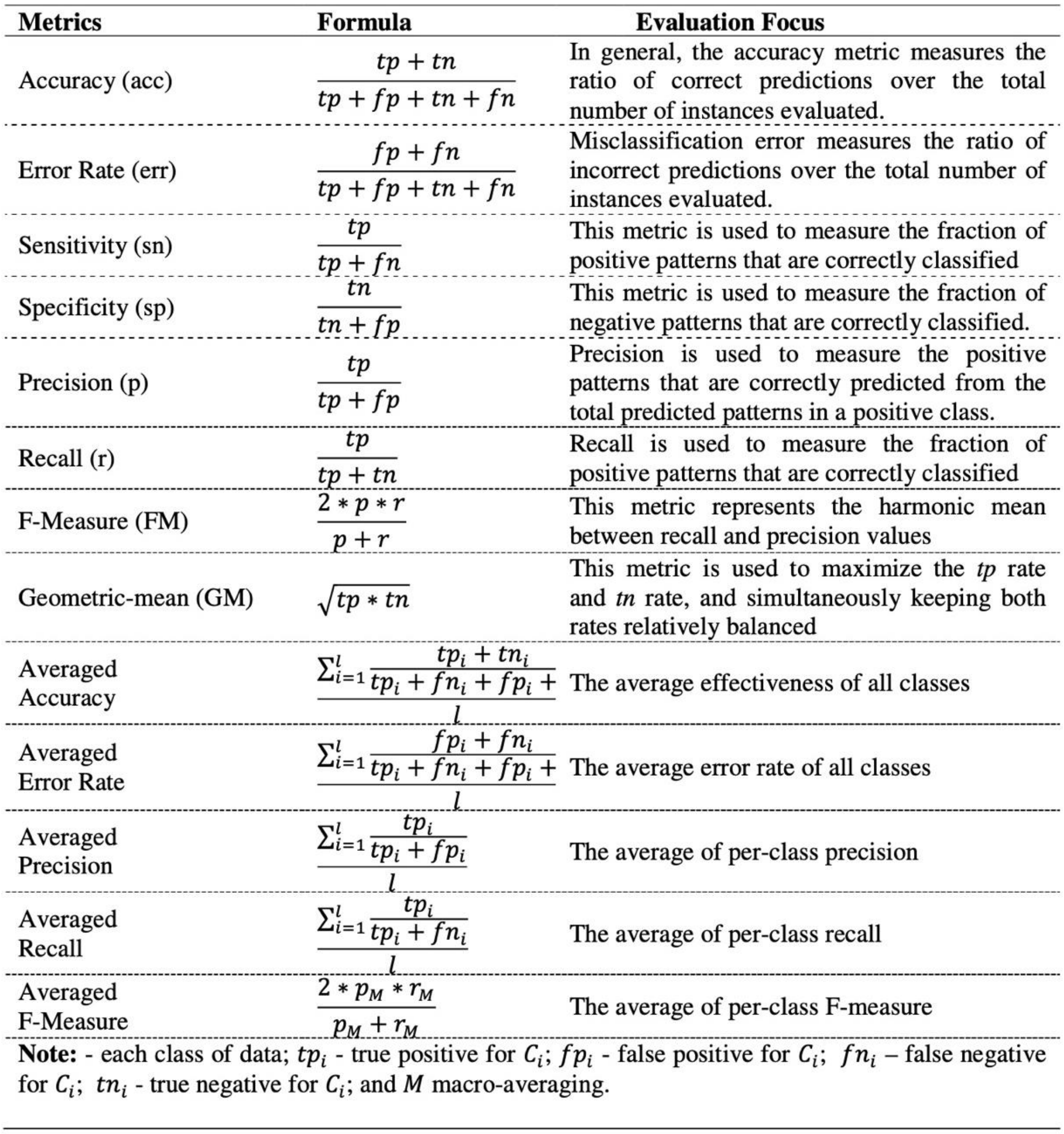


		Actual Condition		
	Total Samples	Actual Positive	Actual Negative	
Output of Classifier	Classify Positive	TP	FP	PPV (Precision)
	Classify Negative	FN	TN	
		TPR (Recall)	TNR (Specificity)	ACC F-measure MCC

Figure 2.4: Confusion matrix for a standard binary classification problem. The output of the classifier is compared to actual class memberships. Predictions are separated into true positives, false positives, false negatives, and true negatives. Metrics such as the true positive rate, true negative rate, precision and f-measure are also shown. This figure has not been altered from the original source.

Metrics of accuracy which use elements of the confusion matrix are detailed further in Table 2.1

Table 2.1: Classification metrics for prediction problems (M and M.N, 2015).



Metrics	Formula	Evaluation Focus
Accuracy (acc)	$\frac{tp + tn}{tp + fp + tn + fn}$	In general, the accuracy metric measures the ratio of correct predictions over the total number of instances evaluated.
Error Rate (err)	$\frac{fp + fn}{tp + fp + tn + fn}$	Misclassification error measures the ratio of incorrect predictions over the total number of instances evaluated.
Sensitivity (sn)	$\frac{tp}{tp + fn}$	This metric is used to measure the fraction of positive patterns that are correctly classified
Specificity (sp)	$\frac{tn}{tn + fp}$	This metric is used to measure the fraction of negative patterns that are correctly classified.
Precision (p)	$\frac{tp}{tp + fp}$	Precision is used to measure the positive patterns that are correctly predicted from the total predicted patterns in a positive class.
Recall (r)	$\frac{tp}{tp + tn}$	Recall is used to measure the fraction of positive patterns that are correctly classified
F-Measure (FM)	$\frac{2 * p * r}{p + r}$	This metric represents the harmonic mean between recall and precision values
Geometric-mean (GM)	$\sqrt{tp * tn}$	This metric is used to maximize the *tp* rate and *tn* rate, and simultaneously keeping both rates relatively balanced
Averaged Accuracy	$\frac{\sum_{i=1}^{l} \frac{tp_i + tn_i}{tp_i + fn_i + fp_i +}}{l}$	The average effectiveness of all classes
Averaged Error Rate	$\frac{\sum_{i=1}^{l} \frac{fp_i + fn_i}{tp_i + fn_i + fp_i +}}{l}$	The average error rate of all classes
Averaged Precision	$\frac{\sum_{i=1}^{l} \frac{tp_i}{tp_i + fp_i}}{l}$	The average of per-class precision
Averaged Recall	$\frac{\sum_{i=1}^{l} \frac{tp_i}{tp_i + fn_i}}{l}$	The average of per-class recall
Averaged F-Measure	$\frac{2 * p_M * r_M}{p_M + r_M}$	The average of per-class F-measure

Note: - each class of data; tp_i - true positive for C_i; fp_i - false positive for C_i; fn_i – false negative for C_i; tn_i - true negative for C_i; and M macro-averaging.

Table 2.1: Metrics, formulas and descriptions of accuracy metrics formed from the confusion matrix. This has not been altered from the original source.

The simplest and perhaps the most widely used metric is accuracy. This is calculated as the number of correct predictions, divided by the total amount of predictions made. Predictions can be defined as correct by comparing model predictions to actual observations. Despite its simplicity and popularity, research has recommended that accuracy should not be used in ML evaluation. This is due to the scenario known as the 'accuracy paradox’ (Uddin, 2019). This relates to the effect that imbalances in datasets have on the values of accuracy. A dataset is imbalanced if one of the present classes is dominant over the other. This relationship skews the calculated accuracy (Francisco J Valverde-Albacete and Peláez-Moreno, 2014).

The drawback of the accuracy method has led to the development of more robust methods. An example of these is receiver operating characteristics (ROC), which in turn leads to the area under the curve (AUC) metric. AUC can be defined as the probability of a classifier ranking a positive sample above a negative example. This metric can be calculated from the ROC curve. An example of AUC is given in Figure 2.5:

Figure 2.5: The ROC-AUC curve, in which the TPR and FPR are plotted against each other over different thresholds (Majnik and Bosnić, 2013).

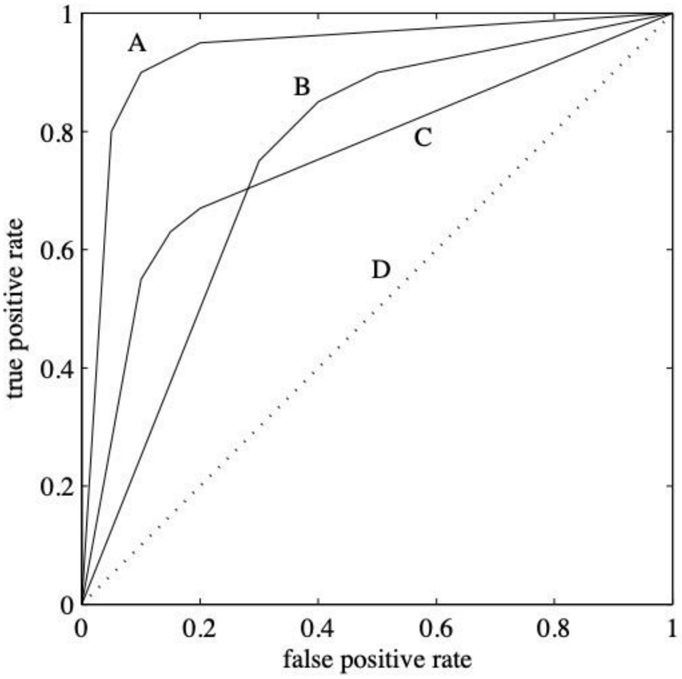


Figure 2.5: ROC curves for four classifiers are plotted. Classifier A shows the best performance, with classifier D displaying performance no better than chance (AUC < 0.5). This figure has not been altered from the original.

The ROC curve is calculated by plotting the false positive rate versus (FPR) the true positive rate (TPR). The TPR can be defined as the probability of an actual positive being classified as positive, calculated by dividing the number of TP by the sum of TP and FN. The FPR is the likelihood of a negative instance being classified as negative. This is calculated by dividing the number of FP by the sum of FP and TN. ML algorithms will often output a prediction probability for each sample belonging to a class. The predicted labels of each sample can therefore be determined by choosing an arbitrary threshold. Usually, a value of 0.5 is used, with all predicted probabilities below this value labelled as negative, whilst all those above delegated as positive (Hajian-Tilaki, 2013). These predicted labels are then compared to real labels to calculate both the true positive and false positive rates. The ROC curve is plotted by calculating these two values at a range of probability thresholds for FPR and TPR, from 0 to 1. AUC is the area between the X-axis and the ROC curve. The diagonal signifies a baseline classifier of 0.5 AUC (Kumar and Indrayan, 2011). For a model to be deemed successful, AUC must be ≥ 0.5, as a classifier below this threshold has not performed better than random chance. However, the general standard for an algorithm to be an effective diagnostic tool is ≥ 0.8 (Nahm, 2022).

2.4.3 Calibration

Predictions from ML models are used in a wide variety of disciplines. Some of these involve high risk decisions, such as diagnosing conditions in healthcare. Prediction errors in such a high-risk area could cause harm to individuals. Therefore, assurances in model performance must be met prior to usage (Van Calster et al., 2019). Calibration can be defined as a measure of the degree to which predicted probabilities for each class match actual observations. This can be put in more simple terms as the 'confidence' of a ML algorithms predictions (Nixon *et al.*, 2019). Despite its importance due to high-risk decision making, calibration has received little attention in prediction studies. Greater emphasis has often been placed on reporting discrimination statistics, such as AUC. However, it is possible for an algorithm to achieve high AUC, but still be poorly calibrated (Van Calster et al., 2019).

There are several reasons why an ML algorithm could be poorly calibrated. For example, the variation in variables not related to model development. Quantities such as disease incidence can vary significantly between cohorts. Larger hospitals maybe more likely to treat patients with a certain disease than localised units. Models trained on areas with higher incidence are more likely to overestimate the likelihood of disease when used in a less likely setting (Van Calster et al., 2019).

To address the issues of poor calibration, probabilities can be adjusted following the modelling process. A common technique is to process the output of a classifier without retraining the algorithm. The two most used methods are Platt scaling and isotonic regression (Guo *et al.*, 2017). Platt scaling, also known as sigmoid scaling, is often used when classifier probabilities follow a sigmoidal relationship:

Figure 2.6: An example of sigmoidal relationship between two variables (Tsikliras and Froese, 2019).

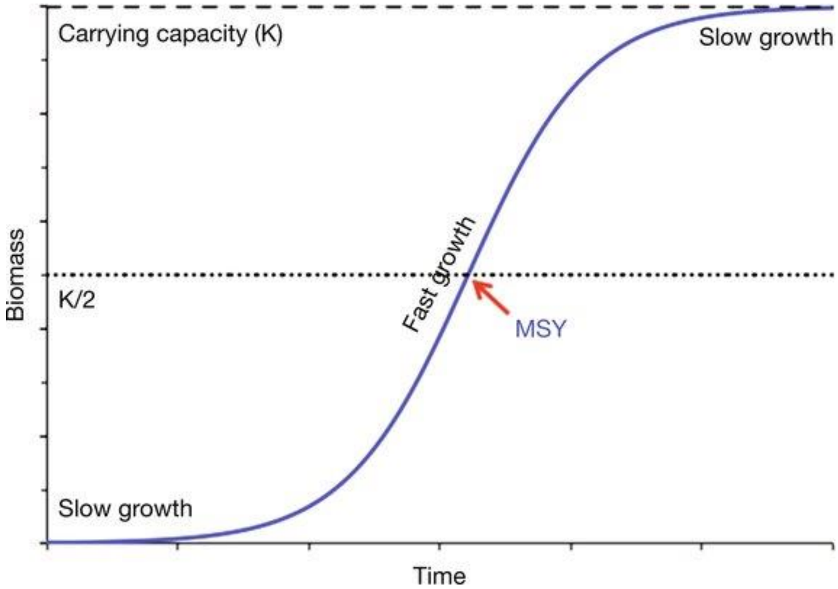


Figure 2.6: This image provides an example of a sigmoidal relationship. The logistic (sigmoid) curve of population growth for a species of fish, with maximum sustainable yield (MSY) the level of fishing required to maintain the population. This figure has not been altered from the original source.

For a real-valued function f and probability P, the platt scaling process can be defined as:

$$P(y = 1|f) = \frac{1}{1+exp(Af+B)} \qquad (2.6)$$

The parameters A and B are learned using maximum likelihood estimation from a training set. The fitting of this function is achieved using gradient descent:

$$argmin_{A,B}\{\sum_i y_i \log(p_i) + (1 - y_i)\log(1 - p_i)\}, where\ p_i = \frac{1}{1+\exp(Af+B)} \qquad (2.7)$$

To avoid the possibility of overfitting, these parameters should be learned on a separate dataset to the one used for model fitting. This can be achieved by using CV (Fonseca and Lopes, 2017).

Isotonic regression is a non-parametric form of regression. Predictions from a classifier f_i and real targets y_i are fitted to the following regression:

$$y_i = m(f_i) + \epsilon_i \qquad (2.8)$$

The parameter m is a non-increasing function. In comparison with platt-scaling, isotonic regression is more prone to overfitting (Fonseca and Lopes, 2017). For calibration purposes in thesis, both platt scaling and isotonic regression were used, with the method which realigned predicted probabilities to observations most effectively chosen.

2.4.4 Tuning hyperparameters

In ML, hyperparameters can be defined as parameters which have a controlling effect on the training process. Most ML models comprise hyperparameters, however, these will often differ between algorithms. Software packages in programming languages such as *Python* provide default values (these will be implemented if user values are not specified). However, users may wish to tune hyperparameters in order to determine optimum values for the given dataset. These optimal parameters can be defined as the set of values which minimise the generalisation error for the chosen ML model (Probst, Bischl and Boulesteix, 2018).

Several methods to tune hyperparameters are generally used in ML development. A manual search involves a user selecting values for hyperparameters. Generally, this requires a good knowledge of both the ML algorithm and hyperparameters in question. This can in some circumstances prove to be a quick solution, but in general even experienced users will struggle to select appropriate values (Yu and Zhu, 2020). A more *methodical* approach is to use the grid search procedure. All possible combinations from a range of pre-specified values are tested, with the set of values which minimise model error chosen. Advantages of this process lie in its automation and ease of implementation. However, the extensive nature of testing all possible combinations can lead to significant resource usage and time. A background knowledge of hyperparameters is still also required (Yu and Zhu, 2020).

A more efficient way to tune hyperparameters is by the random search method. This process draws random combinations of parameters from a range of pre-defined values. The search process continues until either the desired accuracy is achieved, or a predetermined time or memory is reached. This selective nature reduces the computational burden experienced during grid-search, as not all possible combinations are explored (Shekhar, Bansode and Salim, 2022). This method has proven to out-perform grid search in a number of studies (Elgeldawi *et al.*, 2021).

A more complex approach to hyperparameter tuning is Bayesian optimisation. This can be termed as an informed algorithm, meaning that each iteration uses information from previous rounds to make decisions. Similarly, to the random search algorithm, Bayesian optimisation samples combinations of hyperparameters from a predefined space. The process uses a 'surrogate' model to achieve an optimal set of values. The most common type of surrogate model used is the Gaussian Process (GP). Initially, sets of hyperparameter values are chosen randomly. These combinations are tested on the chosen ML model, with accuracy used as the metric for choosing best performing values. A decision is then reached on whether to continue to search for a superior combination in this region of parameter values. If not, the algorithm draws a new set of random combinations from a different range of values. This method does not conduct an exhaustive search of all possible combinations and does use prior information to select values. Therefore, it is a more viable option than both random and grid search (Elgeldawi *et al.*, 2021). However, the random search method was used to derive hyperparameters for algorithms in this thesis, due to its proven ability to outperform the grid search method and reduced complexity of implementation when compared to the Bayesian approach.

2.5 Common ML modelling challenges

2.5.1 Overfitting

In terms of ML, generalisability refers to the ability of a ML algorithm to predict unseen data. A poorly generalised model may lead to a poor performance, which could have adverse effects in various domains (Ying, 2019). There are two general terms used when assessing generalisability, these are 'underfitting' and 'overfitting'. Underfitting occurs when a ML algorithm fails to learn most of the nuances within a training set. This leads to an inability to make inferences on the validation data. Overfitting, results from an algorithm becoming too reliant on training data. The algorithm's ability to deal with slight differences in validation data is therefore weakened, resulting in poorer generalisation (Salman and Liu, 2019).

The relationship between underfitting and overfitting can be defined further in terms of both variance and bias. This is shown in Figure 2.7.

Figure 2.7: The Bias-Variance trade-off in ML (Neal et al., 2018).

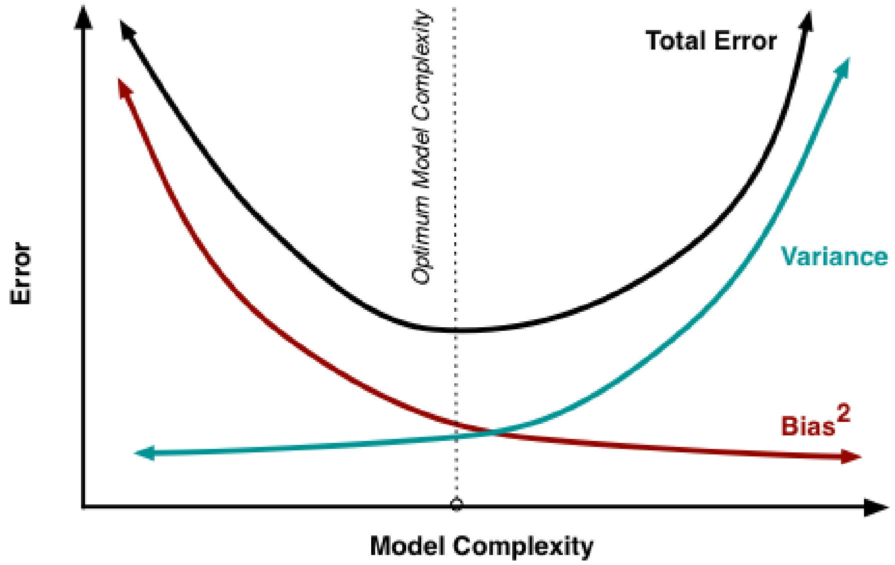


Figure 2.7: The relationship between bias and variance in ML, with the demonstration of increased model complexity and variance. In circumstances where model bias is high the model has higher likelihood over underfitting, however increased model variance raises the likelihood of overfitting. This image has not been modified from the original source.

Model bias is defined as the difference between the predicted value and expected value of a single observation. The variance of prediction errors in a model is the change in prediction performance when independent datasets are used for training. For example, a model with high variance will produce a wide range of accuracies when given alternative datasets (Ghojogh and Crowley, 2019). In ideal circumstances, an ML algorithm should minimise both factors to create a stable and accurate model. However, a trade-off exists between the two elements; if one factor is altered, it has a direct impact on the other. For instance, increasing the bias of a model will decrease the level of variance, whilst increasing the variance will decrease the bias (Ghojogh and Crowley, 2019).

When considering the bias-variance trade off, underfitting is caused by low variance and high bias, also known as over-generalisation. In contrast, overfitting occurs due to low bias and high variance (Ghojogh and Crowley, 2019). The reasons for high variance and in turn overfitting can be broadly categorised into two areas. One is the presence of noise in a training set, which is not representative of the underlying target relationship. This is most likely when the training dataset contains too fewer observations, however this can also occur in larger datasets. A well-functioning algorithm should be able to distinguish between this unwanted noise and desired data (Ying, 2019). A further cause of overfitting can be the abundance of features within a large dataset. The presence of many variables leads to a range of possible hypotheses during analysis, which in turn can lead to high performance, but poor consistency across multiple datasets. Figure 2.7 also demonstrates that the complexity of a model is related to the possibility of overfitting. A more complex model contains a greater number of parameters, which increases the likelihood of an algorithm becoming too reliant on the training data (Ghojogh and Crowley, 2019).

The curse of dimensionality is a further obstacle in the realm of ML and is linked to overfitting. Despite advances in computing technologies, modern computers still encounter obstacles when analysing large datasets (Fan, Han and Liu, 2014). Such datasets can be termed as 'high-dimensional', meaning that each sample has a high number of features. This is relevant in the field of genome wide association studies (GWAS), as each case/control can have many thousands of SNPs (Marttinen *et al.*, 2013). This increase in dimensionality has several effects on ML development. Datasets richer in features require greater resources in terms of compute power and memory. This in-turn leads to longer time required for algorithm training (Debie and Shafi, 2019).

A statistic related to the curse of dimensionality is events per variable (EPV). The number of events is defined as the number of instances for the minority class in a binary variable. EPV is calculated by dividing the number of events by the total number of predictors in the model. The acceptable threshold for EPV in ML has been a topic of debate amongst researchers. A minimum threshold of 10 is generally accepted for modelling, however, if possible, a presence of 20 observations is preferred (Austin and Steyerberg, 2017). The reason that EPV is an important aspect of ML development is the risk of overfitting if its value falls below the recommended value. EPV's below 10 increase the likelihood of random noise in the dataset and reduces the likelihood of an algorithm generalising well to unseen data. However, it has been suggested that the threshold of 10 is more appropriate to less complex algorithms such as regression techniques. Algorithms such as deep learning and other more complex ML algorithms may require values of 100 or greater (Austin and Steyerberg, 2017).

2.5.1.1 Methods to overcome overfitting.

Several different methods have been developed to reduce the possibility of overfitting; these are outlined below.

2.5.1.1.1 Ensemble learning

Ensemble learning is a paradigm in which multiple ML models are used in parallel for prediction. Ensemble learning is based upon the theory of the 'weak learner'. A weak learner is defined as an ML algorithm which achieves prediction accuracy of just higher than chance (AUC greater than 0.5) (Vaghela, Ganatra and Thakkar, 2009). An example of a weak learner is a single decision tree, whose performance is susceptible to variances within training data, leading to overfitting. The combined use of multiple algorithms has been shown to reduce this effect of variance, converting multiple weak learners to a 'strong learner' (Vaghela, Ganatra and Thakkar, 2009). Two aspects which impact the performance of an ensemble method are accuracy and diversity. The accuracy of a model is related to its error rate, a model becomes more accurate as this rate is minimised. Two classifiers are said to be diverse if they produce different errors on unseen data. Research has established that the performance of ensemble methods improves as models become more diverse (Fawagreh, Gaber and Elyan, 2014).

There are three main methods for training algorithms in ensemble learning. These are bagging, boosting and stacking. Bootstrap aggregation (bagging) functions by choosing random samples of a dataset with replacement. Every algorithm within this ensemble method is provided with one of these samples, with each learner used to make predictions on unseen data. The method arrives at a final decision via a voting system. Perhaps the most common of these is known as 'majority voting', in which each learner within an ensemble framework is asked to predict one observation. The decisions are then summed, with the option receiving the most votes chosen as the overall decision (Fawagreh, Gaber and Elyan, 2014).

For boosting, a model is fitted to a random sample of samples. Further models are then fitted sequentially with the aim of improving upon the previous algorithm. This process continues until model accuracy can improve no further. Stacking is a less commonly used method for ensemble learning. This involves combining the predictions of multiple machine learning models, these are then provided to a separate learner known as a ‘meta-learner’. Boosting has been shown to outperform bagging in data when low levels of noise are present in training data, however this relationship reverses as noise increases (Fawagreh, Gaber and Elyan, 2014).

2.5.1.1.2 Regularisation methods

An area of research developed to improve the generalisability of ML models is regularisation. This is achieved by applying constraints to the minimised loss function. These constraints are in the form of ‘penalties’, in which the complexity of a model is reduced (Tibshirani, 1996). Techniques outlined in this section focus on regularisation methods employed on linear estimators. When considering a regression, with a set of explanatory and response variables, coefficients are learned through minimising the residual squared error, known as the ordinary least squares (OLS). There are several reasons why this process can often lead to inadequate results, such as overfitting and a reduction in interpretability (Tibshirani, 1996).

A common approach to improving OLS estimates is the least absolute shrinkage and selection operator (LASSO). In which L1 regularisation is used to shrink coefficients towards zero. When defining LASSO, there is a set of variables (Xi, yi), in which Xi and yi are predictor and responses variables, respectively, alongside of set of feature coefficients termed betas β. When the assumptions of linear regression are met, such as independence of observations, the minimisation problem becomes:

$$\beta = (\beta_1, \beta_2, \dots, \beta_p) \; are \; estimated \; by \qquad (2.9)$$

$$(\hat{\beta}, \hat{\beta}^{lasso}) = argmin\left\{\sum_{i=1}^{N}\left(Y_i - (\beta_0 + \beta X_i^T)\right)^2 + \lambda \sum_{j=1}^{P} |B_j|\right\} \qquad (2.10)$$

The λ parameter is a non-negative tuning parameter used to control the strength of the L1 penalty. If this parameter is set to zero, none of the model’s parameters (betas) are reduced to zero. In this instance, the resulting model is an ordinary regression model (Musoro *et al.*, 2014). As the value of λ increases, greater numbers of betas are set to zero. The greater the value of λ, the more variance is reduced. However, this leads to an increase in bias of model predictions due to the bias-variance trade-off. LASSO can be used either as a stand-alone prediction model, or a form of feature selection prior to further prediction. Non-penalised features are passed onto an additional algorithm algorithms or what algorithm (Musoro *et al.*, 2014).

However, despite the ability to reduce a feature space, the LASSO algorithm has several disadvantages. If the number of features (p) is larger than the number of samples (n), i.e., p>n), LASSO can only select at most n features before saturating. The presence of pairwise correlations within a dataset can also introduce disadvantages, as LASSO tends to only select one of the correlated variables. This can lead to the rejection of features which may have predictive value (Zou and Hastie, 2005).

Another commonly used technique for penalised regression is ridge regression. This is a tuning method, which aims to address multicollinearity, in which features in a dataset are correlated in such a manner that one can be predicted from another. The presence of such a characteristic can increase the likelihood of overfitting (Shariff and Ferdaos, 2017). Similarly, to LASSO, ridge regression uses a penalty term, known as L2. This penalty term is added the loss function and is a squared value of the beta coefficient for the variable. For the following regression model:

$$Y = Xb + \mathrm{e} \tag{2.11}$$

where, as above, X and Y are predictor and responses variables, respectively. The values denoted by b are coefficients for the explanatory variables, with e representing an error term.

$$\hat{b} = (X^t X)^{-1} X^T Y \tag{2.12}$$

To address the issue of multicollinearity, a constant k is added to the XtX This reduces the dependency in explanatory variables and results in the following:

$$\widehat{b_R} = (X^t X + K I_n)^{-1} X^T Y \tag{2.13}$$

The value of k is an important part in how ridge regression performs. If the value of k is zero, the loss function represents a linear regression. However, if this value becomes too large, the model will most likely underfit the data. Therefore, optimising the k parameter is advisable. Despite the main advantage of ridge regression in reducing model complexity and the likelihood of overfitting, the method has several disadvantages (Shariff and Ferdaos, 2017). For example, for the purposes of feature selection, ride regression only penalises the coefficients of inputs, it does not remove their presence. This contrasts with LASSO, which reduces certain effect sizes to zero. Given this, ride regression cannot be used if the aim is to reduce the feature space.

A technique developed to overcome the disadvantages of both ridge regression and LASSO is the elastic net (Zou and Hastie, 2005). We have a set of features X for n observations, with a corresponding group of response variables y. For any two positive values for λ1 and λ2, the elastic net criterion can be defined as:

$$L(\lambda_1, \lambda_2, \beta) = |y - X\beta|^2 + \lambda_2|\beta|^2 + \lambda_1|\beta|_1 \tag{2.14}$$

$$|\beta|^2 = \sum_{j=1}^{p} B_j^2, \tag{2.15}$$

$$|\beta|_1 = \sum_{j=1}^{p} |\beta_j|. \tag{2.16}$$

Where the estimator of the elastic net problem can be realised by minimising:

$$\hat{\beta} = argmin_\beta\{L(\lambda_1, \lambda_2, \beta)\}. \tag{2.17}$$

By letting α be defined as:

$$\alpha = \frac{\lambda_2}{\lambda_1 + \lambda_2} \tag{2.18}$$

The minimisation function becomes:

$$\hat{\beta} = argmin_\beta|y - X\beta|^2, \tag{2.19}$$

$$subject\ to\ (1 - \alpha)|\beta|_1 + \alpha|\beta|^2 \leq t\ for\ some\ t \tag{2.20}$$

The elastic net penalty is a convex combination of both the LASSO and ridge penalties. For the scenario where α = 1, the elastic regression behaves as a ridge regression. Conversely, a penalty of zero causes the estimator to act as LASSO. Similarly, to the LASSO algorithm, the elastic net method achieves feature space reduction. However, correlated variables can also be selected in groups, avoiding the issue of random selection within LASSO (Shariff and Ferdaos, 2017).

2.5.1.1.3 Cross-validation as an approach to overcome overfitting.

Overfitting occurs when a model becomes too reliant on training data. As CV separates the original data into different training sets, training multiple ML algorithms on different sections of the data can average out the variance and random noise within the dataset across all models. Therefore, the combination of all ML algorithms reduces the likelihood of overfitting (Brodeur, Herman and Steinschneider, 2020).

2.5.1.1.4 Dimensionality reduction methods

The curse of dimensionality is a further obstacle in the realm of ML and is linked to overfitting. Despite advances in computing technologies, modern computers still encounter obstacles when analysing large datasets (Fan, Han and Liu, 2014). Such datasets can be termed as 'high-dimensional', meaning that each sample has a high number of features. This is relevant in the field of GWAS, as each case/control can have many thousands of SNPs (Marttinen *et al.*, 2013). This increase in dimensionality has several effects on ML development. Datasets richer in features require greater resources in terms of compute power and memory. This in-turn leads to longer time required for algorithm training (Debie and Shafi, 2019).

The increasing burden of dimensionality issues has led to the development of techniques to reduce their impact. These come under the umbrella term of dimensionality reduction, whereby the number of features is reduced to a size where the effects of high dimensionality are suppressed (Xie, Li and Xue, 2017). This reduction can be achieved in two ways: feature selection and feature extraction. During the process of feature selection, features deemed redundant by statistical tests are removed (Velliangiri, Alagumuthukrishnan and Thankumar joseph, 2019). This results in a reduced set of features which are likely to comprise variables most appropriate for predicting the target class. On the other hand, feature extraction involves extracting important information from the set of original features. This information is then used to create a new set of variables with fewer dimensions (Velliangiri, Alagumuthukrishnan and Thankumar joseph, 2019).

2.5.1.1.5 Feature selection

The advantages of feature selection include removing detrimental features and reducing the possibility of overfitting (Hira and Gillies, 2015). Features can be defined as detrimental if they have the potential to reduce the performance of an algorithm. Such features could include outliers, large amounts of missing data or may not be statistically useful. The inclusion of such features will increase the level of noise in a dataset. This could reduce the algorithm's ability to learn the desired patterns within the dataset (Ying, 2019) and to perform well on previously unseen data (Salman and Liu, 2019).

Methods of feature selection can be broadly split into two categories: supervised and unsupervised techniques. Supervised feature selection is used on datasets with labelled data, whilst unsupervised methods focus on data without labels. Supervised methods can be further subdivided into three classes: filter methods, wrapper methods and embedded feature selection. A filter method is where univariate models are built between each feature and the target variable. Features are then either retained or deemed redundant depending on the acceptance criteria used. Acceptance criteria could be derived from for example, information gain, the chi-squared statistic and correlation coefficient. This method is computationally efficient and therefore useful in situations with high dimensionality. However, correlations between different features are not taken into consideration (Sánchez-Maroño, Alonso-Betanzos and Tombilla-Sanromán, 2012).

Wrapper methods involve selecting a subset of the original feature set which produces the best prediction accuracy. This can be achieved in two ways: either forward or backwards feature selection. Forward feature selection starts with an empty set of features, predictors are then added sequentially. These new predictors are then assessed, with emphasis on whether the new variable hinders or improves ML performance. The feature is retained if its effect is positive and rejected otherwise (Talavera, 2005). This recursive process continues until a feature set of a pre-defined size is reached. Backwards feature elimination functions in a similar manner, however the algorithm begins with all possible features and recursively removes one at a time. The effect of removing a feature is then tested to assess the impact on

ML performance. Depending on whether the feature improved or reduced performance, the feature is either kept or removed. Wrapper methods usually produce more predictive feature sets than filter methods, however the process is more computationally intensive and so is not suitable in high dimensional datasets (Talavera, 2005).

Embedded feature selection is a paradigm in which the feature selection stage is built into the modelling process. An example of this is the least absolute shrinkage and selection operator (LASSO). A method used to reduce training errors known as regularisation is used to penalise the regression coefficients of features that are deemed to be contributing to overfitting. This penalisation reduces regression coefficients of the targeted features to zero, removing their role in modelling. Embedded methods combine the advantages of both filter and wrapper methods, due to assessing the correlations between features, whilst also promoting computational efficiency (Jovic, Brkic and Bogunovic, 2015)

2.5.1.1.6 Feature extraction

Feature extraction differs from feature selection as it is an unsupervised technique. Information regarded as important is extracted from the original feature set and is used to create a reduced feature space. One of the most common examples of this is principal component analysis (PCA) (Karamizadeh *et al.*, 2013). In this process, orthogonal vectors named principal components (PCs) are computed. These are linear combinations of the original features, which capture the variance within the dataset (Jolliffe and Cadima, 2016). The number of PCs created is predefined by the user. The advantages of this method are that it can retain important information and reduce dimensionality. This lowers the possibility of overfitting due to the reduction of both dimensionality and unwanted noise in the feature set. Additionally, smaller sets of inputs reduce both the training times and computational burden for algorithms (Karamizadeh *et al.*, 2013). However, some information is always lost in the process of creating PCs. This loss of information can lead to a reduced performance of ML algorithms if an appropriate number of principal components for the data in question is not chosen (Karamizadeh *et al.*, 2013).

2.6 Missing data

Most pre-processing techniques and ML models will be affected by missing data. One common method to deal with this issue is to remove features with missing values (Kang, 2013). However, if missing values are spread along multiple features, samples with missing data may be removed instead. Despite having the option to choose between methods, both have the drawback of removing information from the dataset. This may hinder the performance of a ML algorithm, therefore other methods which do not remove data have been developed (Kang, 2013).

Data imputation is the process where missing values are replaced with predicted values. One method of data imputation is the replacement of missing values with the expected value of that variable. This can be achieved by using a measure of central tendency (mean, median, mode). The advantage of using averages for imputation is that it is easy to implement, prevents data loss and does not change the distribution of the variable (Kang, 2013). However, the method has been shown to have unwanted consequences. These consequences are related to the type of missing data within a dataset, where the missingness can be defined in terms of its random nature. If the missing values are defined as missing completely at random (MCAR), then the probability of a value missing is only dependent on itself. If the missing values are statistically related to the observed features, these missing values are known as missing at random (MAR). If the missing values are dependent on both the observed data and other missing values, these are known as missing not at random (MNAR). The type of randomness is important as it dictates which imputed method will work most effectively (Kang, 2013).

When missing values are defined as MCAR and MAR, the reasons for missing values can be ignored and any method of imputation can be used. However, this does not imply that all methods have equal effectiveness (Kang, 2013). For instance, the use of expected (or central tendency) values can reduce the variability in a dataset. This will lead to a reduction of estimated errors during modelling. Alongside this, when imputing features with central tendency measures, the correlations with other features could be altered. This could hinder modelling performance. A method which has been deemed more reliable than central tendency imputation is the regression imputation (Kang, 2013). In this instance, the missing data variable is defined as the target variable, with other features in the dataset being used as explanatory variables. The regression model is originally trained on observed data points. The regression coefficients from this model are then used to predict missing values and impute them. The advantage of this method over central tendency methods is that correlations between variables are preserved (Kang, 2013).

2.7 Imbalanced classes

A common issue concerning classification tasks is the imbalanced nature of datasets. This occurs when the quantity of samples in one class is greater than the other. For most cases, the negative class outweighs the positive class (Yadav and Bhole, 2020). The presence of these imbalances can be detrimental for ML classification purposes. The majority class tends to bias the algorithm, leading to poor prediction accuracy in the positive class. This could have detrimental implications in areas such as disease classification (Johnson and Khoshgoftaar, 2019).

Several methods have been developed to resolve the issue of class imbalances. Resampling techniques adjust the class ratio prior to prediction. These can be broadly split into two different categories, namely under and over sampling. Random under-sampling reduces the number of majority samples by removing instances at random. This occurs until the number of samples in each class are equal. However, the removal of samples can reduce the amount of predictive information in the dataset (Susan and Kumar, 2021). To avoid this loss, the method of oversampling has been developed. An example of this is the synthetic minority over sampling technique (SMOTE). In which, synthetic samples of the minority class are created using interpolation. At first, random samples of the minority class are chosen. A k-nearest neighbour algorithm is then used to select nearest neighbours for the chosen samples. Synthetic samples are then generated whose values lie between the chosen random samples and nearest neighbours. The aim of the technique is not to alter the variance of the dataset, however issues with random noise and overfitting have been reported (Guo *et al.*, 2008).

Another method designed to overcome the issue of poorly distributed classes is reweighting. This is simpler process than resampling and works by altering the influence the minority class has during classification. One example of this from the programming language *Python* is the *class_weight* function in *scikitlearns RandomForestClassifier*. The '*balanced*' input option calculates the inverse relationship between the two classes, this is achieved by dividing the number of samples in the minority class by the number of majority samples. This relationship is then inversed, and the minority samples are reweighted to balance the relationship before training. This process does not involve removing samples or introduce random noise and can be used across many ML algorithms (Aljohani, Fayoumi and Hassan, 2021).

2.8 Types of learners

A variety of ML algorithms shall be used for analyses in this thesis. Section 2.8 gives an in-depth overview of all of these classifiers.

2.8.1 Decision tree

The popularity of decision tree-based algorithms has increased due to several advantages, such as ease of use, interpretability and robustness when encountering missing values (Song and Lu, 2015). Decision trees are also flexible in nature, as both continuous and discrete variables can be used for either features or class values. Decision trees can be used for feature selection, as an intermediate step during the classification process, in which redundant variables are removed prior to further modelling. They can also be used as stand-alone predictors, for both classification and regression purposes (Song and Lu, 2015).

The structure of a decision tree comprises nodes, branches and splitting lines. Nodes consist of three types, a root (decision) node is a choice which results in the division of samples into two mutually exclusive sets. Internal nodes can be defined as chance nodes, which represent choices at a point in the tree. This node is connected to its parent node, whilst also connected to leaf nodes. Leaf nodes represent the end of the tree, which display the final result of the decision tree. Branches represent the outcomes that result from root and internal nodes. These can be defined as if-then rules, which result in splitting (Song and Lu, 2015)

Figure 2.8: An example of a decision tree, including a root node, internal nodes and decisions (Ma and Qin, 2009).

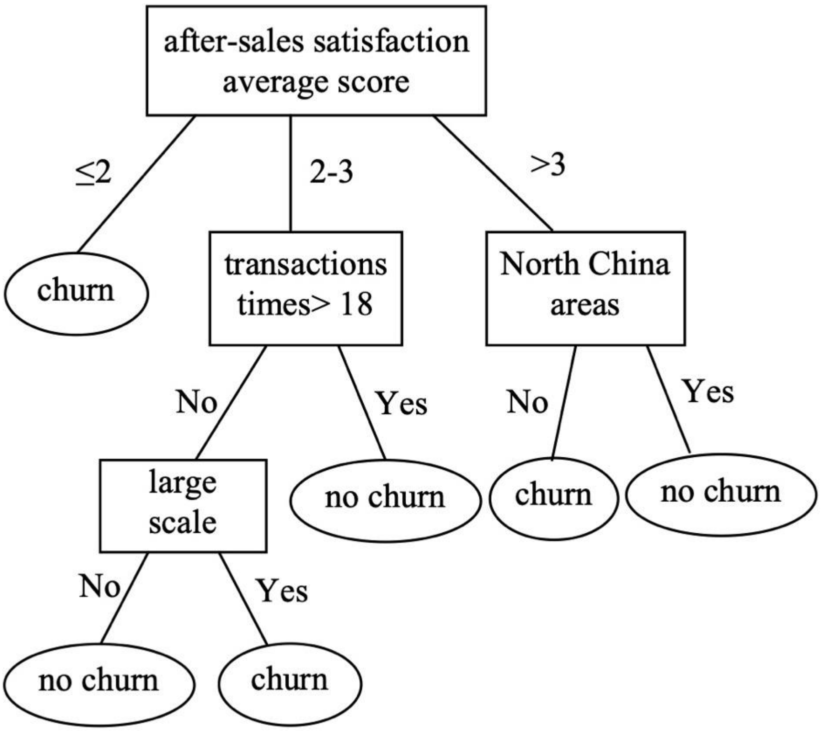


Figure 2.8: A depiction of a decision tree model. Customer churn can be defined as the rate for loss of customers or subscribers for businesses. Exponents typically use decision tree models to root causes for customer decision making. This figure has not been altered from the original source.

2.8.2 Risk of overfitting

Decision trees are prone to overfitting, this is due to the continual splitting nature of the algorithm. As the depth of the tree increases, the model learns to predict training data more accurately (Amro *et al*., 2021). Accuracy will eventually reach 100%, as the algorithm learns to split the dataset perfectly. This leads to a reduction in the model's generalisability, as the decision tree becomes more reliant on training data. Aspects such as random noise are learned, which reduce the algorithm’s ability to predict unseen samples (Amro *et al*., 2021).

2.8.3 Random forests

The Random Forest (RF) is an ensemble which combines the predictive power of decision trees (weak learners) to achieve either classification or regression. The algorithm is trained using the bagging method, as described in Section 2.5.1.1.1. For each tree, a sample of features are used to continually split until a stopping criterion is reached. This sequential splitting is known as the CART (Classification and Regression Trees), in which a greedy algorithm splits on features using a binary method (Sarica, Cerasa and Quattrone, 2017).

Figure 2.9: An example of the RF process (Kirasich, Smith and Sadler, 2018).

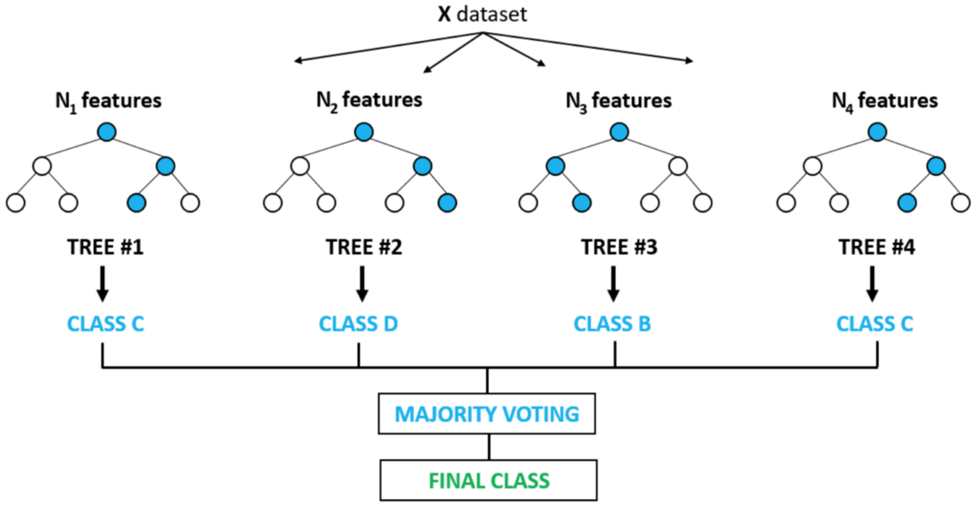


Figure 2.9: An example of an ensemble of decision trees (RF). N sets of features from the dataset X are used, with blue dots representing decision nodes. A prediction is made by each decision tree, with a majority decision made for final prediction. This image has not been altered from the original source.

Single decision trees are defined as high-variance estimators due to variance exhibiting a large impact on predictions. The bagging method of using random sections of data to train multiple trees with replacement reduces the impact of variance on prediction (Altman and Krzywinski, 2017). A final decision for an unseen instance is made using the majority voting technique as seen in Figure 2.9.

2.8.3.1 Splitting criterion

The aim of a decision tree when splitting at a node is to produce homogenous subsets, in which each subset contains samples from one class only. In practice, this is difficult to achieve, as resulting subsets will likely contain a mixture of classes. This notion has been termed the ‘goodness of split criterion’, which is derived from the idea of impurity. A decision tree whose splitting leads to a greater mixture of classes, is deemed to be more impure. Therefore, splitting criteria look to minimise this impurity. Two of the commonly used methods are GINI index and information gain (Tangirala, 2020).

Information gain is defined as a measure of how much information a feature can provide about a class variable. This is an estimation of the effectiveness of a feature for classification. The feature that maximises information gain is considered the best option for splitting (Tangirala, 2020). The GINI index measures the purity of a class following splitting on a particular feature. A split is determined the best if the purity of the resulting classes increases. When comparing the performance for both splitting criteria, studies have identified that models using either criterion perform to a similar degree. Therefore, it is difficult to choose between a method when considering methodologies (Raileanu and Stoffel, 2004).

2.8.3.2 Hyperparameter tuning.

The RF contains a range of hyperparameters which can be tuned to optimise the algorithm (Probst, Wright and Boulesteix, 2018). The depth of a decision tree has an impact on the possibility of overfitting. The deeper a tree becomes, the better it can fit training data. However, this will reduce the performance of the algorithm on unseen data. Therefore, establishing an optimum cut off value for tree depth can achieve acceptable performance on training data whilst producing a generalisable model (Biau, 2010). Packages within the programming language *Python* allow users to specify values for hyperparameters. For instance, the *max_depth* parameter controls the number of splits a decision tree is permitted to make. Another hyperparameter which can influence overfitting is the number of samples at each leaf node. In general terms, low minimum leaf samples can promote overfitting, whilst values above a certain threshold can lead to underfitting (Wickramasinghe, 2020). The number of samples at a leaf node can be specified by the parameter *min_samples_leaf*. This specifies the number of samples which must be present in a leaf node after splitting. If this criterion is not met, the leaf node cannot be used for decision purposes.

When building a RF, the width of the forest is another criterion to consider. Initially, adding further decision trees to the RF can improve prediction performance. However, a point of saturation is then reached, in which the addition of further learners does not contribute to performance. At this point, including more trees will increase the computational resources required (Oshiro, Perez and Baranauskas, 2012). The number of decision trees in a RF can be specified by the hyperparameter *n_estimators*. Another factor which can influence the performance of a RF is the number of samples provided to each decision tree. Similarly, to the number of learners, performance of a RF will increase as the percentage of the dataset provided increases. However, a point of saturation is reached, in which a larger proportion of samples will increase training time for the RF (Contreras *et al.*, 2021). This can be controlled by the hyperparameter called *max_samples*.

2.8.4 Gradient boosted trees

Similarly, to bagging, boosting is an ensemble technique used in ML. It combines multiple weak learners into a more robust algorithm. Unlike bagging, models are not trained in unison, rather learners are fitted sequentially until no further improvement for prediction can be achieved. Many different types of learners can be boosted. The algorithm adaptive learning (AdaBoost) provides a background of understanding for GB. AdaBoost is an ensemble method which uses decision trees as learners. However, these trees are known as 'stumps' as they only contain one split (Chengsheng, Huacheng and Bing, 2017). AdaBoost begins by assigning all observations within a dataset an equal weighting. A single 'stump' is then fitted to this dataset. The results of this model are then analysed, with particular attention paid to misclassified instances. The algorithm then employs a reweighting scheme to all samples. Instances which were previously classified correctly are down weighted, whilst data points who were misclassified are upweighted. Prediction is then tested with the addition of a new 'stump', this process continues until the classification error can be reduced no further (Chengsheng, Huacheng and Bing, 2017).

Gradient boosting is similar to AdaBoost as it combines the performance of weak learners into a more accurate algorithm. However, GB is more flexible, as it allows for the use of a range of cost functions, whilst also being available to use on different ML algorithms (Natekin and Knoll, 2013). Similarly, to AdaBoost, GB works by reducing prediction error using the addition of further learners. However, the two methods differ on how this is achieved. For GB, the error rate is reduced by fitting the next model on the residuals of the previous. These residuals are calculated from the difference between the predicted outcomes of the base model and observed frequencies. Residuals are then used as the target variable for the next sequential model. The aim of this process is to minimise the cost function until no further improvement can be made (Natekin and Knoll, 2013). The process of gradient boosting is depicted in Figure 2.10.

Figure 2.10: The GB tree process, in which individual trees are used in an ensemble fashion (Deng *et al.*, 2021).

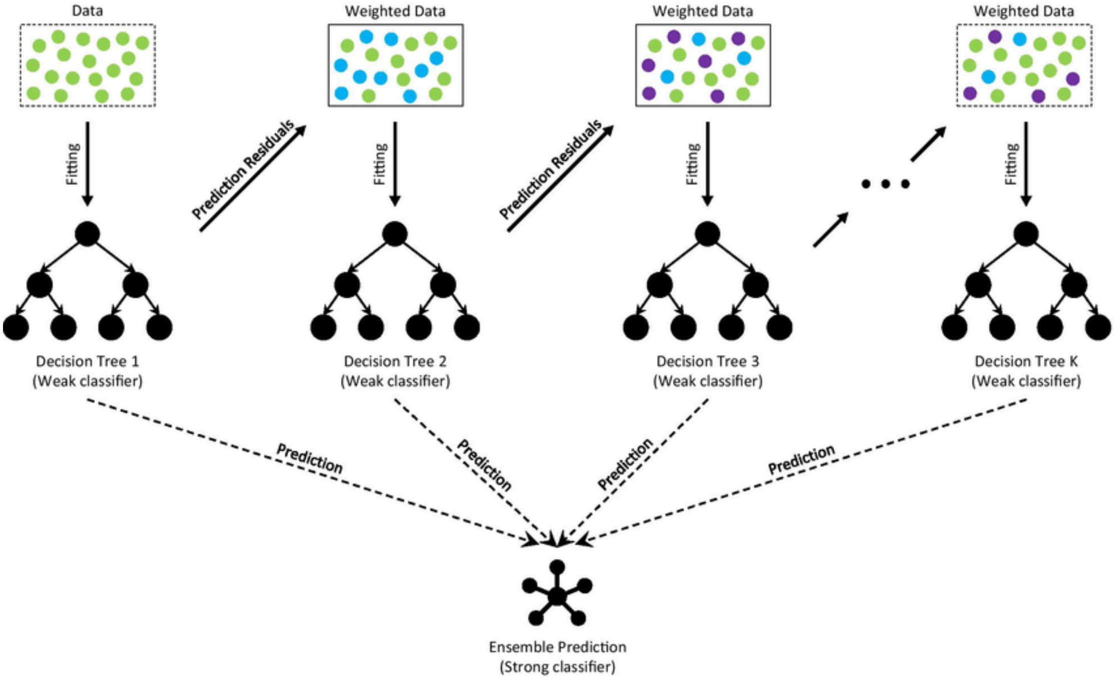


Figure 2.10: An example of the gradient boosting process for decision trees. The changing sets of dots represent the updating of features using the residuals from the previous decision tree. Similarly, to the bagging process, the predictions of each decision tree are combined into a final ensemble prediction. This image has not been altered from the original.

2.8.4.1 Hyperparameter tuning

There is an overlap between the hyperparameters used in RFs and GB trees, due to both using decision trees. However, some parameters are unique to GB. The parameter learning rate effects the contribution of each tree on prediction. A small learning rate results in a small step when minimising a cost function. It has been shown that using small alterations improves the chances of effective prediction (Touzani, Granderson and Fernandes, 2018). This can be implemented by the hyperparameter *learning_rate* in *Python*. A further factor which aims to prevent overfitting is earlystopping. Prior to algorithm training, the maximum number of sequential learners to be used is defined. However, if this is set too high, the model could begin to overfit on the training data. Therefore, the earlystopping algorithm stops the addition of sequential learners. The hyperparameter *earlystopping* can be tuned by the user (Zhang and Yu, 2005).

2.9 Support vector machines

The support vector machine (SVM) is a classification algorithm, it was developed by Cortes and Vapnik, 1995. The idea was based loosely upon the separating nature of the perceptron, which can be defined as a single layer neural network or linear classifier (Collobert and Bengio, 2004). In which a set of weights are multiplied by inputs, corresponding values are then passed to an activation function which determines classification. This formed the basis of the SVM, in which a hyperplane is used to separate instances. The algorithm aims to maximise the distance between data points and this hyperplane, hence achieving effective classification (Collobert and Bengio, 2004). However, the algorithm is computationally expensive due to the complex mathematical calculations involved, this leads to longer training times when compared to other algorithms. Therefore, this algorithm is better suited to smaller datasets (Yu, Yang and Han, 2003).

2.9.1 Optimal hyperplanes

As previously stated, the SVM classifier is based upon a separating hyperplane. The simplest version of this is the optimal hyperplane, in which all points in a dataset can be separated without error.

In order to derive the optimum hyperplane for separation, the distance between the dividing line and the surrounding points is maximised (James *et al*., 2013). For linearly separable datasets, there exists an infinite number of hyperplanes which can separate the two classes. The 'margin' can be defined as the distance between any hyperplane and the surrounding points. Therefore, an algorithm which employs this method is known as the maximal margin classifier (James *et al*., 2013). This process is illustrated in Figure 2.11.

Figure 2.11: An example of the maximal margin classifier with demonstration of maximising the distance between the hyperplane and nearest data points (James et al., 2013).

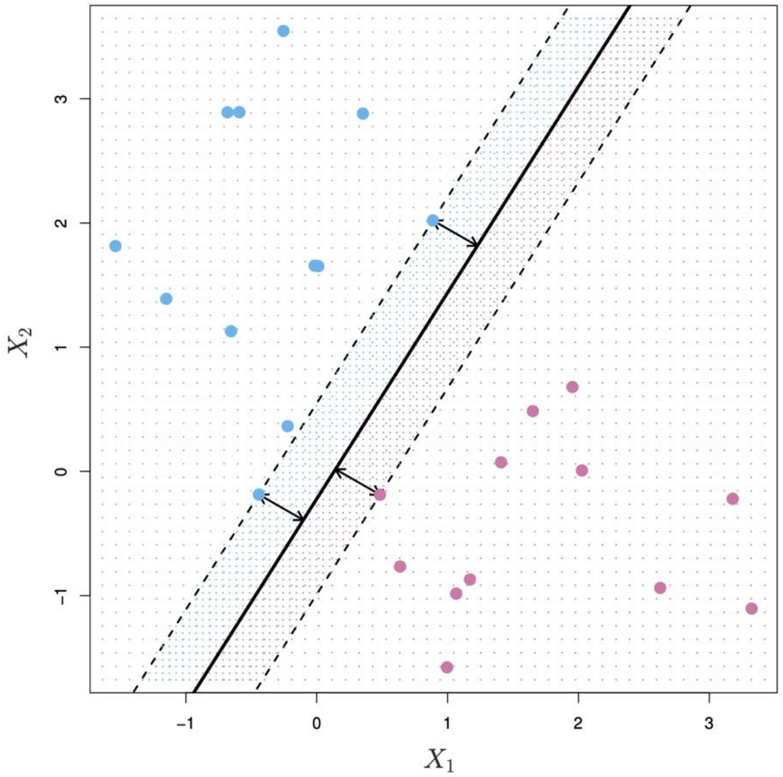


Figure 2.11: A maximal margin classifier for two features (X1, X2). Two classes of observations are represented by blue and purple dots. The maximal margin hyperplane is detailed by the solid line, with the margin represented by the distance between the solid and dashed lines. Those blue and purple dots which lie on the dashed lines are support vectors. This figure has not been altered from the original source.

To calculate a maximal separating plane, the margin of 2M is maximised, where M can be defined as follows:

$$M = \frac{1}{||w||} \tag{2.30}$$

, where ||w|| is described as the Euclidean norm of the set of weights w1, ………, wp where p is the number of features in the dataset. The resulting hyperplane is termed the maximal margin hyperplane. Those points which lie closest to the hyperplane have the greatest influence on its placement, these points are known as support vectors (James *et al.*, 2013).

2.9.2 Support vector classifier

The maximal classifier relies on a dataset to be linearly separable, in practice many datasets are not separable in this manner (James *et al.*, 2013). To classify these non-linearly separable datasets, an element of tolerance can be implemented. In this instance, a small minority of points lie on the incorrect side of the hyperplane. This element of tolerance can be described as using a 'soft margin' (James *et al.*, 2013). To achieve this, a set of errors for all support vectors known as 'slack' variables are defined. The inclusion of this variable alters the optimisation problem to inclusion of a hyperparameter *C*, that can be described as an upper bound to the sum of the 'errors' to the hyperplane. This value is pre-determined and defined as the acceptable level of error (Prosvirin, Duong and Kim, 2019). Whilst T is the number of samples in the dataset.

2.9.3 Support vector machine

The SVM uses a combination of both the maximal margin hyperplane and a soft margin to separate classes. This allows for the classification of non-linearly separable datasets. The optimal way to classify a non-linear dataset is to transform the feature set to a higher dimensional space. This is achieved by using a technique called a kernel function. In which the original dataset can be considered linearly separable in its projected state (James *et al.*, 2013). There are several choices when considering a suitable kernel. A commonly chosen first choice is the radial basis function (RBF). When implementing a SVM with an RBF kernel, focus is given to optimising two hyperparameters. The first of these, gamma, controls the impact of each training sample on the projection into the 3rd dimension. If the value of gamma is chosen to be too low, model accuracy will be poorer. This is also the case for values which are too high for the given dataset. Therefore, the value must be tuned in order to achieve maximum performance (James *et al.*, 2013).

As introduced in Section 2.9.2, the hyperparameter *C* determines the tolerance of the model to classification error. The toleration of misclassified samples is increased with high values of C, in which the resulting model will represent a maximal-margin classifier. Decreasing the value of C reduces the tolerance of errors, which achieves greater generalisation than the maximal-margin classifier. A higher value of C results in an increased possibility of overfitting, whilst a value considered too low could result in a loss of training accuracy.

Therefore, similarly to gamma, the correct estimation of C is important for model performance (James *et al.*, 2013). Another used kernel function is the polynomial kernel with a hyper-parameter (p) representing the order of the polynomial. A value of p=1 renders the kernel equivalent to a linear kernel. In general, the flexibility of the decision boundary increases as the order of the polynomial also increases (Savas and Dovis, 2019).

2.10 The naïve bayes classifier

The naïve bayes (NB) classifier receives its naïve title due to the underlying assumption of independence for all input features. In practice, absolute independence between predictors is extremely rare (H. Chen et al., 2021). The independence assumption is included as it simplifies the calculations required for estimating the conditional probabilities. This simplification exists due to the way in which the NB method calculates the joint probability between a target variable y and a set of features X.

Due to the condition of feature independence being rarely met, probability estimates from NB are usually poor. However, NB has often been known to perform as well as other classification algorithms (Zolnierek and Rubacha, 2005). This is due to relative prediction values driving classifications rather than absolute values. If correlations between features support certain classification results, then the classifier is likely to perform well. This is also true if relationships between features cancel themselves out during prediction (Zolnierek and Rubacha, 2005). The NB classifier is easy to implement and can perform well in noisy datasets. It also performs robustly when data points are missing (Zolnierek and Rubacha, 2005).

3 Machine learning approaches for the life-time risk prediction of Alzheimer's disease (literature review)

3.1 Introduction

Advancements in biotechnology have resulted in various aspects of human biology being reliably recorded, including genetic data and other commonly used biomarkers, e.g., cerebral blood flow, brain imaging. This has led to the accumulation of large biological datasets which ML algorithms can learn from, with the aim of classifying the participants or predict the membership of predefined classes (Cho *et al.*, 2019). The combination of genetic data with other data modalities often leads to complexity, which cannot be processed easily by humans in an un-biased way (Sivarajah *et al.*, 2017).

Risk prediction modelling is an approach to assist diagnosis of a disease or a condition. To accomplish this, statistical models are used to make informed decisions using disease relevant predictors. For the case of AD, the ability to predict the likelihood of disease early can not only prevent misdiagnosis but also assist treatment if detected early (Iddi *et al.*, 2019). In conjunction with an individual's age, genetics has been shown to be a strong risk factor for developing AD. However, genome wide association studies (GWAS) have failed to explain the level of heritability shown in twin studies (Sierksma, Escott-Price and De Strooper, 2020). GWAS-based heritability estimates assume an additive model, which, in statistical terms, is equivalent to looking for the main effects of common variants contributing to disease risk. However, for the genetics of complex diseases, it is unknown whether and to what extent non-additive genetic interaction effects contribute to risk (Hardy and Escott-Price, 2019). This inability to assess non-linear relationships between loci might explain the missing heritability between GWAS and twin studies (Escott-Price and Hardy, 2022).

This chapter reviewed the ability of ML methods to predict lifetime risk for AD using primarily genetic data in the form of single nucleotide polymorphisms (SNPs), however, studies in which SNPs had been combined with other forms of data were also considered. A systematic literature review was employed, where initially all forms of dementia were examined, however searches returned publications focused on AD only. The review was written in line with the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) guidelines (Liberati *et al.*, 2009). Databases were searched for relevant scientific articles, followed by an assessment on how prediction models were developed. Reviews in this area have been conducted previously (Mishra and Li, 2020), however this review is unique in its assessment for the possibility of bias for prediction models in this subject area, as well as in the number of ML methods that it includes. The risk of bias was assessed by using the prediction model risk of bias assessment tool (PROBAST) (Wolff *et al.*, 2019).

3.2 Materials and methods

3.2.1 Search strategy

The online article databases Scopus, PubMed and Google Scholar were used to identify relevant publications for this review. Search terms used were exclusively machine learning, genetics, dementia, Alzheimer's, Single Nucleotide Polymorphism (SNP), polymorphism, mutation, variant and marker. It was decided not to use the names of ML algorithms such as 'Random Forest' directly as publications would use the phrase 'machine learning' in either the title or abstract. These terms were used to retrieve studies published between December 2009 – June 2020. An initial search and screening for relevant publications was conducted by assessing both abstracts and titles. Based on eligibility criteria (listed below), publications from the initial search were then further assessed by two independent reviewers. Any discrepancies were then resolved by a third reviewer.

Inclusion Criteria

- Written in the English language.
- Subject matter of Alzheimer's disease.
- The use of SNP data only unless it was combined with other forms of non-genetic information.
- Supervised ML techniques.
- Prediction resulting in a binary outcome (i.e., case/control).

Exclusion Criteria

- Prediction of Alzheimer's disease related sub-phenotypes (e.g., MCI vs. controls).

- The use of genetic variants other than SNPs as predictors. The search was deliberately broad (see Search Strategy section) to capture papers from non-genetic fields, which do not apply a refined definition of genetic variants.

Articles published between December 2009 and June 2020 were identified. ML techniques have been used in studies prior to this time frame. However, interest in ML in biological research has increased mostly in the last decade (Camacho *et al.*, 2018). Therefore, studies previous to this were sparse and hence a recently defined window was used. SNPs were the only form of genetic variation accepted to facilitate comparisons between studies, therefore articles focusing on gene expression data or other forms of genetic data (e.g., rare variants) were not included. Instances where authors had combined SNP data with other forms of predictive biological variables were included, e.g., Magnetic Resonance Imaging (MRI) and Positron Emitting Tomography (PET). Only models which predicted a binary outcome between cases and controls were included, with instances of mild cognitive impairment (MCI) excluded. For those studies which assessed a binary event and also developed models predicting between MCI individuals and AD cases, information for models used to predict the binary relationship referenced were extracted only. This was due to historic difficulties for clinicians to distinguish between MCI and AD status (Forlenza *et al.*, 2013). Therefore, accepting models which discriminated between case and control status allowed a clearer assessment of the predictive performance.

For the purpose of assessing the suitability and comparability of ML approaches, prognostic and diagnostic models are usually considered separately. Prognostic models are defined as those which focus on future events and use longitudinal data, whereas diagnostic models are based upon current events using cross-sectional data. The search limited to binary outcomes only revealed no prognostic models.

3.2.2 Data extraction

The Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies (CHARMS) (Moons *et al.*, 2014) was used as a tool to perform data extraction. CHARMS provides two tables of check points to be considered by the reviewer. The first table provides guidelines on how to frame the aim of a review, including how to search and filter extracted publications. The second table lists aspects to be extracted from each study for comparison, including predictor type, sample size and the amount of missing data. CHARMS also gives guidance on assessing how certain aspects were reported such as model development, model performance and model evaluation. Advantages of using CHARMS includes replicability across different types of reviews, its ease of use, and assisting reviewers in producing transparent publications (Moons *et al.*, 2014).

The ability of ML methods to discriminate between two classes was extracted independently from all studies by two authors. Accuracy (ACC) describes the performance of a classifier with respect to all samples, it is calculated as the number of correct predictions divided by the total number of predictions made. However, it does not provide information on how well the model performs within the positive and negative classes (Flach, 2019). Sensitivity is calculated by using observed positive outcomes to determine the proportion of classifications correctly made in the positive class, while specificity measures the same statistic in the negative class. Area under the receiver operating characteristic curve (AUC) represents the trade-off between these two measurements at different thresholds, aiming to find the optimal balance (Flach, 2019). AUC was extracted in order to draw comparisons between the studies. Confidence intervals for AUC were also extracted if provided, otherwise these were calculated using the Newcombe method (Debray *et al.*, 2019). Precision can be defined as the ratio of correct predictions in the positive class, divided by the total number of positive predictions. Measures of performance such as accuracy, sensitivity, specificity and precision were also recorded alongside AUC if present. As the true positive rate and recall are different terms used for sensitivity, while specificity is also known as the true negative rate, they were categorised under sensitivity or specificity (if reported). Statistics such as age and gender for participants, types of predictors and ML models were also extracted, as per the CHARMS checklist guidance. Figures in this study were created using Microsoft Word (Fig. 3.1) and the programming language *Python* (Fig. 3.2 and 3.3).

Studies were analysed in order to determine whether they reported the calibration of their models. Calibration is defined as the accuracy of risk estimates and demonstrates how well predicted and observed probabilities of the class membership line up. Previous systematic reviews conducted for prediction models across a number of research areas have shown that calibration is rarely reported (Van Calster et al., 2019b). Poor calibration could lead to healthcare professionals or patients having false expectations for certain events (Van Calster et al., 2019b).

3.2.3 Data analysis

When assessing a number of studies in a review, meta-analyses are often conducted. A meta-analysis produces a weighted average of the reported measures, where the heterogeneity between studies is taken into consideration. If studies overlap, e.g., contain (partially) the same individuals, the resulting correlation between the studies will bias the results of the meta-analysis (Bom and Rachinger, 2020), unless taken into account. Since the majority of the extracted publications used the same dataset, a meta-analysis was not performed in this review.

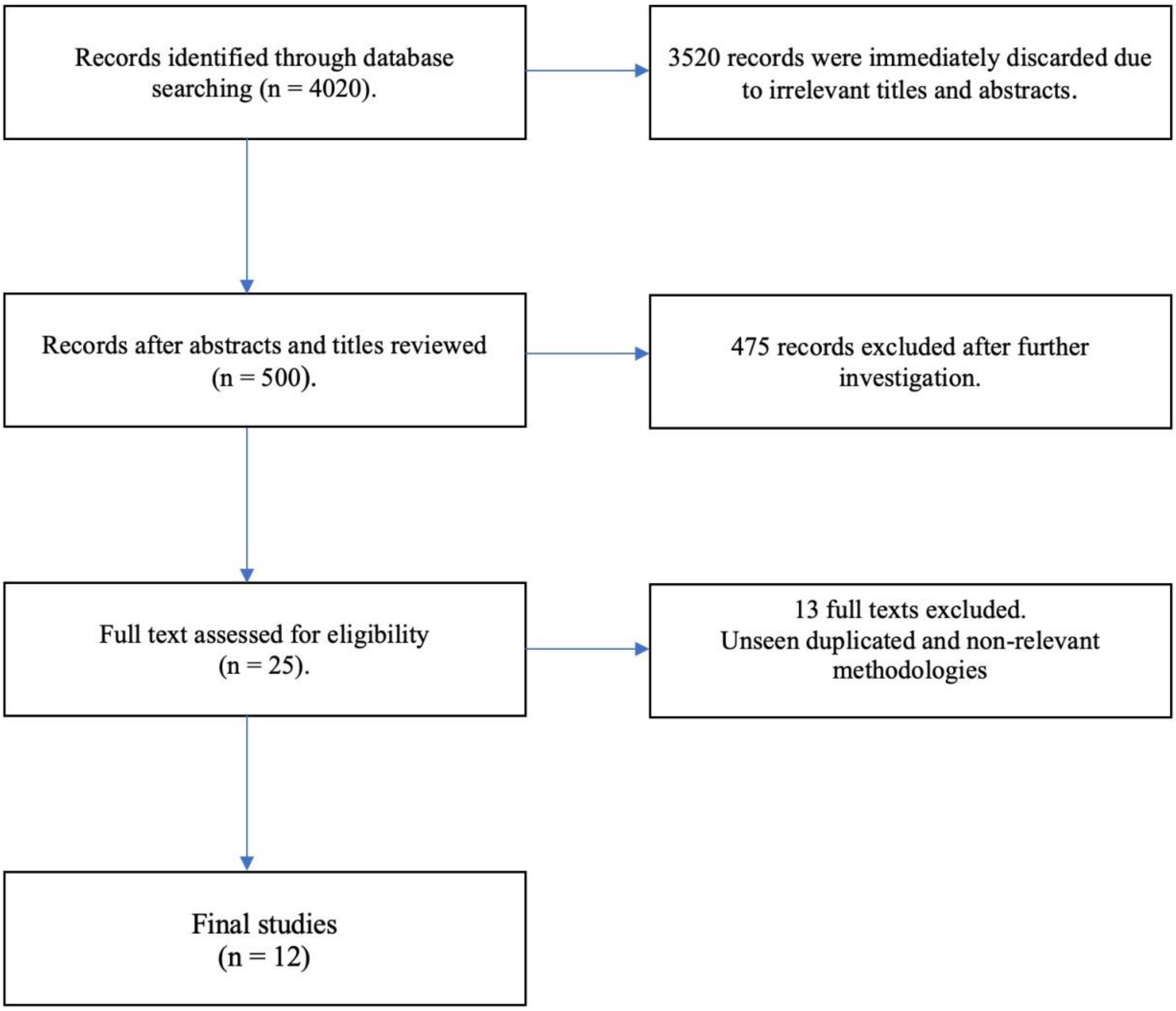
Risk of Bias (ROB) is another component to critically assess when conducting a systematic review of prediction models within studies. PROBAST uses a system of questions split over four categories: participants, predictors, outcome and analysis. Each category contains multiple choice questions assessing an occurrence of shortcomings in that category (with choice of answers from: “yes”, “probably yes”, “no”, “probably no” and “no information”). If any question is answered with no or probably no, this flags the potential for the presence of bias, however assessors must use their own judgement to determine whether a domain is at ROB or not. An answer of no does not automatically result in a high ROB rating. PROBAST does offer assistance on how to reach an overall conclusion on the level of bias in that category. All included studies were assessed for ROB.

3.3 Results

3.3.1 Search results

Following an initial search, a total of 4,020 publications were returned. This number was reduced by assessing whether both titles and abstracts aligned with the inclusion criteria, resulting in 500 studies. A more in-depth analysis was then conducted on the full texts, removing publications which did not pass the inclusion criteria upon a detailed inspection, 25 texts remained at this stage. These were further reduced to 21 due to the presence of duplicates, comprising both pre-prints and conference abstracts. Nine further publications were then removed due to non-relevant methodologies, leaving a final set of 12 studies to be included. A visual representation of the selection process is given in Fig. 3.1.

Figure 3.1: Visual breakdown of publication selection based on a similar diagram found in PRISMA.


Records identified through database searching (n = 4020).
3520 records were immediately discarded due to irrelevant titles and abstracts.
Records after abstracts and titles reviewed (n = 500).
475 records excluded after further investigation.
Full text assessed for eligibility (n = 25).
13 full texts excluded. Unseen duplicated and non-relevant methodologies
Final studies (n = 12)

The majority of publications (10/12) used the publicly available Alzheimer's Disease Neuroimaging (ADNI) (R. C. Petersen *et al.*, 2010a) dataset. ADNI is a longitudinal study measuring various biomarkers in both AD cases and healthy age-matched controls. However, all studies reported here analysed a particular subset of the cohort at a fixed timepoint only. Therefore, only cross-sectional format data were used, and hence models throughout publications were classed as diagnostic rather than prognostic. Out of the publications using

ADNI, four used the initial five-year study (ADNI-1), whilst the remaining studies did not specify which cohort was used. There were two studies that did not use ADNI. Wei, Visweswaran and Cooper (Wei, Visweswaran and Cooper, 2011a) used a combination of three datasets (Reiman *et al.*, 2007) in which biomarkers were collected at a fixed time point, therefore data were cross-sectional. Romero-Rosales et al., 2020 used a longitudinal source of data known as the National Institute on Aging-Late-Onset Alzheimer's Disease Family Study (NIA-LOAD) (Lee, 2008). Again, values for predictors were taken at a fixed time point, thus the data used was cross-sectional. All models across the included studies were classified as diagnostic.

A range of ML approaches were used across the 12 reviewed studies. Table 3.1 outlines all types of models used and their frequency across the publications. The most commonly used ML approach across the analysed publications was Support Vector Machines (SVM), followed by Naïve Bayes (NB) and penalised regression. The number of tested models was also the highest for SVMs. This approach allows the most flexibility when adapting models via kernel functions (Auria and Moro, 2008). Penalised regression was commonly used in the form of the Least Absolute Shrinkage and Selection Operator (LASSO). This type of regularisation shrinks coefficients closer to zero when compared to their maximum likelihood estimates and simultaneously reduces variance in predictions and performs predictor selection. These aspects make penalised regression a popular method in prediction analysis (McNeish, 2015). Random forests (RF) were also used across three studies, these algorithms are intuitive in their use of decision trees, are invariant to scaling, and provide an in-built measure of predictor importance, which likely explains their favour in biology (Chen and Ishwaran, 2012).

Table 3.1: Summary of ML methods used in the analysed publications.

ML approach[a]	Number of publications[b]	Number of models reported across publications[c]	Additional Information[d]
Support Vector Machine (SVMs)	8	44	Linear kernels (22 models, 5 studies). Quadratic polynomials (4 models, 2 study). Cubic Polynomials (4 models, 2 study). Radial basis functions (3 models, 2 studies). Pearson kernel function (2 models, 1 study). Unreported kernels (9 models, 3 studies). A supervised method which uses distance-based calculations to separate samples into groups.
Penalised Regression (LASSO)	4	15	All 15 LASSO regressions across 3 studies. A regression analysis which performs both feature selection and regularisation.
Naïve Bayes (NB)	4	10	Six ordinary NB models, three tree-augmented NB and one model averaged NB. A probabilistic classifier which uses bayes theorem to make predictions.
Random Forest (RF)	3	5	Five classification RFs used, two of which used the RPART package. These are an ensemble of decision trees which produce aggregated classifications.
Bayesian Networks (BN)	2	4	2 BNs with K2 learning algorithm, one markov blanket and one minimal augmented markov blanket. A graphical model which calculates conditional dependencies between variables using Bayesian statistics.
Linear Models	2	4	Bootstrapping Stage-Wise Model Selection (BSWiMS). A supervised model-selection algorithm which uses a combination of linear models for prediction.
K Nearest Neighbour (KNN)	2	3	This is a distanced based algorithm which uses similarities in features to classify.
Ensemble Methods	1	2	Ensembles are the use of a number of ML models, these arrive at a collective prediction result.
Logistic Regression (LR)	1	1	A form of linear regression whereby the outcome is a categorical variable.
Multi-Factor Dimensionality Reduction (MFDR)	1	1	A technique used to detect combinations of independent variables that influence a dependent variable.

Random Forest (RF), Bayesian Networks (BN), K Nearest Neighbour (KNN), Logistic Regression (LR), Multi-Factor Dimensionality Reduction (MFDR) a – Type of machine learning model; b – The number of publications models were used in; c – The number of publications these models occurred in.

3.3.2 Risk of bias (ROB)

For diagnostic models, data sources with the lowest risk of ROB for participants are of the cross-sectional form. The publications which used the ADNI dataset assessed it in a cross-sectional format. This assertion is reinforced in Gross et al., 2016, where ADNI is described as a cross-sectional study with longitudinal follow up. A similar decision was reached when considering the two studies which did not use ADNI, Wei, Visweswaran and Cooper, 2011 and Romero-Rosales et al., 2020. After considering this, ROB was deemed low for participants.

The focus of PROBAST for predictors is to assist the researcher in determining whether the procedures for measuring biomarkers were equal for all members of the study. ADNI provides publicly available documents which outline the methods for biomarker collection. Predictors derived from blood samples or MRI scans were collected using the same protocols for all participants. Therefore, the process of collecting predictors was deemed to be of low ROB. Genotyping of SNPs for the NIA-LOAD dataset (Lee, 2008) was performed in the same way across all samples, therefore ROB for predictors was low for Romero-Rosales et al., 2020. Procedures for collecting predictors in Wei, Visweswaran and Cooper, 2011 were not provided. This was also the case when assessing the original source of the data by Romero-Rosales et al., 2020, therefore ROB for predictors for these publications was stated as not known.

Blinding is the process whereby samples from patients are collected without prior knowledge of their disease status. Such knowledge has been shown to introduce bias to collection procedures (Karanicolas, Farrokhyar and Bhandari, 2010). According to the ADNI data generation policy, samples were collected using blinding and only unblinded when uploaded to databases. Imaging data were collected and processed using standardised automated pipelines, thereby reducing the possibility of multiple clinicians using different methods when collecting predictors (Davis-Turak *et al.*, 2017). ROB was deemed low for blinding in ADNI. Policies for blinding were not provided by either Wei, Visweswaran and Cooper, 2011 or Romero-Rosales et al., 2020. Therefore, a judgement could not be made for either publication.

ROB in the PROBAST category "outcome" was considered to be low for the majority of studies. PROBAST's questions regarding this section focus on how the outcome was determined and whether this determination was applied equally to all participants. ADNI used a range of clinically accepted methods to determine an individual's AD status, including the Mini Mental State Examination and the Clinical Dementia Rating. The use of multiple methods of cognitive performance reduced the possibility of misdiagnosis, which in turn reduced the ROB. Diagnosing the outcome for participants in the NIA-LOAD (Lee, 2008) study was also achieved using a range of stringent methods. NINCD-S-ADRDA (Varma *et al.*, 1999) criteria were used for AD diagnosis at recruitment, while diagnosis was pathologically confirmed for participants who were deceased. Controls were determined using neuropsychological tests in which memory function was examined, coupled with examination for any previous history of neurological disorders. As methods for both controls and cases were applied uniformly across the study participants, with the exception of deceased and alive AD individuals, the ROB for Romero-Rosales et al., 2020b was deemed low for outcome. In Wei, Visweswaran and Cooper, 2011 all brain donors for cases satisfied clinical and neurobiological criteria for cases of late onset AD, while clinical cases satisfied criteria for probable AD (McKhann *et al.*, 2011). Also, brain donor controls did not have significant cognitive impairment at the time of death and clinical controls exhibited no cognitive impairment. However, the methods used to determine these diagnoses were not elaborated upon. For instance, whilst there was a mention of using clinical criteria, these were not defined. Therefore, ROB for outcome was unclear.

The fourth and final category in which PROBAST aids investigation is in the analysis phase of a study. All studies exhibited high ROB for this section, with a consistent lack of reporting for calibration; additionally, 5 out of 12 publications did not report possible missing values in their data and how these were dealt with if present. To assess whether sample sizes used in modelling are adequate, PROBAST suggests the use of the metric Events per Variable (EPV). EPV is defined as the number of events in the minority class (i.e., the smaller of either cases or controls), divided by the number of candidate predictors used. In cases where more in depth algorithms (e.g., Neural Networks (NNs)) are used, model parameters are also included in the calculation of EPV. We evaluated ROB using a value of at least 10 EPVs, following common recommendations (Austin and Steyerberg, 2016). However, this threshold may be tailored more to the accurate estimation of regression coefficients in a logistic regression model. More complex algorithms which require the tuning of hyperparameters (RFs, SVMs, NNs) may require a value of over 100 (van der Ploeg, Austin and Steyerberg, 2014). Values across all studies were assessed to be below this threshold. The study with the highest EPV of 9.43 was Chang et al., 2020. The lowest EPV, 0.0018, was found for Wei, Visweswaran and Cooper, 2011.

Values of EPV below the recommended threshold of 10 introduce the possibility of overfitting, which in turn could result in spurious results (Austin and Steyerberg, 2017b). However, efforts were made by most studies to overcome the problem of overfitting, mostly in the form of Cross-Validation (CV) (11/12 studies). During this process, the data is divided into k partitions, with k-1 partitions used as training data and the remaining partition used as the test set. This process is then repeated k times. It has been demonstrated that using CV is a viable method for authors to address overfitting (Hosseini *et al.*, 2020). Despite this, the possibility of bias could still be present if the correct form of CV is not used. To investigate the importance of CV type selection, several methods of CV were used on datasets with low EPV values (Vabalas *et al.*, 2019). The simplest form of CV (k-partitioning) was shown not to counteract the issue of overfitting in some instances and could even exacerbate the problem. Nested-CV has been shown to achieve the best performance of all methods (Varma and Simon, 2006) and it operates by using an outer and inner loop of CV. The outer loop splits k times to perform model validation while hyperparameters and feature selection are conducted in the inner loop. This method was only reported by two of the included studies (Zhou, Liu, *et al.*, 2019).

3.3.3 Machine learning performance

As discussed in Section 3.3.1, 12 studies were accepted for inclusion in this review. For each publication, a range of statistics were extracted, some of which are outlined in Table 3.2.

Table 3.2: Summary of the reviewed publications.

Publication title and authors/publication date[a]	Machine learning approaches[b]	AUC for models[c]	Accuracy for models[d]	Data source used[e]	Sample size[f]
1. Benchmarking machine learning models for late-onset Alzheimer's disease prediction from genomic data 52	2LASSO, 2RF, 2RPART, 2KNN, 2SVM (no-kernel),2 ensemble of all methods, 2BSWIMS = Linear models	(0.494-0.719)	N/A	ADNI SNPs only	Discover dataset: 230 Cases 241 Controls Validation dataset: 37 Cases 130 Controls
2. Effective Diagnosis of Alzheimer's Disease via Multimodal Fusion Analysis Framework 51	5 SVMS kernel unspecified	N/A	Accuracy (0.70-0.87)	ADNI MRI and SNPs	37 Cases 35 Controls
3. Latent Representation Learning for Alzheimer's Disease Diagnosis with Incomplete Multi-Modality Neuroimaging and Genetic Data 41	9 SVMs kernel unspecified	(0.62-0.65)	Accuracy (0.59 – 0.67)	ADNI 1 MR images and SNPs	171 cases 204 Controls
4. Discovering Alzheimer Genetic Biomarkers Using Bayesian Networks 48	NB, TAN NB, Markov blanket, minimal augmented markov blanket	N/A	Accuracy (0.62-0.66)	ADNI – SNPs only	282 Controls 48 Cases

5. The application of naive Bayes model averaging to predict Alzheimer's disease from genome-wide data 24	NB, FSNB, MANB	(0.59-0.72)	N/A	GWAS collected and analysed originally by Reiman (M Reiman *et al*, 2008) LOAD GWAS – SNPs only	550 Controls 861 Cases
6. A Hierarchical Feature and Sample Selection Framework and Its Application for Alzheimer's Disease Diagnosis 42	5 SVMS all linear	(85.5-0.97) SNPs only model = 85.5 MR + SNP model = 97.4		ADNI 1 – MRI and SNPs	204 Controls 171 Cases
7 Integrated higher-order evidence-based framework for prediction of higher-order epistasis interactions in Alzheimer's disease 79	NB, RF, KNN, LR, SVM (rbf), multi-factor dimensionality reduction	N/A	Accuracy (0.62-0.78)	ADNI – SNPs only	306 Cases 125 Controls
8. Integrative analysis of multi-dimensional imaging genomics data for Alzheimer's disease prediction 80	4 SVMS - all linear	N/A	Accuracy (0.88-0.95)	ADNI MRI and SNPs	49 Cases 47 Controls
9. Identifying genetic biomarkers associated to Alzheimer's disease using Support Vector Machine 45	10 SVMS 2 linear, 2 quadratic polynomial, 2 cubic polynomial, 2 RBF, 2PUK	N/A	Accuracy (0.62-0.77)	ADNI 1 SNPs only	214 Controls 177 Cases

10. Improving predictive models for Alzheimer's disease using GWAS data by incorporating misclassified samples modelling 26	2BSWiMS-logistic, 2GALGO-SVM (no-kernel). 2LASSO 8 LASSOs	(0.68-0.844)	N/A	National Institute on Aging—Late-Onset Alzheimer's Disease SNPs only	2000 Controls 1856 Cases
11. GenEpi: Gene-based Epistasis Discovery Using Machine Learning 37	3 Lasso regressions	N/A	Accuracy (0.83-0.94)	ADNI SNPs only	241 Controls 123 Cases
12. Developing an early predictive system for identifying genetic biomarkers associated to Alzheimer's disease using machine learning techniques 46	2 SVM(Linear-kernel), 2 SVM(Quadratic-polynomial), 2 SVM(Cubic-polynomial), 2 Naïve Bayes, 2 Naïve Bayes(tree-augmented) and 2 Bayesian networks (K2)	N/A	Accuracy (0.95-0.99)	ADNI 1 data set. Also, a separate ADNI whole genome sequencing data set, genotyped using Illumina Omni 2.5 M. SNPs only	214 Controls 177 Cases 321 Controls 49 Cases

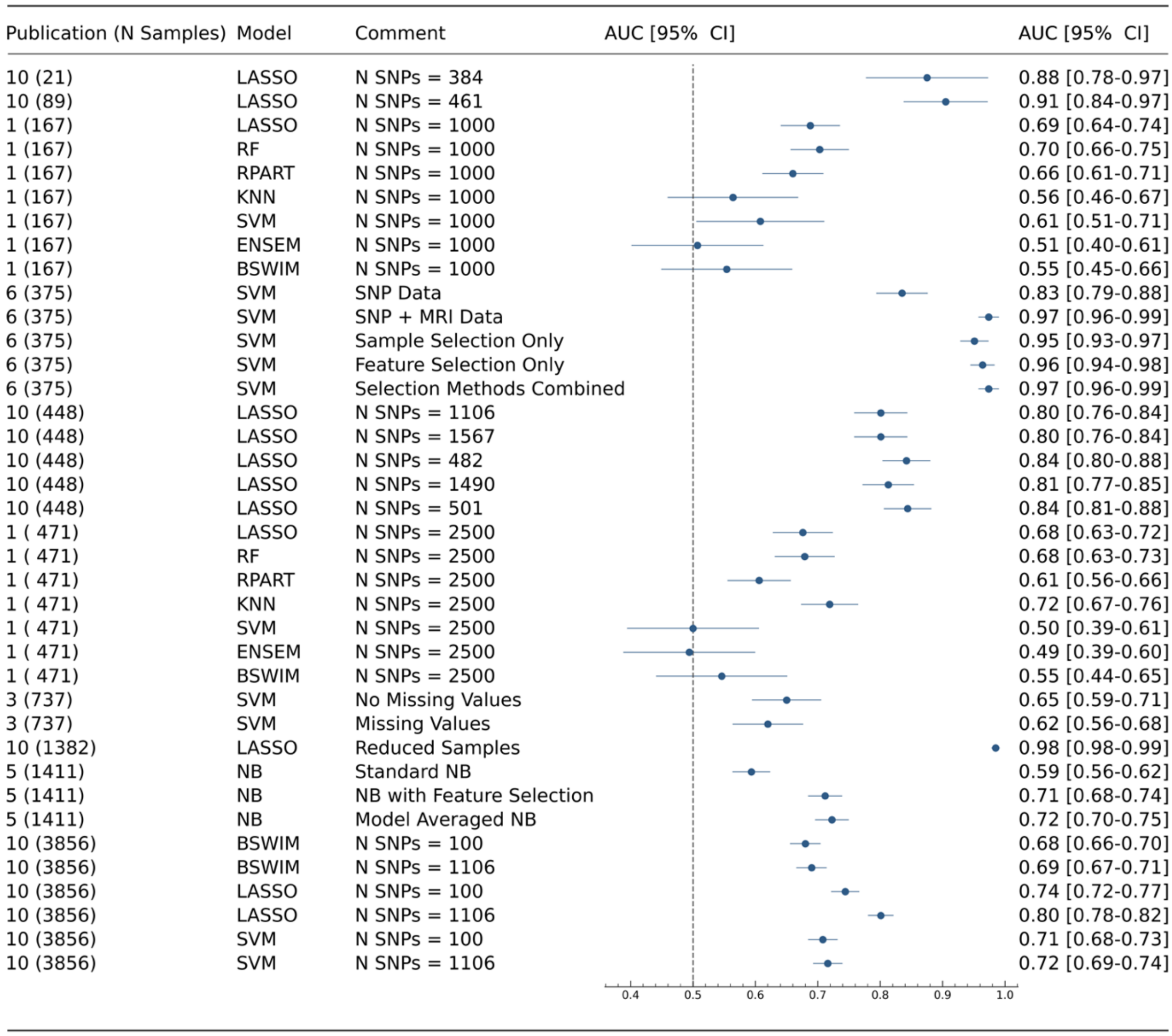
a – Publication title; b – The range of sensitivity values detailed within each study. – c AUC values reported by models; - d ACC values reported for models; - e The source of data used in the study; - f the sample size used by cases and controls.

Five studies recorded AUC for the performance of models, ranging from 0.49 to 0.97. The remaining seven studies reported mainly ACC, sensitivity and specificity The highest AUC value was achieved by (An *et al.*, 2017), where the authors used a hierarchal method to find the optimal set of features for the prediction of AD. Manifold regularisation was used to combine both genetic and MRI data in a semi-supervised hierarchal feature and sample selection framework. This method utilised both labelled and unlabelled data in order to maximise the amount of information for prediction. For classification purposes, SVMs were used to discriminate between controls and cases. However, the EPV score was 0.919 and this is below the recommended threshold of 10. This could introduce the possibility of overfitting which can in turn lead to spurious results (Austin and Steyerberg, 2017b). The authors used CV to alleviate the potential for overfitting.

A single study reported calibration statistics (Wei, Visweswaran and Cooper, 2011a) (Publication 5 in Table 3.2). The authors compared the predictive capability of a model using averaged NB with both standard NB and NB with feature selection. The method used to report calibration was calibration curves. The results highlighted that the model using averaged NB achieved better calibration than the standard NB model and achieved similar performance to the NB with feature selection. The prediction accuracy of these models was 0.59-0.72. Further conclusions which can be drawn from Table 3.2 include the consistent use of the Alzheimer's Disease Neuroimaging Initiative (ADNI) (R C Petersen *et al.*, 2010) dataset, with 10/12 publications using the publicly available source. However, despite the majority of studies using the ADNI, cohort sizes differed between analyses. This is most likely due to different quality control procedures used across publications.

The prediction performance of ML methods in each study are further summarised further in Figures 3.2 and 3.3. The first column shows the reference number of the publication as listed in Table 3.2 along with the sample size used in the respective ML model. ML approaches used are shown in the second column. The third column displays information which assists the reader in distinguishing between models in the same study, this includes factors such as number of SNPs used, and methodologies implemented. Studies were sorted by sample size in ascending order. The vertical dashed line shows the accuracy of 0.5, which indicates a 50% chance of the result being correct. The last column shows the actual values of the accuracy achieved. Confidence intervals of AUC values in Figure 3.2 were calculated using the Newcombe method (Debray *et al.*, 2019). These confidence intervals reflect the variability of AUC controlling for sample size. This allows for comparison between studies with large sample size differences. If the intervals overlap between studies, then the AUCs are not significantly different between models.

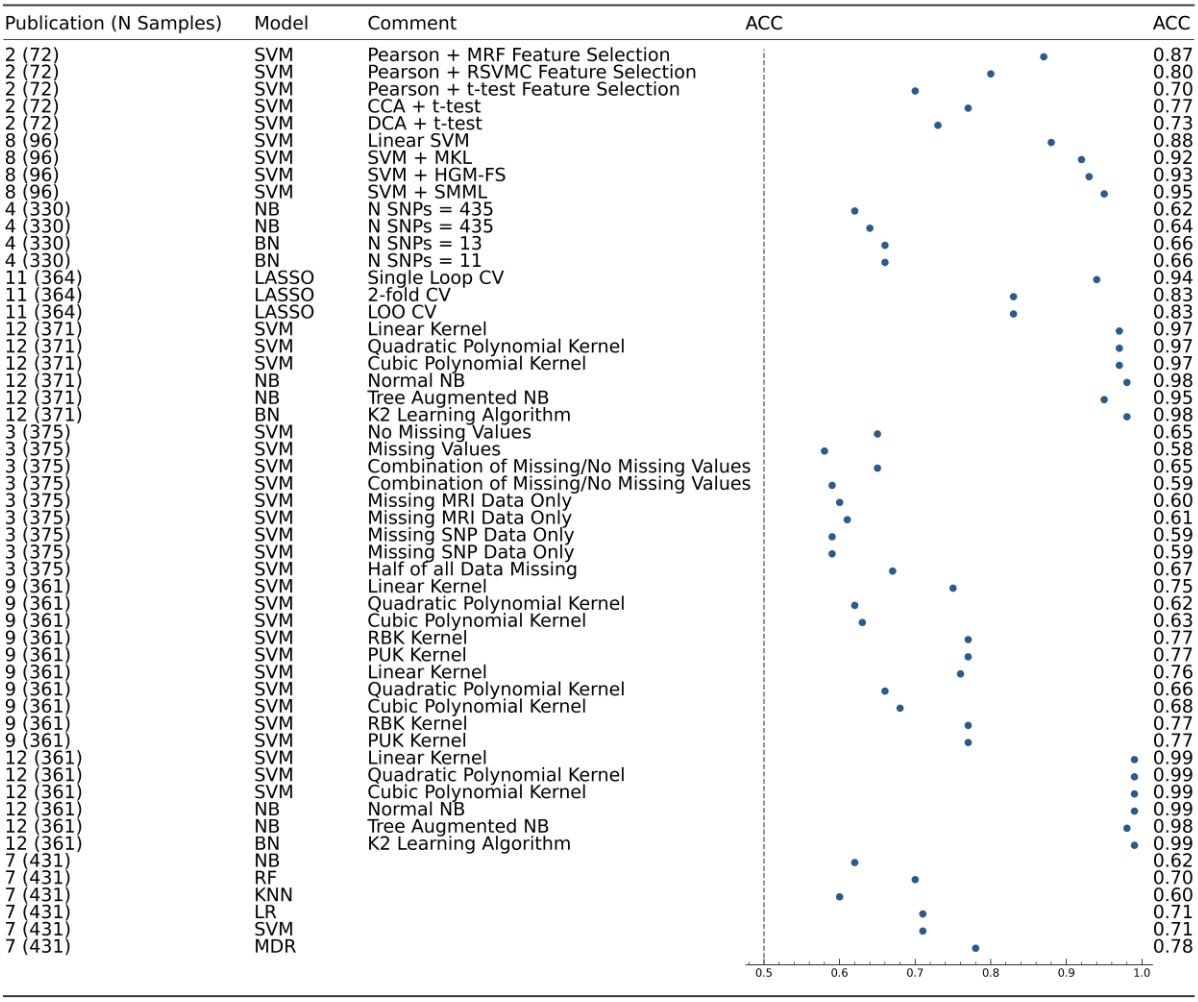
Fig 3.2: A forest plot displaying models used across publications which reported AUC, with the addition of confidence intervals derive using the Newcombe method



Publication (N Samples)	Model	Comment	AUC [95% CI]
10 (21)	LASSO	N SNPs = 384	0.88 [0.78-0.97]
10 (89)	LASSO	N SNPs = 461	0.91 [0.84-0.97]
1 (167)	LASSO	N SNPs = 1000	0.69 [0.64-0.74]
1 (167)	RF	N SNPs = 1000	0.70 [0.66-0.75]
1 (167)	RPART	N SNPs = 1000	0.66 [0.61-0.71]
1 (167)	KNN	N SNPs = 1000	0.56 [0.46-0.67]
1 (167)	SVM	N SNPs = 1000	0.61 [0.51-0.71]
1 (167)	ENSEM	N SNPs = 1000	0.51 [0.40-0.61]
1 (167)	BSWIM	N SNPs = 1000	0.55 [0.45-0.66]
6 (375)	SVM	SNP Data	0.83 [0.79-0.88]
6 (375)	SVM	SNP + MRI Data	0.97 [0.96-0.99]
6 (375)	SVM	Sample Selection Only	0.95 [0.93-0.97]
6 (375)	SVM	Feature Selection Only	0.96 [0.94-0.98]
6 (375)	SVM	Selection Methods Combined	0.97 [0.96-0.99]
10 (448)	LASSO	N SNPs = 1106	0.80 [0.76-0.84]
10 (448)	LASSO	N SNPs = 1567	0.80 [0.76-0.84]
10 (448)	LASSO	N SNPs = 482	0.84 [0.80-0.88]
10 (448)	LASSO	N SNPs = 1490	0.81 [0.77-0.85]
10 (448)	LASSO	N SNPs = 501	0.84 [0.81-0.88]
1 (471)	LASSO	N SNPs = 2500	0.68 [0.63-0.72]
1 (471)	RF	N SNPs = 2500	0.68 [0.63-0.73]
1 (471)	RPART	N SNPs = 2500	0.61 [0.56-0.66]
1 (471)	KNN	N SNPs = 2500	0.72 [0.67-0.76]
1 (471)	SVM	N SNPs = 2500	0.50 [0.39-0.61]
1 (471)	ENSEM	N SNPs = 2500	0.49 [0.39-0.60]
1 (471)	BSWIM	N SNPs = 2500	0.55 [0.44-0.65]
3 (737)	SVM	No Missing Values	0.65 [0.59-0.71]
3 (737)	SVM	Missing Values	0.62 [0.56-0.68]
10 (1382)	LASSO	Reduced Samples	0.98 [0.98-0.99]
5 (1411)	NB	Standard NB	0.59 [0.56-0.62]
5 (1411)	NB	NB with Feature Selection	0.71 [0.68-0.74]
5 (1411)	NB	Model Averaged NB	0.72 [0.70-0.75]
10 (3856)	BSWIM	N SNPs = 100	0.68 [0.66-0.70]
10 (3856)	BSWIM	N SNPs = 1106	0.69 [0.67-0.71]
10 (3856)	LASSO	N SNPs = 100	0.74 [0.72-0.77]
10 (3856)	LASSO	N SNPs = 1106	0.80 [0.78-0.82]
10 (3856)	SVM	N SNPs = 100	0.71 [0.68-0.73]
10 (3856)	SVM	N SNPs = 1106	0.72 [0.69-0.74]

Column1 – Publication number as found in Supplementary Table 1, along with sample size. Column2 – Type of machine learning model. Column3 – Information to help distinguish between models in publications, including differing SNP numbers and methodologies.

Fig 3.3: A forest plot displaying all models used across publications which reported ACC



Publication (N Samples)	Model	Comment	ACC
2 (72)	SVM	Pearson + MRF Feature Selection	0.87
2 (72)	SVM	Pearson + RSVMC Feature Selection	0.80
2 (72)	SVM	Pearson + t-test Feature Selection	0.70
2 (72)	SVM	CCA + t-test	0.77
2 (72)	SVM	DCA + t-test	0.73
8 (96)	SVM	Linear SVM	0.88
8 (96)	SVM	SVM + MKL	0.92
8 (96)	SVM	SVM + HGM-FS	0.93
8 (96)	SVM	SVM + SMML	0.95
4 (330)	NB	N SNPs = 435	0.62
4 (330)	NB	N SNPs = 435	0.64
4 (330)	BN	N SNPs = 13	0.66
4 (330)	BN	N SNPs = 11	0.66
11 (364)	LASSO	Single Loop CV	0.94
11 (364)	LASSO	2-fold CV	0.83
11 (364)	LASSO	LOO CV	0.83
12 (371)	SVM	Linear Kernel	0.97
12 (371)	SVM	Quadratic Polynomial Kernel	0.97
12 (371)	SVM	Cubic Polynomial Kernel	0.97
12 (371)	NB	Normal NB	0.98
12 (371)	NB	Tree Augmented NB	0.95
12 (371)	BN	K2 Learning Algorithm	0.98
3 (375)	SVM	No Missing Values	0.65
3 (375)	SVM	Missing Values	0.58
3 (375)	SVM	Combination of Missing/No Missing Values	0.65
3 (375)	SVM	Combination of Missing/No Missing Values	0.59
3 (375)	SVM	Missing MRI Data Only	0.60
3 (375)	SVM	Missing MRI Data Only	0.61
3 (375)	SVM	Missing SNP Data Only	0.59
3 (375)	SVM	Missing SNP Data Only	0.59
3 (375)	SVM	Half of all Data Missing	0.67
9 (361)	SVM	Linear Kernel	0.75
9 (361)	SVM	Quadratic Polynomial Kernel	0.62
9 (361)	SVM	Cubic Polynomial Kernel	0.63
9 (361)	SVM	RBK Kernel	0.77
9 (361)	SVM	PUK Kernel	0.77
9 (361)	SVM	Linear Kernel	0.76
9 (361)	SVM	Quadratic Polynomial Kernel	0.66
9 (361)	SVM	Cubic Polynomial Kernel	0.68
9 (361)	SVM	RBK Kernel	0.77
9 (361)	SVM	PUK Kernel	0.77
12 (361)	SVM	Linear Kernel	0.99
12 (361)	SVM	Quadratic Polynomial Kernel	0.99
12 (361)	SVM	Cubic Polynomial Kernel	0.99
12 (361)	NB	Normal NB	0.99
12 (361)	NB	Tree Augmented NB	0.98
12 (361)	BN	K2 Learning Algorithm	0.99
7 (431)	NB		0.62
7 (431)	RF		0.70
7 (431)	KNN		0.60
7 (431)	LR		0.71
7 (431)	SVM		0.71
7 (431)	MDR		0.78

Column1 – Publication number as found in Supplementary Table 1, along with sample size. Column2 – Type of machine learning model. Column3 – Information to help distinguish between models in publications, including differing SNP numbers and methodologies.

Alongside both AUC and accuracy, additional measures of model performance were used across publications. These are detailed in Table 3.3.

Table 3.3: Sensitivity and specificity values per publication.

Publication title	Sensitivity	Specificity	Precision
Benchmarking machine learning models for late-onset Alzheimer's disease prediction from genomic data	0.033 – 0.719	0.62 – 0.981	N/A
Effective Diagnosis of Alzheimer's Disease via Multimodal Fusion Analysis Framework	N/A	N/A	N/A
Latent Representation Learning for Alzheimer's Disease Diagnosis with Incomplete Multi-Modality Neuroimaging and Genetic Data	N/A	N/A	N/A
Discovering Alzheimer Genetic Biomarkers Using Bayesian Networks	0.59-0.89	0.16 – 0.66	N/A
The application of naïve Bayes model averaging to predict Alzheimer's disease from genome-wide data	N/A	N/A	N/A
A Hierarchical Feature and Sample Selection Framework and Its Application for Alzheimer's Disease Diagnosis	0.75 – 0.86	0.85 – 0.96	N/A
Integrated higher-order evidence-based framework for prediction of higher-order epistasis interactions in Alzheimer's disease	0.62 – 0.75	0.55 – 0.82	N/A

Integrative analysis of multi-dimensional imaging genomics data for Alzheimer’s disease prediction	0.90 – 0.94	0.85 – 0.96	N/A
Identifying genetic biomarkers associated to Alzheimer's disease using Support Vector Machine	0.62 – 0.77	N/A	0.59 – 0.67
Improving predictive models for Alzheimer’s disease using GWAS data by incorporating misclassified samples modelling	0.61-0.83	0.73 – 0.86	N/A
GenEpi: Gene-based Epistasis Discovery Using Machine Learning	0.66 – 0.85	N/A	0.77 – 0.96
Developing an early predictive system for identifying genetic biomarkers associated to Alzheimer’s disease using machine learning techniques	0.57 – 0.98	N/A	0.59 – 1.00

a – Publication title; b – The range of sensitivity values detailed within each study. c – The range of specificity values detailed within each study. D – Values for precision if reported

Validation methods are an important factor in the development of ML models. The type of method chosen can influence algorithm performance, with emphasis on reducing the possibility of overfitting (Maleki et al., 2020). All types of validation methods used across the included publications are outlined in Table 3.4.

Table 3.4: Methods of validation used.

Methods[a]	Number of Studies[b]	Number of Models[c]
Cross-validation – Number of folds not specified.	1	14
Cross-validation – 20 Folds	1	14
Cross-validation – 10 Folds	7	50
Cross-validation – 5 Folds	1	3
Cross-validation – 2 Folds	1	2
Leave-one-out (LOO) CV	1	1
Training/Test split 60:40	1	5

a – Publication title; b – The range of sensitivity values detailed within each study. c – The range of specificity values detailed within each study

Ten-fold CV was the most common form of validation used, however a range of other values of k were also documented. One further study used a nested CV approach to optimise both model performance and hyperparameter tuning. Leave one out CV was also used by one study, this functions by creating a number of folds equal to the number of data points in the training set. Within each fold a single data point is removed to be used as the test set, the algorithm is then trained on the remaining points. Prediction performance is calculated by averaging over the results for all folds. Also, one publication explored a different approach of dividing the data into training and test datasets called a split sample. In this process, a model is trained using a training set and is subsequently tested on a validation (test) set, where the test dataset contains the remainder of the original data not included in the training dataset. All of these methods are known as internal validation, where model optimisation and hyperparameter tuning is achieved using a single dataset. External validation involves using a completely separate cohort to validate an already trained model, usually this cohort has been independently gathered and assessed to the initial training data (Ramspek *et al.*, 2021). This method was not used by any study in this review.

3.3.4 Sample size

Sample sizes ranged from 72 to 3,856 individuals, with the largest cohort being the NIA-LOAD dataset (Vardarajan *et al.*, 2014). As discussed in Section 3.2.1, it was decided to only include those studies who focused on AD prediction. All other forms of dementia were purposely excluded. Another aspect of criteria used was the exclusion of any models which predicted progression from MCI to AD, with the classification of those with AD (cases) and cognitively healthy individuals (controls) included only. The breakdown of cohorts by cases and controls is outlined in Table 3.5, as well as Imbalances between classes, as a ratio between controls over cases. These ranged from 0.408-6.55, with a median value of 1.193. The accuracy for the study with the highest-class imbalance (6.55) was 0.95-0.99 ACC (Abd El Hamid, Mabrouk and Omar, 2019).

Table 3.5: Class imbalances for each study.

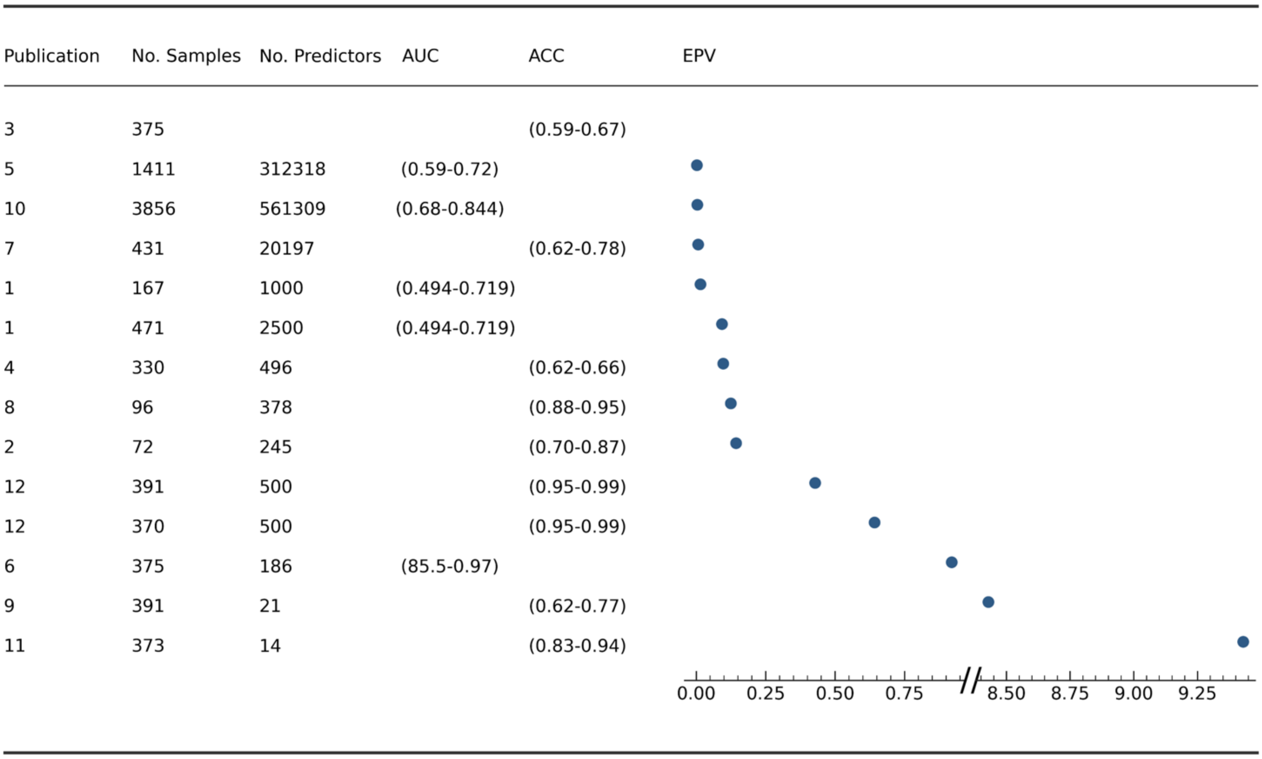
Publication title[a]	Number of Controls[b]	Number of Cases[c]	Class Imbalance[d]
Benchmarking machine learning models for late-onset Alzheimer's disease prediction from genomic data	Discovery dataset: 230 Validation dataset: 130	Discovery dataset: 241 Validation dataset: 37	0.954 3.514
Effective Diagnosis of Alzheimer's Disease via Multimodal Fusion Analysis Framework	35	37	0.946
Latent Representation Learning for Alzheimer's Disease Diagnosis with Incomplete Multi-Modality Neuroimaging and Genetic Data	204	171	1.193
Discovering Alzheimer Genetic Biomarkers Using Bayesian Networks	282	48	5.875
The application of naïve Bayes model averaging to predict Alzheimer's disease from genome-wide data	550	861	0.639
A Hierarchical Feature and Sample Selection Framework and Its	204	171	1.193

Application for Alzheimer’s Disease Diagnosis			
Integrated higher-order evidence-based framework for prediction of higher-order epistasis interactions in Alzheimer’s disease	125	306	0.408
Integrative analysis of multi-dimensional imaging genomics data for Alzheimer’s disease prediction	47	49	0.959
Identifying genetic biomarkers associated to Alzheimer's disease using Support Vector Machine	214	177	1.209
Improving predictive models for Alzheimer’s disease using GWAS data by incorporating misclassified samples modelling	2000	1856	1.078
GenEpi: Gene-based Epistasis Discovery Using Machine Learning	241	132	1.826
Developing an early predictive system for identifying genetic biomarkers associated to Alzheimer’s disease using machine learning techniques	214 321	177 49	1.209 6.551

a – Publication title; b – The number of controls in the study; c – The number of cases in the study; d – The ratio of cases versus controls – Imbalance.

The number of SNPs used in models varied between studies, with numbers ranging from 21 to 561,309 SNPs. The large range in the number of SNPs used was due to differences in the used methodologies. The study which used the greatest number of SNPs (Romero-Rosales et al., 2020) investigated improving AUC by reintroducing initially misclassified samples to the final models. The study which used the least number of SNPs focused only on the top 10 genes associated with AD (Mostafa Abd El Hamid, Omar and Mabrouk, 2016a), thereby limiting the number of SNPs included in the study. EPV ranged from 0.0018-9.43 for eleven studies, with one study not providing enough information to calculate EPV. These values are displayed in Figure 3.4, this also includes the number of samples, number of predictors used and values of either ACC or AUC for each study. The publication number corresponds to those used in Figures 3.2 and 3.3. Due to the large difference between two values and the rest, two scales were used to allow for all points to be plotted on the same figure.

Fig 3.4: A forest plot displaying all available EPV values across the included studies



Publication	No. Samples	No. Predictors	AUC	ACC	EPV
3	375			(0.59-0.67)	
5	1411	312318	(0.59-0.72)		●
10	3856	561309	(0.68-0.844)		●
7	431	20197		(0.62-0.78)	●
1	167	1000	(0.494-0.719)		●
1	471	2500	(0.494-0.719)		●
4	330	496		(0.62-0.66)	●
8	96	378		(0.88-0.95)	●
2	72	245		(0.70-0.87)	●
12	391	500		(0.95-0.99)	●
12	370	500		(0.95-0.99)	●
6	375	186	(85.5-0.97)		●
9	391	21		(0.62-0.77)	●
11	373	14		(0.83-0.94)	●

Column1 – Publication number as found in Supplementary Table 1 Column2 – Number of samples. Column3 – Number of predictors used, Column 4 – AUC of models if reported, Column 5 – ACC of models if reported, Column 6 – values of EPV.

3.3.5 Predictors

Criteria used for inclusion specified that SNPs were the only form of genetic data used as predictors. However, other predictors were also considered, whereby other forms of predictive material were used alongside SNPs. Table 3.6 outlines the different types of data modalities used across publications, as well as further statistics.

Table 3.6: Information extracted for the type of predictive materials used and methods for the pre-processing of SNPs.

Publication title[a]	Types of data modality used[c]	SNPs QC General	MAF	Missing value rate
Benchmarking machine learning models for late-onset Alzheimer's disease prediction from genomic data	SNPs	Marker call rate - ≤ 99% Hardy Weinberg Equilibrium test - ≤ 0.05 LD based clumping – p-value ≤ 0.01 and r2 ≤ 0.05.	≤ 0.01	N/A
Effective Diagnosis of Alzheimer's Disease via Multimodal Fusion Analysis Framework	SNPs/MRI	Sample call rate – 95% Genotyping – 99.9% Hardy Weinberg test 0.0001 %	4%	N/A
Latent Representation Learning for Alzheimer's Disease Diagnosis with Incomplete Multi-Modality Neuroimaging and Genetic Data	SNPs/MRI/PET (Positron emitting tomography)	Selected SNPs were imputed to estimate missing genotypes. Illumina annotation information was used to select a subset of SNPs		
Discovering Alzheimer Genetic		Hardy Weinberg test ≤ 0.001	0.01	

Biomarkers Using Bayesian Networks	SNPs			
The application of naïve Bayes model averaging to predict Alzheimer’s disease from genome-wide data	SNPs	N/A	N/A	N/A
A Hierarchical Feature and Sample Selection Framework and Its Application for Alzheimer’s Disease Diagnosis	SNPs/MRI	Gender Check Hardy Weinberg Equilibrium test Population Stratification	Percentage not specified	
Integrated higher-order evidence-based framework for prediction of higher-order epistasis interactions in Alzheimer’s disease	SNPs	N/A	N/A	N/A
Integrative analysis of multi-dimensional imaging genomics data for Alzheimer’s disease prediction	SNPs/MRI/PET/CSF (Cerebrospinal fluid)	Call rate check per subject, gender check, The Hardy Weinberg Equilibrium test, Population stratification	Percentage not specified	
Identifying genetic biomarkers associated to Alzheimer's disease	SNPs	Removing individuals with discordant gender information, LD pruning, subjects with high IBD are removed, Hardy	≤ 0.001	≤ 10%

using Support Vector Machine		Weinberg Equilibrium test ≤ 0.000005 %.		
Improving predictive models for Alzheimer's disease using GWAS data by incorporating misclassified samples modelling	SNPs	Marker call rate removal ≤ 98%, monomorphic markers also removed, Hardy Weinberg Equilibrium rate ≤ 0.000.	N/A	N/A
GenEpi: Gene-based Epistasis Discovery Using Machine Learning	SNPs	Missing data imputation according to the 1000 genome haplotypes.		
Developing an early predictive system for identifying genetic biomarkers associated to Alzheimer's disease using machine learning techniques	. SNPs	Removing individuals with discordant gender information, LD pruning, subjects with high IBD are removed, Hardy Weinberg Equilibrium test ≤ 0.000005 %.	≤ 0.01	≤ 10%

a – Publication title; b – Hyperparameter tuning methods. c – Types of data modality used; d – Pre-processing steps for SNPs.

The most common form of secondary data used was MRI, included in four publications. PET imaging data was also used in two studies. Additionally, Cerebrospinal Fluid (CSF) was used in one publication. Pre-processing techniques for SNPs were reported in the majority (10/12) of studies. All these studies excluded SNPs which did not satisfy Hardy-Weinberg equilibrium (Namipashaki, 2015). SNPs were selected with a variety of AD association significance thresholds (0.00007 – 0.05), leading to different numbers of SNPs being retained across studies. Seven of the studies which discussed pre-processing for SNPs also used minimal minor allele frequency (MAF), i.e., rare variants were removed from a SNP set based on their allele frequency. Thresholds used for MAF varied (0.01 – 0.04) across studies.

Two studies did not report steps taken to pre-process SNPs; this could lead to questions regarding data quality.

Common steps required when conducting ML analyses include handling missing data and the tuning of hyperparameters for algorithm optimisation. Methods used to approach these factors for the included publications are outlined in Table 3.7.

Table 3.7: Methods to deal with missing data and to optimise algorithms.

Publication title	Were methods used to deal with missing data.	Missing data Methods used.	Hyperparameter Tuning Methods[b]
Benchmarking machine learning models for late-onset Alzheimer's disease prediction from genomic data	No	N/A	N/A
Effective Diagnosis of Alzheimer's Disease via Multimodal Fusion Analysis Framework	No	N/A	Model aspects such as number of trees in RF and tree inputs were optimised.
Latent Representation Learning for Alzheimer's Disease Diagnosis with Incomplete Multi-Modality Neuroimaging and Genetic Data	Yes	The formulation of a latent representation learning method, which used incomplete samples.	Hyperparameters were tuned using CV
Discovering Alzheimer Genetic Biomarkers Using Bayesian Networks	Yes	Missing values imputed by Expectation Maximization algorithm.	N/A
The application of naïve Bayes model averaging to predict Alzheimer's disease from genome-wide data	No	N/A	N/A

A Hierarchical Feature and Sample Selection Framework and Its Application for Alzheimer’s Disease Diagnosis	Yes	Missing genotypes were imputed, no method given.	Parameters for feature selection method were optimised using CV
Integrated higher-order evidence-based framework for prediction of higher-order epistasis interactions in Alzheimer’s disease	No	N/A	N/A
Integrative analysis of multi-dimensional imaging genomics data for Alzheimer’s disease prediction	Yes	Missing genotypes were imputed using MaCH software.	Parameters of SVM tuned using grid search
Identifying genetic biomarkers associated to Alzheimer's disease using Support Vector Machine	Yes	Those samples who had greater than 10% of samples missing were discarded.	N/A
Improving predictive models for Alzheimer's disease using GWAS data by incorporating misclassified samples modelling	Yes	Missing values replaced with median of nearest neighbours	CV was used for the tuning of lambda hyperparameter
GenEpi: Gene-based Epistasis Discovery Using Machine Learning	Yes	Missing genotypes imputed according to the 1000 genome haplotypes	N/A
Developing an early predictive system for identifying genetic biomarkers associated to Alzheimer’s disease using machine learning techniques.	Yes	Those samples who had greater than 10% of samples missing were discarded.	N/A

a – Publication title; b – Whether methods were employed to deal with missing data. c – Methods used to deal with missing data.

Eight out of 12 studies used methods to address missing data values. Two studies excluded samples with >10% missing predictor values. A further four publications described processes for the imputation of missing genotypes. For instance, Sherif, Zayed and Fakhr, 2015 imputed missing SNP values by using the expectation maximisation algorithm. Another study Romero-Rosales et al., 2020b imputed missing genotypes by using the median value of the nearest neighbours, this was the only example of using a measure of central tendency. Zhou et al., 2019 did not remove or impute missing data, rather they designed a method in which samples with missing values were incorporated in the models. All complete samples were used to develop a latent representation space. Samples with missing values were used to learn independent modality specific latent specifications. These latent representations were then used as an input for the AD classifier. This process allowed these authors to produce models which outperformed comparable methods of dealing with missing data and selecting features.

None of the analysed studies which reported the use of imputation methods specified whether this process was undertaken before CV or afterwards, which may be prone to the issue of data leakage (Samala *et al.*, 2021).

3.3.6 Hyperparameter search

Hyperparameter tuning is a common step in developing prediction models, it is implemented to ensure the optimisation of AUC (Probst, Bischl and Boulesteix, 2018). Reporting of techniques for hyperparameter optimisation was inconsistent across studies as detailed in Table 3.7, with seven publications not providing values or the process of tuning. For the remaining five studies, a range of differing techniques were used. Zhou et al., 2019 used a nested approach to optimise model parameters. Ten-fold CV was used to fit models, whilst an inner loop of five-fold CV trained model hyperparameters. However, this was only the case for some hyperparameters, as some were fixed at pre-determined values to reduce training times. This arbitrary fixing of values could introduce bias. Hao et al., 2016 also used a nested approach for hyperparameter tuning. Five-fold CV was used to optimise parameters for regularisation, with a separate loop of five-fold CV used for model validation. These were the only two studies which reported the use of nested CV for hyperparameter tuning. The remaining 3 studies reported hyperparameter optimisation but did not specify whether a nested approach was used.

Bi et al., 2019 used an iterative process to determine the optimum number of decision trees to use in their RF approach. Furthermore, grid search and CV techniques were employed to optimise varying hyperparameters across the studies. In this process, CV is used to test different combinations of hyperparameter values, with the aim of producing the set which leads to the highest value of AUC. Seven publications did not report optimisation methods. Of these seven studies, four used NB methods, which do not require hyperparameter tuning. For the remaining three studies, hyperparameter tuning was required but not reported.

3.3.7 Descriptive statistics

Descriptive details for cohorts used in the included studies are outlined in Table 3.8. Eight studies did not report values regarding both age and gender for study participants. The remaining four reported the age and gender distributions in both classes (cases and controls). De Velasco Oriol et al., 2019 reported age and gender for both the discovery and validation sets. Values for the mean age for both cases (75.4-75.5) and controls (76.1-77.4) were similar across studies. This similarity is due to the consistent use of the ADNI dataset throughout the analysed studies. The proportion of males to females in controls ranged from 0.59-1.22; in cases this proportion ranged from 1.05-1.22.

Table 3.8: Descriptive statistics if reported.

Publication title[a]	Age[b]	Gender[c]
Benchmarking machine learning models for late-onset Alzheimer's disease prediction from genomic data	Discovery dataset: Mean age – 75.57 Validation dataset: Mean age – 72.17	Discovery: Males – 252, Females – 219 1.15 Validation: Males – 92, Females - 75 1.23

Effective Diagnosis of Alzheimer’s Disease via Multimodal Fusion Analysis Framework	Cases: Mean age - 75.35 Controls: Mean age – 77.14	Cases: Males – 19, Females – 18 1.05 Controls: Males – 13, Females – 22 0.59
Latent Representation Learning for Alzheimer's Disease Diagnosis with Incomplete Multi-Modality Neuroimaging and Genetic Data	Cases: Mean age – 75.5 Controls: Mean age – 76.1	Cases: Males – 94, Females – 77 1.22 Controls: Males – 112, Females – 92 1.33
Discovering Alzheimer Genetic Biomarkers Using Bayesian Networks	N/A	N/A
The application of naive Bayes model averaging to predict Alzheimer's disease from genome-wide data.	N/A	N/A
A Hierarchical Feature and Sample Selection Framework and Its Application for Alzheimer's Disease Diagnosis	Cases: Mean age – 75.5 Controls: Mean age – 76.1	Cases: Males – 94, Females – 77 1.22 Controls: Males – 112, Females – 92 1.22
Integrated higher-order evidence-based framework for prediction of higher-	N/A	N/A

order epistasis interactions in Alzheimer’s disease.		
Integrative analysis of multi-dimensional imaging genomics data for Alzheimer’s disease prediction	N/A	N/A
Identifying genetic biomarkers associated to Alzheimer's disease using Support Vector Machine	N/A	N/A
Improving predictive models for Alzheimer's disease using GWAS data by incorporating misclassified samples modelling	N/A	N/A
GenEpi: Gene-based Epistasis Discovery Using Machine Learning	N/A	N/A
Developing an early predictive system for identifying genetic biomarkers associated to Alzheimer’s disease using machine learning techniques.	N/A	N/A

a – Publication title; b – The mean age of study participants, recorded either by case/control split or by different datasets.

c – The breakdown of males/females for participants, by case/control split or by dataset. With the ratio of males to females.

3.4 Conclusion

This review assessed a selection of studies which used ML to predict AD from mainly genetic data. Using a systematic approach (PRISMA), 12 studies were identified which met inclusion criteria. This could be perceived as a low number of studies; however, this amount is consistent with other ML reviews (Bracher-Smith, Crawford and Escott-Price, 2021). A potential reason for this small number is that ML is a relatively novel technique in AD prediction. Also, the disease risk associated with SNP data in complex genetic disorders has gained recent interest due to the appearance of GWAS, followed by prediction using polygenic risk scores (Escott-Price, Sims, Bannister, Harold, Vronskaya, Majounie, Badarinarayan, Morgan, et al., 2015). In addition, difficulties exist in accessing datasets with sufficient sample size for prediction. These manuscripts were reviewed to identify the type of models used, model development and the validity of the reported results.

AUC results in the included studies (5 out of 12) varied (0.49-0.97) for AD risk prediction. The most accurate models were shared across two studies, with the authors recording AUC >0.8, which could be considered as high (e.g., approved clinical prediction models in cardiovascular disease and diabetes typically achieve AUCs of 0.8-0.85 (Lewis and Vassos, 2020)). Given that genetic prediction for complex traits is bounded by heritability and the disease prevalence (Wray *et al.*, 2010), these results match and outperform the theoretical maximum prediction accuracy in AD using Polygenic Risk Scores (AUC=0.82, assuming SNP-based heritability h2=0.24 and life-time disease prevalence of 2% (Escott-Price *et al.*, 2017). Seven out of 12 publications did not report AUC for their models, with accuracy and sensitivity being the preferred choices. The most common measure of performance used other than AUC was ACC. Four studies reported ACC >0.8, which is considered important when attempting to reduce the possibility of miss-communicating risk to clinicians and the public. However, ACC can be skewed by the presence of class imbalances (Sun, Wong and Kamel, 2009). In addition, ACC is calculated from all predictions against all observed outcomes, although this does not clarify how the model performs per class. For these reasons we advocate that AUC should be used as a standard measure for reporting performance.

Continued research and development in the field of ML has led to an increasing number of algorithms available for use in risk prediction (Sun et al., 2019). This is reflected in the use of

10 different types of approaches across all studies, the most popular of these being Support Vector Machines. SVMs are known for their simple application and predictive accuracy and are therefore used regularly in prediction modelling (Cervantes *et al.*, 2020). Other notable algorithms used in the assessed studies were Random Forests (RFs) and Naïve Bayes (NB). Similar to SVMs, NB is known for its easy implementation. However, its performance can be hindered due to correlations between features used for prediction, which negates the naïve assumption that all input features are independent (Langley and Sage, 2013). If correlation between features is present, the importance of these features will be overemphasised during modelling (Misra and Li, 2020). Random Forests (RFs), used in three studies, are a popular classifier due to their ability to negate overfitting. However, applying RFs to prediction problems can be challenging due to the need for hyperparameter tuning (Wyner *et al.*, 2015). Given the success of the forementioned algorithms in a range of application areas, it is perhaps not surprising that these three algorithms were the most used across all publications (Pretorius, Bierman and Steel, 2016).

None of the included studies used Neural Networks (NNs) to predict AD. NNs are powerful predictive algorithms, with the ability to learn non-linear patterns in complex datasets. In some scenarios, they can infer relationships in the data which are beyond the scope of other ML techniques (Kumar, 2014). A possible explanation for their absence could be the structure of datasets used across the selected models, where the number of predictors often outnumbered individuals. In the scenario where a dataset has many more predictors than individuals, a prediction algorithm is more susceptible to overfitting (Pavlou et al., 2015). NNs are known for being complex to implement, as well as difficult for hyperparameter tuning and susceptible to overfitting (Pavlou et al., 2015). This could explain why they were not present in the reviewed studies.

Another potential reason for the absence of NNs in this review is the omission of the term from our keyword search, that is we searched for the term machine learning, rather than specific ML techniques. This could be purported as the main limitation of this review as some research papers might have been mistakenly excluded. A subsequent search for the use of NNs for AD prediction returned a study (de Velasco Oriol, 2019), which used deep NNs to predict AD from SNP data. Using the ADNI dataset, the authors conducted several experiments to predict case-control status. A standard architecture was implemented for the NN, along with 5-fold CV for model validation. Results for the NN across experiments centred around 65% AUC. However, this paper would not have been included in the review due to it being a pre-print, and therefore lacking a peer review. A secondary study using NNs was also found, that used SNPs and MRI data from ADNI (Zhou, Thung, *et al.*, 2019). The authors developed a novel stage-wise deep learning framework, which fused multimodal data in stages. This method achieved a classification accuracy of 64.4%.

Greater focus in recent years has been given to the possibility of bias when authors introduce novel concepts. For instance, authors may aim to achieve the best prediction accuracy possible in order to supersede previous publications. This may have been achieved by choosing datasets which produce the best accuracy only, leading to a lack of generalisation in the research area. This possibility has led to comparative studies which draw comparisons between novel techniques and historic models (Hand, 2006).

A number of consistent issues were highlighted across the included studies. One of the main focus points was the widespread usage of the ADNI dataset, where 10 of the 12 included studies used this as a data source. Methods used to demonstrate model performance were reported inconsistently. The combination of low EPV values and inconsistent model performance reporting led to the possibility of bias in the analysis phase of modelling. In terms of model implementation, the main aspects scrutinised were the use of any hyperparameter tuning, as well as the methods used for model validation. Hyperparameter tuning has become an increasingly important part of ML development. The majority of algorithms require certain values for hyperparameters which are specified by the user. If these values are not optimised, then the model is susceptible to overfitting and inaccurate predictions (Weerts, Mueller and Vanschoren, 2020). Five out of the 12 studies referenced the use of hyperparameters, the remaining 7 studies did not outline any tuning methods. Greater transparency about the use of hyperparameters and their tuning allows the reader to understand whether issues such as overfitting were accounted for. Therefore, researchers should report both hyperparameter values and methods used to obtain them.

Model validation is also an important aspect of predictive analysis. Correct methods of validation reduce the likelihood of overfitting, whereby algorithms become too reliant on the training/test data and cannot perform sufficiently when tested on unseen data (Vabalas *et al.*, 2019). The most commonly used method among the selected studies (11/12) was CV. This method has become increasingly popular in prediction models, due to its ability to counteract overfitting (Ghojogh and Crowley, 2019). Eleven of the 12 studies which reported CV used a varying number of folds, whilst one of these publications used a technique called leave one out CV. In the majority of cases, the higher the number of folds, the greater the accuracy from CV. However, increasing the number of folds leads to a higher chance of overfitting (Ghojogh and Crowley, 2019). Therefore, leave one out CV is only suitable for small datasets, where the number of samples is <100 (Yadav and Shukla, 2016). Nested CV was used by two studies only. These were the only evidence of using separate validation folds for both model optimisation and hyperparameter tuning throughout all included studies. Using the same CV split for both of these tasks can introduce overfitting (Varma and Simon, 2006), therefore we recommend the use of nested CV for future analysis. The only publication which did not report CV used a train and test split method for internal validation. The model is trained only once, increasing the chance of a model becoming too reliant on the training data and thereby reducing its ability to replicate in independent datasets. Since the split of the data is conducted randomly, an argument could be made that the derived results could be influenced by this single split (Ibrahim and Bennett, 2014). Therefore, methods which use a form of CV are recommended.

Calibration compares the similarity of probabilistic predictions with observed outcomes. This metric was only reported in one study (Wei, Visweswaran and Cooper, 2011a). As described in Chapter 2 (methods chapter), calibration is of high importance when assessing ML performance, this is especially true when considering models which may be implemented in the medical sector (Steyerberg *et al.*, 2010). The implications of incorrectly communicating the risk of developing AD to an individual could cause considerable harm, by means of both physical and psychological trauma. With the potential of causing death due to incorrect treatment in the most serious of circumstances (Park and Ho, 2020). Therefore, we recommend that authors aim to produce highly calibrated models and also report calibration statistics.

Another aspect investigated in this review was the sample size used in the training of models. These were relatively small with most studies using between 300-900 individuals (due to the common use of the ADNI dataset). Different quality control techniques also resulted in the number of predictors (SNPs) to vary across publications, ranging between tens of SNPs to over 100,000. The combination of small number of samples and large number of predictors led to low EPV scores, the highest of which was 9.43 in (Chang *et al.*, 2020). The common use of ADNI also contributed to low EPV values due to the consistent implementation of small numbers of participants and high numbers of predictors. A more commonly known term for low EPV values is the 'curse of dimensionality'. This refers to the requirement for more training data when the number of features is increased. If the number of samples is not sufficient with respect to the number of features present, an ML algorithm is more likely to overfit. The number of samples therefore must increase at a certain rate in order to balance this relationship. Low EPV values suggest this balance has not been achieved (Verleysen and François, 2005a).

One method for dealing with a large number of features and the issues that this could cause, is feature selection. An example of this is Minimum Redundancy Relevance (mRMR). This method is widely used in genetic studies (Radovic *et al.*, 2017). In mRMR, features which are significantly correlated with the target variable are identified and this subset is then filtered further based upon correlations between features, with heavily correlated features being discarded. However, this method was used in only one (de Velasco Oriol, Edgar E. Vallejo, et al., 2019) of the 12 studies reviewed. To summarise, all EPV scores were below the threshold recommended by PROBAST. Small sample size may be a difficult issue to overcome therefore, it is advisable to use CV to reduce the impact of possible overfitting. Further techniques, such as nested CV have been shown to mitigate overfitting more effectively (Vabalas *et al.*, 2019). We therefore encourage authors to investigate which type of validation technique would be suitable for their models.

This review aimed to assess ML models which used SNP data for AD prediction. Of the 12 studies reviewed, eight used SNPs only, and the remaining four combined SNPs with other data modalities. In terms of AUC, it appears that using a multimodal approach may lead to better prediction performance. For example, An et al., 2017 have shown that AUC was

85.5% for SNPs alone and 97.4% when both SNP and MRI data were considered together. However, for the studies that reported ACC only, there appears to be little difference in predictive performance between those which used SNPs only and those which used a multimodal approach.

When considering other factors which may cause differences in prediction performance, class imbalances appeared to have a negligible effect. Extreme values of class imbalance did not lead to largely different accuracy results. Class imbalances can lead to poorer prediction due to the model favouring the majority class. Techniques such as under/over sampling can be used in order to overcome this issue. Between the two methods, under sampling has been found to be more effective in addressing predictive bias (Blagus and Lusa, 2013). This is due to a common issue amongst over sampling algorithms, in which the creation of synthetic minority samples can introduce noise to the data (Jiang *et al.*, 2021). The issue of class imbalance was not of major concern in the reviewed papers, however with the availability of large population cohorts (e.g., UK Biobank), care should be taken when analysing diseases with small prevalence, which includes AD and other dementias.

Data leakage is another issue to be considered. It occurs when an algorithm's performance is artificially inflated due to information being leaked from the training to test dataset. Manipulating data before training and validation may inadvertently leak information and boost performance. A way in which this can occur is pre-processing on the entire dataset before data is split. This is relevant to imputation of missing values, derivation of and adjustment for population structure. In order to avoid this, any pre-processing steps should be carried out separately in both the training and test datasets (Samala *et al.*, 2021). To achieve non-biased results, an ML algorithm should always be validated on data separate to training data. Nested CV can be used to ensure pre-processing is carried out per fold, as this reduces the risk of data leakage (Parvandeh *et al.*, 2020).

ROB in the remaining three sections of PROBAST (participants, predictors and outcome) was considered to be low for all publications. The usage of cross-sectional data reduced the ROB for the study participants. The use of a well-documented dataset (ADNI) provided details in areas such as predictor collection, the determination of disease status and inclusion of individuals in these studies. These areas could not be assessed in the two studies which did not use ADNI. The widespread use of ADNI also provided the possibility of comparison between studies due to the common data samples, however this prevented the possibility of performing a meta-analysis. The use of a range of data sources in future studies would be beneficial for the development of ML models and is likely to improve their robustness and replicability. In particular, the continued use of the same resource does not provide insight into the performance of ML in different populations. If used in frontline medicine, models will have to be able to predict upon individuals from different genetic backgrounds (Martin *et al.*, 2017). For instance, 93% of the participants of ADNI are Caucasian (R. C. Petersen *et al.*, 2010b). It has been shown that GWAS results from primarily Caucasian subjects do not replicate well in other races, which may also impact the prediction success of ML algorithms trained on them (Haga, 2010). Overall, despite ROB being low for the first three sections of PROBAST, issues within the analysis phase of modelling introduced possibilities of bias. This could bring the validity of the results into question.

Reviews in the field of ML for AD prediction have been previously conducted. Tanveer et al., 2020 conducted a comparison between three different ML techniques (SVMs, NNs and ensemble methods). The type of data used was imaging only, leading to a greater number of included texts. Comparisons were drawn between the methods but further detail on ROB was not included. Khan, 2015 also conducted a review into ML prediction for dementia which included models using imaging data. In their review a large percentage of studies used ADNI as their data source, and their results and conclusions follow a similar pattern to this review, however the authors did not formally assess ROB.

A potential limitation of this review is the exclusion of models predicting between MCI individuals and AD cases. It is known that not all those who experience mild cognitive decline develop AD (Knopman and Petersen, 2014). However, it could be argued that future diagnostic tools will be used to predict disease likelihood in both cognitively healthy individuals and MCI examples. Therefore, models will be required to distinguish between cognitively healthy, MCI and AD patients. When considering MCI however, differentiating between this and AD status has proven to be difficult (Jekel *et al.*, 2015). It was for this reason that attention was only given to examples of prediction between cognitively normal and AD patients.

This review has highlighted a number of areas which require improvement in the field of ML for AD prediction using genetic data. Some areas require greater attention than others, namely the reporting of model performance and development. Reporting these measures thoroughly will allow for an accurate comparison between studies and provide better clarity for the performance of the models. More detailed description is also required when explaining model implementation, with special emphasis on hyperparameter tuning. This will provide greater understanding of how authors have attempted to maximise performance and reduce the possibility of overfitting. Furthermore, the majority of studies in this review used the publicly available ADNI dataset, which demonstrated a clear overreliance on one particular data source of Caucasian origin. Using a wider range of data sources would enhance the validity of results and also develop understanding of the applications of ML for AD prediction in more diverse populations.

3.4.1 Conclusion

In conclusion, ML will continue to be used more extensively in both academia and the industry due to its ability to analyse complex patterns in datasets, which will allow users to achieve better risk prediction as compared to more classical statistical methods . The continued usage of ML will boost the development of feature selection techniques and lead to improvements for classification and model optimisation algorithms. These models have great potential to improve clinical risk prediction for AD, and many other complex genetic diseases. Since genetic data is classed as sensitive data under General Data Protection Regulation, most of the large genetic datasets require strict permissions and exact description of usage. UK Biobank is one of the largest cohorts, however it may not be suitable for application of ML to AD, as it is a population-based cohort with relatively young participants. The Dementias Platform UK (DPUK) (Bauermeister *et al.*, 2020) is an attempt to provide a secure computational platform collecting genomic data from UK cohorts suitable for dementia research. The future of artificial intelligence (AI) applied to large genomic data lies with specifically designed secure computing facilities to store and analyse the sensitive data.

This chapter was published as a systematic review (Rowe et al., 2021a). Analyses conducted in this thesis will be guided by some of the issues highlighted in this review. For instance, model performance shall be reported using AUC in all cases. Emphasis will also be given to providing calibration statistics for algorithms, as this review revealed a consistent lack of reporting of these. Another important factor which shall be implemented is the use of nested CV when tuning algorithm hyperparameters, with the intention of avoiding inflated prediction results. This process will be clearly outlined, as the majority of studies included in this review failed to detail whether models had been developed in this manner.

4 Assessment of machine learning versus PRS for the prediction of AD

4.1 Introduction

Analyses conducted in Chapter 4 focused on evaluating initial comparisons between polygenic risk score (PRS) and a selection of machine learning (ML) algorithms to predict Alzheimer's Disease (AD). As discussed in the Introductory chapter, AD exists in a number of different forms. Analyses in this chapter and throughout the thesis predicted examples of the sporadic version only, however both late and early onset cases were present. Genetic variants used as predictors were those deemed GWAS significant in Kunkle 2019 (Kunkle *et al.,* 2019). Methods to control for confounders such as population stratification, age and sex were also investigated. Calibration statistics for ML models were also reported, with methods to improve calibration also explored.

A commonly used method for genetic prediction of a phenotype or a disease risk is the PRS (described in Section 1.10.1, Chapter 1) (Collister, Liu and Clifton, 2022b). Disease risk predictions derived from a regression model are used to discriminate between cases and controls, with accuracy often measured by means of the area under the curve (AUC). When assessing the neurodegenerative disorder AD, empirical evidence has suggested that AD is a polygenic disease (Escott-Price et al., 2015) with its genetic component most likely the result of many mutations, with potential gene-gene interactions (Zhou *et al.*, 2021). Accuracies differ across psychiatric and neurodegenerative disorders. For instance, studies have produced AUCs of 70% for schizophrenia (Calafato *et al.*, 2018), up to 82% for AD (Leonenko, Shoai, *et al.*, 2019) and ~58% for major depressive disorder (Fullerton and Nurnberger, 2019). A potential drawback of linear modelling (and in turn PRS) is the inability to assess complex non-linear relationships between SNPs, leading to the possibility of reduced accuracy when predicting AD by genetic variants (Slunecka *et al.,* 2021).

Machine learning classifiers function by learning complex patterns between inputs and outcomes, with in turn could lead result in the assessment of complex non-linear relationships between SNPs. Similarly, to PRS, ML models using SNPs as inputs have been previously used to make predictions of disease risk. For example, Wei *et a*l., 2013, used penalised regression to predict disease status for Crohn's disease, a chronic bowel condition. A total of six thousand SNPs were used for predictors, with a discrimination value of 83% AUC achieved. SNPs have also been used for the prediction of celiac disease, in which six European cohorts were combined for analysis. Prediction accuracy of 90% AUC was reported when using a penalised support vector machine. In terms of neuroscience, genomic variants have been used to make predictions for various diseases. Yang *et al.,* 2010 combined SNPs with neuroimaging features to predict schizophrenia, with a support vector machine achieving a prediction of 87% AUC. A number of studies have used ML to predict AD from genetic data. Algorithms including the random forest, support vector machine, penalised regression and k-nearest neighbours achieved between 60-70% AUC. Despite the growing interest in ML applications, the number of research articles investigating ML for prediction of AD using SNPs is still relatively small. The majority of studies focus on the use of other data sources such as brain imaging data (Rowe *et al*., 2021b).

As mentioned previously, interest in disease prediction by both PRS and ML has increased recently. However, methodological issues have also been discovered. When assessing the genetic prediction of AD, population stratification and the difference in age of cases and controls have been identified as a potential limitation for ML. These differences may lead to false positives if not controlled for, as ML algorithms might predict based upon age (or population stratification) and not the underlying genetic structure (Le Guen *et al*., 2021). Techniques to control for the effect of confounders on AD prediction will be explored in this chapter.

Efforts have been made to compare the disease prediction capabilities of ML and PRS. As stated previously, a limitation of PRS is its use of linear modelling, which renders the method unable to assess complex non-linear patterns between SNPs. In contrast, ML algorithms can learn complex data interactions between features (if these exist). This advantage should allow ML to achieve greater prediction accuracy for complex disorders (Mena Mamani, 2020). This has proven to be correct for a number of diseases (Elgart *et al*., 2022), however published results have also demonstrated superior prediction for PRS in some instances (Gola, Erdmann, Müller-Myhsok, *et al*., 2020). A reason for this might be the requirement of large amounts of training data for ML to assess non-linear relationships. Such volumes of data are not always available due to difficulties in collecting samples and costs associated with sampling (Medina-Gomez *et al.*, 2015). A definitive conclusion cannot yet be reached as to which is the superior method, as results appear to be dependent on the phenotype in question (Gola, Erdmann, Müller-Myhsok, *et al.*, 2020).

Bracher-Smith *et al.*, 2022 compared the predictive capability of ML versus PRS using logistic regression (LR) for Schizophrenia. Detailed analyses identified that ML did not improve prediction accuracy over and above LRs. However, comparisons of PRS based LRs and ML algorithms trained using SNPs are yet to be drawn for AD prediction. Given this, the central aim of this chapter is to compare the accuracy of PRS-based regression models and ML algorithms for AD prediction, with the intention of providing novel insights into whether ML can outperform the use of PRS. This investigation is motivated by the variability of accuracy when using PRS for disease prediction in the medical sector. Poor accuracy can result in incorrect decisions being made when treating individuals, increasing the possibility of harm for patients (Kelly *et al.*, 2019). Therefore, health services only employ predictive algorithms which have passed stringent testing (Kumar *et al.*, 2022). Given this, it is hoped by many in the genetics community that ML methods will be able to outperform PRS in predicting life changing disorders such as AD (Ho *et al.*, 2019b).

Main aims:

- To compare the performance of PRS and ML for the prediction of AD.
- Explore techniques to correct for the confounding of disease risk.

4.2 Materials and methods

4.2.1 Data

Data used in this chapter were taken from the GERAD consortium GWAS (Harold *et al.*, 2009). The study comprised 6980 cases and 12022 controls. The Illumina 610-quad chip (610 array) was used to genotype 4,113 cases and 1,602 controls. A further 844 cases and 8,080 controls were previously genotyped using either the Illumina HumanHap550 (550 array) or Illumina HumanHap300 chips (300 arrays). These previously genotyped samples were from seven different studies, making eight studies in total: 1) 610; 2) Mayo; 3) 1958 birth cohort (sanger); 4) 1958 birth cohort (T1DGC); 5) ALS control; 6) Coriell control; 7) Heinz Nixdorf Recall (HNR) study; 8) KORA F4. Due to the use of different genotype arrays, stringent quality control (QC) methods were applied, in order to minimise the difference in genotyping error rates between groups. None of these steps were conducted by myself and were carried out prior to this thesis.

Only autosomal SNPs were included in the study, and can be split into four different categories: 1) 266,714 SNPs common to all three arrays and all genotyped individuals; 2) 202,516 SNPs common to the 610 and 550 arrays, but not present in those genotyped using the 300 arrays; 3) 7,744 SNPs common to the 610 and 300 arrays, but not present in those genotyped on the 550 arrays; 4) 105,614 SNPs genotyped only on the 610 arrays. Further QC steps carried out in (Harold *et al.*, 2009) on both samples and variants are described in greater detail in Table 4.1.

Table 4.1: A description of the QC procedures used on both samples and SNPs.

QC Stages-Samples	Individual Sample Exclusions	QC Stages-SNPs	SNP Exclusions
Missing Genotypes	1,469 samples were removed due to missing genotype rate > 0.01.	MAF and Hardy Weinberg Exclusion	SNPs were excluded with a MAF of < 0.01 and Hardy Weinberg P-value < 1x10-5.
Mean autosomal heterozygosity	578 Individuals removed	Further MAF exclusion	For SNPs with MAF > 0.05, these were excluded if their genotype missing rate was > 0.03 in both case and controls. For those SNPs with MAF between 0.01 and 0.05, variants were removed if their genotype missing was less than 0.01. These steps resulted in 43,542 SNPs being excluded.
Inconsistent Gender report	A further 71 samples were removed due to incorrect gender.	Addressing the usage of inter-chip and inter-cohorts	To achieve this, minor allele frequencies were compared between controls in the different groups.

			Logistic regression analysis was used, with the previously determined four PCs used as covariates. These comparisons were made only on individuals from the same geographical areas. Quantile – quantile plots were used to compare cohorts, with chi squared statistics used to determine whether SNPs were to be excluded or not. A further 9,828 SNPs were removed due to this
Inconsistent Gender by genotypes	93 individuals were removed due to genotypes not matching specified gender.	Final number of SNPs following all QC stages.	529,218 autosomal SNPs for analysis
Identity By Descent (IBD), Hardy Weinberg rate and MAF > 0.01	An individual from a pair was removed if a value of 0.125 IBD or above was calculated. This combined with a Hardy-Weinberg rate of p> 1x10-5 and a minor allele frequency of > 0.01, led to the removal of a further 506 samples.		
Detecting individuals of non-European ancestry	Genotypic data from the remaining individuals were merged with the same SNPs from 210 unrelated individuals from the HapMap project. IBS distances were again computed for all pairs of samples. The values from this were then used as an input matrix for multi-dimensional scaling (MDS). Plots from this identified three clusters of European, Asian, and Yoruba samples. A further twelve individuals were identified as outliers from this and removed.		

Population Structure assessment using PCA	PCs were computed on 57,966 SNPs common to all arrays used. PCs were calculated and the EIGENSTRAT program was used to identify outliers. 188 individuals were identified as outliers and subsequently removed.		
Resulting number of Samples after all QC stages	Following these QC measures, a total of 3,491 AD cases and 7,488 controls remained.		

Table 4.1. The first column related to QC stages carried out for samples. The second column describes these stages in detail. The third column outlined QC stages for SNPs, whilst the final column explains these steps in detail.

As discussed during the introduction chapter, AD can be separated into early onset (EOAD and late onset (LOAD) forms of the disease depending on the age of the patient. EOAD can further be split into the sporadic and mendelian versions of the disease, whilst late onset comprises the sporadic form only. Most cases present in the GERAD dataset are examples of the late onset sporadic form of the disease. However, early onset sporadic cases were also included. In order to be defined as a case, patients were required to pass one of the national institute of neurological and communicative disorders and stroke (NINCDS) (Jack *et al.*, 2011), the diagnostic and statistical manual of mental disorders (DSM-IV) (GUZE, 1995) or the consortium to establish a registry for Alzheimer's disease (CERAD) (Fillenbaum *et al.*, 2008) criteria.

Although the original source of data for analyses in this thesis was Harold et al., 2009, the exact dataset used was from Leonenko, Sims, *et al.*, 2019. In which samples from the GERAD dataset were used to predict lifetime risk of AD development. The following table provides an outline of all included cohorts and a range of descriptive statistics.

Table 4.2: A description of the cohorts present within the dataset.

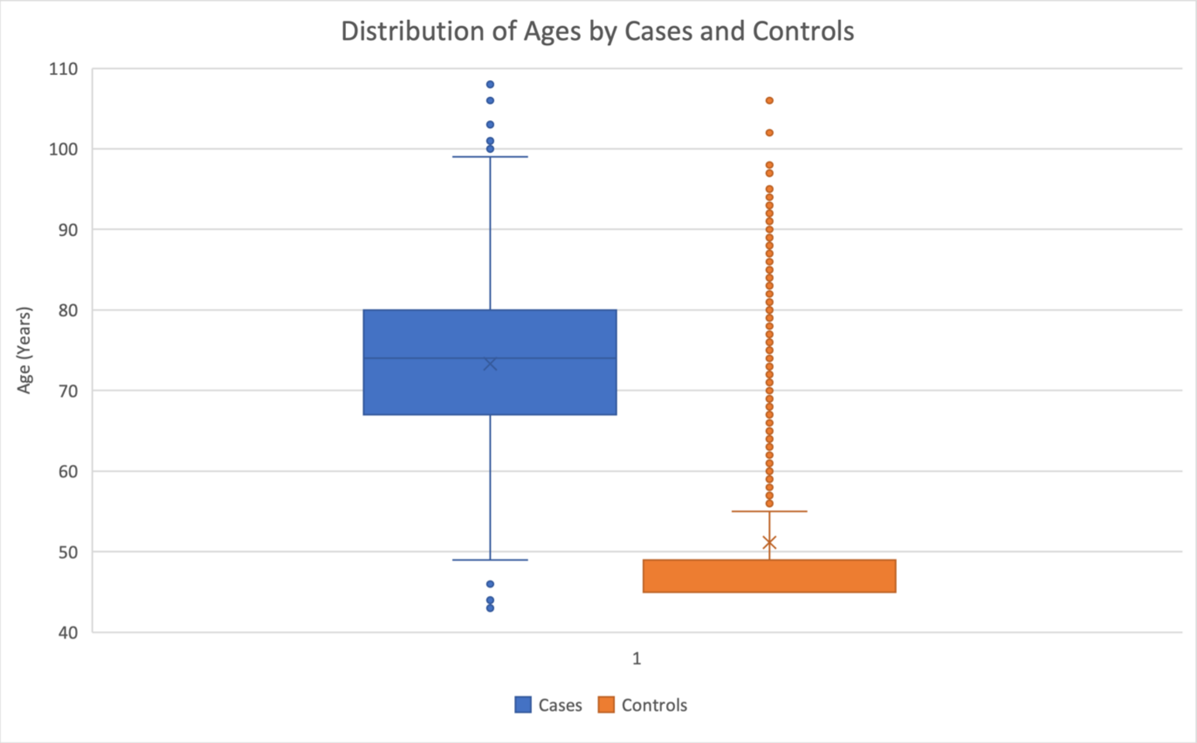
Statistics/Cohort	MRC	ART	BONN	WASHU	NIMN	UCL-Laser	UCL-PRION	HNR	1958BC	KORA	Total
Cases	1009	960	555	423	127	47	211				3332
Female (%)	63.7	60.9	63.8	56	63	74.5	58.3				
Mean Age	77.9	76.6	76.6	82.1	80.1	80.6	63.6				
Controls	873	82	37	233				353	5343	434	7355
Female (%)	51.6	59.8	64.9	66.5				53	49.8	49	
Mean Age	51	77.9	79.5	78.5				54.6	45	56	
Geographical Region	UK/Ire	UK	Germany	USA	USA	UK	UK	Germany	UK	Germany	
Illumina Chip	610	610	610	610	610	610	610	550	550	550	

Table 4.2. Statistics are separated by cases and controls, with values for the different cohorts given in each column.

4.2.2 Sample selection

This sample size used for analyses comprised 10687 people, of which 3332 are AD cases and 7355 are controls. Box and whisker plots shown in Figure 4.1 demonstrate the distribution of age within both cases and controls.

Figure 4.1. Distribution of age within GERAD.



When observing the box and whisker plots in Figure 4.1, it is clear the distribution of ages for cases is different to that of controls. The mean age of cases is 75.3 years, with a median age of 77. In contrast, the mean age of controls is 51.2 and a median value of 45. The diagram also outlines that controls over the age of 55 are considered outliers. The difference in age distribution between the two sets can be explained by the presence of the 1958 birth cohort which contains 5443 controls aged 45 years.

The following figure demonstrates the breakdown of gender within the GERAD data set.

Figure 4.2. Distribution of gender within GERAD.

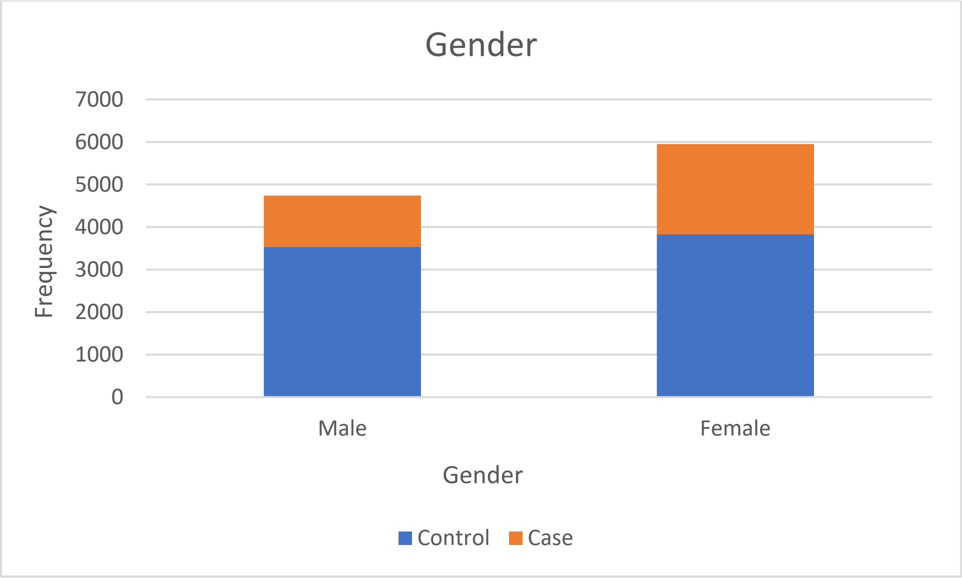
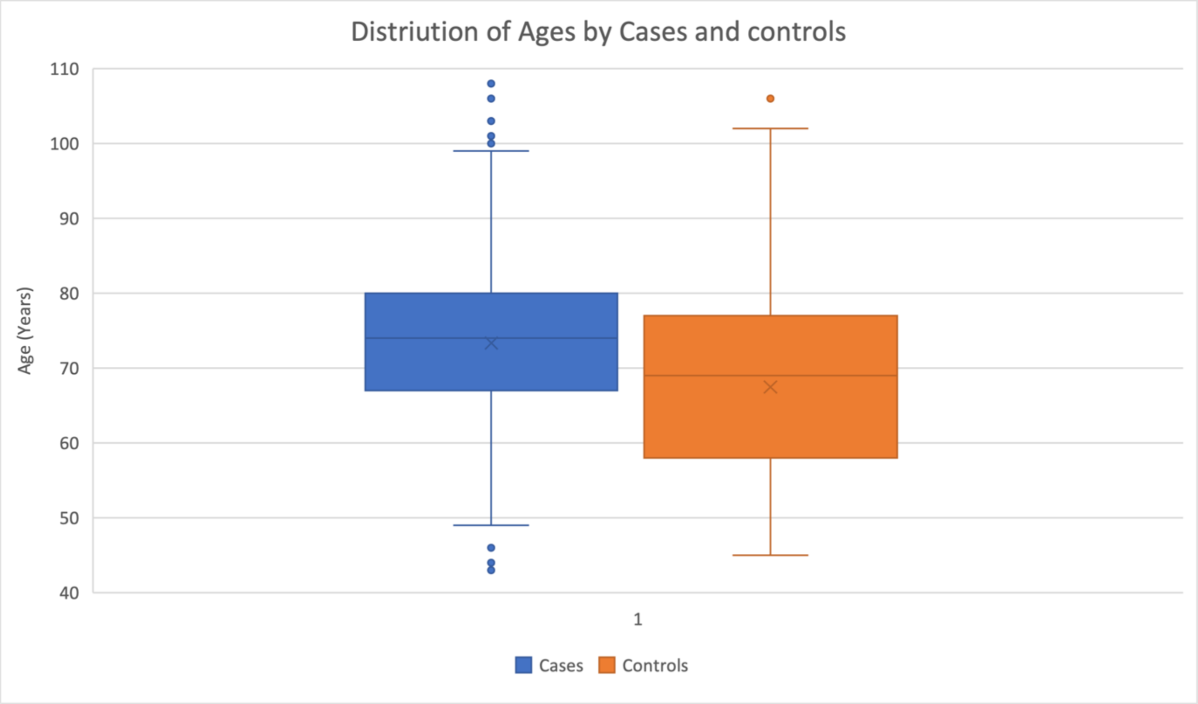


Figure 4.2 shows the greater number of females than males within GERAD. When assessing the breakdown of cases and controls by gender, controls consist of 48.0% males and 52.0% females, whilst cases comprise 36.3% males and 63.7% females.

For all analyses in this thesis, individuals within the 1958 birth cohort were also removed in order to make the age of distributions for cases and controls more similar. The result of this on the distribution of ages in both cases and controls is shown in Figure 4.3.

Figure 4.3. Distribution of age within GERAD, with the 1958 birth cohort removed.



The box and whisker plots shown in Figure 4.3 demonstrate a change in the distribution of ages in controls when compared to Figure 4.1. The spread of age for controls is now more in line with that of cases, with a lesser number of outliers. The mean age of controls is now 67.5 compared to 51.2 previously. This final dataset consisted of 3332 cases and 2012 controls (this is discussed further in Section 4.5.2).

4.3 Methodology

4.3.1 Selecting SNPs for analysis

Independent genome-wide significant SNPs (p-value $< 5x10^{-8}$) (N SNPs=23) as reported in Kunkle et al., 2019 are likely to be strong predictors of AD due to being the most statistically significantly associated. Therefore, these 23 SNPs were used as predictors for all classifiers. None of these 23 SNPs were present in the GERAD (Harold *et al.*, 2009) dataset. This absence is likely be due to the genotype data used in this chapter not being imputed. In order to overcome this mismatch, each of the 23 SNPs were used as index SNPs and were subsequently used to find a set of proxy SNPs, which were available in GERAD (Harold *et al.*, 2009).

The SNP in GERAD most closely correlated with each GWAS index SNP was identified. Two separate methods were used to achieve this. In method one, R-squared (r^2) values between each index SNP and all variants within a 1MB window around the index SNP were calculated using an online calculator https://grch37.ensembl.org/Homo_sapiens/Tools/LD. The SNP in each window with the highest r^2 with the index SNP was chosen as the proxy variant. The second method for selecting a set of proxy SNPs included the use of clumping. Clumping is a method which calculates 'clumps' (blocks of SNPs in linkage disequilibrium (LD)) in a predefined genomic window. These blocks are formed around index SNPs, who are deemed significant based on an initial p-value threshold (p-values provided by summary statistics). Those SNPs surrounding the index SNP are then pruned based upon an r^2 threshold with relation to the index SNP. Variants whose r^2 values are greater than this threshold are removed, with those below the threshold remaining. Index SNPs are also retained within each clump.

The GERAD (Harold *et al.*, 2009) dataset was clumped using the Kunkle-noGERAD summary statistics (18,805 cases and 34667 controls), in which all samples from GERAD were excluded to achieve independence. Clumping was performed twenty-three separate times, in which the chromosomal position of a different Kunkle SNP was used on each occasion. The GERAD SNP with the lowest p-value in each of the 23 sets was then chosen as the proxy variant. For each separate action of clumping a p-value threshold of 0.1 was used. This was alongside an r^2 threshold of 0.1 and window of 500 kilobases. Clumping was achieved using the genome association analysis toolset *PLINK* (Purcell, Neale, Todd-Brown, Thomas, Manuel A.R. Ferreira, et al., 2007). This software provides the function --*clump*, which achieves the clumping process.

4.3.1.1 Further SNP selection

The *APOE* gene has been strongly associated with AD risk (Safieh, Korczyn and Michaelson, 2019). To assess how the PRS and ML algorithms would perform without the *APOE* influence, SNPs in the associated region (19:45409039-45412650) were removed. Therefore, four sets of SNPs were included in the analysis, two sets including the *APOE* region and two without.

Alongside these sets of variants, a less stringent p-value threshold was used to select a larger set of SNPs for analysis. Reasoning for this centred on the possibility that the genome-wide significant SNPs reported by Kunkle et al., 2019 may not represent the entire genetic component of AD. In order to achieve this, GERAD SNPs (Harold *et al.*, 2009) were clumped using Kunkle-noGERAD (Kunkle et al., 2019). To generate a larger SNP set for analysis in this chapter, a p-value threshold of 0.01 was used. This was accompanied by an r^2 threshold of 0.1 and window of 500 kilobases (Privé *et al.*, 2019). A total 422 SNPs resulted following this process.

4.3.2 Cross-Validation

Cross-Validation (CV) is a resampling technique used in ML and other statistical models to reduce the chances of overfitting. Analysis in this chapter used the validation technique nested CV, using the *StratifiedKFold* function from the *Python* package *sklearn*. The function generates k partitions, with a defined training and test set. This method was chosen as it allows for manipulation of both the training and test folds, for the purposes of pre-processing data. For ML algorithms to achieve a true estimate of prediction accuracy, all data within the test/validation set must be kept separate from model training (Hillel *et al.*, 2021). One of the most common practices which can lead to data leakage in ML, is the manipulation of data before splitting into training and test sets, with a common example being the standardisation of values (Hannun, Guo and van der Maaten, 2021). The ability to manipulate both training and test folds separately decreases the risk of data leakage, as pre-processing can be carried out separately in each fold.

The ‘stratified’ nature of the process ensures a consistent number of both cases and controls throughout folds of CV. In practice, any number of k folds can be used, however the most used values are five and 10 (Krstajic *et al*., 2014). Research has shown that little difference occurs in classifier performance when using either value, however five rounds of CV require less computational resources than 10 (Krstajic *et al*., 2014). Therefore, it was determined that five rounds of CV were sufficient for modelling. The ‘nested’ aspect of this method refers to the tuning of hyperparameters.

4.3.2.1 Imputation of missing genotypes

The stratified nature of the CV method used for analyses allowed for pre-processing steps to be carried out separately in training and test folds. Missing genotypes in both sets were imputed by calculating the modal genotypic value for each SNP respectively. This was achieved using the *mode* function from the *Python* package statistics. Critics of this technique state however, that imputed values are an incorrect representation of the population values thereby reducing variance within the data set (Khan and Hoque, 2020). Given the small number of missing values, it was decided that such an impact would be minimal due to the adjustment of variation being minor.

4.3.2.2 Correction for population stratification

It is widely accepted that SNP associations in GWAS should be adjusted by the inclusion of principal components (PCs) in regression models to account for population differences and other potential confounders e.g., genotyping platforms. This approach can also be used for adjusting PRS values as used in Escott-Price et al., 2015. For the purposes of ML, it is not clear how to adjust genotypes for confounders as the inclusion of PCs alongside SNPs will be considered as extra predictors by the algorithm. Therefore, PCs were used to adjust genotypes of each SNP (predictor) prior to ML (Price *et al.*, 2006). Regression modelling was used, in which genotypes for each SNP were response variables and three PCs were used as explanatory variables. The residuals from each model were extracted and standardised, with the resulting. Z-scores used as input values instead of genotypes to ML algorithms. The same method was used to adjust PRS values, with the respective Z-scores used for prediction.

PCs were generated within *Python* scripts with values derived for samples within the training and test set respectively. The function *PCA* from the package *sklearn* was used to achieve this. The *PCA* function was fitted on the training set, this same function was then used to create PCs in both the training and test sets. It was decided to use three PCs for adjustment purposes, as this has been the amount used in previous studies using the GERAD dataset (Leonenko, Sims, *et al.*, 2019). In line with other pre-processing techniques, the adjustment of genotypes for population stratification was carried out independently in the training and test sets. A linear regression was fitted using the *LinearRegression* function from *sklearn* in *Python*. Each SNP was adjusted on an individual basis. The response variable was the genotype values of the respective SNP, with explanatory variables comprising PCs. The residuals of each regression were then extracted for use in ML.

4.3.2.2.1 Adjusting genotypes for age and sex

As age and sex may be (statistically) related to AD status, increasing the likelihood of a confounding relationship, the adjustment of both genotypes and PRS for these factors might remove the impact of confounding from analyses. Age information was not available for some individuals (N= 233). We imputed the missing values using the mean age of existing values, with cases and controls not imputed on a separate basis. Similarly to missing genotype values in Section 4.3.2.1, the use of mean imputation was used due to the small percentage of missing values and subsequent minimal effect on variation. Following the imputation of missing age, both age and sex were added as further variables to the adjustment similar to adjustment for PCs.

4.3.2.3 Scaling values

The last pre-processing step was to standardise the resulting residuals using the *Python* function *StandardScaler*. The function was fitted on the training set for adjustment, with the same function was then used to adjust the test set. The resulting values were then used for the training and validation of ML algorithms. PRSs generated from genotypes in both the training and test sets were also regressed using respective PCs, age and sex. Residuals from these adjustments were then standardised separately. Overall accuracy for ML algorithms was averaged across performance in the five CV test sets. This was also done for PRS, in which a logistic regression was used to assess case/control discrimination.

4.3.2.4 Calculating polygenic risk scores

To generate risk scores from the SNPs for samples in GERAD, score files from Kunkle-noGERAD summary statistics were compiled. The data set included four columns for each variant, namely the SNP ID, reference allele, effect size and p-value. *PLINK* was used to generate risk scores. The *PLINK* parameters used were *–bfile* and *–score*, which indicate the original dataset and score (GWAS summary statistics) files, respectively. The output of these functions are PRSs generated at a pre-defined set of p-value thresholds. The PRS values calculated at the optimal threshold are then used (this is explained further in the results section). Once calculated, the PRS values were then adjusted for PCs and standardised, as described previously. Once PRS values were calculated and adjusted, these were then used for disease prediction purposes.

4.4 Machine learning methods

To develop a broad understanding of how effectively ML predicts AD, a range of algorithms were used. These included Random Forests (RFs), Support Vector Machines (SVMs) with two different kernel methods, Gradient Boosted Trees (GB) and Naïve Bayes (NB).

4.4.1 Random forests

The *Python* function RandomForestClassifier from the package *sklearn.Ensemble* was used to build RFs, with the classifier instance chosen due to the need to predict between cases and controls. Hyperparameters are values which are specified prior to the training process. *Python* packages use default values; however, these can be optimised to maximise classifier performance. Four hyperparameters were chosen to be adjusted, these were *max_depth*, *random_state*, *min_samples_leaf* and *min_samples_split*. The hyperparameter *max_depth* is the maximum number of levels in each tree allowed and random state controls the random seed used by the algorithm. The parameters *min_samples_leaf* and *min_samples*_split are the minimal number of samples required to be at a leaf node and the minimum number of samples required to split an internal node respectively. To address class imbalances, the RF hyper-parameter *class_weight* was used to redistribute class weights. The option balanced was used, which determines class distributions in the training set and inversely upweights for the minority class. This assigns both classes equal weighting for prediction purposes.

The function *RandomizedSearchCV* was used to determine the optimum values for all five hyperparameters. This functions by assessing the performance of all combinations of parameter values from a user defined sample. The hyperparameters values which produce the RF with the highest AUC are chosen. These parameters were derived using the nested CV. Once the optimum set of hyperparameters are identified using five rounds of inner CV, the RF is then validated using the outer loop of CV.

4.4.2 Gradient boosting

Python's function XGBClassifier was used to develop GB trees. *RandomizedSearchCV* was used to find the optimal values for both *max_depth* and *n_estimators*. GB includes a further parameter known as *learning rate*, determining how quickly the model corrects training error from one tree to another. This parameter was also learned through the *RandomizedSearchCV* method. Similarly, to RFs, a hyperparameter was used to counter the effect of class imbalances. The function *XGBClassifier* includes a hyper-parameter called *scale_pos_weight*. A weight must be passed to the function to redress the imbalance. This is calculated by dividing the number of minority class instances in the dataset by the number of majority samples, with the result multiplied by 100. In the case of the dataset used in this chapter, the result was 66. Models were trained and validated in the same manner as RFs.

4.4.3 Support vector machines

The package used to utilise SVMs was *sklearn.svm*, which provides the function *SVC*. Three parameters were passed to the function. The linear kernel and the radial basis function (gaussian kernel) were tested in order to determine the best performing option. These methods require values for the hyperparameter *C*, which controls the size of the dividing margin between classes The function *RandomizedSearchCV* was used to determine the optimum values. The use of a radial basis function kernel results in the need to define the second parameter *gamma*, which controls the curvature of a decision boundary. The setting for this was chosen as automatic. The effect of class imbalances on predictions was accounted for in the same manner as RFs, using the hyperparameter *class_weight* with the option balanced. Hyperparameter values were tuned using the nested CV approach.

4.4.4 Naïve Bayes

Software used for NB was *sklearn. naive_bayes*, which provides the function *GaussianNB*. In this instance no hyperparameters were used. The NB package *sklearn. naive_bayes* does not provide a method for dealing with unbalanced classes. The algorithm was trained and validated using stratified CV.

4.4.5 Logistic regression for PRS

Following the derivation and PC adjustment of PRS, the next step is to use these to assess discrimination between cases and controls. To achieve this the function *LogisticRegression* from the *sklearn* package in *Python* was employed. To ensure fair comparison, the same individuals in each test set used for ML models were also used in the LR. Therefore, PRS generated for the samples in each test fold were inputs, whilst disease status was the target variable. For the purposes of this chapter and thesis, this method will be known as 'PRS-LR' when comparing to ML algorithms.

4.4.5.1 Further data balancing methods

A further method tested to address the issue of age-based confounding in AD prediction was case-control matching (de Graaf *et al.*, 2011). The matching process reduces the effect of confounding by ensuring an equal distribution of the confounding variables throughout both cases and controls, where cases are matched to controls based on suspected confounding variables. In the case of this study, they were matched on both age and sex.

The process of matching was carried out in *Python*, in which custom functions were developed. Firstly, cases and controls were separated into two data frames. Following this, the ages of controls were taken and matched to each case. Once a match was determined both the case and control were appended to a new data frame. The matched case was then removed from the data frame of cases, this was to ensure the same sample could not be chosen twice. Once this was completed, the new data frame was then passed to a second function. Within this function, the same process was conducted, however this time individuals were matched on sex. The result of application of these two functions, was a data set with an equal number of controls and cases, whilst also being matched in sex and age. Prediction performance of both ML and PRS-LR was then assessed on this balanced dataset. The balancing hyperparameters for ML explained in previous sections were not employed for ML algorithms, as their use would be redundant due to the number cases and controls being equal. This was also true for adjusting PRS values and genotypes by age and sex.

4.4.6 Comparison of ML Performance

Prediction performance for ML algorithms and PRS-LR in this chapter was reported by calculating the mean of all AUC values across test sets corresponding to each cross-validation fold. However, when assessing classifier performance, comparing means visually does not provide statistical evidence to determine comparative relationships. Therefore, the t-test was used to examine whether any differences in performance could be derived. As five rounds of CV were used, the number of paired observations will consist of five values of AUC (equal to the number of CV folds), with tests conducted on two classifiers at a time. The function *stats.ttest* from the package *scipy* was used to employ the paired t-test.

In this and future chapters, the false discovery rate (FDR) controlling method Benjamini-Hochberg was used to adjust for possible false positives. The adjustment was made using the statistical programming language R, with the function *p.adjust* used. Corrections were made for the number of pairwise comparisons between algorithms on an analyses-wide basis (for each supplementary table at one time).

4.4.7 Model calibration

Initial prediction probabilities were calculated using the *predict_proba function*. These were then plotted using a histogram from *matplotlib*. A loess smoother was used to aid in assessing the relationship between predicted and observed probabilities. The initial classifier probabilities were then calibrated using the function *CalibratedClassifierCV* with five rounds of CV. Probabilities were calibrated using the isotonic option. These calibrated probabilities were then plotted using the same method for uncalibrated probabilities.

4.5 Results

The below table demonstrates SNPs which have been chosen using the r^2 method (see Section 4.3.1):

Table 4.3: GERAD Proxy SNPS's selected using the r^2 method from Index SNP's from Kunkle *et al.*, 2019.

Original Marker code from Kunkle[a]	Chr[b]	Effect Size[c]	P-value[d]	Most related SNP In GERAD[e]	r^2 between index and proxy SNP[f]	Proxy SNP chromosome/ position[g]
rs4844610	1	0.1466	8.246e-16	rs3818361	0.896	1:207784968
rs6733839	2	0.1693	4.022e-28	rs744373	0.508	2:127894615
rs10933431	2	0.1001	2.552e-07	rs11678851	0.558	2:233825947
rs9271058	6	0.094	5.136e-08	rs1063355	0.515	6:32627714
rs75932628	6	0.6989	2.948e-12	rs9367085	0.027	6:40848013
rs9473117	6	-0.0823	2.323e-07	rs9381563	0.722	6:47432637
rs12539172	7	-0.0674	2.093e-05	rs5015756	0.511	7:100013457
rs10808026	7	-0.1018	3.058e-08	rs11767557	0.971	7:143109139
rs73223431	8	0.0936	8.342e-10	rs755951	0.800	8:27226790
rs9331896	8	0.1269	3.624e-16	rs11136000	0.911	8:27464519
rs3740688	11	0.0935	9.702e-11	rs10769258	0.546	11:47391039
rs7933202	11	0.1165	2.150e-15	rs7926344	0.916	11:59962166
rs3851179	11	-0.1198	5.809e-16	rs7941541	0.769	11:85858538
rs11218343	11	0.2053	2.633e-08	rs3781834	0.618	11:121445940
rs17125924	14	-0.1222	6.621e-07	rs17125944	0.960	14:53400629
rs12881735	14	0.088	4.876e-07	rs12878418	0.322	14:92923032
rs3752246	19	-0.124	6.621e-10	rs2072102	0.548	19:1073073
rs429358	19	-1.2017	0	rs8106922	0.127	19:45401666
rs6024870	20	-0.1279	1.102e-06	rs6064392	0.913	20:54984768
rs7920721	10	-0.0782	1.942e-07	rs7094380	0.637	10:11723257
rs138190086	17	0.2535	7.463e-06	rs2440139	0.174	17:61285198
rs190982	5	0.0564	0.0002809	rs304132	0.885	5:88215594
rs4723711	7	0.0538	0.0002727	rs2718058	0.813	7:37841534

a – Marker code for the original 23 significant SNPS from Kunkle; b – Kunkle SNP's chromosome; c – Each Kunkle SNP's effect size; d – The p-value of each SNP e –; The SNP from GERAD which is most related with each Kunkle SNP using the r-squared method. f – R-squared value between Kunkle SNP and GERAD SNP; g – The GERAD SNP's chromosome position.

When assessing the values of r^2 for each proxy variant, it can be deduced that some of these were below 0.5. It could be argued that such values represent a small amount of association between the index and proxy SNP. However, it was decided to accept these low values as they were still larger than any other variants association in the genomic region.

The following table displays SNP's chosen by the method of the best p-value, also described in Section 4.3.1.

Table 4.4: GERAD Proxy SNP's selected using a p-value (GERAD) method with Index SNP's from Kunkle *et al.*, 2019.

Original Marker code from Kunkle[a]	Chr[b]	Chr Position in kb[c]	Effect Size[d]	P-value[e]	Most Significant SNP in GERAD[f]	P Value[g]
rs4844610	1	1:207802552	0.1466	8.246e-16	1:207786289	2.84E-13
rs6733839	2	2:127892810	0.1693	4.022e-28	2:127889637	2.00E-17
rs10933431	2	2:233981912	0.1001	2.552e-07	2:233977318	0.000172
rs9271058	6	6:32575406	0.094	5.136e-08	6:32224388	6.28E-05
rs75932628	6	6:41129252	0.6989	2.948e-12	6:41150591	7.9E-05
rs9473117	6	6:47431284	-0.0823	2.323e-07	6:47432637	6.84E-07
rs12539172	7	7:100091795	-0.0674	2.093e-05	7:99633385	2.30E-05
rs10808026	7	7:143099133	-0.1018	3.058e-08	7:143109139	9.15E-08
rs73223431	8	8:27219987	0.0936	8.342e-10	8:27226790	2.69E-06
rs9331896	8	8:27467686	0.1269	3.624e-16	8:27464519	9.95E-13
rs3740688	11	11:47380340	0.0935	9.702e-11	11:47391039	2.17E-10
rs7933202	11	11:59936926	0.1165	2.150e-15	11:59975078	3.72E-14
rs3851179	11	11:85868640	-0.1198	5.809e-16	11:85868640	5.82E-13
rs11218343	11	11:121435587	0.2053	2.633e-08	11:121436270	0.0000168
rs17125924	14	14:53391680	-0.1222	6.621e-07	14:53400629	3.09E-06
rs12881735	14	14:92932828	0.088	4.876e-07	14:92344244	0.000526
rs3752246	19	19:1056492	-0.124	6.621e-10	19:1051214	1.90E-05
rs429358	19	19:45411941	-1.2017	0	19:45382034	1.78E-81
rs6024870	20	20:54997568	-0.1279	1.102e-06	20:54984768	5.03E-06
rs7920721	10	10:11720308	-0.0782	1.942e-07	10:11720308	3.32E-06
rs138190086	17	17:61538148	0.2535	7.463e-06	17:61560763	0.000654
rs190982	5	5:88223420	0.0564	0.0002809	5:88215594	0.00031
rs4723711	7	7:37844263	0.0538	0.0002727	7:37882317	0.00019

a – Marker code for the original 23 significant SNPS from Kunkle; b – Kunkle SNP's chromosome; c – Each Kunkle SNP's chromosome position; d – The effect size of each Kunkle SNP e – Each Kunkle SNPs p-value; f – SNP from GERAD which is most related with each Kunkle SNP after clumping. g – P-value of GERAD SNP after clumping.

These methods resulted in two sets of SNPs, with an overlap of 30%. A larger set of SNPs was also created using the second method explained in Section 4.3.1.1, this led to a set of 422 independent (r^2=0.1) SNPs from GERAD at a p-value threshold of 0.01.

4.5.1 Removal of the *APOE* region

For SNPs chosen using the r^2 method (from Table 4.3), rs2072102 and rs8106922 were removed. For the second method of SNPs chosen using the p-value method, rs6859 and rs3752240 were extracted (from Table 4.4). SNPs in the *APOE* region were not removed from the larger SNP set of 422 variants, with the intention of assessing whether SNPs in this region plus other variants would outperform the smaller datasets. Therefore, in total, five sets of SNPs were formed for analysis in this study.

4.5.2 The removal of the 1958 birth cohort

The presence of the 1958 birth cohort, consisting of 5343 controls all 45 years of age, skews the age distribution of the dataset in favour of controls. One of the aims of analyses in this chapter was to investigate methods to control for the confounding effects of both age and sex on AD prediction. However, it soon became apparent the presence of the 1958 birth cohort would result in spurious results even when accounting for these factors. This is displayed in results shown in Supplementary Tables 1 & 2. Analyses in Supplementary Table 1 represent the use of age and sex when adjusting both PRS values and genotypes. This resulted in high values of AUC for the two-decision tree-based algorithms (RFs and GB). Values for age and sex were then added as separate predictors to both ML and PRS-LR in Supplementary Table 2, resulting in very high AUC for all classifiers. It was decided that these values were too high and biased due to the presence of the 1958 birth cohort. Therefore, it was decided to remove all birth cohort samples from GERAD. After removing 5343 controls, 5344 individuals remained for all analyses, with 3332 cases and 2012 controls. Therefore, the class balancing techniques discussed in this chapter were still required due to the unequal numbers of cases and controls.

4.5.3 Selection of the optimal threshold for PRS

In addition to the set of 23 genome-wide significant SNPs, the clumping and thresholding method will be used to generate PRSs throughout analyses in this thesis. Following the selection of LD pruned SNPs through clumping, a range of PRS values are generated at different p-value thresholds. The next stage of the process is to determine which of these p-value thresholds explains the highest phenotypic variance. This is achieved by fitting PRSs versus phenotype values in a series of LRs. The p-value threshold which returns the highest r^2

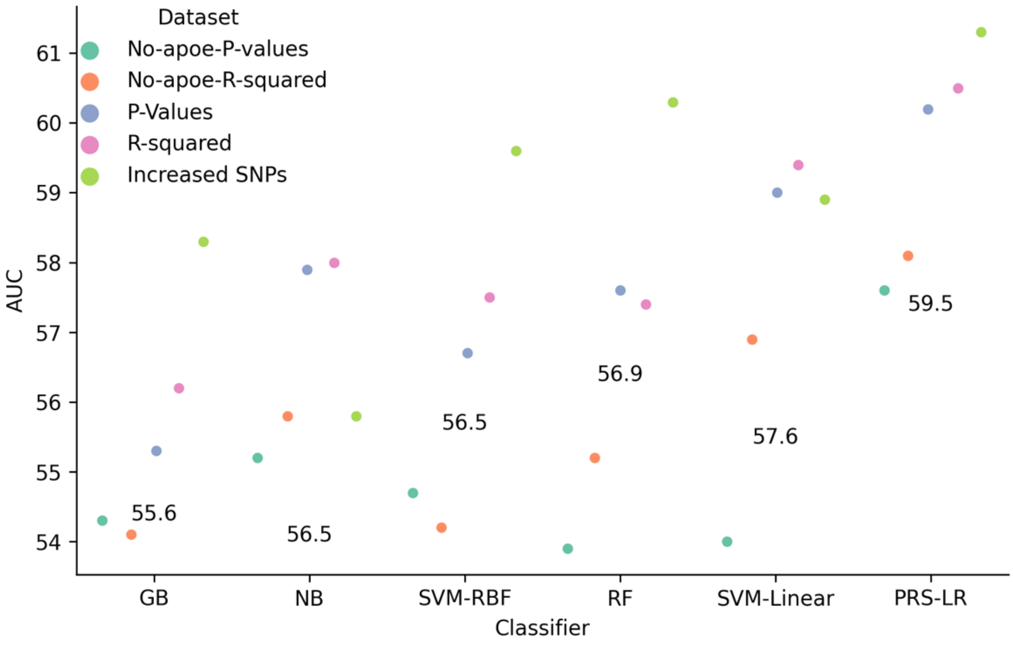
value is considered to explain the highest phenotypic variance (Choi, Mak and Paul F O'Reilly, 2020b). The p-value thresholds compared in this thesis were 1e-8, 1e-6, 1e-4, 0.01, 0.05, 0.1, 0.5. The use of PRS combined with prediction through a LR will be denoted as PRSS-LR for analyses.

4.5.4 Results of analyses

Results are detailed in figures, in which prediction performance is grouped by classifier (x-axis). These results are given as AUC (y-axis), with the mean performance across five rounds of CV denoted by a coloured dot. Each dot represents one of the five types of datasets used (varying by number of SNPs). The mean performance of each classifier across these five sets is given. Classifier performance was compared statistically using the paired t-test, with test fold AUC values from two classifiers compared on each occasion. Statistics for each t-test calculated can be found in Supplementary Tables 7-9. Calibration statistics for each plot were also reported in accompanying figures. In this instance, each calibration plot represents the RF trained using the SNPs chosen by p-values (including *APOE*).

Figure 4.4 displays results in which genotypes and PRS were adjusted by PCs only. For each classifier, the mean AUC across five folds of CV is shown for each dataset.

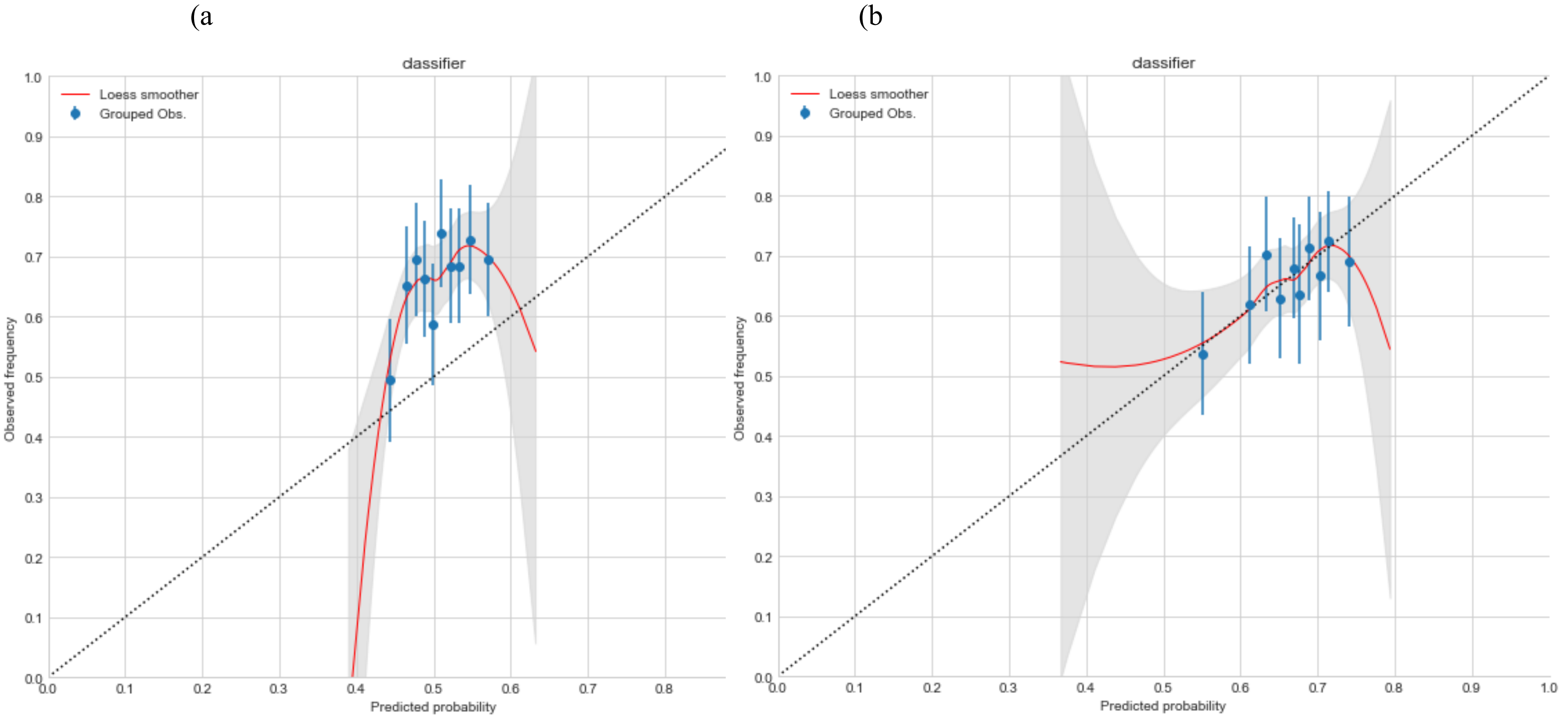
Figure 4.4: Results of PRS versus all ML algorithms: genotypes and PRS have been adjusted by PCs.



Y-axis represents AUC in %; with classifiers placed on the X axis. Each dot represents the mean score for the prediction algorithm across SNP sets, with an accompanying 95% CI bar. The numbers placed centrally are the mean of the three p-value threshold scores; GB Gradient Boosting; RF Random Forest; PRS-LR Polygenic Risk Scores Logistic Regression; AUC Area Under the Curve. Datasets described in the legend relate to how SNPs were chosen, including the larger SNP set denoted as 'Increased SNPs.

When assessing results in Figure 4.4, PRS-LR achieved higher accuracy than all ML algorithms by 2-4% AUC. This is supported by the majority of t-tests conducted between PRS-LR and ML classifiers returning a significant p-value (<0.05) (Supplementary Table 7). However, little variability occurred between ML methods. Results show that all classifiers achieved higher accuracy when using a p-value threshold of 0.01 to derive the PRS. This is not a surprising outcome as the algorithms were provided with more information to predict upon.

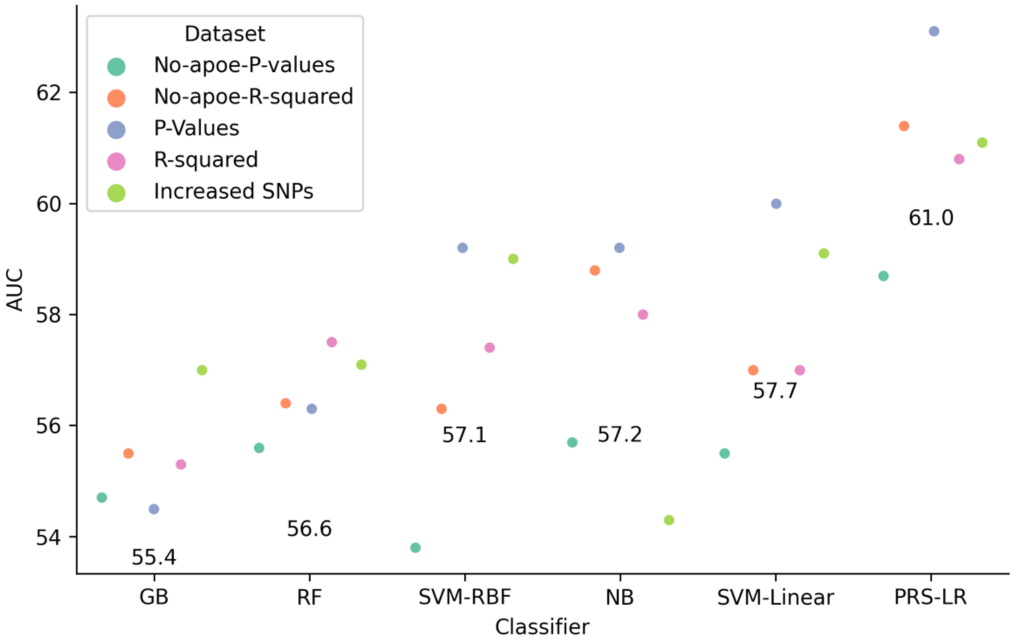
Figure 4.5: The comparison of non-calibrated vs calibrated prediction probabilities



These figures represent pre a) and post b) calibration plots for the related RF algorithm (Figure 4.4) (p-value 0.0001). The x-axis represents the prediction output of the classifier in terms of the probability of being a case. With the y-axis denoting observed class frequencies. Perfect calibration in which predicted probabilities match observed accuracies is denoted by the diagonal dotted line. The blue dots represent the mean probability values within each quantile and are accompanied by a 95% confidence interval (blue bar). The overall relationship between predicted probabilities and observed frequencies (calibration curve) is given by the fitted loess smoother (red line), with a 95% (grey shaded area) used.

Plot 4.5a represents the uncalibrated output of the RF from Figure 4.4. The red line indicates a mixture of both under and over-estimation of case probabilities. This is due to the red line sitting underneath the diagonal between (0.4-0.5) and residing above the diagonal thereafter. Plot 5.b represents all probabilities post calibration, with some alteration in predicted probabilities. Probabilities between 0.4-0.5 are now underestimated, however a greater amount of predicted probabilities now lie on the diagonal, suggesting an improved level of calibration.

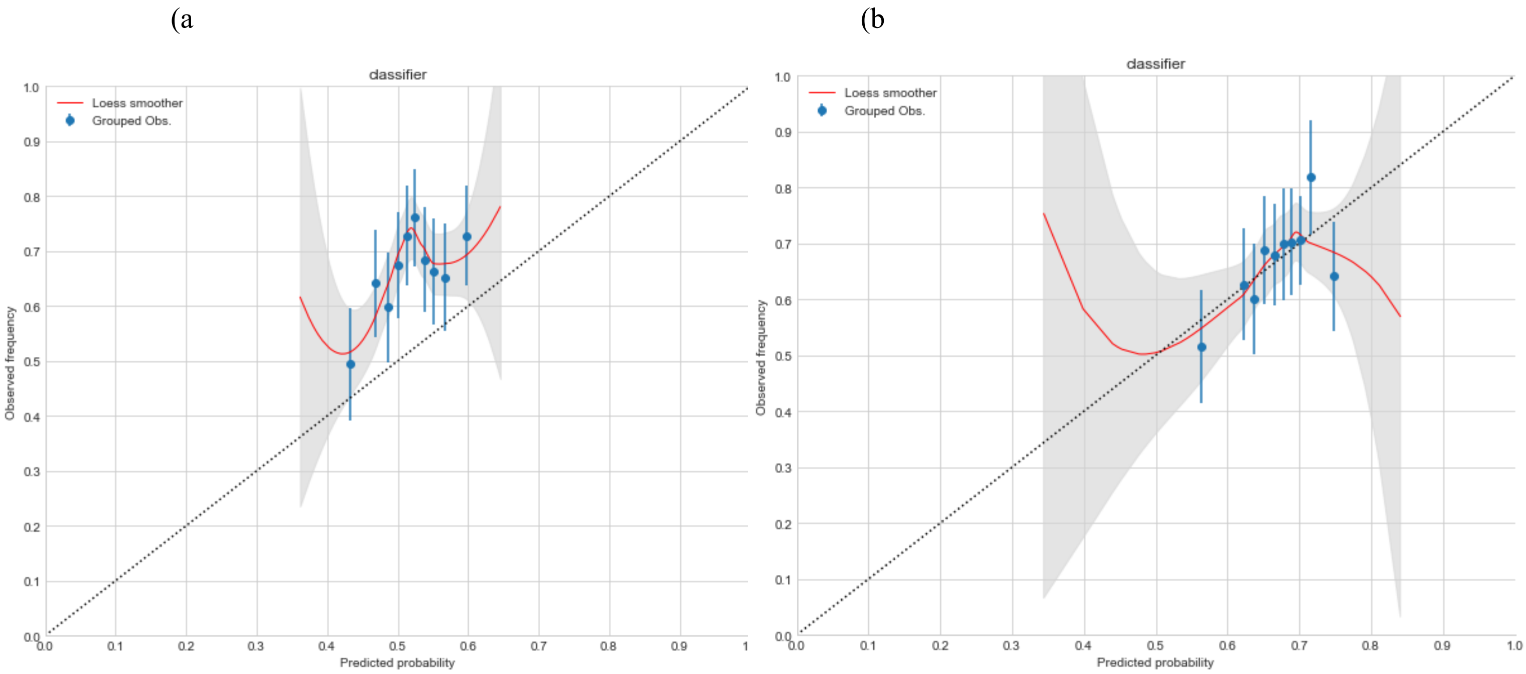
Figure 4.6: Results of PRS versus all ML algorithms with the 1958 birth cohort removed from GERAD, genotypes and PRS values have been adjusted by PCs, age and sex



Y-axis represents AUC in %; with classifiers placed on the X axis. Each dot represents the mean score for the prediction algorithm across SNP sets, with an accompanying 95% CI bar. The numbers placed centrally are the mean of the three p-value threshold scores; GB Gradient Boosting; RF Random Forest; PRS-LR Polygenic Risk Scores Logistic Regression; AUC Area Under the Curve. Datasets described in the legend relate to how SNPs were chosen, including the larger SNP set denoted as 'Increased SNPs.

Prediction accuracies displayed in Figure 4.6 are similar to those in Figure 4.4. Adjusting both genotypes and PRS by age, sex and PCs lead to no discernible difference to adjusting by PCs only. T-statistics generated again showed a significant difference between PRS-LR and ML classifiers, the only exception to this was SVM-Linear, which returned non-significant p-values for some datasets (Supplementary Table 8). Comparisons between ML classifiers were largely non-significant, however algorithms such as SVM-Linear and NB returned some significant values when compared with GB.

Figure 4.7: The comparison of non-calibrated vs calibrated prediction probabilities



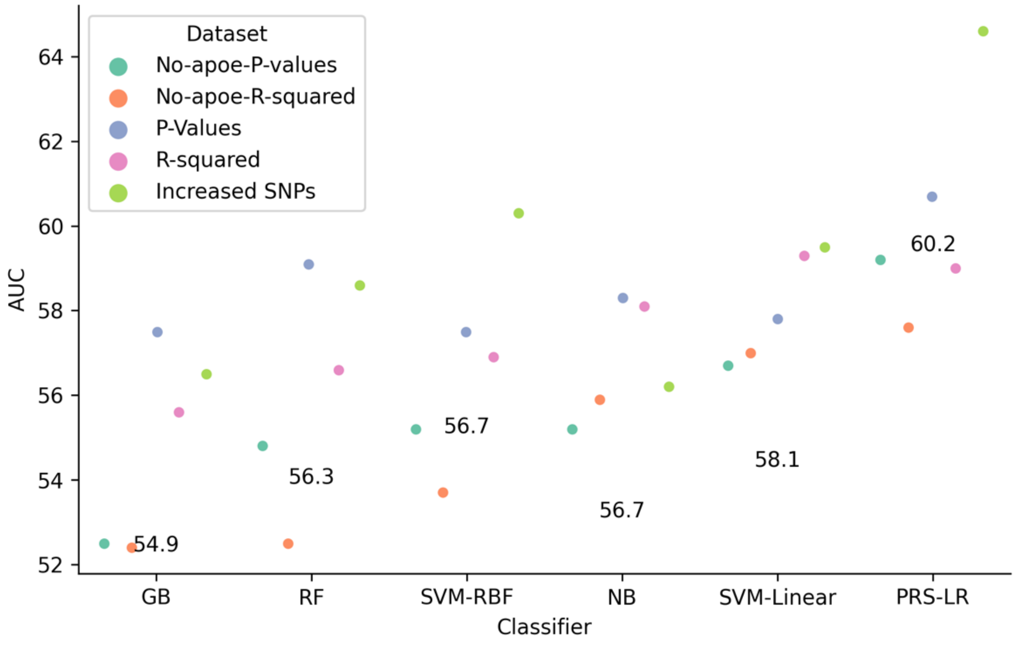
These figures represent pre a) and post b) calibration plots for the related RF algorithm (Figure 4.6) (p-value 0.0001). The x-axis represents the prediction output of the classifier in terms of the probability of being a case. With the y-axis denoting observed class frequencies. Perfect calibration in which predicted probabilities match observed accuracies is denoted by the diagonal dotted line. The blue dots represent the mean probability values within each quantile and are accompanied by a 95% confidence interval (blue bar). The overall relationship between predicted probabilities and observed frequencies (calibration curve) is given by the fitted loess smoother (red line), with a 95% (grey shaded area) used.

When assessing pre-calibrated probabilities in Figure 4.7.a, an under-estimation of risk was between for all predicted probabilities. This under-estimation was reduced in Figure 4.7.b, as probabilities lay closer to the diagonal between 0.5 and 0.7. However, under estimation and over estimation of risk occurred at either end (0.35-0.5, 0.7-0.85).

4.5.5 Predictions with balanced data set

All previous analyses only balanced the numbers of cases and controls prior to training and validation. A further method of balancing using both sex and age was used in this section. This was to counteract their confounding effect.

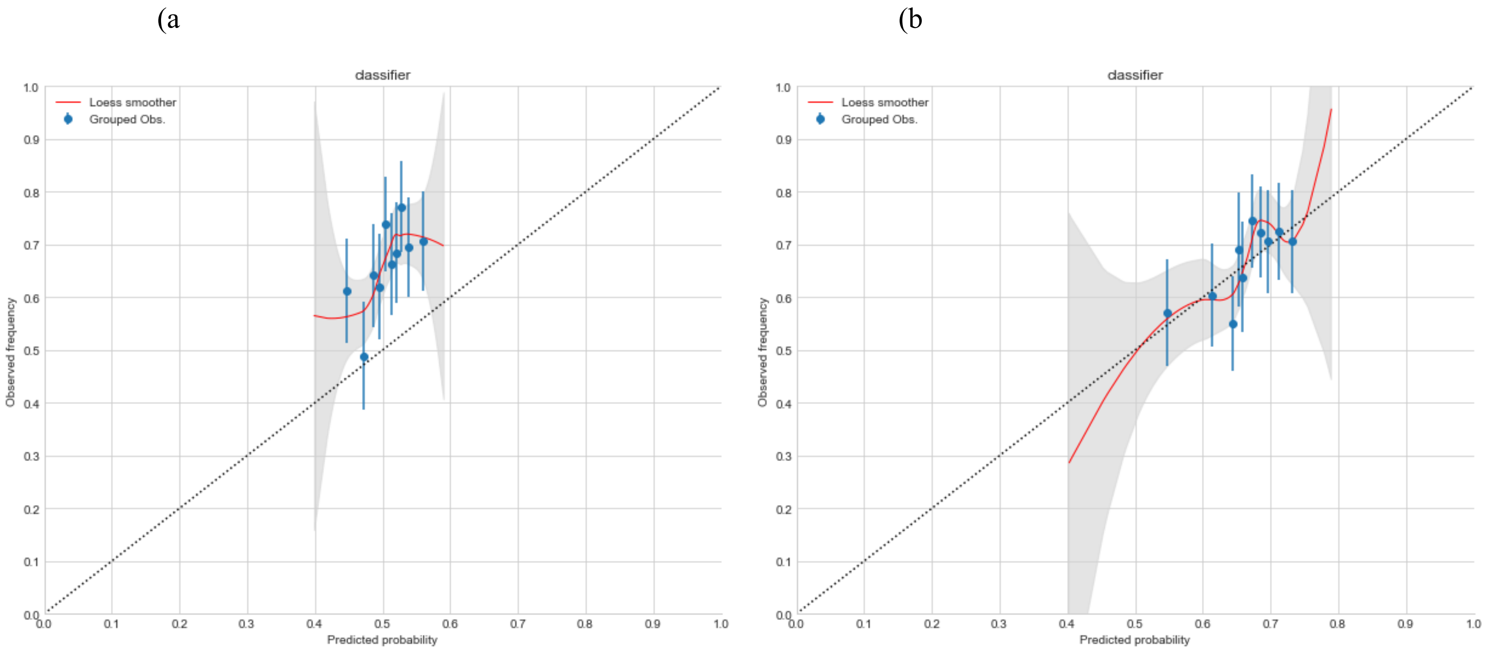
Figure 4.8: Results of PRS versus all ML algorithms with CV balanced for sex and age, with genotypes and PRS have been adjusted by PCs



Y-axis represents AUC in %; with classifiers placed on the X axis. Each dot represents the mean score for the prediction algorithm across SNP sets, with an accompanying 95% CI bar. The numbers placed centrally are the mean of the three p-value threshold scores; GB Gradient Boosting; RF Random Forest; PRS-LR Polygenic Risk Scores Logistic Regression; AUC Area Under the Curve. Datasets described in the legend relate to how SNPs were chosen, including the larger SNP set denoted as 'Increased SNPs.

Results displayed in Figure 4.8 can be compared with those in Figure 4.4, with prediction performance for all algorithms similar between the two plots. As with all previous analyses, PRS-LR achieved higher AUC across the five datasets than all ML algorithms. This was supported by significant statistics generated when comparing AUCs across CV (Supplementary Table 9). Similarly, to previous analyses, little variability existed when comparing the performance of ML algorithms, with the only discernible pattern being the superior performance of SVMs and NB when compared to GB for some datasets.

Figure 4.9: The comparison of non-calibrated vs calibrated prediction probabilities



These figures represent pre a) and post b) calibration plots for the related RF algorithm (Figure 4.8) (p-value 0.0001). The x-axis represents the prediction output of the classifier in terms of the probability of being a case. With the y-axis denoting observed class frequencies. Perfect calibration in which predicted probabilities match observed accuracies is denoted by the diagonal dotted line. The blue dots represent the mean probability values within each quantile and are accompanied by a 95% confidence interval (blue bar). The overall relationship between predicted probabilities and observed frequencies (calibration curve) is given by the fitted loess smoother (red line), with a 95% (grey shaded area) used.

Similarly, to Figure 4.7a, pre-calibrated probabilities in Figure 4.9a show a complete under-estimation of risk for all predicted probabilities. Following calibration, this over-estimation was reduced as demonstrated in Figure 4.8b.

4.6 Conclusions

The central aim of this chapter was to compare ML and PRS approaches for the prediction of AD from genetic data, age and sex. Alongside this, a further aim was to assess methods to counteract the possible effect of age confounding.

4.6.1 Results of analyses

All analyses performed in this chapter were conducted on GERAD (Harold *et al.*, 2009) following the removal of the 1958 birth cohort. This was due to the bias leading to the high prediction accuracies achieved, with AUC reaching 80-90% for all classifiers (Supplementary Table 2) when the 1958 birth cohort was part of the dataset. It was determined the most likely cause for these suspiciously high results was the presence of the 1958 birth cohort within

GERAD, resulting in the controls being on average younger than the cases. The consequence of this led to prediction accuracy ignoring the genetic component of AD. Thus, the decision was made to remove the 1958 birth cohort from GERAD.

Following the exclusion of the birth cohort, results of all analyses demonstrated PRS-LR generally outperformed all ML algorithms. When excluding SNPs within the *APOE* region, AUC was between 57-59%. This increased to 60-61% when reintroducing the two *APOE* related SNPs, with prediction increasing by 1-2% further when using the increased SNP set (p-value threshold = 0.01). On occasion, SVMs achieve 60% AUC, which is similar to with PRS-LR. However, in most cases the AUC of PRS-LR is 2-5% greater than that of all ML methods. The superior performance of PRS-LR when compared to ML algorithms was confirmed by comparing AUC values using the t-test. Most p-values were less than the significance level threshold of 0.05 following correction for comparisons between multiple classifiers and SNP sets (Supplementary Table 7). In terms of ML methods, little variability occurred when considering prediction performance. All algorithms achieved a prediction performance of around 55% when excluding *APOE* related SNPs, with AUC increasing to 56-59% when including the removed SNPs. AUC again increased for all ML classifiers except NB when using the larger SNP set. Despite the small amount of variability in ML performance, NB and SVMs did outperform GB for some datasets. This is outlined by the t-test statistics shown in Supplementary Table 7. Following the use of PCs only for genotype and PRS adjustment, both sex and age were introduced into the regression (Supplementary Table 5). The results of this analysis did not alter from the previous analysis, demonstrating that the inclusion of both variables had no effect on classifier performance.

The last section of analysis in this chapter investigated the use of a dataset balanced on both age and sex. As previously discussed, age could be defined as a confounding variable for AD prediction (Falahati *et al.*, 2016). Methods such as genotype adjustment and stratification have been used to counteract its effect, however, debate still exists surrounding the most efficient technique (Dukart *et al.*, 2011). Cases and controls were balanced on both age and sex, leading to a dataset of around 1500 samples, with size depending on the number of SNPs used in analysis. Nested CV was used for the training and validation of models, with both genotypes and PRS adjusted by PCs only. It was determined that adjusting variables by age and sex would be unnecessary due to using a balanced dataset, as this would be accounting for both variable twice. Performance of for all prediction algorithms remained similar with and without the use of balanced dataset (Supplementary Table 6).

4.6.2 Assessment of calibration statistics

Calibration statistics assess the confidence with which ML predictions have been made. The presence of well-calibrated algorithms in scenarios where high importance decisions are made is important. Calibration statistics in this chapter were plotted using calibration curves, with values belonging to the corresponding RF algorithm. This was trained and validated using SNPs chosen by the p-value method (including *APOE*). Calibration figures display a consistent underestimation of risk across analyses for pre-calibrated probabilities. This underestimation is of concern, as this may result in undiagnosed individuals, which could lead to harm if appropriate treatment methods are not administered. However, following the use of calibration, predicted probabilities were more in line with observed frequencies, resulting in greater confidence for prediction.

4.6.3 Comparison of classifiers

When assessing ML performance across all analyses, little variability existed between the chosen methods. One noticeable difference was the reduction in prediction accuracy for NB when increasing the number of SNPs. As discussed, a reason for this could be the lack of a regularisation method in NB methodology, which is often used to reduce the possibility of overfitting (Sánchez García and Cruz Rambaud, 2022). No reduction in AUC when increasing the number of SNPs occurred for any other algorithm. SVMs achieved levels of AUC similar to those observed for both decision tree-based methods (RFs, GB), as well as occasionally matching PRS-LR. This is not surprising as SVMs have been shown to be effective prediction models across many domains (Ben-Hur *et al.*, 2008).

However, one drawback of the SVM algorithm is the longer time required for training when compared to other methods, due to the more complex mathematical calculations required (Ghaffari, 2021). Training periods increase substantially as the number of features used rise. Training times in this chapter remained manageable despite using up to 422 SNPs, however since the number of SNPs will increase to thousands in following chapters, only the decision tree-based algorithms will be used moving forward. In general, PRS-LR outperforms all ML algorithms by a margin of 4-5%. A possible reason for this difference in AUC is that PRS was calculated using an external dataset comprising a cohort of around 53,000 samples. Therefore, PRS-LR may have an advantage over ML, as effect sizes used were derived from a larger population than used for ML training. This will be taken into consideration for analyses in further chapters.

4.6.4 Summary

In summary, the core aim of this chapter was to compare the predictive accuracy of PRS-LR to several ML algorithms, with the secondary aim of assessing whether methods to minimise the confounding nature of age on AD risk might alter prediction performance. Results demonstrated superior performance for PRS-LR on most occasions. In some instances, SVMs returned similar prediction performance to PRS, however this was inconsistent with the general trend. A possible reason for this greater performance by PRS-LR is the use of external data (Kunkle-nogerad summary statistics) to generate PRS. However, in some instances ML was able to achieve 60% AUC. Techniques used to supress the confounding effect of age had no effect on prediction accuracy. The balancing method reduces the number of samples available for prediction by around three quarters. This could be considered a limitation and was not used for analyses in later analyses.

5 Assessment of feature selection methods

5.1 Introduction

Analyses conducted in this chapter explored the use of variants on a genome wide scale. Datasets used in Chapter 4 comprised a small number of statistically informed single nucleotide polymorphisms (SNPs), however it was hypothesised that providing a larger range of variants might enhance prediction performance. The methodology for creating PRS was also adjusted in this chapter. Previous scores were generated by clumping GERAD Harold *et al*., 2009 using Kunkle no-gerad summary statistics Kunkle *et al*., 2019, however concern was raised as to whether PRS gained an advantage over ML due to a larger cohort size. Therefore, both ML models and PRSs were generated within CV in this chapter. A GWAS was conducted within the training set (80% of samples) with subsequent PRSs calculated within the test set (20% of samples). Techniques to overcome potential dimensionality issues such as feature selection and extraction were tested. Whilst further methodological approaches such as removing clumping and varying pruning thresholds were also explored to assess the impact on prediction performance.

The emergence of 'Big Data' in recent years has led to larger datasets being compiled in both industry and academia (Gupta and Rani, 2019). In the case of biological data, this has not only led to the expansion of the number of subjects (people) but also an increase in the number of predictor variables (features) (Ching *et al.*, 2018). Despite advances in computational technology, issues have arisen when analysing ever-increasing datasets. Examples of these include the requirement of greater resources such as memory and computational power (Tsai *et al.*, 2015). Whilst more predictors may provide a better understanding of variability of the outcome variable, large amounts of features can also result in a reduction in predictive performance when assessed in an independent dataset, known as the curse of dimensionality (Verleysen and François, 2005b). Previous analyses in this thesis used a relatively small number of predictors (SNPs) for Alzheimer's Disease (AD) risk prediction. Analyses in this chapter focused on using larger number of predictors with techniques employed to reduce the issues caused by increased dimensionality. These techniques are termed feature selection methods, which involve choosing features based upon certain statistical criteria.

There have been various attempts of using such methods for selecting SNPs in AD prediction. Osipowicz et al., 2021 investigated the use of the ‘Boruta’ algorithm for feature selection. This is a random forest-based approach which uses feature importance measures to retain important variables and remove redundant variables. In this instance, SNPs from the combined dataset of Alzheimer’s Disease Neuroimaging Initiative (ADNI) https://adni.loni.usc.edu, the Religious Orders Study and the Rush Memory and Aging Project (ROSMAP) https://www.radc.rush.edu were used for predictors. To assess the impact of possible data leakage on prediction performance, two separate methodologies for SNP selection were compared, with feature selection conducted prior to splitting samples into training and test folds and subsequent selection within the training set only. Selected features were used for the training of RF classifiers, with results varying between the two methods. Classifiers trained on SNPs prior to sample split achieved 98% AUC, whilst conducting feature selection within the training set only resulted in 67% AUC. Therefore, recommendations were made to conduct selection in the training set only, in order to avoid over optimistic models due to data leakage.

Muhammed Niyas and Thiyagarajan., 2022 investigated the use of feature selection when predicting AD from the Alzheimer’s disease prediction of longitudinal evolution (ADNI-TADPOLE) https://adni.loni.usc.edu/tadpole-challenge-dataset-available/ and Australian Imaging Biomarkers (AIBL) https://adni.loni.usc.edu/aibl-australian-imaging-biomarkers-and-lifestyle-study-of-ageing-18-month-data-now-released/ datasets. These are publicly available longitudinal data sources comprising of a range of AD related biomarkers including magnetic resonance imaging (MRI), positron emitting tomography (PET), cerebrospinal fluid (CSF) and genetic information. Feature selection was achieved using the Fisher score (FS), a scoring metric which selects features based upon their ability to maximise between class distance. The use of FS was combined with a greedy search algorithm. All features were ranked in terms of their FS score, with the most significant variable in terms of FS was selected to test classifier performance. Other features were then added sequentially in the order of FS importance, with each feature either retained or removed depending on if it improved classifier performance. The final set of features were then used to assess AD prediction, using both the support vector machine (SVM) and k-nearest neighbour (KNN) classifier methods. The use of feature selection resulted in 97 and 91% AUC for the SVM and KNN approaches respectively.

5.1.1 Methods for feature selection

The effectiveness of dimensionality reduction techniques to reduce overfitting was assessed in this chapter. A range of techniques commonly used in ML development were used and their performance was compared. The methods used are listed below.

5.1.1.1 Linkage disequilibrium

To reduce dimensionality due to LD, methods have been developed to remove SNPs in high LD, creating sets of independent variants at a chosen LD (r^2) threshold and further informed by the trait association (p-value)) threshold. The clumping algorithm was used throughout the analyses in this chapter. It could be argued that clumping is a form of feature selection, as SNPs are removed based upon their correlations with one another whilst the most statistically relevant SNP remains. Therefore, whilst it was not directly tested as a feature selection method, it was still necessary to consider its influence.

5.1.1.2 Random forests for feature selection

Random Forests (RFs) conduct a form of feature selection when classifying data. Each decision tree within the RF is trained on a bootstrap of samples, with the tree sequentially splitting on features deemed significant for discrimination. Therefore, across all bootstraps and decision trees, those features which are deemed to be of little importance are used less often or not at all (Hasan *et al.*, 2016). The RF is fitted on the training set, with the algorithm assessing the importance of all features in the dataset. Features which pass certain pre-defined criteria are then passed to the final ML algorithm. Due to the effectiveness of RFs as feature selection techniques, they have been used in varying fields. Sylvester et al., 2018 used this method to identify SNPs proficient in assigning Salmon to different populations. The RF algorithm identified related SNPs more effectively than the traditional method of fixation (Sylvester *et al.*, 2018). Reasoning suggested for this was ML's ability to consider relationships between loci, rather than individual importance considered when using the traditional method.

5.1.1.3 Extra trees classifier for feature selection

Similarly, to RFs, ExtraTrees are used as an intermediate step in feature selection. Features deemed significant due to a predefined criteria are passed to the final ML model. The algorithm selects features based upon information gain and entropy. Those features with the greatest information gain are deemed to be the most useful for classification. This is due to their ability to separate the class variable (Latha and Mohanasundaram, 2019).

5.1.1.4 LASSO regression and Elastic net for feature selection

The least absolute shrinkage and selection operator (LASSO) is a technique for feature selection and optimisation in linear regression models. Due to the need for prediction between classes defined by a binary variable, a logistic regression (LR) will be used in this chapter. Despite the ability to reduce the dimensionality of feature sets, the LASSO algorithm has disadvantages, performing inadequately in the presence of correlated features, with features deemed important omitted incorrectly (Freijeiro-González, Febrero-Bande and González-Manteiga, 2022).

Elastic net is a further form of regularised regression and is related to the LASSO technique. The algorithm performs both variable selection and shrinkage, whilst also assessing groups of correlated features.

5.1.1.5 Feature selection based on biological relevance of SNPs

Thus far, discussed feature selection techniques have been based on statistical methods only. However, known properties of AD biology can also be used to select SNPs from genes relevant to the disease development and progression.

5.1.1.6 Microglia

Microglial degeneration has been linked to the formation of plaques within the brain, one of the two neuropathological signs of AD (Edler, Mhatre-Winters and Richardson, 2021). SNPs in genes which have been associated with microglial function (Gosselin *et al.*, 2017)(Tansey, Cameron and Hill, 2018) were used as features and the prediction by this set of SNPs was compared to prediction by the SNPs selected genome-wide.

5.1.1.7 Synapses

Synapses can be defined as intercellular junctions, which are designed for fast information transfer from a presynaptic neuron to a postsynaptic cell (Südhof, 2018). Similarly, to microglia, SNPs in genes associated with synaptic function were used as features for ML and PRS (Koopmans *et al.*, 2019).

5.1.2 Modelling the *APOE* region

SNPs which lie within the *APOE* gene have been consistently shown to be the most significant genetic factor for AD risk. Leonenko et al., 2021 investigated different methods for modelling the *APOE* region for PRS prediction analysis. The most accurate prediction was achieved when the *APOE* region was removed from the set of SNPs and PRS was calculated without these variants.

For discrimination between case/control status the PRS and *APOE* ε2 and ε4 alleles were added to the LR as two variables. Instead of using effect sizes derived in Kunkle et al., 2019, coefficients were generated within each CV fold through the use of a LR. Disease status was used as the target variable, whilst genotypes for all selected SNPs including the ε2/ε4 alleles were features. The coefficients for the ε2/ε4 alleles were then extracted and used as weights for the allele counts. This provided values of greater accuracy as those derived from Kunkle et al., 2019 were calculated on a larger cohort. The weighted counts for each individual were then added as an extra predictor to the PRS-LR model.

The *APOE* genotypes (ε2/ε2, ε2/ε3, ε3/ε3, ε3/ε4 and ε4/ε4) were constructed from the two SNPs within the *APOE* gene, r429358 and rs7412 (Correa *et al.*, 2014). Table 5.1 shows which SNP combinations lead to the different haplotype.

Table 5.1: ***APOE*** **Alleles (Crawford** ***et al.*****, 2022).**

rs429358	**rs7412**	***APOE*** **name**
C	T	ε1
T	T	ε2
T	C	ε3
C	C	ε4

Description of APOE SNPs and corresponding ε values

Analyses used in this chapter will use a similar method to represent the *APOE* region.

5.1.3 Parallel programming

Despite advances in computing, certain tasks which require large resources can take long periods of computing time. A facility developed in recent decades to improve upon these long run times, is high performance computing (HPC) (Lee *et al.*, 2011). One of the elements of HPC which developers use to improve efficiency is termed parallel programming, which separates a large task into multiple subtasks run in parallel. This can be achieved in two different ways, either using multiple computers concurrently, or utilising multiple cores within a central processing unit (CPU). This functionality was used to assist with analysing large datasets. The process of CV was made more efficient by using parallelisation, this functioned by separating folds over multiple cores and run in parallel. The result of this is then concatenation of all folds.

5.1.4 Aims and objectives

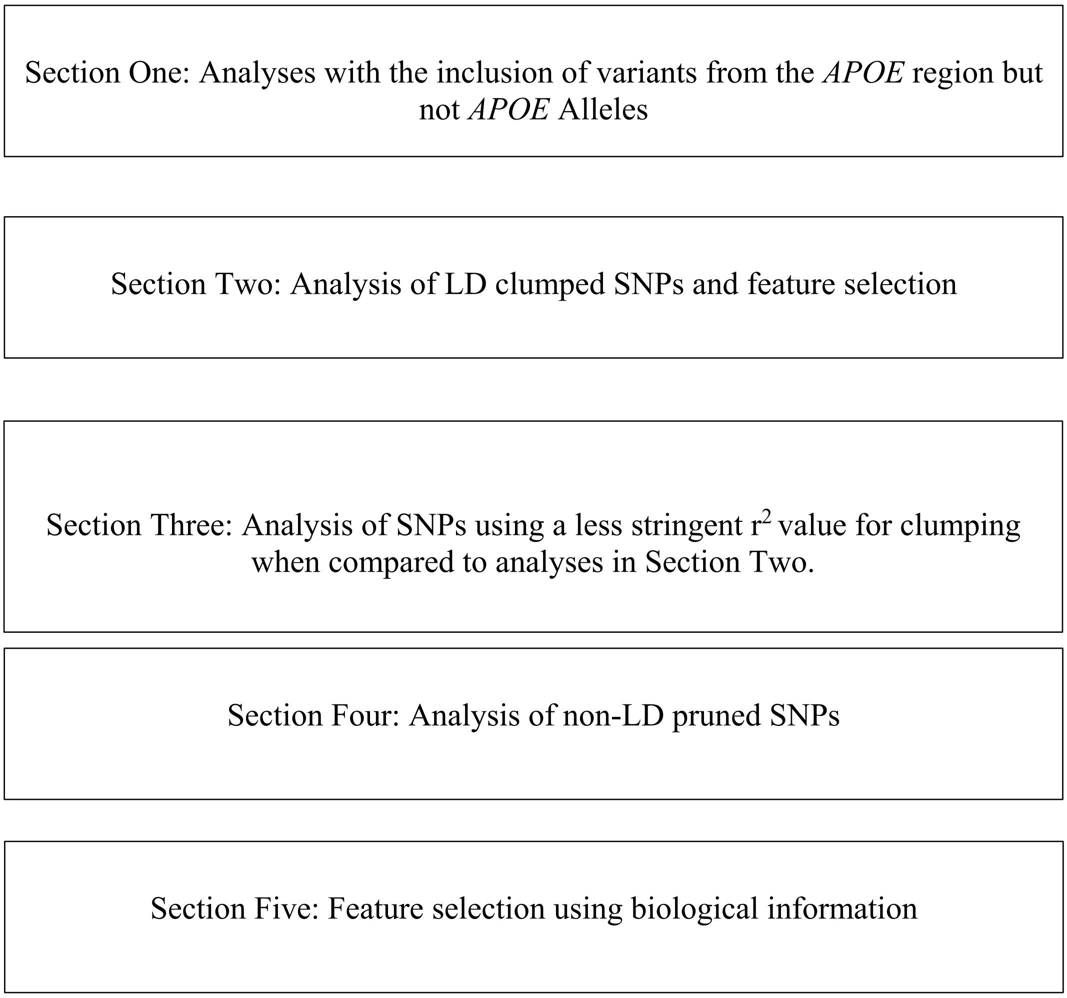
In this chapter I will build upon analysis conducted in Chapter 4 in which a small sample of genome wide significant (GWS) variants were analysed. Larger numbers of SNPs will be used here to assess a possible improvement of predictive utility of AD status by SNPs. The inclusion of more predictors may incur both computational and dimensionality issues. Therefore, this chapter will also test a range of feature selection methods (RFs, LASSO, ElasticNet, Extra Trees and biological information). All methods will be compared in terms of AUC in order to establish their effectiveness in mitigating dimensionality issues. The levels of calibration for algorithms will also be assessed, with methods to recalibrate predictive probabilities also tested. Further, all ML methods will be again compared to the predictive performance of PRS.

5.2 Materials and methods

5.2.1 Data analyses

Analyses conducted in this chapter were separated into five main sections. Figure 5.1 gives an outline of these.

Figure 5.1: Outline of analyses across the sub-sections in this chapter



Section One: Analyses with the inclusion of variants from the *APOE* region but not *APOE* Alleles

Section Two: Analysis of LD clumped SNPs and feature selection

Section Three: Analysis of SNPs using a less stringent r^2 value for clumping when compared to analyses in Section Two.

Section Four: Analysis of non-LD pruned SNPs

Section Five: Feature selection using biological information

Analyses differed through the alteration of parameters used for SNP selection, along with the use of biological information to select specific set of variants. The below section named "SNP Selection Process", explains the methodology for selecting SNPs. Following this, common methodology between each section of analysis is outlined.

5.2.1.1 SNP selection process

The *Python* function *StratifiedKFold* was used to create five folds of cross-validation (CV) for analysis. This function was used due to being able to access both training and test folds within each round of CV. This enabled any pre-processing to be carried out per training and test fold, to avoid data leakage. The ratio of data between train and test folds was 80:20, with the number of cases and controls evenly distributed between the two datasets.

Despite the use of different parameters when selecting SNPs, the method used to select SNPs was consistent. Within each CV fold, a GWAS was conducted for the "training" individuals only. This was preferred to clumping SNPs prior to CV, in which all samples would have been clumped using summary stats derived from (Kunkle *et al.*, 2019) as the latter may give PRS an advantage over ML due to effect sizes being calculated on a much larger sample size. To run a GWAS, *PLINK* was called within *Python*. This was achieved using the *Python* package *subprocess* which allows separate software processes to be run within *Python*. This process was run per CV fold and the output from the LR stored. Several parameters are required for clumping. A range of p-value threshold values were used to compare between analyses, 0.0001, 0.001, 0.01, 0.05, 0.1, 0.2, 0.3, 0.4 and 0.5. These values were chosen as they provided SNP sets in which the number of variants were wide ranging. The value for r^2 was also specified at 0.1, with a clumping window of 500 kilobases (Privé *et al.*, 2019). SNPs from each output were then saved for each level of p-value threshold.

5.2.1.1.1 The analysis of LD clumped SNPs and modelling of *APOE* region ($r^2 < 0.1$)

The *APOE* region in this chapter was modelled in two ways. Firstly, the analysis was run for the whole genome including the *APOE* region prior to clumping (but excluding *APOE* alleles). Secondly, to match the method outlined in Section 5.1.2, SNPs within the chromosomal region for *APOE* (19:45409039-45412650) were removed before analysis, and the ε4 and ε2 alleles were added separately and the aggregate scores for the ε2/ε4 alleles for each individual were added to the SNP set for both PRS-LR and ML. Then feature selection methods were applied for ML only.

5.2.1.1.2 The analysis of LD clumped SNPs ($r^2 < 0.5$)

Analysis in this section focused on using a different level of r^2 when clumping SNPs to that used in Section 5.2.1.1.1. This was to assess the hypothesis that too stringent values of r^2 may remove useful SNPs for AD prediction. The threshold used for r^2 was altered from 0.1 to 0.5. Analyses were conducted with and without the use of feature selection, with the selection techniques which achieved the highest AUC in Section 5.2.1.1 used.

5.2.1.1.3 Selecting SNPs using biological information

16,095 SNPs related to the function of microglia were used as predictors. Further analysis was also conducted using 35,997 variants related to synapses within the brain. These two SNP sets were clumped using the same p-value thresholds as in Section 5.2.1.1. Clumped SNPs were then passed to both ML algorithms and PRS.

5.2.1.2 Population stratification

To adjust SNPs for population stratification, PCs were calculated for both the training and test sets by using the *PCA* package within *Python*. Three PCs were used for adjustment, as to be consistent with analyses conducted in previous studies using the GERAD dataset (Leonenko, Shoai, et al., 2019). The same fit-*transform* function was used to generate PCs in both the training and test sets.

For deconfounding, a custom *scikit-learn* transformer was used. For adjustment, n regressions were performed, whereby n was equal to the number of SNPs in the chosen data set. These regressions were performed using the *statsmodels* package, in which the Ordinary Least Squares *OLS* function was used, regressing PCs from the SNP variables. Variants within the training and test sets were adjusted in separate functions, with the intention of minimising the risk of data leakage. The standardised residuals from each regression were then used instead of SNP genotypes for inputs to both ML algorithms and PRS calculation.

5.2.1.3 Machine learning methodology

Chapter Four investigated the performance of the following ML approaches: Random forests (RFs), gradient boosting (GB), Naïve Bayes (NB) and support vector machines (SVMs). Results showed little difference between the decision tree-based algorithms RFs and GB and SVMs for the prediction of AD. However, it was also noted that the use of SVMs resulted in longer training times than their counterparts. Therefore, only the decision tree-based algorithms were used thereafter.

5.2.1.3.1 Random forest methodology

The RF algorithm was implemented using the function *RandomForestClassifie*r from *Pythons' sklearn* package. Values for hyperparameters were specified for the following parameters *max_depth*, *min_samples_leaf*, *min_samples_split*. The package *RandomizedSearchCV* was used for tuning. This fits several models with a range of values for the defined hyperparameters, with the set of values which produce the best AUC chosen as the final inputs.

The RF was trained using the standardised residuals (as described in Section 5.2.1.2) to discriminate between cases and controls. These residuals were standardised using the *Python* function *StandardScaler*. This was carried out in training and test folds separately to avoid possible data leakage. Once the algorithms were trained, prediction performance was assessed in terms of AUC. This was calculated using the *roc_auc_score* function, which is also part of the *sklearn* package in *Python*. The probabilities of being a case are calculated using the sub-function *predict_proba*. The overall score of each ML algorithm was computed as the mean AUC across the five CV folds.

5.2.1.3.2 Gradient boosting methodology

The *Python* package *xgboost* was used to implement this technique. The function *XGBClassifie*r was used to fit the model, with residuals of genotypes used as inputs. The hyperparameters were tuned using the same *RandomizedSearchCV* function (similar to RFs). Once hyperparameters were tuned, AUC was computed in the same manner as RFs. To calculate AUC, both the *predict_proba* and *roc_auc_score* functions were used. Both predicted and observed values were passed to the *roc_auc_score* function to calculate AUC. As with RFs, the overall value of AUC was computed as the mean across the five CV folds.

5.2.1.4 Polygenic risk score calculation

SNPs used for PRS calculation were taken from variants deemed significant at the clumping stage (p-value thresholds = 0.0001,0.001,0.01,0.05,0,1,0.2,0.3,0.4,0.5). As discussed in Section 5.2.11, SNPs were selected using an in built GWAS within each round of CV. This allowed for the creation of SNP effect sizes and p-values in training samples only. PRS values were then subsequently calculated using allele scores in test samples only, ensuring separation of samples. This method was used due to the concerns of advantages afforded to PRS when using the Kunkle-nogerad summary statistics in Chapter 4. The same SNPs were used for both ML and PRS calculation to ensure a direct comparison between prediction methods. However, when feature selection techniques were employed, only a fraction of the same SNPs were used for ML. This was done intentionally to compare the traditional method of clumping and thresholding for PRS, with ML and the use of embedded feature selection. Therefore, more SNPs were used for the calculation of PRS on these occasions. LR is used to discriminate between case and control status. AUC for the LR was calculated in the same manner as the ML techniques. Due to the use of a LR, PRS modelling will be defined as PRS-LR from now on.

5.2.1.5 Missing data imputation

Missing genotypes were imputed using the same process outlined in Chapter 4 Section 4.3.2.1. This process was carried out per CV fold and after training/test splitting.

5.2.1.6 Parallel computing

To optimise run times, the *Multiprocessing* package in *Python* was used. This package enables users to parallelise computations, so they can be run simultaneously. For analysis in this chapter, the *Multiprocessing* package allowed CVs to be run in parallel, which is faster than running each fold sequentially. To achieve this the function *sub-class Pool* was used. A pool of five CPUs were used to allow for five rounds of CV to be run in parallel.

5.2.1.7 Calibrating prediction probabilities

Prediction probabilities were calculated using the *predict_proba* function and then plotted using a histogram from the *matplotlib* and *seaborn* libraries. A loess smoother was used to aid in assessing the relationship between predicted and observed probabilities. The initial classifier probabilities were then calibrated using the function *CalibratedClassifierCV* with five rounds of CV. Probabilities were calibrated using the *sigmoid* technique and then plotted against uncalibrated probabilities. All AUC values reported for analyses in this chapter were updated using the calibrated probabilities.

5.2.1.8 Feature selection methods

The central aims of this chapter was to use and compare a range of feature selection techniques. The implementation of these methods is outlined below.

5.2.1.8.1 Random forests for feature selection

RFs were used both as classifiers and as a form of feature selection in this chapter. For feature selection, the function *RandomForestClassifier* from *Python* was used to define a model. This model was fitted on the entire set of SNPs. A RF assigns weights to features based upon on their importance in prediction. The *Python* function *SelectFromModel* was used to assess these weights and select those features whose weights are greater than a defined threshold. If a threshold value is not pre-defined, the mean value of all weights is used. However, to ensure the optimum number of predictors were chosen, a for loop was used to select the threshold which returned the highest AUC for the subsequent ML algorithm.

5.2.1.8.2 ExtraTree classifier

The Extra Trees classifier was used as an alternative option for feature selection. For this, the *Python* function *ExtraTreesClassifier* from *sklearn* library was used. First, an Extra Trees model was developed. This model was then fitted using training data specified by the CV fold. Following this, the optimum set of features were selected using the same method for RFs outlined in Section 5.2.1.3.1.

5.2.1.8.3 LASSO regression

The function *SelectFromModel* from the package *sklearn.feature*_selection was used. This model was fitted on the adjusted SNPs in each CV fold. Those features whose coefficients were reduced to zero were then identified and removed from both the training and test sets. These reduced datasets were then passed to the ML algorithm for training and testing.

5.2.1.8.4 Elastic net algorithm

Lastly, the Elastic Net algorithm was tested for feature selection. To implement this, the *ElasticNet* function from the *Python* package *sklearn. linear_model* was used. The model was instantiated with two specified hyperparameters, these were *alpha* and *l1_ratio*. Where *alpha* is the constant with which the penalty term is multiplied and *l1_ratio* is the penalty. Those features which were penalised were then removed from the original dataset and then used for ML as inputs.

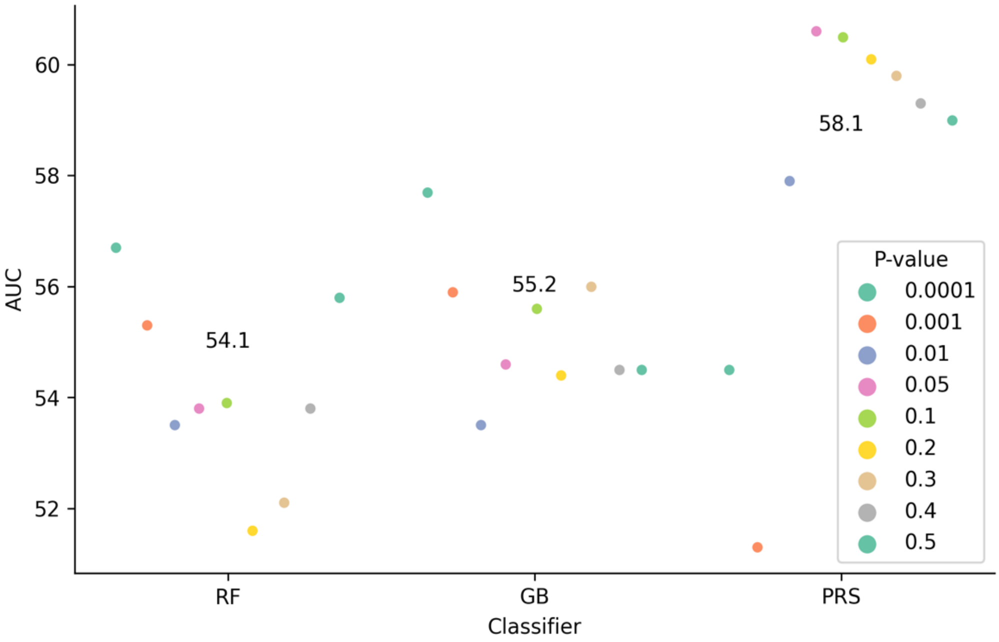
5.3 Results

Results of all analyses are supplied in Supplementary Tables 10-28, with analyses separated into five sections (detailed in Section 5.2). Those analyses described in Supplementary Tables 10-23 which achieved the highest overall AUC across methods were chosen to be displayed in plots in this chapter, created using *Python* packages *seaborn* and *matplotlib*. Calibration plots are also present displaying both non-calibrated and calibrated probabilities at a p-value threshold for SNP selection of 0.0001, with the classifier chosen to a RF from the previous figure.

5.3.1 Analysis without *APOE* alleles

Analyses in this section included SNPs within the *APOE* region. The use of feature selection was also omitted, results are shown in Figure 5.2.

Figure 5.2: The comparison of PRS-LR vs selected classifiers (RF, GB) for LD pruned SNPs, without the inclusion of *APOE* alleles.



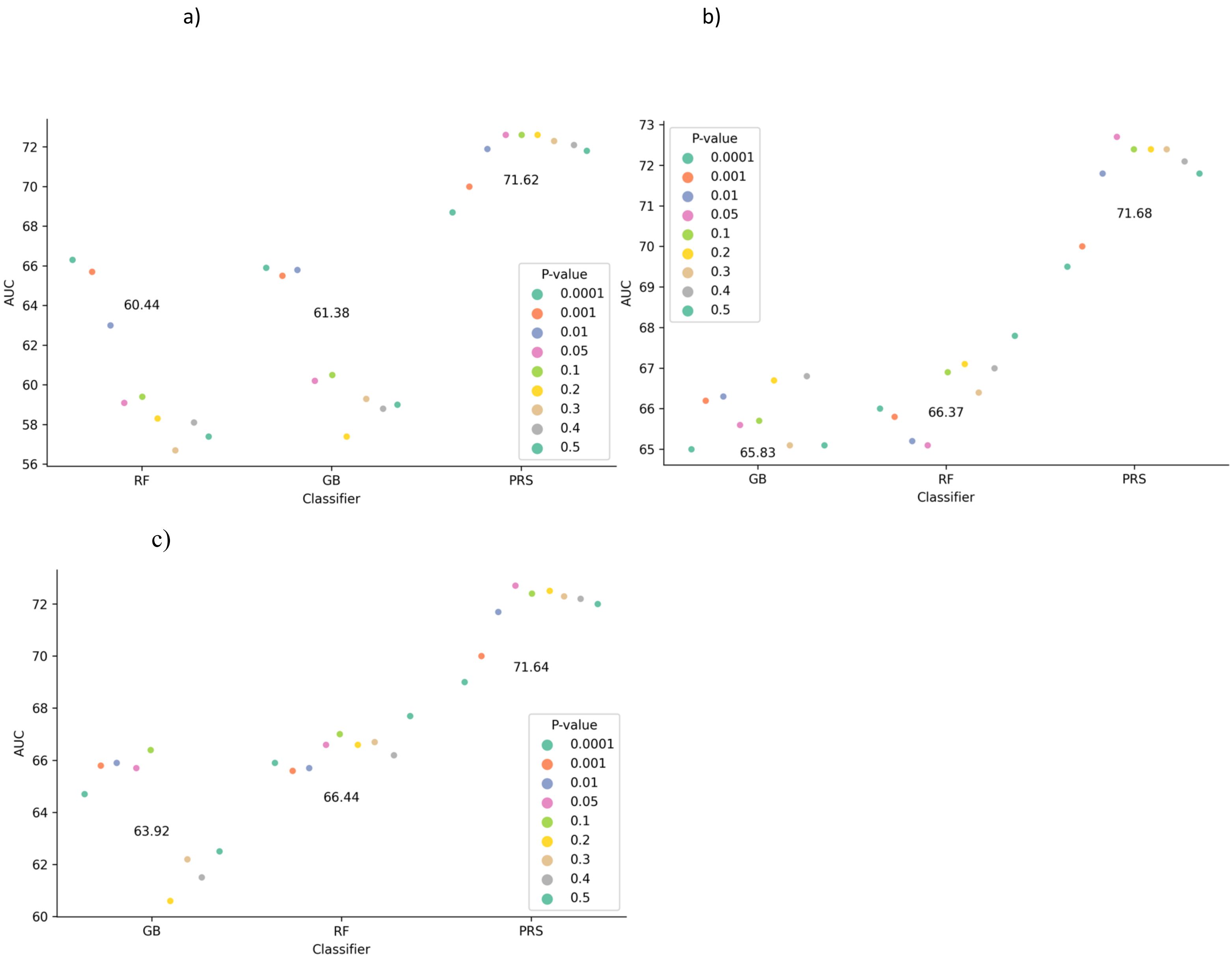
Y-axis represents AUC in %; with classifiers placed on the X axis. Each dot represents the score for the prediction algorithm for all p-value thresholds. The numbers placed centrally are the mean score across p-value threshold values; GB Gradient Boosting; RF Random Forest; PRS-LR Polygenic Risk Scores Logistic Regression; AUC Area Under the Curve.

Analysis represented in Figure 5.2 compared prediction performance of PRS-LR against both RFs and GB. The set of SNPs used for prediction included the *APOE* region but not *APOE* alleles. This is the only analysis in which SNPs within this region are included. For subsequent analyses, this region will be represented by allele counts for the ε2 and ε4 *APOE* variants. Results demonstrate that PRS-LR outperformed both RFs and GB, with mean AUC across all p-value thresholds 3-4% higher. This difference in prediction performance is shown to be statistically significant by analyses in Supplementary Table 24, where PRS-LR outperformed both ML algorithms in some of the p-value thresholds reported (PRS-LR vs GB = 0.0001, 0.01, 0.1, 0.3, 0.5) (PRS-LR vs RFs = 0.01, 0.1, 0.3, 0.5). AUC for PRS-LR improved from 52% towards 60% as p-values become less stringent (best performing p-value threshold of 0.05), whereas discrimination for both RFs and GB followed a different pattern as performance worsens as SNPs increase with an optimal p-value cut-off of 0.0001. This suggests that both ML algorithms are susceptible to overfitting as predictors number into the tens and hundreds of thousands.

5.3.2 Analysis of LD clumped SNPs ($r^2 = 0.1$) including *APOE* counts and feature selection

This section of analysis assessed the use of feature selection techniques on LD pruned SNPs. Three experiments were chosen to be portrayed in plots, the first of these being analysis where no feature selection was used, followed by the two feature selection algorithms which achieved the highest overall AUC across algorithms.

Figure 5.3: The comparison of PRS-LR vs chosen classifiers (RF, GB) for LD pruned SNPs, with a) no feature selection method used, b) the use of a RF for feature selection and c) the use of the ExtraTrees algorithm for feature selection.



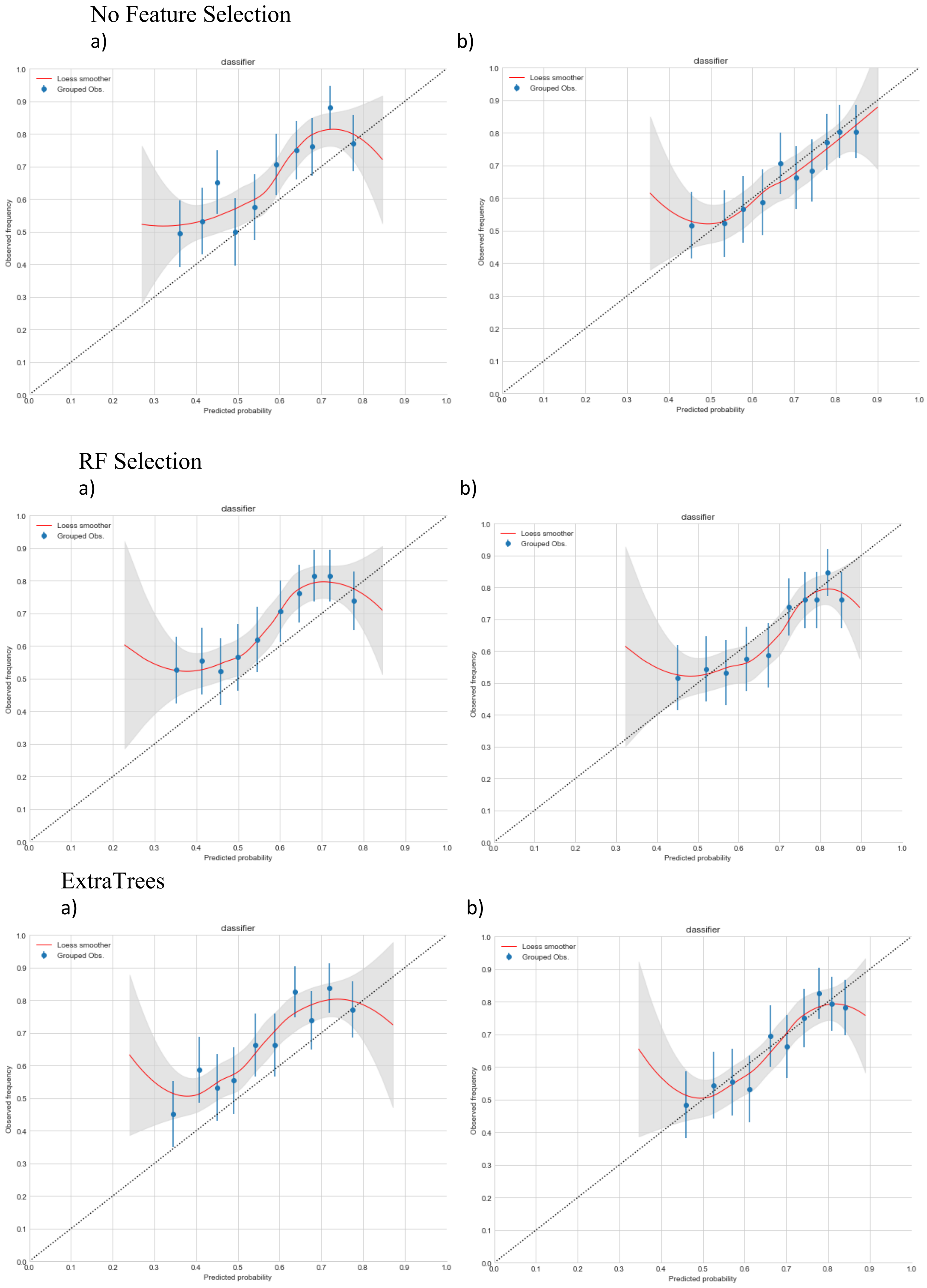
Y-axis represents AUC in %; with classifiers placed on the X axis. Each dot represents the score for the prediction algorithm for all p-value thresholds. The numbers placed centrally are the mean score across p-value threshold values; GB Gradient Boosting; RF Random Forest; PRS-LR Polygenic Risk Scores Logistic Regression; AUC Area Under the Curve.

Results shown in Figure 5.3a demonstrate the introduction of the *APOE* alleles ε2 and ε4, where ML algorithms were trained without the use of feature selection. Mean AUC across all p-value thresholds indicates an increase in prediction performance when compared to Figure 5.2, suggesting that the introduction of both *APOE* alleles benefited disease prediction.

However, similarly, to results in Figure 5.2, ML performance slightly worsened as the number of SNPs increased (typically after including SNPs with association p-values>0.05), with AUC for PRS-LR increasing and then stabilising as p-values become less significant. Unlike performance shown in Figure 5.2, AUC for PRS-LR remained more consistent as the number of variants increased. Significance of t-tests shown in Supplementary Table 25 also demonstrate that PRS-LR outperformed both ML algorithms across the majority of tested p-value thresholds (PRS-LR vs GB = 0.0001, 0.01, 0.1, 0.3, 0.4, 0.5) (PRS-LR vs RFs = 0.01, 0.1, 0.3, 0.4, 0.5).

Prediction performance for both GB and RFs increased further following the introduction of feature selection algorithms. Results shown in Figures 5.3b and 5.3c demonstrate an increase of 3-6% mean AUC across all p-value thresholds when compared to displayed results in Figure 5.3a. The grouping of coloured dots suggests that the use of feature selection reduced the impact on AUC of increasing numbers of SNPs, thereby enabling ML algorithms to deal efficiently with large numbers of predictors. However, despite the increase in prediction performance for both ML algorithms, PRS-LR still achieved superior mean AUC (71.6%). The better performance of PRS-LR is shown to be statistically significant when compared to using RFs for feature selection (PRS-LR vs GB = 0.0001, 0.01, 0.1, 0.3, 0.5) (PRS-LR vs RFs = 0.0001, 0.01, 0.1, 0.3, 0.5). This was also true for the use of ExtraTrees for feature selection (PRS-LR vs GB = 0.0001, 0.01, 0.1, 0.3, 0.5) (PRS-LR vs RFs = 0.0001, 0.01, 0.1, 0.3, 0.5).

Figure 5.4: The comparison of non-calibrated vs calibrated prediction probabilities for RFs using LD pruned SNPs



No Feature Selection

a)

b)

RF Selection

a)

b)

ExtraTrees

a)

b)

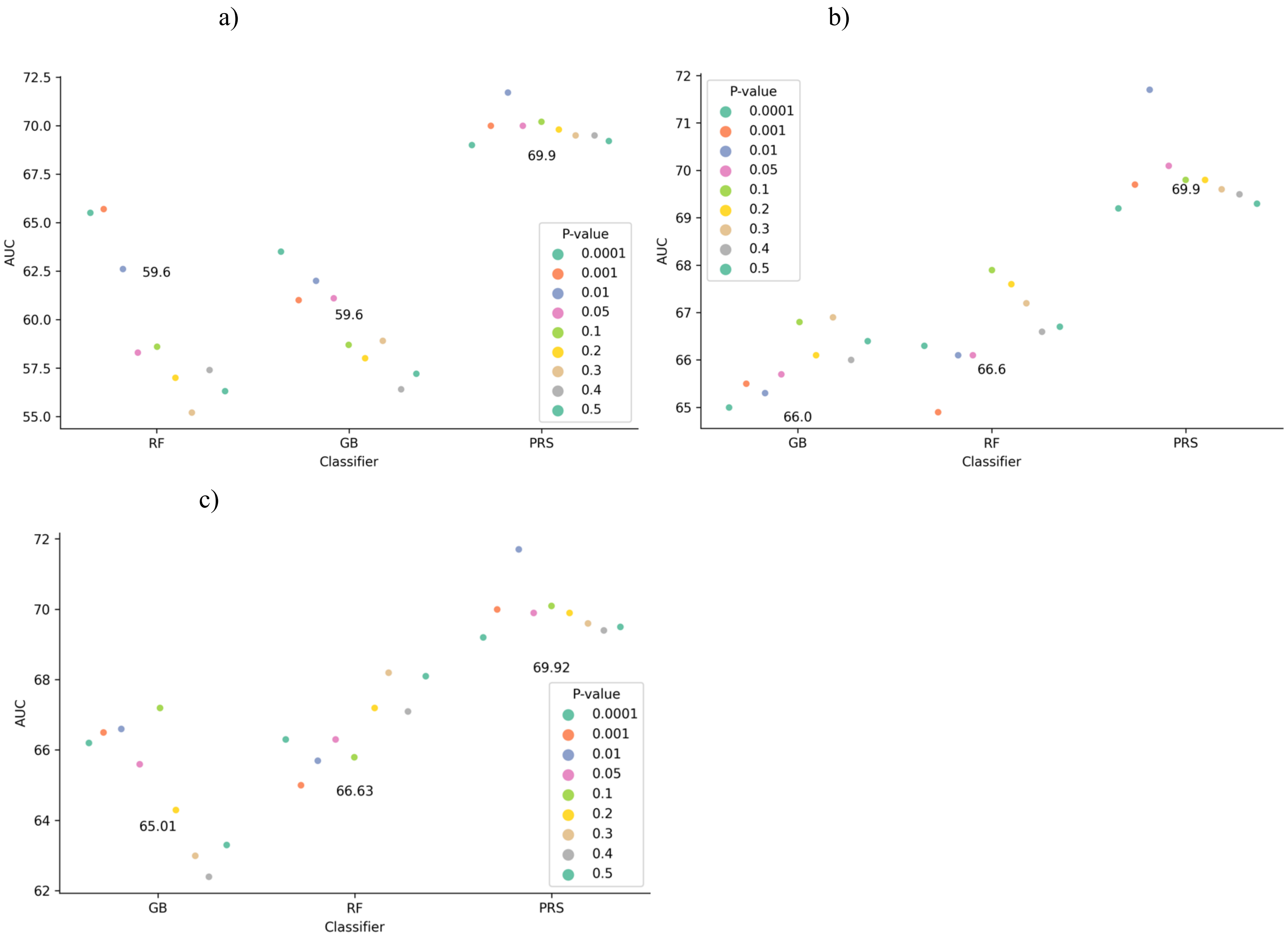
These figures represent pre a) and post b) calibration plots for the related RF algorithm (Figure 5.3) (p-value 0.0001). The x-axis represents the prediction output of the classifier in terms of the probability of being a case. With the y-axis denoting observed class frequencies. Perfect calibration in which predicted probabilities match observed accuracies is denoted by the diagonal dotted line. The blue dots represent the mean probability values within each quantile and are accompanied by a 95% confidence interval (blue bar). The overall relationship between predicted probabilities and observed frequencies (calibration curve) is given by the fitted loess smoother (red line), with a 95% (grey shaded area) used.

The calibration plots in Figures 5.4 suggest that predictions were initially underestimating risk for RFs, this is evidenced by loess smoothers lying above the diagonal. Following calibration, predictive probabilities are now more aligned with the diagonal line. However, this is not true for initial predictions in which the loess smoother still lies above the line.

5.3.3 Analysis of LD clumped SNPs ($r^2 = 0.5$) including *APOE* and feature selection

Analyses in Section 5.3.2 used an r^2 of 0.1 when clumping SNPs, this was altered to 0.5 in this section. This resulted in set of variants with greater levels of LD and therefore a greater volume then in previous analyses.

Figure 5.5: The Comparison of PRS vs RF, GB for LD Pruned SNPs using a Value of 0.5 for r^2, with a) no Feature Selection Method, b) used the use of a RF for Feature Selection and c) ExtraTrees.

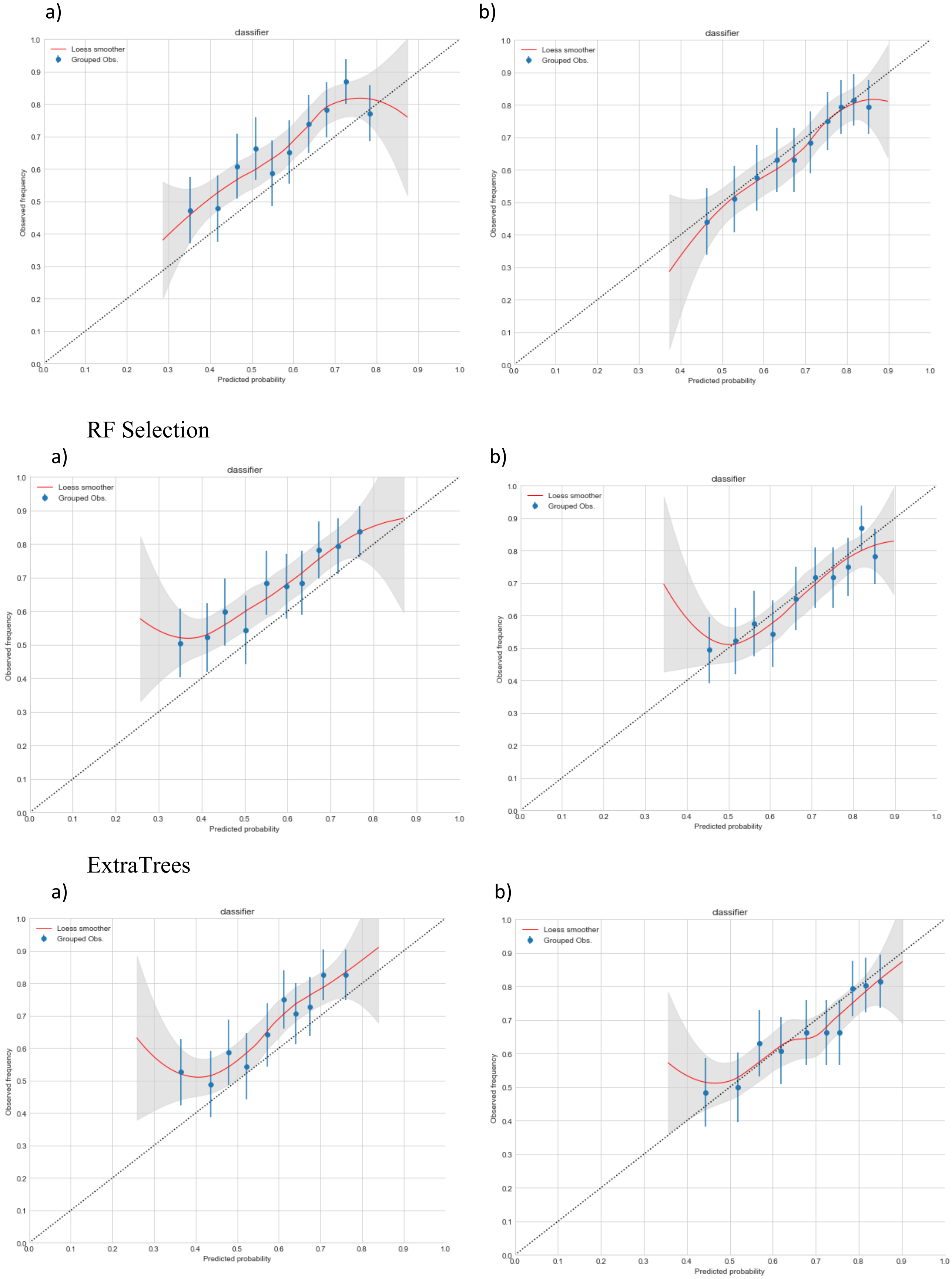


Y-axis represents AUC in %; with classifiers placed on the X axis. Each dot represents the score for the prediction algorithm for all p-value thresholds. The numbers placed centrally are the mean score across p-value threshold values; GB Gradient Boosting; RF Random Forest; PRS-LR Polygenic Risk Scores Logistic Regression; AUC Area Under the Curve.

The three plots shown in Figure 5.5 display similar analyses to those displayed in Figures 5.3, however a more lenient value off r^2 (0.5) was used when clumping. Prediction performance for both ML classifiers when not using feature selection was lower than those displayed in Figure 5.3a when using an r^2 value of 0.1 (Figure 5.5a vs Figure 5.3a) (GB = 59.6-61.4, RFs = 59.6-60.4, PRS-LR = 69.9-71.6). A reason for the reduced AUC for ML in Figure 5.5a could be the increase in the number of SNPs provided to algorithms by using a higher value of r^2 when clumping. An increased number of features will lead to increased dimensionality, which results in higher likelihood of overfitting.

The introduction of feature selection (Figures 5.5b, 5.5c) improved prediction performance for both GB and RFs. However, mean AUC across all p-value thresholds was similar to those in Figures 5.3, in which analyses were conducted using an r^2 of 0.1 for clumping. Therefore, the use of a more lenient value of r^2 did not result in a difference in classifier performance, with PRS-LR again outperforming ML algorithms. This is true without the use of feature selection (PRS-LR vs GB = 0.0001, 0.01, 0.1, 0.3, 0.5) (PRS-LR vs RFs = 0.0001, 0.01, 0.1, 0.3, 0.5), when using RFs for feature selection (PRS-LR vs GB = 0.01) (PRS-LR vs RFs = 0.3) and using ExtraTrees (PRS-LR vs GB = 0.01) (PRS-LR vs RFs = 0.01, 0.3).

Figure 5.6: The Comparison of non-Calibrated vs Calibrated Prediction Probabilities for RFs where a More Lenient Value of r^2 (0.5) for Clumping



No Feature Selection

a)

b)

RF Selection

a)

b)

ExtraTrees

a)

b)

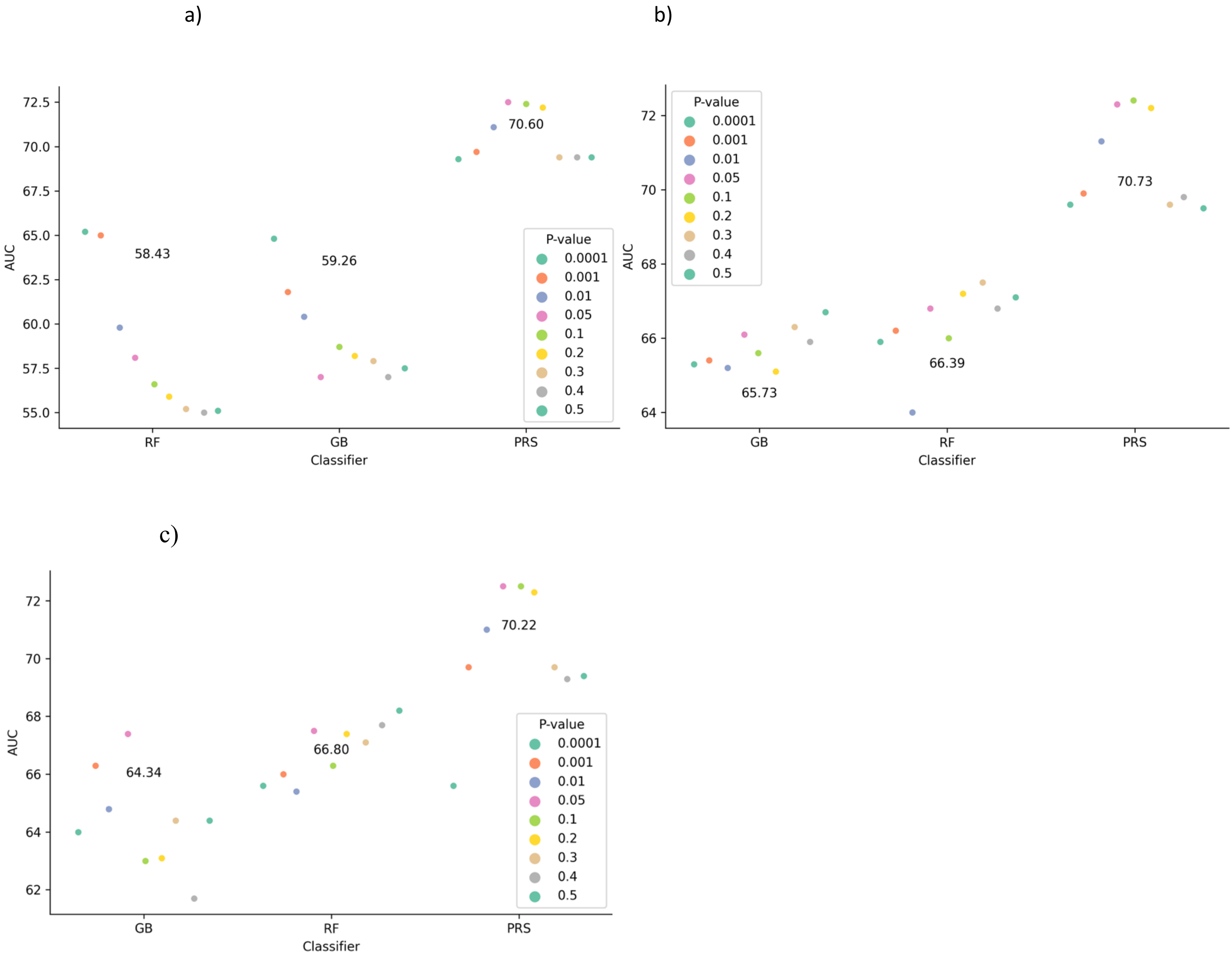
These figures represent pre a) and post b) calibration plots for the related RF algorithm (Figure 5.5) (p-value 0.0001). The x-axis represents the prediction output of the classifier in terms of the probability of being a case. With the y-axis denoting observed class frequencies. Perfect calibration in which predicted probabilities match observed accuracies is denoted by the diagonal dotted line. The blue dots represent the mean probability values within each quantile and are accompanied by a 95% confidence interval (blue bar). The overall relationship between predicted probabilities and observed frequencies (calibration curve) is given by the fitted loess smoother (red line), with a 95% (grey shaded area) used.

Calibration plots in Figure 5.6 demonstrate a consistent over estimation of risk for classifiers as loess smoothers lie above the diagonal prior to calibration. However, prediction probabilities realign with the diagonal following calibration, suggesting that the model was now estimating risk accurately.

5.3.4 Analysis of non-LD pruned SNPs.

This section displays results for three different analyses, a) no feature selection, b) RFs and c) ExtraTrees used for feature selection. Here SNPs were not clumped before being passed to feature selection techniques. The first analysis compared the performance of PRS and ML algorithms, with no feature selection techniques employed.

Figure 5.7: The comparison of PRS-LR vs chosen Classifiers (RF, GB) for non-LD pruned SNPs, with a) no feature selection method used, b) the use of a RF for feature selection, and c) the use of a ExtraTrees for feature selection.



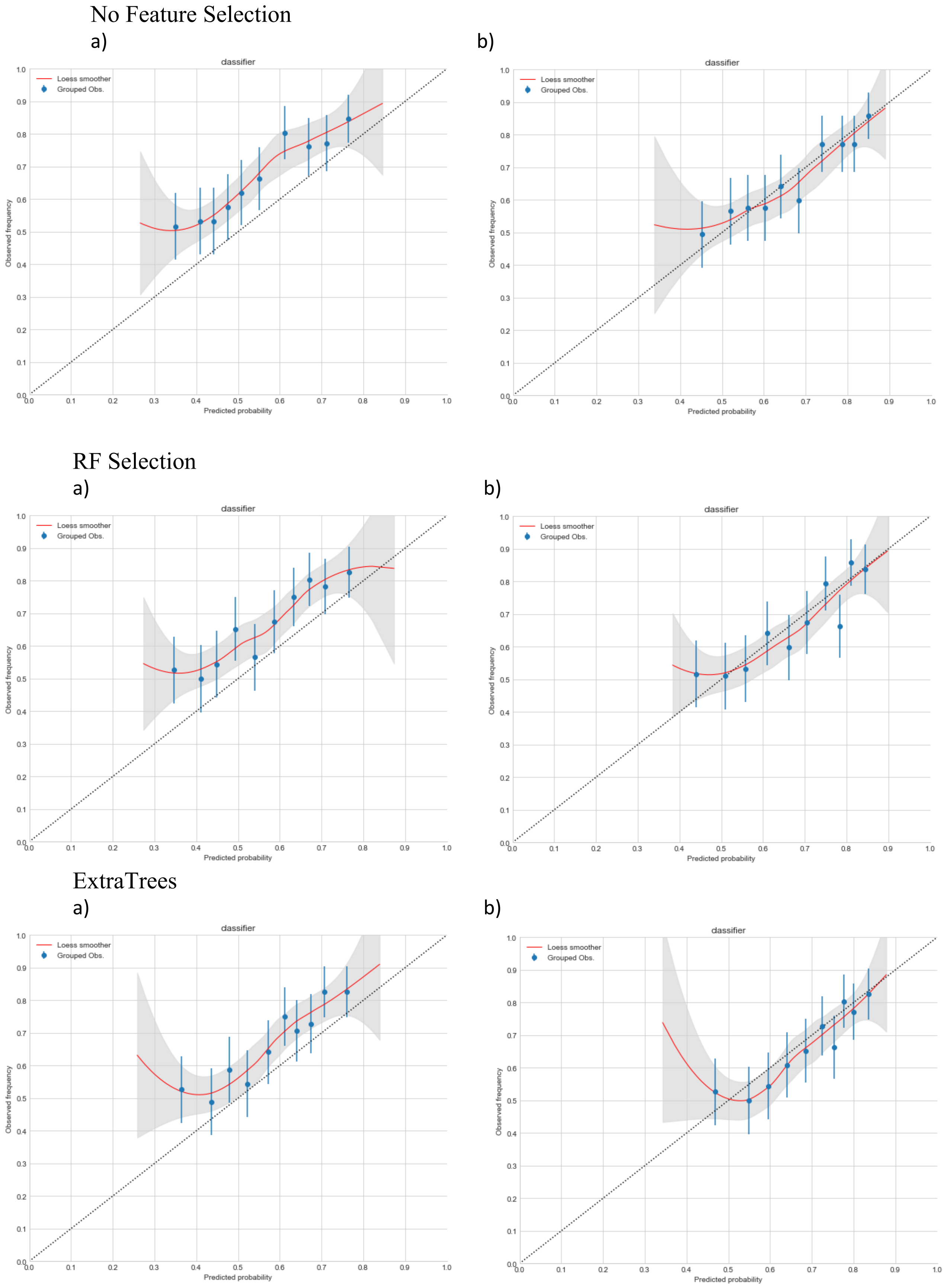
Y-axis represents AUC in %; with classifiers placed on the X axis. Each dot represents the score for the prediction algorithm for all p-value thresholds. The numbers placed centrally are the mean score across p-value threshold values; GB Gradient Boosting; RF Random Forest; PRS-LR Polygenic Risk Scores Logistic Regression; AUC Area Under the Curve.

Plots represented in Figure 5.7 display the same analyses conducted in Figures 5.3 & 5.5, however on this occasion the clumping process was removed. When not using feature selection, AUC for both GB and ML algorithms detailed in Figure 5.7a were similar to those in Figure 5.5a. This suggests that increasing the number of SNPs through removing the clumping phase did not result in alterations in performance. When comparing classifier performances, statistics detailed in Supplementary Table 26 demonstrate that PRS-LR achieved higher prediction than both ML algorithms (PRS-LR vs GB = 0.0001, 0.01, 0.1, 0.3, 0.5) (PRS-LR vs RFs = 0.0001, 0.01, 0.1, 0.3, 0.5).

The use of feature selection represented in Figures 5.7b and 5.7c resulted in an increase in prediction performance for both GB and RFs when compared to AUC in Figure 5.7a, with mean AUC increasing by 7-8% for both ML algorithms. Mean AUC values using both RFs and ExtraTrees algorithm were similar to those in Figures 5.3b, 5.3c, 5.5b and 5.5c. Similarly, to analyses in Section 5.3.2, the use of feature selection did not alter higher performance of PRS-LR over ML algorithms. This was true for the use of both RFs (PRS-LR vs GB = 0.0001, 0.01, 0.1, 0.3, 0.5) (PRS-LR vs RFs = 0.0001, 0.01, 0.1, 0.3, 0.5) and the ExtraTrees algorithm (PRS-LR vs GB = 0.0001, 0.01, 0.1, 0.3, 0.5) (PRS-LR vs RFs = 0.0001, 0.01, 0.1, 0.3, 0.5).

When summarising findings from Figures 5.2 – 5.7, two main conclusions can be drawn. The linear method PRS-LR consistently outperforms the more complex ML algorithms throughout analyses. The margin between methods can be reduced following the use of feature selection, however PRS-LR remains the superior prediction method. These trends are observed irrespective of the pruning method used, either through the removal of clumping or alterations in the value of r^2 used.

Figure 5.8: The comparison of non-calibrated vs calibrated prediction probabilities for RFs when removing clumping.



No Feature Selection

a)

b)

RF Selection

a)

b)

ExtraTrees

a)

b)

These figures represent pre a) and post b) calibration plots for the related RF algorithm (Figure 5.7) (p-value 0.0001). The x-axis represents the prediction output of the classifier in terms of the probability of being a case. With the y-axis denoting observed class frequencies. Perfect calibration in which predicted probabilities match observed accuracies is denoted by the diagonal dotted line. The blue dots represent the mean probability values within each quantile and are accompanied by a 95% confidence interval (blue bar). The overall relationship between predicted probabilities and observed frequencies (calibration curve) is given by the fitted loess smoother (red line), with a 95% (grey shaded area) used.

Similarly, to previous plots, the RFs appear to be underestimating risk in Figure 5.8. However, following calibration, predictive probabilities lie more closely to the perfect calibration line. However, initial predictions between 0.3-0.5 remained above the diagonal in all tree plots, suggesting an underestimation of risk.

5.3.5 Feature selection using biological information.

Previous sections used statistical techniques to reduce dimensionality. Further investigation was also conducted in which biological information was used to select SNPs for analysis. The results of which are shown in Supplementary Figure 1. In comparison with Figure 5.3a where SNPs were selected through clumping all SNPs within GERAD, mean AUC for RFs was similar, although the performance of GB was reduced. Discrimination between classes here for PRS was lower, with AUC falling by around 2%. Therefore, the use of biologically informed SNP sets reduced the prediction performance of both GB and PRS-LR, which is expected as other genes/SNPs associated to AD are not included in these analyses. Despite this reduction, PRS-LR still outperformed both ML algorithms, this is evidenced in statistics provided in Supplementary Table 28. This was true for the use of SNPs related to the function of Microglia (PRS-LR vs GB = 0.01, 0.1, 0.3, 0.5) (PRS-LR vs RFs = 0.01, 0.1, 0.3, 0.5), as well as those variants related to the function of synapses (PRS-LR vs GB = 0.001, 0.01, 0.1, 0.3, 0.5) (PRS-LR vs RFs = 0.001, 0.01, 0.1, 0.3, 0.5). In contrast with previous analyses, RFs outperformed GB for some p-value thresholds, this was the case for microglia related variants (RFs vs GB = 0.01, 0.1, 0.3) and synapse related SNPs (RFs vs GB = 0.0001, 0.01, 0.1, 0.3, 0.5).

As can also be seen in Supplementary Figure 2, both RFs were underestimating risk initially. After calibration, predicted probabilities lie closer to perfect calibration for Microglia related SNPs. However, calibration has not improved the distribution of probabilities for Synapse related variants.

5.4 Discussion

Work conducted in this chapter focused on two main aims. The first of these was to examine whether using larger number of SNPs could improve AD prediction beyond analyses conducted in Chapter 4. The second aim was to assess the performance of several feature selection techniques, with the purpose of determining whether any of these reduced the issues associated with high dimensionality.

5.4.1 Machine learning performance

Analyses conducted in Chapter 4 focused on testing the performance of a series of ML algorithms on a small selection of AD related SNPs. Prediction algorithms achieved AUC of 57-60% when using 23 and 422 SNPs (including SNPs in the *APOE* region). Analyses in this chapter used variants selected on a genome-wide scale, with the intention of assessing whether using increased amounts of SNPs could improve upon this level of AUC. Results shown in Supplementary Table 10 demonstrate that AUC fell to 55% and 52% AUC for RFs and GB respectively when increasing the number of SNPs to ~50,000 (including SNPs in the *APOE* region without special consideration of the *APOE*-specific effects). Thus, increasing the number of SNPs resulted in decreased performance for all algorithms. A possible reason for this could be the issue of increased dimensionality, whereby increasing the number of features alongside a fixed number of samples can reduce prediction performance. This issue is known as overfitting and occurs when an ML algorithm becomes too reliant on training data, reducing performance when using unseen samples (Ying, 2019). However, increased dimensionality would not have the same outcome for PRS-LR as the same number of predictors are used irrespective of the number of SNPs. Therefore, the decrease in AUC for PRS could be linked to the calculation of GWAS statistics per CV fold, where effect sizes for less significant SNPs might become less reliable due to sample sizes.

All subsequent analyses differed in how the *APOE* region was represented. To reiterate, all variants within this region were removed from the original SNP set, with direct allele counts for both the ε2 and ε4 *APOE* variants used instead. Results demonstrated a difference in prediction performance (see Figures 5.2 and 5.3), with the inclusion of the ε2/ε4 alleles improving prediction by around 10% AUC when compared with results shown in Supplementary Table 10. This effect was observed throughout all subsequent analyses, with the increase in performance potentially explained by the significant association between ε2/ε4 alleles and AD risk. Therefore, we conclude that this approach may provide greater prediction utility than when including SNPs within the *APOE* region and not alleles. A further possible reason for the difference in AUC could be the poor coverage of the *APOE* region by the genotyping array used. This lack of coverage might result in missed variation attributed to AD, in addition to ε2/ε4 alleles.

5.4.2 The performance of statistical feature selection techniques

Another aim of this chapter was to assess the capability of feature selection algorithms to reduce the effect of the curse of dimensionality for ML. A range of different techniques were explored. Analysis represented in Figures 5.3 and 5.4 display the impact of dimensionality issues on prediction. As the r^2 used for clumping becomes less stringent, the number of SNPs used for classification increases. Initially, this may provide more predictive information to classifiers, which might explain increases in AUC until a certain number of SNPs is reached. However, AUC decreases from 65% to 55% for p-value threshold <0.1. This decrease in discrimination may be related to the increased number of predictors (SNPs) relative to the fixed number of samples used during training.

The best performing feature selection techniques were RFs and the Extra Trees algorithm. These consistently selected variables across all p-value thresholds which retained AUC with a mean 65%. For the more stringent of p-value thresholds (0.0001), most often only the ε4 *APOE* allele was chosen. Therefore, the overlap in features selected between p-value thresholds was often 100%, as only the *APOE* allele was selected. As p-value thresholds became less lenient and the number of SNPs provided to the feature selection increased, further SNPs were selected in addition to the *APOE* allele. However, despite this increase in selected SNPs, the overlap in selected features between p-value thresholds was low, with overlap ranging between 1-3%. Therefore, prediction performance appears to be mostly driven by the *APOE* e4 allele, with additional SNPs providing little utility, proven by inconsistent selection across thresholds.

Both the LASSO and Elastic Net algorithms did not perform as well as the decision tree-based algorithms with respect to prediction accuracy. These approaches selected SNPs at each p-value threshold but did not retain AUC as well when compared to the decision tree-based algorithms. Whilst the number of SNPs selected by these two algorithms were less than when no feature selection method was used, greater amounts were still selected when compared to both RFs and ExtraTrees. Feature selection techniques were effective in retaining AUC across p-value thresholds, however none of the methods resulted in ML performing above PRS-LR. None of the reduced sets of SNPs from feature selection resulted in discrimination >70%. The dimensionality reduction approaches could only replicate the performance of the most significant threshold for clumping with a p-value threshold of 0.0001. Increasing the number of SNPs through less stringent p-values and selecting SNPs using statistical techniques provided no greater prediction utility.

However, there is one obvious advantage of feature selection, which is the reduction in training times for ML algorithms. Fewer features are passed to the ML, which in turn reduces the number of calculations and computational memory required. Both LASSO and Extra Trees feature selection reduced computation times as compared to not using any technique. Another advantage of using feature selection techniques is the freedom to employ more extensive hyperparameter searches for ML algorithms. This stems from the reduced training times and strain on computational resources due to reduced amounts of features. These reductions increase available resources for hyperparameter searches, which might result in more accurate algorithms.

5.4.3 Biological information for feature selection

16,095 SNPs relating to the function of microglia and 35,997 relate to the synapse were used as features. Selecting SNPs based on their relatedness to two potential neuropathological aspects of AD, microglia and the synapse, reduced prediction performance for both GB and PRS-LR, with only RFs retaining similar prediction accuracy. Restricting the available set of SNPs might remove SNPs from other relevant to neurodegeneration biological functions, which provide better predictive utility (in combination with microglia and synapse SNPs sets).

5.4.4 LD-based pruning as feature selection

To ascertain whether LD clumping for removing SNPs was important for prediction, I analysed datasets which had not been clumped prior to classification. Here, SNPs were selected based upon their association to AD (p-values) after each AD GWAS was computed per CV fold. The removal of the clumping phase resulted in a larger number of SNPs used for analysis. For instance, p-value thresholds of 0.5 returned >200,000 SNPs. Results show that omitting the clumping phase made no difference to the performance of ML. AUC for RFs and GB were similar, the only difference occurring in the number of SNPs present in each p-value threshold. Interestingly PRS-LR without clumping performed similarly to when clumping was used, with AUC again greater than any ML algorithm.

It was also hypothesised that using a more stringent value of r^2 during LD-clumping may remove SNPs important for prediction. In the first section of analysis, an r^2 value of 0.1 was used. Using a less stringent value (r^2 of 0.5), resulted in larger numbers of SNPs after LD pruning. Results show that this had no significant impact on classifier performance. This was also true for the two chosen methods of feature selection: RF selection and ExtraTrees. In line with the previous analysis, PRS-LR again outperformed all ML techniques.

5.4.5 The use of parallel computing

Analyses conducted in this chapter involved the use of up to hundreds of thousands of SNPs. The parallel computing through the *Python* package multiprocessing was used to quicken analysis. Analysis of feature sets containing 100,000 or more SNPs which were previously taking up to a week to complete, were now running in up to 24 hours. Parallel Computing allowed greater flexibility in making alterations to analysis and also enabled larger amounts of SNPs to be processed in time.

5.4.6 Calibration statistics

Following calibration, predictions tended to lie more closely to the diagonal line of perfect calibration, in which predicted probabilities match class values better. AUC values for all ML algorithms in this chapter were adjusted using calibrated probabilities, resulting in a marginal increase for prediction performance.

5.4.7 PRS versus ML

The results presented in this chapter show that the PRS-LR outperformed ML across all sections of analysis. This was evidenced by statistics generated from paired t-tests (Supplementary Tables 24-28), in which the mean difference between values of AUC were tested. The superior performance of PRS-LR occurred despite the use of effect sizes and p-values calculated from a GWAS within CV, rather than using statistics from an external source (Kunkle *et al.*, 2019). The only occasion in which the performance of PRS-LR reduced was the use of biological information for selecting SNPs, instead of using variants from the whole genome. Reasoning for the superior performance of PRS-LR may be due to the simplicity of the PRS model, thus reducing overfitting and resulting in more robust risk prediction. In addition, ML uses individual genotypes adjusted using population stratification and not accounting for the effect sizes (B-coefficients from logistic regression). These coefficients might be providing PRS-LR with additional predictive information.

Several studies have attempted to compare the predictive capabilities of PRS-LR and ML in other diseases, with results demonstrating mixed outcomes, with ML outperforming PRS-LR in some cases, whilst PRS-LR was superior in others (Attaran and Deb, 2018a). The reasons for this are not yet clear, factors such as the quality of data, genetic disease architecture and ML models used may influence this (Attaran and Deb, 2018a).

5.4.8 Limitations

Analyses conducted in this chapter were subject to some limitations. The first limitation is the imputation of missing values (SNP genotypes), in which the overall modal value of the variant was used to fill missing values (Section 5.2.1.5). The use of the modal value may reduce the variance of genotypes for each imputed SNP, which in turn might have reduced effect sizes of SNPs (Das, Nayak and Pani, 2019). A technique for imputation which avoids skewing variances for imputed features is the use of ML, however this requires greater computational resource than using averages. (Das, Nayak and Pani, 2019). Considering the large number of SNPs used in analyses in this chapter, the use of ML methods for imputation would result in longer training times. Therefore, the use of modal values was continued.

A second limitation is the use of thresholding and clumping approaches for PRS-LR. This could be considered a traditional approach for assessing disease risk. There are novel recent approaches to PRS generation, which may be better performing methods for discrimination between cases and controls with PRS (Lewis and Vassos, 2020c). Therefore, other techniques of PRS might outperform ML by a greater margin than clumping and thresholding. This thought process has resulted in the use of more complex methods of PRS (PRS-CS) in subsequent chapters (Chapter 7) of this thesis.

5.4.9 Conclusions

The two main aims of this Chapter were 1) to analyse a larger number of SNPs for disease prediction and compare the results with PRS-LR, and 2) evaluate the performance of feature selection techniques. For the case of ML, the introduction of larger number of SNPs did not provide any further improvement for disease prediction. As the number of SNPs increased, discrimination further reduced. This can most likely be explained by the increase in dimensionality and overfitting. The two statistical feature selection algorithms used (RFs and Extra Trees Algorithm) reduced the effects of increased dimensionality; but they didn't improve discrimination beyond SNPs chosen at the most significant p-value thresholds. However, feature selection enabled ML to be trained in shorter periods of time and so provided an efficiency advantage. PRS-LR consistently outperformed ML across all analyses, with prediction performance increasing until plateauing with increasing number of SNPs, suggesting that ML's ability to analyse complex patterns in this data is still outperformed by the linear method. The importance of efficient programming was also highlighted in this chapter, as the use of parallelisation decreased run times from days to hours. This is important message for the field of genetics as analysis can often involve the use of many thousands of predictors (Kalina, 2014).

6 Assessment of the predictive capability of machine learning using imputed genotype data

6.1 Introduction

Investigations in this chapter focused on the use of imputed sets of variants when conducting comparisons between machine learning (ML) and polygenic risk score (PRS). Analyses in previous chapters used non-imputed variants, however it was theorised that increasing the number of single nucleotide polymorphisms (SNPs) might increase prediction performance. The same methodology used to develop PRS in Chapter 5 was used for analyses in this chapter, with a genome wide association study (GWAS) run within each fold of cross validation (CV).

Datasets consisting of non-imputed genotypes typically comprise 10,000-1,000,000 single nucleotide polymorphisms (SNPs) (Li et al., 2009b). Whilst genome wide association studies (GWAS) using this number of common variants (MAF > 1%) have published significant results in AD (Harold *et al.*, 2009), a large proportion of the genetic heritability of the disease is yet to be explained (Ridge et al., 2013). Analyses in this chapter will use imputed genotypes for AD prediction, with original SNPs (Harold *et al.*, 2009). Imputed best-guess genotypes and dosages are used for prediction, with the aim of comparing the performance between them. The same machine learning (ML) algorithms used in Chapters 5 and 6 were also used, with the intention of evaluating prediction in both imputed and non-imputed SNPs. Two of the feature selection algorithms tested in Chapter 5 were also employed (Random Forest, ExtraTrees algorithm). This was to establish whether their use resulted in a better prediction performance in comparison to non-imputed genotypes. Alongside this analysis, the predictive performance of ML was also compared to PRS.

Aims:

1. To compare the performance of both ML and PRS on predicting AD using imputed variants.
2. To assess whether there were differences in prediction performance for both ML and PRS between allelic dosage values and best-guess genotypes.

3. To compare the results of both ML and PRS to those achieved when using non-imputed SNPs (Chapters 4&5).
4. To assess the use of feature selection on ML performance, with the intention of comparing whether feature selection using imputed variants leads to different results than non-imputed SNPs (Chapter 4&5).

6.2 Methods

6.2.1 Data

Data used in this chapter originated from the GERAD consortium (Harold *et al.*, 2009). Following quality control (QC) and exclusion of the 1958 birth cohort (described in Chapter 4), 4603 samples remained for prediction, with 1554 controls and 3049 cases.

6.2.2 Predictors

Analyses in this chapter used imputed SNPs. Two different imputation formats were analysed: genotypic dosages and best-guess genotypes, with the intention of comparing the performance of PRS and ML on both. Further comparisons were also drawn prediction performance between imputed variants and non-imputed SNPs in previous chapters.

6.2.2.1 Imputed dosages

The non-imputed version of the GERAD (Harold *et al*., 2009) dataset was imputed within University by Dr Aura Frizzati. The haplotype reference consortium (HRC) was used as the reference panel of European ancestry, with 39,235,157 SNPs for imputation. The Michigan imputation server was used for this task https://imputationserver.sph.umich.edu/index.html#!. To conduct analyses using the dosage format, I conducted a number of pre-processing and quality control (QC) steps. SNPs were initially in the format of genotype dosages (later converted to allelic dosages) and resided in 22 (one for each chromosome) genset files. These were converted to the pgen format using the genetic software package *PLINK 2.0* (Chen *et al.*, 2019). QC steps were then used to remove SNPs considered poor quality; carried out using *PLINK 2.0*. Imputation quality can be measured using the metric r^2. Variants with r^2 scores <0.7 were excluded (Hanks *et al.*, 2022), with the intention of removing variants with substandard imputation quality. The function *--extract* from *PLINK 2.0* was used to achieve this in each chromosome separately.

Further QC steps then followed, with the *--maf* function used to retain all SNPs with a minor allele frequency (MAF) ≥0.01. The *PLINK 2.0* function *--geno* was then used to exclude all SNPs with missing genotype frequency ≥0.05. Following this, the command *--hwe* was used to test each variant for departure from Hardy Weinberg equilibrium (HWE), with a value of 1e-6 (Hanks *et al.*, 2022) used to accept or reject variants. Following QC, the resulting .chr files were combined into a single file using the command *--pmerge-list.*

6.2.2.2 Imputed genotypes

All pre-processing and QC steps for the imputed genotyped version of the dataset were conducted by a separate individual prior to this thesis. Dosage files were initially converted to VCF format, with subsequent transformation into the standard binary PED format for *PLINK 1.9.* QC steps were then used to remove variants using *PLINK 2.0,* with SNPs whose INFO score was less than 0.4 filtered out. An INFO score can be defined as a measure of uncertainty for imputation, with scores ranging from zero to one. Scores close to zero represent inaccurate imputation, with confidence increasing as values tend towards one (Mitt *et al.*, 2017). SNPs were filtered using *MAF, geno* (SNPs were excluded based on missing genotype rate %) and *HWE,* in which thresholds of ≤0.1, >0.05 and <1e-06, respectively.

6.2.3 Methods

6.2.3.1 ML training and testing

Analyses conducted in this chapter assessed large number of SNPs, which required significant amounts of computational resources. Therefore, all analyses were run using the high-performance computing cluster 'Hawk' (Supercomputing Wales) https://www.supercomputing.wales. Classifiers were trained and evaluated using a nested cross-validation (CV) approach. The *Python* function *StratifiedKFold* was used to achieve this, with five rounds of CV. The stratified nature of this function leads to the class ratio of the entire dataset being preserved in each round of CV. This reduces the possible variation in model performance per CV round, which may occur if class labels are poorly distributed through CV folds (Prusty, Patnaik and Dash, 2022). Another advantage of *StratifiedKFold* is the ability to access training and test folds separately within each round of CV. This allows pre-processing techniques to be applied per data split, reducing the possibility of data leakage (Bey *et al.*, 2020).

6.2.3.2 SNP selection

SNPs were selected for both ML and PRS using an inbuilt GWAS performed per CV fold. This was achieved by calling *PLINK* within *Python*, made possible by a function known as a *subprocess*. For analyses involving imputed genotypes, *PLINK 1.9* was used. Effect sizes and p-values were calculated using a logistic regression (LR) generated from the *PLINK* function *--logistic*. PCs were used to account for population stratification and were computed using *Python's sklearn* package. Only those individuals from the training sets of CV were used in the LR, as well as when computing PCs. This was to avoid possible data leakage between training and test sets. Clumping was performed using *PLINK 1.9's --clump* function, this was called within *Python* using *subprocess*. A window of 1,000 kb was used, with $r^2 = 0.1$ (Privé *et al.*, 2019). Three p-value thresholds were used, 0.0001, 0.1 and 0.5. Similarly, to the GWAS phase, only individuals from the training set were used. Following clumping, the resulting SNP from each p-value threshold were formatted for ML. This was achieved by the function *--recodeA* within *PLINK 1.9,* which codes genotypes in a 0-2 format.

The method of conducting the per-CV GWAS when using dosages performed with *PLINK 2.0* and the .pgen file format. PCs used were calculates in the same manner when using genotypes, with only samples within the training folds of CV used. The output of the GWAS was again formatted into summary statistics. *PLINK 1.9* was used for clumping. As this version of the tool cannot handle pgen format files, these were converted to VCF format. The genomic window and r^2 values were the same as those used for genotypes, with the same three p-value thresholds used (0.0001, 0.1, 0.5). SNPs chosen through clumping were then converted into a format suitable for ML. The *PLINK 2.0* functions *--export* and --*A* were used to achieve this.

6.2.3.3 Imputed SNPs QC

SNPs with more than 5% of genotypes missing were excluded and then were additionally imputed, as further removal of missing values could result in a loss of samples. The modal value of each SNP was calculated from the variant's genotypes using *mode* from *Pythons'* package *NumPy*. Missing values within each SNP were then imputed with the respective modal value. The additional imputation was applied per CV fold to avoid a possibility of data leakage.

6.2.3.4 Population stratification correction

To adjust SNPs for population stratification, PCs were calculated for both the training and test sets separately by using the *PCA* package within *Python*. Three PCs were used for adjustment, as this is the number of PCs used in previous analyses conducted using the GERAD dataset (Leonenko, Sims, et al., 2019). The same function was used to generate PCs in both the training and test sets. To control for population stratification, a custom scikit-learn transformer was used (outlined in previous chapters).

6.2.3.5 Machine learning methodology

Two types of supervised ML algorithms were used in this chapter, Random Forests (RFs) and the gradient boosted decision tree (GB). The *Python* package *RandomForestClassifier* was used to implement RF. SNPs were chosen using a similar approach used in Chapter 5, whereby the imputed SNP sets were clumped at three p-value thresholds (0.0001,0.1,0.5) (Escott-Price, Sims, Bannister, Harold, Vronskaya, Majounie, Badarinarayan, Morgan, *et al.*, 2015). *APOE* alleles were also included and adjusted for population stratification (as described in Section 6.2.3.4). Since the GERAD dataset is imbalanced, comprising almost twice as many cases as controls (1554 controls, 3049 cases), the option *balanced* was used, which determines class distributions in the training set and inversely adjusts the weighting for the minority class. Discrimination was assessed using area under the curve (AUC), calculated using the function *roc_auc_score* from *sklearn*. The function *predict_proba* was used to calculate prediction probabilities for each classifier, which were then passed to *roc_auc_score* to calculate AUC.

To implement GB, the *Python* package *XGBoost* with the classifier *XGBClassifier*. The *Python* function *RandomizedSearchCV* was again used to tune hyper-parameters. These were *max-depth, n-estimators* and *learning rate*. Similarly, to RFs, GB performance can be affected by class imbalances in the training data. The function *XGBClassifier* includes a hyper-parameter called *scale_pos_weight* which functions in the same manner as *class_weight* for RFs. A weight must be passed to the function to redress the imbalance. This is calculated by dividing the number of minority class instances in the dataset by the class number, with the result multiplied by 100. In the case of the GERAD data, the result was 66. For both classifiers, an individual score for AUC was calculated per CV fold, leading to five scores per classifier. The mean of each set of AUCs was reported as the overall performance of each classifier.

6.2.3.6 Polygenic risk score

PRS were generated only for individuals in the test set, determined from the CV split with GWAS statistics being calculated in the training set. Following this, the *APOE* counts alleles (e4 and e2) for each individual were derived and subsequently multiplied by their respective effect sizes and then summed to produce a final variable "*APOE*" to be used as an additional covariate to PRS.

The next stage was to adjust PRS to account for population stratification. This was achieved using a linear regression, whereby risk scores were regressed on PCs. The residuals of the regressions were then normalised to enable comparisons between different risk scores within CV. Following this, the final model for PRS classification was developed. The two variables (PRS_without_*APOE APOE* and "*APOE*" variable) were used as explanatory variables for a LR. The *Logit* function from the *Python* package *statsmodels* was used to fit the model, with the AD phenotype as the response variable. For the purposes of comparisons between ML and PRS, this LR will be denoted as PRS-LR.

6.2.3.7 Discrimination

Discrimination for prediction algorithms was assessed using AUC. For analyses in this chapter, AUC was reported as the mean value across five rounds of CV. Therefore, for each p-value threshold, a single value was reported for both ML algorithms and PRS-LR. Paired t-tests were used to test for differences in prediction performance across the five rounds of CV.

The false discovery rate (FDR) controlling method Benjamini-Hochberg was used, with the programming language *R* employed to make adjustments through the function *p.adjust*. Corrections were made on an analysis wide basis (for each supplementary table at one time), with all p-values corrected using the same function.

6.2.3.8 Calibration

The function *CalibratedClassifierCV* from *Python's sklearn* was used to calibrate results. Inputs to the function were the original classifier, the method isotonic and three rounds of rounds of CV used. The function was fitted on the original training data, with predictions made in the test data. These newly calibrated probabilities were then plotted against observed probabilities for assessment, and further examined.

6.2.3.9 Feature selection algorithms

Following the results of Chapter 5, for the use of RFs for feature selection, the function *RandomForestClassifier* was used. Hyper-parameters were left as the default values from *scikit-learn*, with only the *class_weight* parameter specified as *balanced*. Once the RF had been fitted on the training data provided, relevant features were selected using the function *SelectFromModel*. The second method of feature selection used was the *ExtraTrees* algorithm using function *ExtraTreesClassifier* in the training data. The function *SelectFromMode*l was then used to select features with the specified threshold from the previous *ExtraTrees* model. Features chosen were then passed onwards for classification purposes.

6.2.3.10 Comparing features selected across AD association p-value thresholds

The stability of a feature selection algorithm is a core metric for assessing its performance. To assess the performance of feature selection algorithms used in this chapter, SNPs selected when using the three p-value thresholds (0.0001, 0.1, 0.5) were compared. The overlap between SNPs selected for the p-value threshold of 0.0001 were compared to 0.1, with 0.1

then compared to 0.5. These statistics were calculated for both feature selection algorithms used.

6.3 Results

6.3.1 SNPs

Following several QC steps, 6,756,941 SNPs were available for analyses in dosage format, with 6,107,587 variants also available as best-guess genotypes.

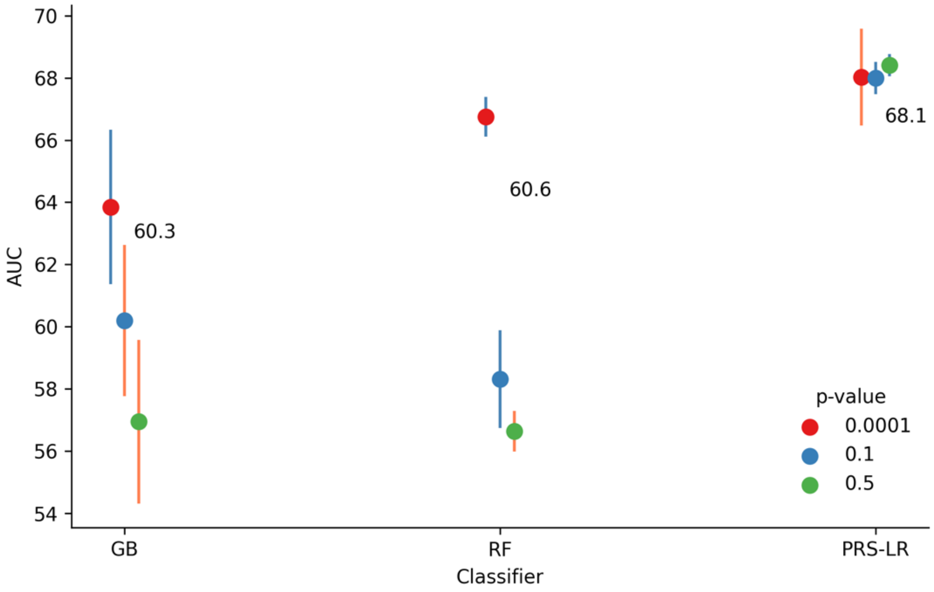
6.3.1.1 Prediction without the use of feature selection

Results for ML and PRS-LR analyses are given Supplementary Tables 29–34. These tables provide the mean AUC of each classifier across five folds of CV for each p-value threshold. The results also include the average number of SNPs used for classification across five folds of CV, as well as the mean number of features used for feature selection. The overlap in features between the p-value thresholds are also provided. Results of the t-tests are detailed in Supplementary Tables 35-40. Results for this chapter are only reported if the comparison between any two algorithms was significant (p-value ≤ 0.05).

6.3.1.1.1 Best-guess genotypes

Initial analysis focused on comparing the performance of ML vs PRS-LR using best-guess genotypes. Three p-value thresholds were used for the clumping process: 0.0001, 0.1 and 0.5. Figure 6.1 displays results for the chosen ML algorithms versus PRS-LR.

Figure 6.1: PRS-LR vs Selected Classifiers (RF, GB) for LD Pruned SNPs in Imputed Genotypes, with the Inclusion of *APOE* Alleles.

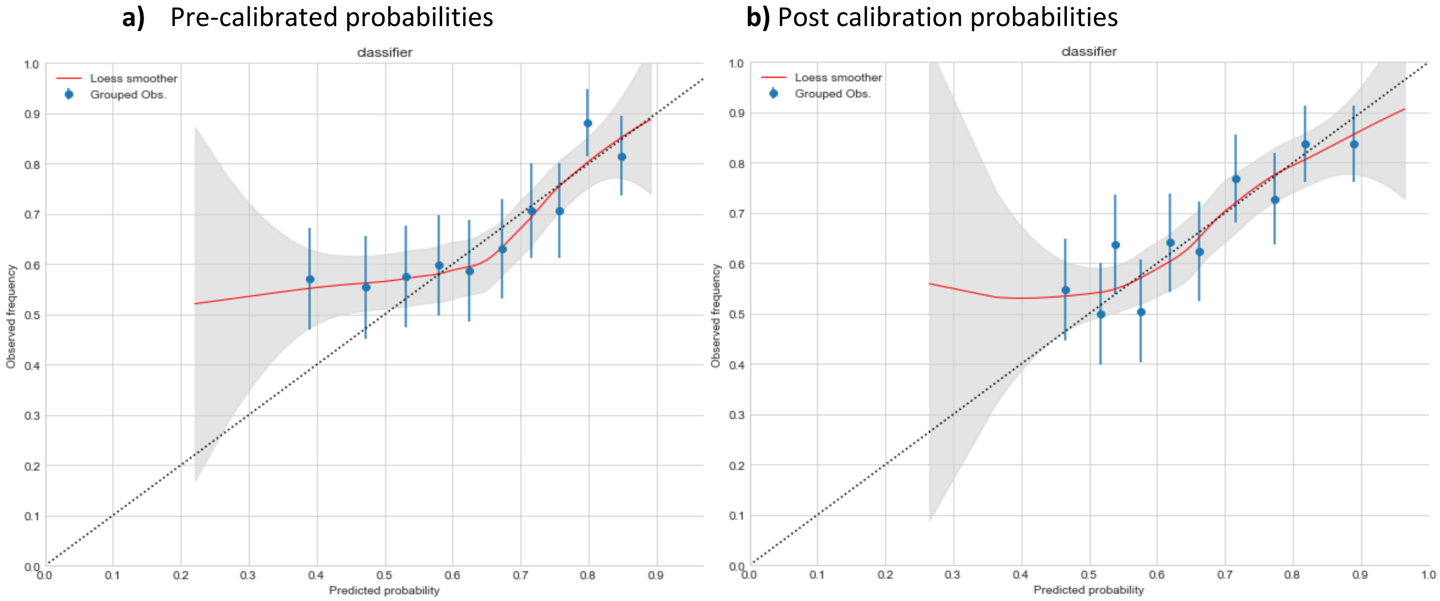


Y-axis represents AUC in %; X-axis represents each classifier's results for a p-value threshold. Each dot represents the mean score for the prediction algorithm across 5 folds of CV, with an accompanying 95% CI bar. The numbers placed centrally are the mean of the three p-value threshold scores; GB Gradient Boosting; RF Random Forest; PRS-LR Polygenic Risk Scores Logistic Regression; AUC Area Under the Curve.

Performance of the decision tree-based classifiers is below that of PRS-LR, with mean AUC at 60.3 and 60.5% for ML (GB, RFs) and 68.1% for PRS-LR. The higher performance for PRS-LR is supported by t-test statistics reported in Supplementary Table 35, where PRS-LR outperformed GB and RFs for two p-value thresholds (0.1,0.5) (PRS-LR vs GB p-values = 1.79e-02, 5.90e-04) (RFs p-values = 8.09e-03, 5.61e-06). Prediction performance for ML reduced with the increase of the p-value threshold, with AUC for the RF reducing from 67% to 54% and performance for GB reducing from 64 to 56%. Confidence intervals for each data point demonstrate that scores for GB were more varied than those of PRS and RFs.

The decrease in ML performance could be attributed to an increase of the number of SNPs used for prediction, from 100 to 128,000. As the p-value used for clumping becomes less stringent and the resulting number of variants increases, it might become more difficult for ML to distinguish the true signal between SNPs and AD from random noise. Random noise may reduce the algorithm's ability to identify the true underlying pattern within the dataset, leading to overfitting. The prediction performance of GB was similar to RFs, with the average prediction performance across the five folds of CV 0.2% less than RFs. This is less of an issue for PRS-LR, as only one or two predictors are used. Performance for PRS-LR was similar across the three p-value thresholds, with mean AUC of 68.1%.

Figure 6.2: Non-Calibrated vs Calibrated Prediction Probabilities for GB



These figures represent pre a) and post b) calibration plots for the related GB algorithm (Figure 6.1) clumped at a p-value of 0.1. The x-axis represents the prediction output of the classifier in terms of the probability of being a case. With the y-axis denoting observed class frequencies. Perfect calibration in which predicted probabilities match observed accuracies is denoted by the diagonal dotted line. The blue dots represent the mean probability/observed values within each quantile and are accompanied by a 95% confidence interval (blue bar). The overall relationship between predicted probabilities and observed frequencies (calibration curve) is given by the fitted loess smoother (red line), with a 95% confidence interval (grey shaded area) used.

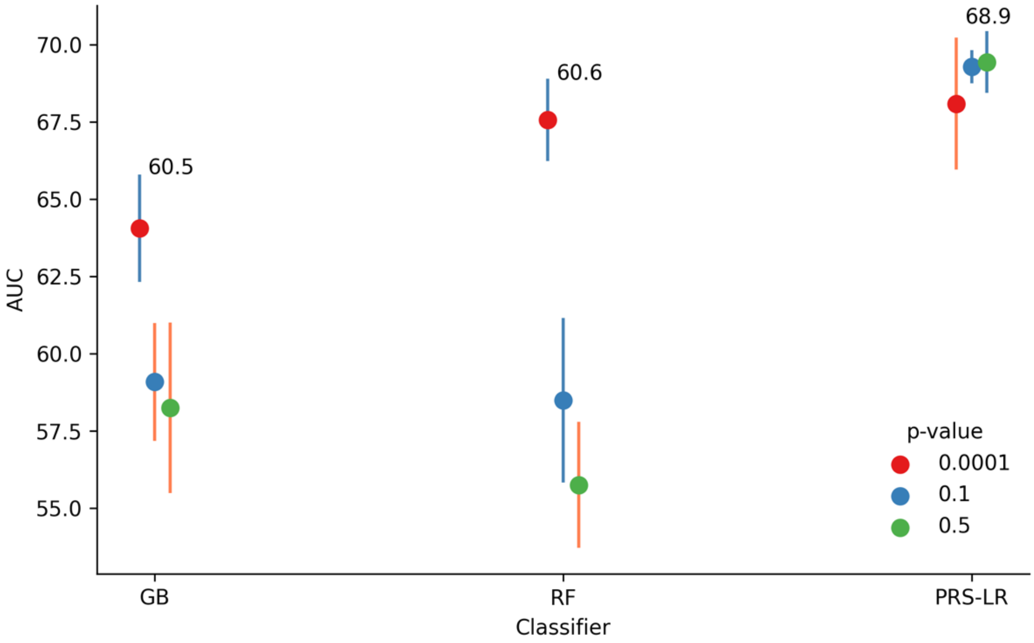
Figures 6.2a and 6.2b demonstrate the calibration plots for the GB (Figure 6.1) at a p-value threshold of 0.1 Model probabilities have been calibrated using the isotonic regression method. Results in Figure 6.2a show that the model is generally underestimating disease risk between 0.25-0.5 predicted probabilities, due to the calibration curve being above the diagonal line. This is due to the algorithm underestimating the likelihood of samples being cases, suggesting that they are controls. However, as the predicted probability moves towards 0.5, probabilities are closer to the diagonal, suggesting that the model is making predictions with greater confidence. Figure 6.2b demonstrates the GB algorithm post calibration. Overall, the calibrated line lies closer to the diagonal, which indicates that calibrating probabilities resulted in more accurate predictions. However, probabilities are still underestimated between 0.25 and 0.5 predicted probabilities.

A consistent observation made for the majority of calibration plots in this chapter is the widening of the confidence interval (grey shaded area) for the loess smoother at the left-hand side. This can be explained by the greater number of cases than controls in the dataset and in turn reduced number of predicted probabilities less than 0.5. This reduction in information results in a reduced confidence for the fitted loess line.

6.3.1.1.2 Dosages

Results shown in Figure 6.3 show analyses when using dosages and no feature selection.

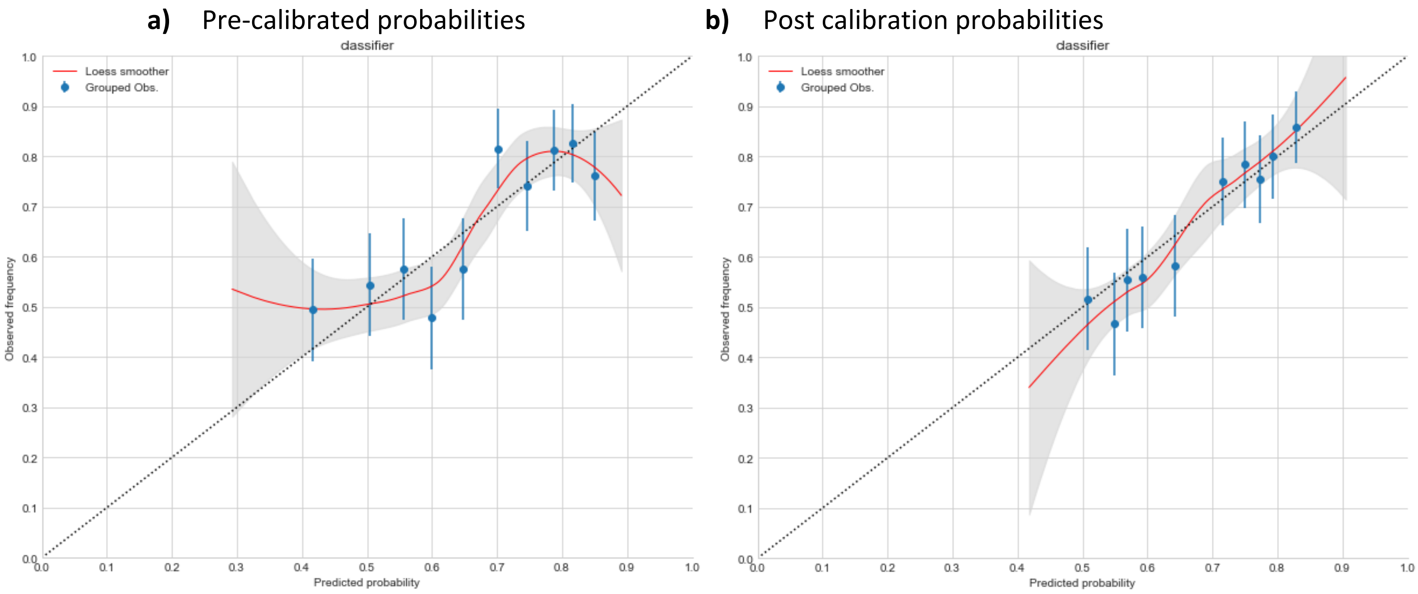
Figure 6.3: PRS-LR vs Selected Classifiers (RF, GB) for LD Pruned SNPs in Imputed Dosages, with the Inclusion of *APOE* Alleles.



Y-axis represents AUC in %; X-axis represents each classifier's results for a p-value threshold. Each dot represents the mean score for the prediction algorithm across 5 folds of CV, with an accompanying 95% CI bar. The numbers placed centrally are the mean of the three p-value threshold scores; GB Gradient Boosting; RF Random Forest; PRS-LR Polygenic Risk Scores Logistic Regression; AUC Area Under the Curve.

Results displayed in Figure 6.3 are similar to those in Figure 6.1. Mean AUC for all three classifiers show that prediction performance between imputed genotypes and dosages was similar (GB 60.3-60.5, RFs 60.5-60.6, PRS-LR 68.1-68.9). Prediction performance for both decision tree-based algorithms worsened as the p-value threshold used for clumping became less stringent. The performance comparison between the decision tree-based ML methods and PRS-LR was also similar when compared to Figure 6.1. Results of paired t-tests in Supplementary Table 36 demonstrate that PRS-LR outperformed both GB and RFs across two p-value thresholds (0.1,0.5) (GB p-values = 1.79e-02, 5.90e-04) (RFs p-values = 8.09e-03, 5.67e-06). When assessing confidence intervals for each classifier, it can be determined that scores were less varied for PRS-LR than both GB and RFs.

Figure 6.4: Non-Calibrated vs Calibrated Prediction Probabilities for GB


a) Pre-calibrated probabilities

b) Post calibration probabilities

These figures represent pre a) and post b) calibration plots for the related GB algorithm (Figure 6.3) clumped at a p-value of 0.1. The x-axis represents the prediction output of the classifier in terms of the probability of being a case. With the y-axis denoting observed class frequencies. Perfect calibration in which predicted probabilities match observed accuracies is denoted by the diagonal dotted line. The blue dots represent the mean probability/observed values within each quantile and are accompanied by a 95% confidence interval (blue bar). The overall relationship between predicted probabilities and observed frequencies (calibration curve) is given by the fitted loess smoother (red line), with a 95% confidence interval (grey shaded area) used.

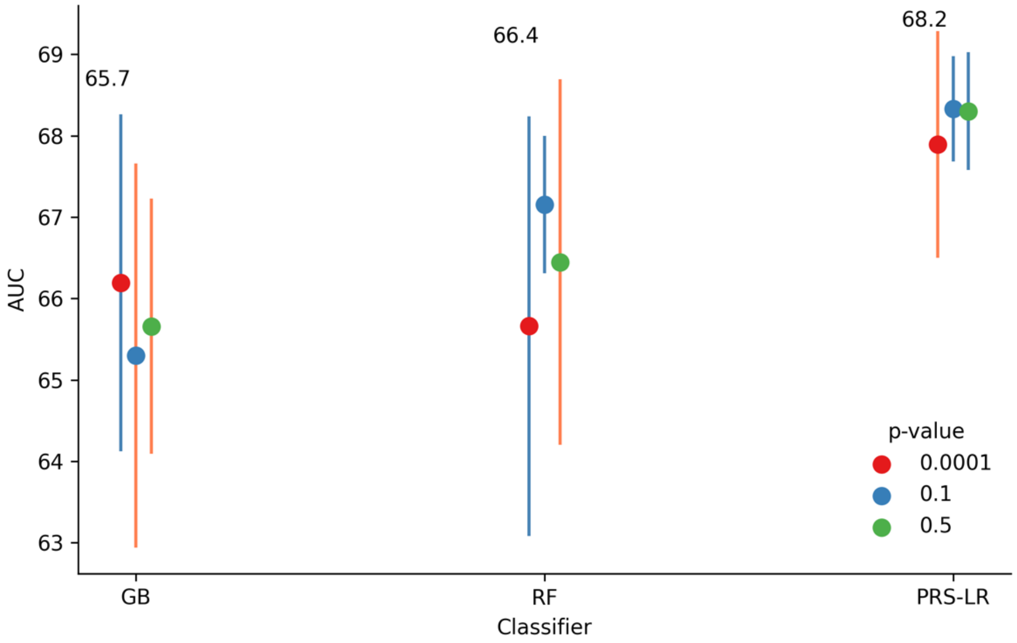
Prediction probabilities shown in Figure 6.4a demonstrate that prior to calibration, the GB in Figure 6.3 was underestimating risk for class memberships predicted probabilities between 0.3 and 0.5. This is due to the algorithm predicting true cases with a greater likelihood of being controls. Following this, the calibrated line moves above and below the line of perfect calibration, demonstrating that prediction accuracy was inconsistent. After calibration using isotonic regression, the calibration line follows more in line with the diagonal in Figure 6.4b. This suggests the recalibrated model estimates risk more accurately.

6.3.1.2 Prediction with the use of an RF for feature selection

6.3.1.2.1 Best-guess genotypes

Results displayed in Figure 6.5 show analyses when using genotypes and RF for feature selection.

Figure 6.5: PRS-LR vs Selected Classifiers (RF, GB) for LD Pruned SNPs in Imputed Genotypes, with the Inclusion of *APOE* Alleles and a RF used for Feature Selection.

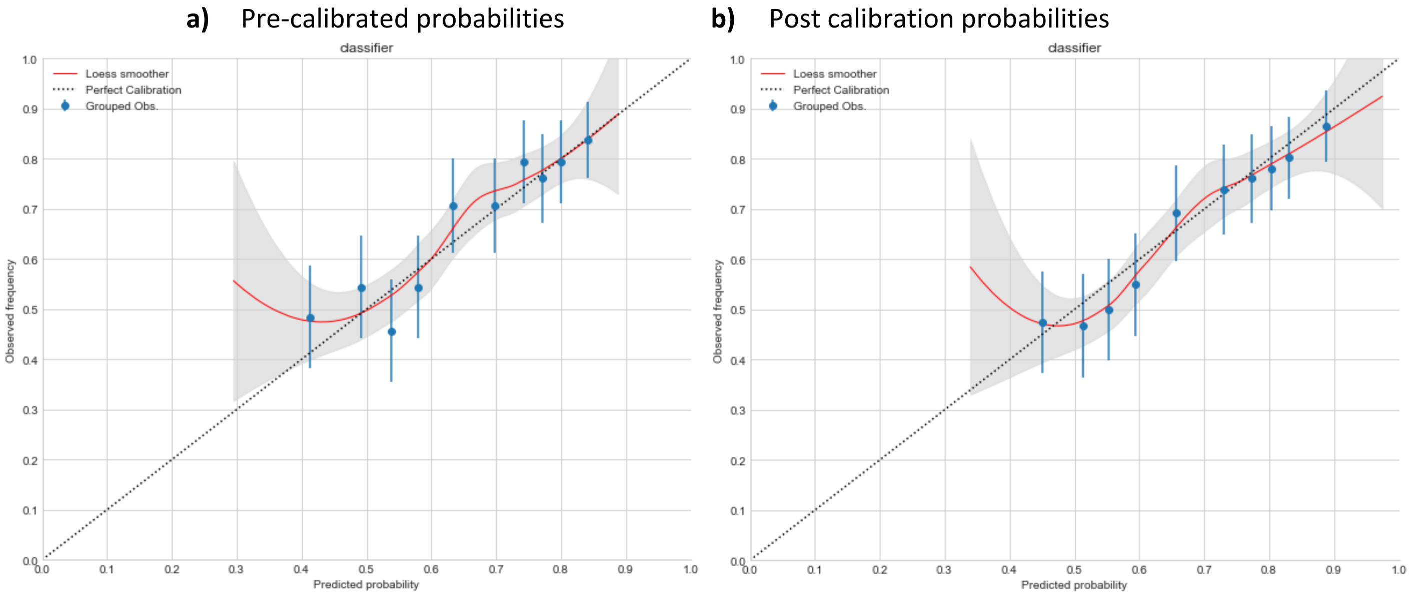


Y-axis represents AUC in %; X-axis represents each classifier's results for a p-value threshold. Each dot represents the mean score for the prediction algorithm across 5 folds of CV, with an accompanying 95% CI bar. The numbers placed centrally are the mean of the three p-value threshold scores; GB Gradient Boosting; RF Random Forest; PRS-LR Polygenic Risk Scores Logistic Regression; AUC Area Under the Curve.

When comparing results shown in Figure 6.5 with those shown in Figure 6.1, a similarity between both sets of analyses is the better performance of PRS-LR when compared to GB and RFs, with mean AUC 2% greater for PRS-LR. However, this difference is less than results observed in Figure 6.1. Therefore, the use of an RF for feature selection resulted in the reduced the loss of AUC for ML as p-values become less significant, as observed for results in Figures 6.1& 6.3. The algorithm may only be selecting the most relevant SNPs, despite the number of available features increasing. Therefore, features deemed redundant are removed. This suggests that the employment of feature selection may have reduced the possibility of dimensionality issues. Despite the use of feature selection, mean AUC for PRS-LR was still higher than both GB and RFs, however these differences were shown not to be statistically significant (Supplementary Table 37). The size of confidence intervals for all three classifiers were larger than in both Figures 6.1& 6.3. Therefore, it would appear that the introduction of feature selection increased the variation of AUC values.

A further point of discussion is the increase in AUC for both RFs and GB from a clumping p-value threshold of 0.0001 to 0.1. This may indicate that hyperparameter optimisation is succeeding in reducing the effect of increased dimensionality on prediction performance. However, AUC falls for both algorithms when using a p-value threshold of 0.5. This suggests that differences in AUC between p-value thresholds might be due to random variation across five folds of CV when using feature selection.

Figure 6.6: Non-Calibrated vs Calibrated Prediction Probabilities for GB



a) Pre-calibrated probabilities

b) Post calibration probabilities

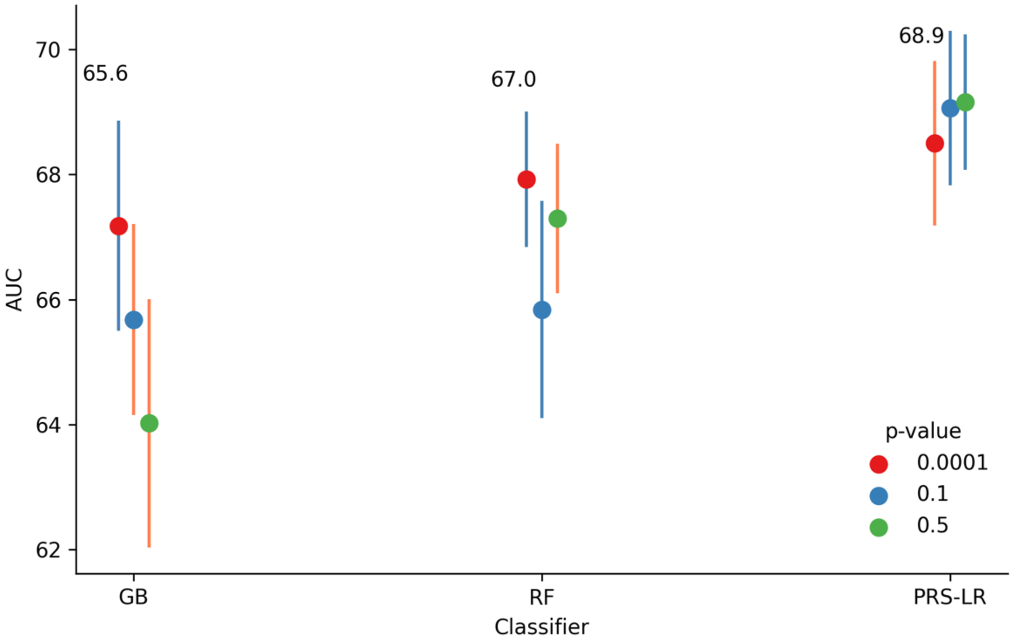
These figures represent pre a) and post b) calibration plots for the related GB algorithm (Figure 6.5) clumped at a p-value of 0.1. The x-axis represents the prediction output of the classifier in terms of the probability of being a case. With the y-axis denoting observed class frequencies. Perfect calibration in which predicted probabilities match observed accuracies is denoted by the diagonal dotted line. The blue dots represent the mean probability/observed values within each quantile and are accompanied by a 95% confidence interval (blue bar). The overall relationship between predicted probabilities and observed frequencies (calibration curve) is given by the fitted loess smoother (red line), with a 95% confidence interval (grey shaded area) used.

Risk is underestimated for predicted probabilities between 0.3–0.5 (Figure 6.6a). Subsequent probabilities follow the perfect calibration line consistently. However, post-calibrated probabilities between 0.3–0.5 were still underestimated. In contrast, subsequent predictions were relatively close to the diagonal.

6.3.1.2.2 Dosage data

Analyses in this section involved the use of dosages and RFs for feature selection.

Figure 6.7: PRS-LR vs Selected Classifiers (RF, GB) for LD Pruned SNPs in Imputed Dosages, with the Inclusion of *APOE* Alleles and an RF used for Feature Selection.



Y-axis represents AUC in %; X-axis represents each classifier's results for a p-value threshold. Each dot represents the mean score for the prediction algorithm across 5 folds of CV, with an accompanying 95% CI bar. The numbers placed centrally are the mean of the three p-value threshold scores (dots); GB Gradient Boosting; RF Random Forest; PRS-LR Polygenic Risk Scores Logistic Regression; AUC Area Under the Curve.

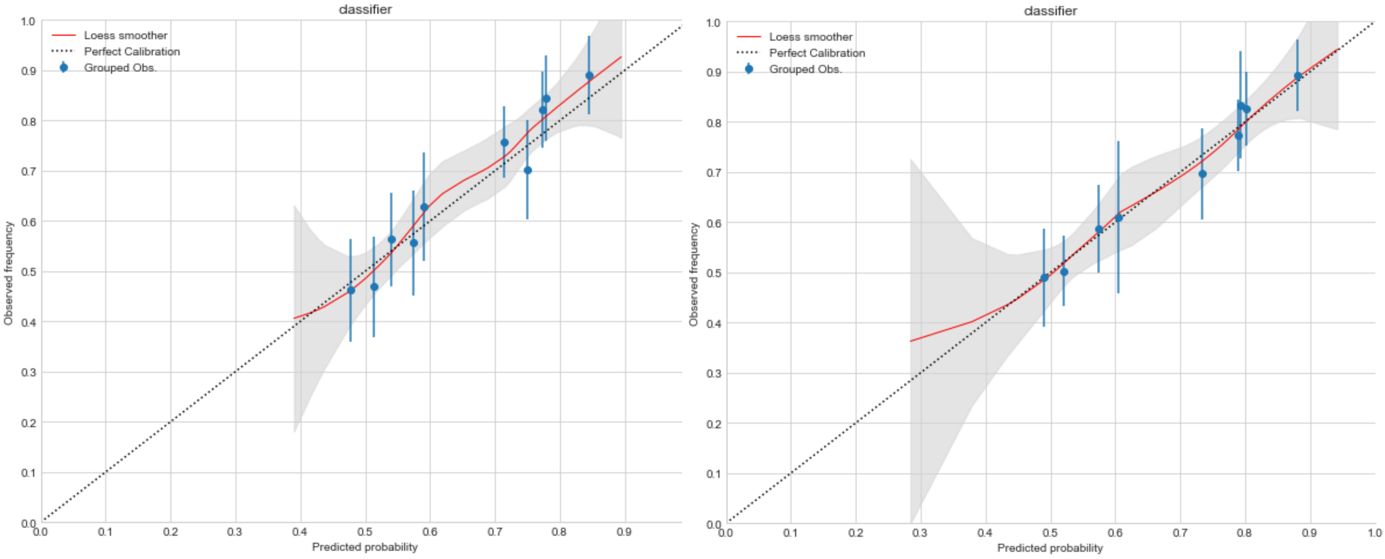
When comparing results when using either best guess genotypes (Figure 6.5) or dosages (Figure 6.7), classifier performance is approximately similar in in both analyses (GB 65.7-65.6, RFs 66.4-67.0, PRS-LR 68.2-68.9). Therefore, the use of RFs for feature selection resulted in no difference when using either genotypes or dosages. However, similarly to genotypes, the use of feature selection increased AUC by 5-6% for ML performance when compared to results shown in Figure 6.3. In line with all previous analyses, mean AUC for PRS-LR was higher as compared to GB or RFs. This superior performance of PRS-LR is supported by the significant results of t-tests (Supplementary Table 38), whereby t-statistics for paired t-tests are significant for PRS-LR versus GB at one threshold (0.1) (GB p-values = 1.4e-02) and RFs at two p-value thresholds (0.1,0.5) (RFs p-values = 5.9e-03, 5.17e-03).

As discussed previously, the difference in AUC between p-value thresholds when using feature selection is generally less than 1% AUC, therefore these alterations are most likely due to random variations across CV. Confidence intervals shown in Figure 6.7 are again larger than those observed when not using feature selection (Figures 6.1& 6.3).

Figure 6.8: The Comparison of non-Calibrated vs Calibrated Prediction Probabilities for GB

a) Pre-calibrated probabilities

b) Post calibration probabilities



These figures represent pre a) and post b) calibration plots for the related GB algorithm (Figure 6.7) clumped at a p-value of 0.1. The x-axis represents the prediction output of the classifier in terms of the probability of being a case. With the y-axis denoting observed class frequencies. Perfect calibration in which predicted probabilities match observed accuracies is denoted by the diagonal dotted line. The blue dots represent the mean probability/observed values within each quantile and are accompanied by a 95% confidence interval (blue bar). The overall relationship between predicted probabilities and observed frequencies (calibration curve) is given by the fitted loess smoother (red line), with a 95% confidence interval (grey shaded area) used.

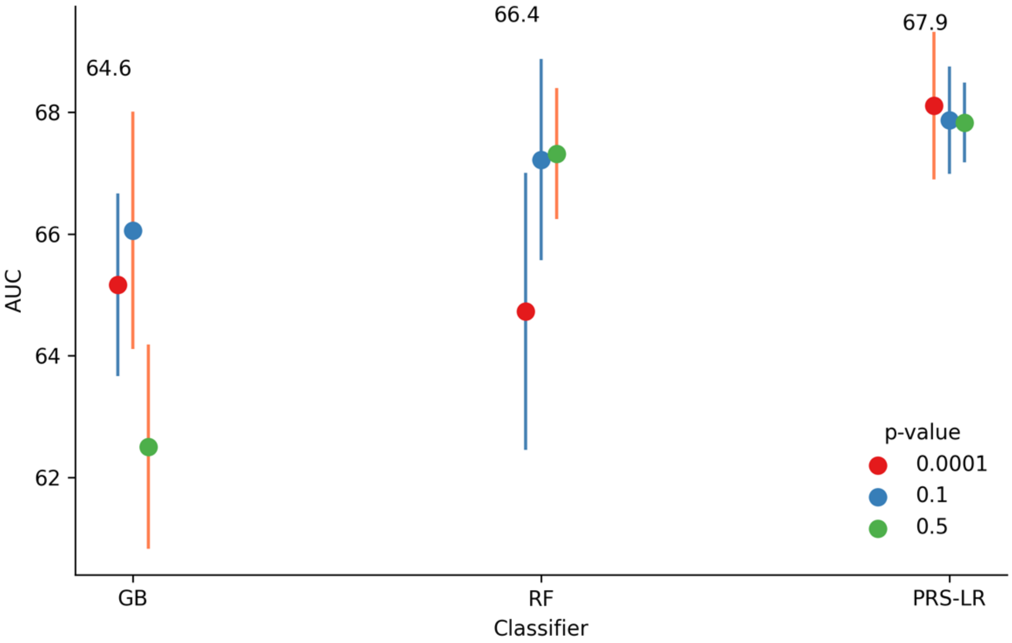
Results shown in Figure 6.8a demonstrate the predicted versus the observed probabilities of the GB tree from Figure 6.7. The plotted red line shown lies close to the diagonal, suggesting that the model was assessing risk well. Figure 6.8b shows the comparison of calibrated probabilities and observed probabilities. The loess smoother again followed the diagonal, suggesting that on this occasion calibration made negligible difference to risk prediction.

6.3.1.3 Prediction with ExtraTrees for feature selection

6.3.1.3.1 Best-guess genotypes

Analyses in this section involved the use of genotypes and Extra Trees for feature selection.

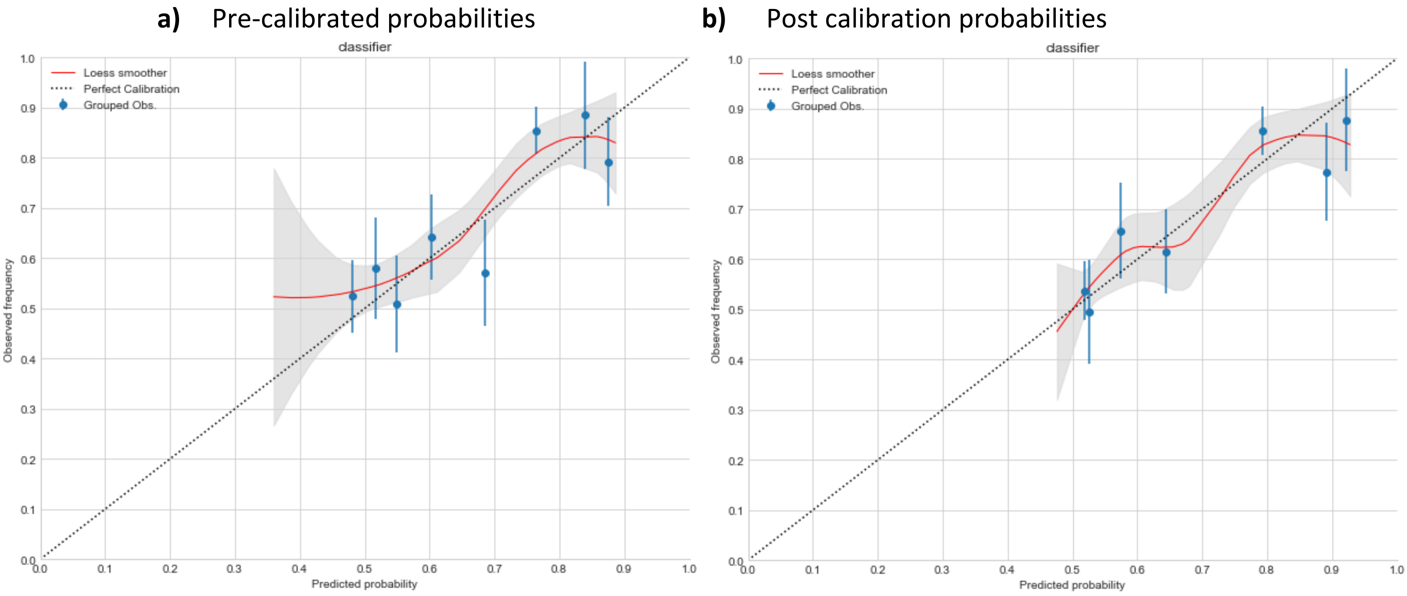
Figure 6.9: PRS-LR vs Selected Classifiers (RF, GB) for LD Pruned SNPs in Imputed Genotypes, with the Inclusion of *APOE* Alleles and an ExtraTrees algorithm used for Feature Selection.



Y-axis represents AUC in %; X-axis represents each classifier's results for a p-value threshold. Each dot represents the mean score for the prediction algorithm across 5 folds of CV, with an accompanying 95% CI bar. The numbers placed centrally are the mean of the three p-value threshold scores; GB Gradient Boosting; RF Random Forest; PRS-LR Polygenic Risk Scores Logistic Regression; AUC Area Under the Curve.

When comparing the performance of classifiers in Figure 6.9 with those in Figure 6.5, the use of the ExtraTrees algorithm for feature selection results in similar results to using RFs. When comparing both ML algorithms and PRS-LR, it can be seen that the use of feature selection does not improve classification above PRS-LR. Supplementary Table 39 displays the results of paired t-tests between classifiers. Results show that PRS-LR significantly outperformed GB for one p-value threshold (0.5) (p-value = 3.43e-02), whilst RFs also outperformed GB for the 0.5 threshold (p-value = 2.48e-02). Similarly, to those observed in Figures 6.5& 6.7, confidence intervals in Figure 6.9 were larger than when not using feature selection.

Figure 6.10: The Comparison of non-Calibrated vs Calibrated Prediction Probabilities for GB


a) Pre-calibrated probabilities

b) Post calibration probabilities

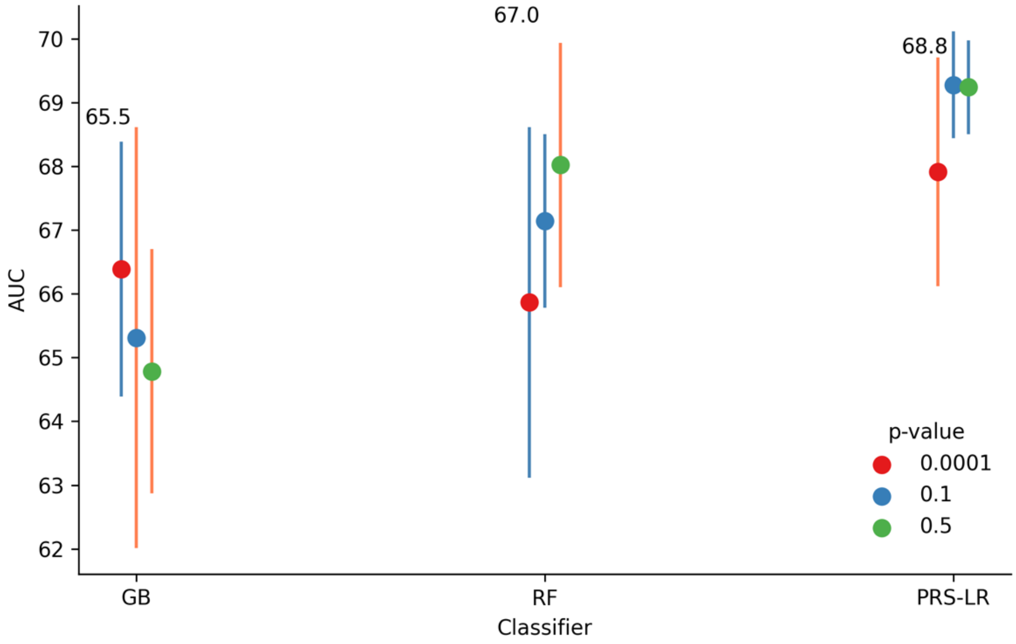
These figures represent pre a) and post b) calibration plots for the related GB algorithm (Figure 6.9) clumped at a p-value of 0.1. The x-axis represents the prediction output of the classifier in terms of the probability of being a case. With the y-axis denoting observed class frequencies. Perfect calibration in which predicted probabilities match observed accuracies is denoted by the diagonal dotted line. The blue dots represent the mean probability/observed values within each quantile and are accompanied by a 95% confidence interval (blue bar). The overall relationship between predicted probabilities and observed frequencies (calibration curve) is given by the fitted loess smoother (red line), with a 95% confidence interval (grey shaded area) used.

The comparison of model predictions versus observed frequency in Figure 6.10a demonstrated the model was underestimating risk in most cases. This is evidenced by the loess smoother lying above the diagonal in most instances. Following calibration in Figure 6.10b, the curve still did not follow the diagonal. This indicates that calibration did not correct the issue of poor risk assessment.

6.3.1.3.2 Dosages

Analyses in this section involved the coding of genotypes as dosages and Extra Trees for feature selection.

Figure 6.11: PRS-LR vs Selected Classifiers (RF, GB) for LD Pruned SNPs in Imputed Dosages, with the Inclusion of *APOE* Alleles and an ExtraTrees algorithm used for Feature Selection.

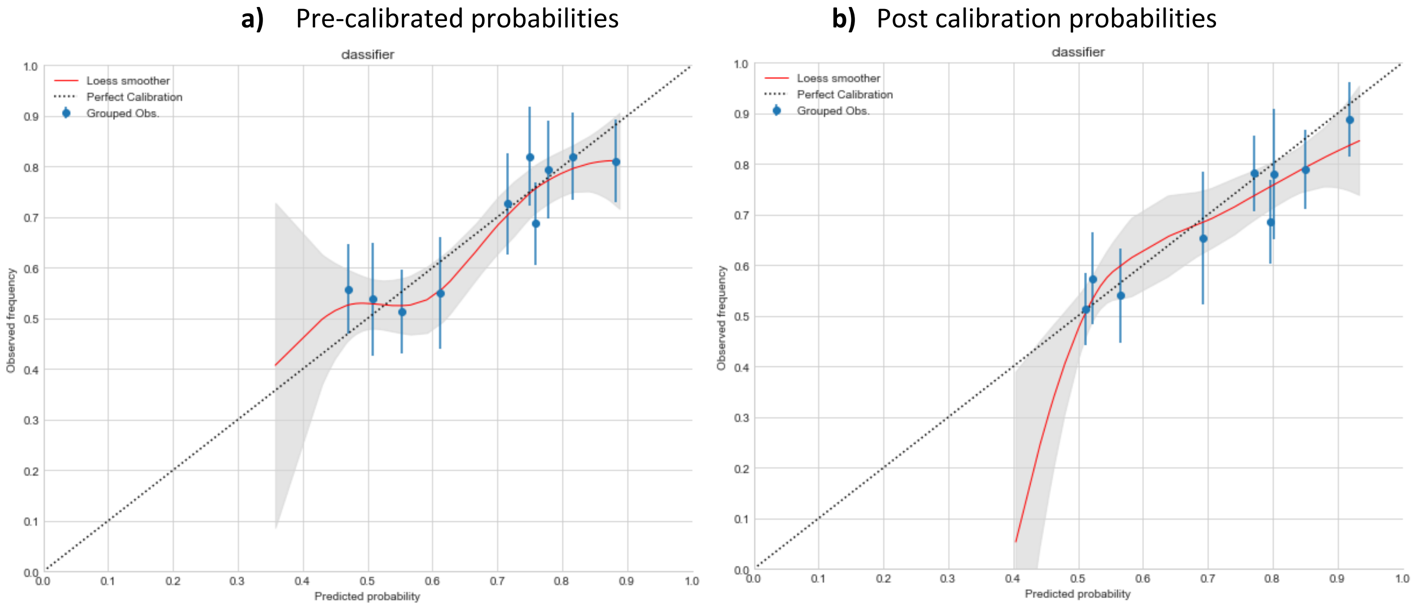


Y-axis represents AUC in %; X-axis represents each classifier's results for a p-value threshold. Each dot represents the mean score for the prediction algorithm across 5 folds of CV, with an accompanying 95% CI bar. The numbers placed centrally are the mean of the three p-value threshold scores; GB Gradient Boosting; RF Random Forest; PRS-LR Polygenic Risk Scores Logistic Regression; AUC Area Under the Curve.

Mean AUC across CV for all three classifiers in Figure 6.11 is similar to those shown in Figure 6.7. Therefore, using either RFs for feature selection or the ExtraTree algorithm resulted in similar levels of ML performance. Mean AUC for PRS-LR was again above that of both ML algorithms, however, results of paired t-tests detailed in Supplementary Table 40 were different to previous analyses. PRS-LR was only shown to significantly outperform GB for one p-value threshold (0.5) (p-value = 2.13e-02), with all other comparisons not returning significant results. In line with previous instances when using feature selection, confidence intervals suggest that scores for ML were more varied following the use of feature selection.

Another point for discussion is the contrast in AUC for both RFs and GB as the number of SNPs increase for prediction. AUC for GB decreases as the p-value threshold used for clumping becomes more lenient, whereas prediction performance for RFs increases. Explanations in previous figures suggested that these apparent trends might be the result of random variation across folds of CV. However, another possibility could be the success of hyperparameter tuning for dealing with the increasing amounts of features. The range of hyperparameters used and subsequent combination might have allowed RFs to overcome dimensionality issues to a better extent to those used for GB.

Figure 6.12: The Comparison of non-Calibrated vs Calibrated Prediction Probabilities for the GB


a) Pre-calibrated probabilities

b) Post calibration probabilities

These figures represent pre a) and post b) calibration plots for the related GB algorithm (Figure 6.11) clumped at a p-value of 0.1. The x-axis represents the prediction output of the classifier in terms of the probability of being a case. With the y-axis denoting observed class frequencies. Perfect calibration in which predicted probabilities match observed accuracies is denoted by the diagonal dotted line. The blue dots represent the mean probability/observed values within each quantile and are accompanied by a 95% confidence interval (blue bar). The overall relationship between predicted probabilities and observed frequencies (calibration curve) is given by the fitted loess smoother (red line), with a 95% confidence interval (grey shaded area) used.

The pre-calibration plot in Figure 6.12a showed the loess smoother followed the diagonal line of perfect calibration reasonably well. There were some areas such as 0.4–0.5 predicted probabilities in which risk was underestimated, and 0.5 -0.65 where risk was overestimated. However, in general, risk assessment was balanced. Figure 6.12b demonstrated the comparison between calibrated probabilities and observed frequencies. An overestimation of risk between 0.4 and 0.5 predicted probabilities occurred post-calibration, suggesting that the algorithm was inaccurately predicting that samples had AD. For other predictions, the loess smoother followed the diagonal, therefore the algorithm was more accurately predicting the probability of AD occurring.

6.4 Discussion

Analyses in this chapter involved the introduction of imputed variants for AD prediction. Two different types of inputs were used in the form of allelic dosages and best guess genotypes. Emphasis was placed on whether the use of imputed variants would result in different outcomes to analyses in Chapter 5, as well as comparing the performance of ML algorithms versus PRS-LR.

6.4.1 Comparison of ML and PRS on imputed variants

One of the central aims of this chapter was to assess the performance of ML vs PRS-LR when using imputed SNPs. The overriding conclusion drawn from the performed analyses was that PRS-LR outperformed ML classifiers in all scenarios. Prior to the use of feature selection, AUC for ML declined as the number of SNPs used increased. This could be further evidence for the existence of the curse of dimensionality, where an increase in features alongside a fixed number of samples will most likely result in a reduction of prediction performance (Verleysen and François, 2005b). The addition of extra SNPs may contribute random noise to the feature set, reducing the ML algorithm's ability to learn the true underlying pattern within the dataset. In a formal sense this is known as 'overfitting', where a classifier fails to generalise to unseen data (Verleysen and François, 2005b). Following the use of feature selection, mean AUC across the three p-value thresholds for both ML algorithms improved. This was similar to results demonstrated in Chapter 5, in which non-imputed genotypes were used. Therefore, using feature selection might assist in reducing the increase in random noise when increasing SNPs through more lenient p-values at the clumping stage.

However, whilst performance was retained when increasing the number of SNPs, AUC did not improve beyond prediction performance when using ML with no feature selection at a p-value of 0.0001 (Figure 6.1). Also, despite the use of feature selection, PRS-LR still achieved higher levels of AUC than ML. This superior performance was evidenced by pairwise t-tests, in which comparisons of AUC between PRS-LR and GB/RFs were significant (Supplementary Table 37). Another observation made when introducing feature selection was the increase in variance of ML scores across CV folds. This was true for all uses of feature selection when compared to confidence intervals in Figures 6.1& 6.3 (no feature selection used). A possible reason for this could be the consistency of features selected between CV folds. Assessments to analyse the stability of feature selection algorithms (discussed at further length in Section 6.4.2) revealed SNPs were inconsistently chosen between folds. This inconsistency might result in differences of prediction performance (AUC) between folds.

6.4.2 The comparison between allelic dosages and best guess genotypes

When comparing classification performance between best-guess imputed genotypes and dosages, PRS-LR and ML algorithms showed little differences. This suggests that using either genotype dosages or best-guess genotypes made no difference in prediction outcome. The only difference between the two data types was the number of SNPs used for prediction, with the dosage format providing over 200,000 SNPs per round of CV for the 0.5 p-value threshold, in comparison to ~128,000 variants for best-guess genotypes. This was the first occasion in this thesis thus far that such a large number of SNPs have been assessed.

6.4.3 Comparison performance between imputed and non-imputed SNPs

Analyses comparing the prediction performance of ML and PRS-LR in Chapter 5 used the non-imputed GERAD dataset. A core aim of this chapter was to use imputed SNPs, with the intention of comparing results to the use of non-imputed genotypes. When comparing ML performance using dosages and best-guess genotypes, AUC was similar across both data types. This is not surprising as despite a loss of SNPs due to the use of a threshold converting dosages to best-guess genotypes the overlap of SNPs between the two sets remained high (90.1%).

When comparing the performance of prediction algorithms in this chapter with results in Chapter 5, the use of imputation did not result in noticeable alteration in AUC. However, AUC for PRS-LR reached 69% only, a reduction of 3-4% from results from when non-imputed genotypes were used (Chapter 5). Logic would suggest that increasing the number of SNPs available for PRS-LR would boost performance, however this is not always the case. Chen et al., 2020, compared the calculation of PRS for several diseases from directly genotyped SNPs and three forms of imputation. Results demonstrated that imputation methods can introduce variations of PRS values at an individual level. The magnitude of variability differed depending on the type of imputation software used, with the method using scholastic elements (error terms) causing the most variation. These variations tend to be small, however in some rare circumstances PRS are substantially different from those computed from non-imputed genotypes. Therefore, it is possible that the imputation of GERAD might have resulted in differences in PRS for some individuals, with the result of a reduction in AUC for PRS-LR.

6.4.4 Feature selection Performance

The stability of a feature selection algorithm relates to the impact of changes in composition of training data through splitting procedures such as CV can have on its performance. A feature selection method can be defined as unstable if a small alteration in training data causes a large change in the features selected. Stability is important as it increases confidence in the viability of the feature selection method and classification performance (Balakrishnan, Dhanalakshmi and Khaire, 2022). To assess the stability of feature selection algorithms used in this chapter, features selected between p-value thresholds were compared. The results of this were recorded in Supplementary Tables 31–34.

Results demonstrate that both RF and ExtraTrees performed poorly in terms of stability. The overlap in SNPs between p-value thresholds was low, with usually only one feature shared between SNP sets. This was most often the ε4 allele, which is unsurprising given its association to AD risk. The small amount of overlap of SNPs between p-value thresholds could be due to in-built GWAS per CV round. The sample used to generate summary statistics changes on each occasion, therefore effect sizes and p-values may also alter. These variations across CV might result in different SNPs being chosen through clumping. However, despite this variation of features, classifier performance remained similar. Therefore, this could suggest that selected SNPs had low effect on AD prediction, with the majority of predictive information coming from the ε4 allele which is consistently selected.

When assessing the overall performance of feature selection for analyses in this chapter, it can be deduced that its use can improve ML performance at less stringent p-values. However, this improvement does not increase AUC above levels achieved for PRS-LR. These patterns were also observed for analyses in Chapter 5, suggesting that the use of non-imputed or imputed variants when using feature selection methods.

6.4.5 Model calibration

Results of calibration varied across all analyses run in the chapter. Figure 6.2 shows the calibration plots for the GB tree analysis in Supplementary Table 29, in which no feature selection was used for genotyped data. The plots demonstrated that GB was under-forecasting risk probabilities between 0.25-0.5, with true cases predicted as controls by the model. Overestimation of risk was observed in only a few instances, with predicted probabilities lying below the diagonal in Figure 6.12 for example. It is worth stating that overestimation mostly occurred following the use of feature selection prior to classification. Whilst the use of feature selection has advantages in terms of reducing possible overfitting and resources required for computation, issues can also arise following its use.

6.4.6 Processing Large Numbers of Variants

Initial analysis in Chapter 4 focused on using a small number of GWAS significant SNPs (N=23) as predictors. In subsequent chapters larger number of SNPs have been used for prediction. ML allows for large number of predictors to be analysed within short run times (Attaran and Deb, 2018b). However, despite advances in modern computing and memory storage, limits still apply on the number of variables which can be analysed (Qiu *et al.*, 2016). Following the increase in large datasets being compiled, these restrictions are becoming more prominent.

Analysis in this chapter used large numbers of SNPs for prediction. This was especially true in the case of dosages, in which the least stringent p-value threshold of 0.5 resulted in 200,000 SNPs being processed per round of CV. Analysing this number of SNPs is not common for ML approaches in AD prediction, as evidenced when reviewing the studies included in our systematic review (Rowe et al., 2021a). Here it was achieved due to a combination of techniques used during the methodology phase. The first of these being the ability to call *PLINK* from within *Python*. The *subprocess* function allowed data intensive processes such as GWAS and clumping to be run by *PLINK* separately to *Python*. It has also meant that large files such as bed, bim and fam files do not have to be loaded into memory for each set of analysis. The *deconfounding* method used also led to reduced run times. This was due to running large numbers of regressions using the *statsmodels* package, whilst also using *NumPy* arrays. These have been shown to be computationally efficient and quicker than other pythonic methods (van der Walt, Colbert and Varoquaux, 2011).

6.4.7 Conclusions

Results in this chapter closely aligned to those in Chapter 5. AUC for both RFs and GB were similar, with PRS-LR also consistently outperforming both algorithms. The only difference between using non-imputed and imputed genotypes was the reduction in PRS performance, with AUC 2-3% less when using imputed SNPs. The analysis of >200,000 SNPs in this chapter has not occurred often for ML prediction in AD. When assessing studies for the systematic review in Chapter 3, only one of the articles analysed a similar number of variants (Wei, Visweswaran and Cooper, 2011b). On this occasion, variations of the ML technique naïve bayes (NB) were used on 312, 318 SNPs. Despite using a larger number of features than analyses in this chapter, the underlying mathematical calculations of the NB algorithm are not as complex as other methods such as decision tree-based classifiers. Therefore, greater computational resources are required to analyse datasets of the same dimension. The ability to process datasets comprising 200,000 SNPs therefore appears to be novel in AD prediction when using decision tree-based approaches. This however is not the case for PRS as other studies have used such numbers of variants.

The use of feature selection methods reduces the random noise but does not improve the predictive performance of ML over when a p-value threshold of 0.0001 is used for clumping. Therefore, using smaller feature sets with higher association with AD appears to result in better disease prediction. This suggests that a different approach may be required to explore whether ML can outperform PRS. One possible option is to use variants linked with biological pathways which have been associated with AD. This will reduce the number of SNPs used but retain variants related to AD development.

7 Assessment of the predictive capability of genetic pathways in Alzheimer's disease

7.1 Introduction

The aim of this chapter is to assess the capability of SNPs to predict AD by Machine Learning (ML) within the 9 pathways detailed in Kunkle et al., 2019. Motivation for the use of pathways was to assess whether using prior biological information could improve predictive ability above that of genome wide variants. Analysis in previous chapters used large sets of SNPs derived on a genomic scale. Analysing large sets of variants may impair the performance of ML algorithms due to excess dimensionality, with use of statistically based feature selection methods required to limit the impact. Therefore, analyses in this chapter will aim to reduce this issue through the use biological information to select relevant variants to AD. This will assess whether prediction performance can be improved by selecting SNPs likely to be associated with disease risk.

The GERAD dataset will be filtered to contain only those SNPs within each pathway. ML will then be used to assess the prediction of disease status obtained from each pathway. Prediction accuracy will be compared to that obtained from a logistic regression (LR) based polygenic risk score (PRS-LR), as well as the use of PRS-CS-LR. The latter is explained further in Section 7.2.2.2.

7.2 Methods

7.2.1 Outline of analysis

Analysis in this chapter was split into three sections: 1) The assessment of predictive accuracy for individual pathways, 2) assessment of the accuracy of risk prediction obtained from analysing the nine pathways simultaneously in a multivariable model, using both genotypes and PRS for inputs to LR and ML, and 3) assessment of risk prediction accuracy given by the amalgamation of all SNPs across pathways into a unified set. Analyses for individual pathways were firstly carried out in non-imputed (genotyped) SNPs, then repeated in imputed variants (as described in Chapter 6). The subsequent multivariable and unified set analyses used imputed variants only. The methodology used to perform the analyses varied by analysis type; an overview of these is given below:

- Fivefold stratified nested cross-validation was used for algorithm development.
- Prediction algorithms used were Gradient Boosting (GB), Random Forests (RFs), PRS-LR (p-value =0.1) and PRS-CS-LR (p-value = 0.1).
- Prediction performance for ML, PRS-LR and PRS-CS-LR was assessed by averaging area under the curve (AUC) across the five test sets from CV.
- Both non-imputed and imputed genotyped SNPs were used individual pathway analysis.
- A GWAS was conducted within each round of cross-validation for samples in the training sets to ensure PRS did not gain an unfair advantage over ML due to the use of external information. Summary statistics generated in training samples were then used to generate PRS in test samples.
- Discrimination between cases and controls was then assessed, with comparisons between LR performed on PRS values and ML.
- ML and LR were also trained and tested using pathways-specific PRS generated with the Kunkle-noGERAD summary statistics in both training and test samples.

7.2.2 Individual pathway analysis

For analysis in this section, pathways were treated as separate entities. RFs, GB, PRS-LR and PRS-CS-LR were trained and validated using SNPs (both non-imputed and imputed) from each pathway. Eighteen sets of SNPs were created (nine non-imputed, nine imputed) using the genomic positions of genes within each pathway including 35kb upstream and 10kb downstream regions around the genes as suggested by (Network and Pathway Analysis Subgroup of Psychiatric Genomics Consortium, 2015). The effect of the *APOE* gene was modelled in three ways: 1) variants within *APOE* were removed from each pathway, 2) variants within *APOE* were included, 3) variants within *APOE* were removed and replaced by counts of the *APOE* alleles ε2 and ε4.

7.2.2.1 Data analysis

We used the same software in this chapter as in previous chapters. The *Python* function *StratifiedKFold* was used to derive five rounds of Cross-Validation (CV). Principal Components (PCs) were produced using the package *PCA* from the *sklearn* package within *Python*. The GWAS was performed by calling *PLINK* within *Python* using the function *subprocess.* Genotypes of chosen SNPs present in training and test sets were then adjusted for population stratification using the *Deconfounding* method observed in Chapters 5 and 6, with subsequent residuals for both the training and test sets were then scaled using the *Python* function *StandardScaler*. The *PLINK* function *--score* was used to generate PRS and then adjusted by PCs using the *Python* package *statsmodels.OLS*, and standardised.

7.2.2.2 PRS-CS

PRS-CS uses a high-dimensional Bayesian framework to derive SNP effect sizes. The novelty of PRS-CS lies with its use of a continuous shrinkage method to adjust SNP effect sizes, thus removing the need for both threshold and LD pruning. The amount of shrinkage applied to a variant's effect size is related to its strength of association within a GWAS. Multiple effect sizes are updated at once, reducing computation time when compared to updating weights on a singular basis (Choi, Mak and Paul F. O'Reilly, 2020). Another advantage of PRS-CS is that it imposes heavy shrinkage on variants with small effect sizes, thereby reducing random noise in the SNP set. Also, SNPs with high effect sizes are lightly pruned, preserving the signal between variants and phenotype (Choi, Mak and Paul F. O'Reilly, 2020). A difference between the clumping and thresholding method and PRS-CS is the absence of clumping which is instead achieved by the shrinkage process in PRS-CS (Choi, Mak and Paul F. O'Reilly, 2020).

To implement PRS-CS, the tool from https://github.com/getian107/PRScs is required. The *Python* function *subprocess* was then used call the tool within each CV fold. Linkage disequilibrium statistics were provided by the file 'ldblk_1kg_eur'. PRS values were then derived in test samples using the effect sizes provided by the online tool with *PLINK* function *–score* was run using *--subprocess.* Similarly, to ordinary PRS, scores were then adjusted for population and the residuals were then standardised.

7.2.2.3 Calibration statistics

The output of all ML algorithms used in this chapter was calibrated, using the same techniques described in Chapters 4,5 and 6. Calibration statistics are then plotted to assess how calibration altered probability distributions.

7.2.3 Multivariable analysis of pathways

The next set of analyses focused on using a multivariable method, in which the PRS for the 9 pathways were used as separate predictors for modelling. Subsequent analyses assessed the predictive capability of using both genotypes and PRS as inputs to ML. Analyses were conducted for imputed SNPs only, with the intent of further establishing whether using imputed variants would result in better prediction performance than non-imputed SNPs.

7.2.4 Multivariable Analysis of pathways using genotypes as inputs

When selecting pathway-specific SNPs, GWAS statistics were generated (training samples only) within each round of CV and then used to select SNPs through clumping. For each CV fold, the nine pathways were clumped separately with two p-value thresholds of 0.1 and 1, a clumping distance of 500 kb and an r^2 of 0.1. These values were used to ensure a reasonable number of variants were present within each pathway. Some genes are present in more than one pathway, suggesting that certain SNPs might occur more than once following the clumping phase. All duplicate copies of SNPs were removed, leaving a set of unique variants. PRS were generated for each of the nine separate sets of SNPs. For the PRS-CS method, PRS values were also generated individually for each pathway. Following the derivation of scores for each pathway, the nine sets of values were adjusted for population stratification using the same method for PRS-LR.

7.2.5 Multivariable analysis of pathways using PRS as inputs

The second form of analysis in this section used PRS values where SNP effect sizes and p-values were taken from an external source (Kunkle-noGERAD summary statistics). As before, both sets of PRS (clumping and thresholding, PRS-CS) were adjusted for population stratification and standardised.

7.2.6 All pathways combined into a single set of SNPs.

The last section of analyses in this chapter focused on using a unified set of SNPs, in which SNPs from all 9 pathways were combined. Duplicate SNPs were removed, producing a set of unique variants. If the multivariable method used nine different datasets, each clumped separately, the unified method used in this section clumped the combined pathway set once, therefore the SNPs used in the for both analyses might differ. The combination of all pathways into one set altered methodology for PRS-LR and PRS-CS more than ML. This is due to the LR used for disease prediction being reduced from a multivariable method to a singular predictor (in the instance when *APOE* allele*s* were not included).

To calculate PRS for PRS-LR and PRS-CS-LR, imputed genotypes from the unified pathway dataset were used. Unlike previous analyses in which ML was provided with multiple features, the use of PRS for training resulted in one variable only. Although seemingly counterintuitive, decision trees can be trained using only one variable. In the case of the RF, each decision tree will take a selection of samples with replacement. As only one feature is present, the decision tree will continue to split on this variable until a decision is reached. PRS values were also generated using the PRS-CS method and were only used to test disease prediction in the test set.

7.2.7 Machine learning methodology

Two ML algorithms were used for classification in this chapter, these were random forests (RFs) and gradient boosted trees (GB). SNPs and PRSs were used as features for ML. RFs were implemented using the *Python* package *RandomForestClassifier.* A further hyperparameter specified was *class_weight,* required due to the class imbalance within the dataset. This is achieved by specifying the *balanced* option. Classification performance was assessed using area under the curve (AUC). Overall performance was calculated by taking the mean AUC across all five folds of CV.

For GB, the *Python* package *XGBClassifier* was used to develop models. A hyperparameter to correct for class imbalance was also passed to XGBClassifier. *Scale_pos_weight* calculates the ratio of the minority to majority class. The minority class is then upweighted during classification. AUC was calculated using the same method as RFs.

7.2.8 Comparing predictive performance of classifiers

As in previous chapters, AUC values for classifiers were compared within each type of analysis using a paired t-test. Values of test set AUCs from each round of CV were compared for each pair of algorithms. The t-test was calculated using the *Python* function *ttest_rel* from the package *stats*. The false discovery rate (FDR) controlling method Benjamini-Hochberg was used to adjust for possible false positives with the function *p.adjust* in R. Corrections were made on an analysis wide basis (for each supplementary table at one time).

7.3 Results

All results for analyses are detailed in Supplementary Tables 41-44, with mean AUC across either 9 pathways or five folds of CV. Classifier comparison statistics from t-tests are also reported only if they were significant (p-value < 0.05) (Supplementary Tables 45-50).

7.3.1 SNPs collected per pathway

7.3.1.1 Non-imputed genotypes

Analysis in this chapter used SNPs from 9 AD associated pathways. Variants within these pathways were extracted from the non-imputed GERAD dataset. The resulting 9 sets of SNPs are detailed in Table 7.1.

Table 7.1: A Breakdown of the Number of non-imputed SNPs within each Pathway

Pathway	Description	Number of SNPs	Number of Genes
1	Protein-lipid complex assembly	277	20 (Including APOE)
2	Regulation of beta-amyloid formation	206	10 (Including APOE)
3	Protein-lipid complex	611	40 (Including APOE)

4	Regulation of amyloid precursor protein catabolic process	232	12 (Including APOE)
5	Tau protein building	145	11 (Including APOE)
6	Reverse cholesterol transport	321	17 (Including APOE)
7	Protein-lipid complex subunit organisation	560	35 (Including APOE)
8	Plasma lipoprotein particle assembly	201	18 (Including APOE)
9	Activation of immune response	6603	432 (No APOE)

The number of SNPs in pathways 1-8 is relatively similar. However, the number of SNPs in pathway 9 is far greater.

Table 7.2: The Overlap of non-imputed SNPs between Pathways given by the Jaccard Index

Pathway	1	2	3	4	5	6	7	8	9
1 (277 SNPs)	X	15	143	15	26	110	271	195	14
2 (206 SNPs)	0.032	X	58	206	5	21	15	15	16
3 (611 SNPs)	0.200	0.079	X	58	26	217	261	122	31
4 (232 SNPs)	0.031	0.888	0.076	X	5	21	15	15	16
5 (145 SNPs)	0.067	0.014	0.037	0.013	X	5	26	5	10
6 (321 SNPs)	0.230	0.042	0.316	0.040	0.011	X	301	110	56
7 (560 SNPs)	0.489	0.020	0.297	0.019	0.039	0.528	X	195	40
8 (201 SNPs)	0.720	0.039	0.185	0.036	0.015	0.274	0.352	X	14
9 (6603 SNPs)	0.002	0.002	0.005	0.002	0.002	0.008	0.006	0.002	X

The overlap of SNPs between pathways is detailed in two ways. Below the diagonal represents the Jaccard index, this is calculated by dividing the intersection (number of SNPs) of two sets by the union of two sets. Values above the diagonal represent the number of SNPs in common between two pathways.

Table 7.2 demonstrates varying degrees of overlap in SNPs between pathways. Some examples of high overlap include pathways 1&7, 6&7, 2&4, and 1&8. The degree of overlap (measured by the Jaccard coefficient) is lowest between pathway 9 and the others, suggesting a greater degree of independence.

Table 7.3: The Number of overlapping Genes between Pathways.

Pathway	1	2	3	4	5	6	7	8	9
1 (277 SNPs)	X	2	11	2	2	8	20	18	0
2 (206 SNPs)	0.071	X	3	10	1	2	2	2	1
3 (611 SNPs)	0.224	0.064	X	3	2	13	17	10	1
4 (232 SNPs)	0.067	0.833	0.061	X	1	2	2	2	1
5 (145 SNPs)	0.069	0.050	0.041	0.045	X	1	16	8	2
6 (321 SNPs)	0.276	0.080	0.295	0.074	0.037	X	16	8	2
7 (560 SNPs)	0.571	0.047	0.293	0.044	0.045	0.444	X	18	1
8 (201 SNPs)	0.9	0.077	0.208	0.071	0.036	0.296	0.514	X	0
9 (6603 SNPs)	0.000	0.002	0.002	0.002	0.002	0.004	0.002	0.000	X

The overlap of genes between pathways is detailed in two ways. Below the diagonal represents the Jaccard index, this is calculated by dividing the intersection (number of genes) of two sets by the union of two sets. Values above the diagonal represent the number of genes in common between two pathways.

The overlap in genes between pathways shown in Table 7.3 follows a similar pattern to the overlap in SNPs demonstrated in Table 7.2.

7.3.1.2 Imputed genotypes

For imputed variants, the imputed version of GERAD was filtered using the same genomic locations as non-imputed variants (Table 7.1).

Table 7.4: A Breakdown of the number of imputed SNPs within each pathway

Pathway	Description	Number of SNPs	Number of Genes
1	Protein-lipid complex assembly	3770	20 (Including APOE)
2	Regulation of beta-amyloid formation	3396	10 (Including APOE)
3	Protein-lipid complex	7871	40 (Including APOE)
4	Regulation of amyloid precursor protein catabolic process	2669	12 (Including APOE)
5	Tau protein building	2318	11 (Including APOE)
6	Reverse cholesterol transport	3413	17 (Including APOE)
7	Protein-lipid complex subunit organisation	6981	35 (Including APOE)
8	Plasma lipoprotein particle assembly	2958	18 (Including APOE)
9	Activation of immune response	87710	432 (No APOE)

The number of variants in each pathway is greater when comparing with non-imputed SNPs. The imputed set of variants contains fifteen times the number of SNPs, therefore greater numbers of SNPs would be expected per pathway. Similarly, to the non-imputed pathways, the final pathway has significantly more SNPs than others.

Table 7.5: The overlap of imputed SNPs between Pathways given by the Jaccard Index

Pathway	1	2	3	4	5	6	7	8	9
1 (3770 SNPs)	X	129	1475	130	394	1160	3770	2743	83
2 (3396 SNPs)	0.018	X	555	2199	52	210	129	129	158
3 (7871 SNPs)	0.152	0.054	X	556	394	1895	2646	1134	184
4 (2669 SNPs)	0.021	0.569	0.058	X	184	211	130	130	158
5 (2318 SNPs)	0.069	0.009	0.042	0.011	X	53	394	53	210
6 (3413 SNPs)	0.202	0.033	0.218	0.038	0.010	X	2984	1160	523
7 (6981 SNPs)	0.557	0.013	0.229	0.014	0.045	0.431	X	2743	365
8 (2958 SNPs)	0.728	0.021	0.125	0.025	0.011	0.246	0.405	X	83
9 (87710 SNPs)	0.001	0.002	0.002	0.002	0.002	0.006	0.004	0.001	X

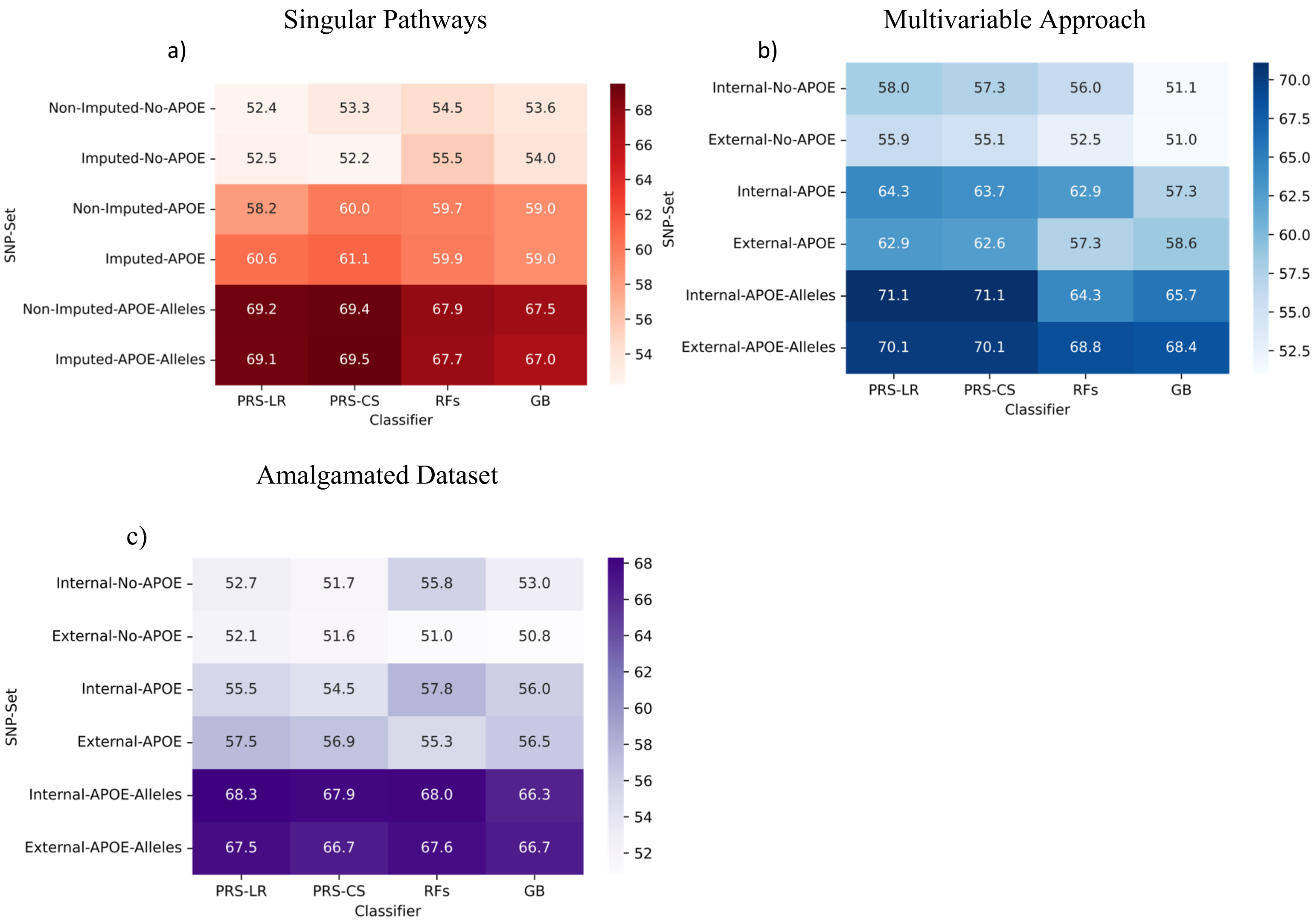
The overlap of SNPs between pathways is detailed in two ways. Below the diagonal represents the Jaccard index, this is calculated by dividing the intersection (number of SNPs) of two sets by the union of two sets. Values above the diagonal represent the number of SNPs in common between two pathways.

Table 7.5 demonstrates overlaps in SNPs between pathways for imputed SNPs were similar to those for non-imputed SNPs (Table 7.2).

7.3.2 General overview of results

An overview of all analyses produced in this chapter is given in Figure 7.1. Results across the three different forms of analyses are shown, these are individual pathways, the multivariable approach and amalgamated dataset. Heatmaps show a consistent increase in classifier performance from the exclusion of *APOE* related SNPs, to the inclusion of *APOE* SNPs and subsequent *APOE* alleles. Results are then investigated further in subsequent sections.

Figure 7.1: Heatmaps to display results for individual pathways analysis, the investigation of a multivariable approach and the combined dataset



Classifier performance is recorded in AUC. Values for singular pathways are mean performance across 9 pathways, whilst multivariable and combined values are computed across five folds of CV. For singular pathways, datasets are separated by whether SNPs were imputed or not. The method of analysing the *APOE* region is also detailed, with no *APOE* related SNPs, *APOE* SNPs included and *APOE* alleles plus other SNPs. Multivariable and combined analyses are separated by types of input to ML methods, genotypes (where summary statistics were generated using an in-built GWAS) (internal information) and PRS (in which the Kunkle-nogerad summary statistics were used to generate PRS) (external information).

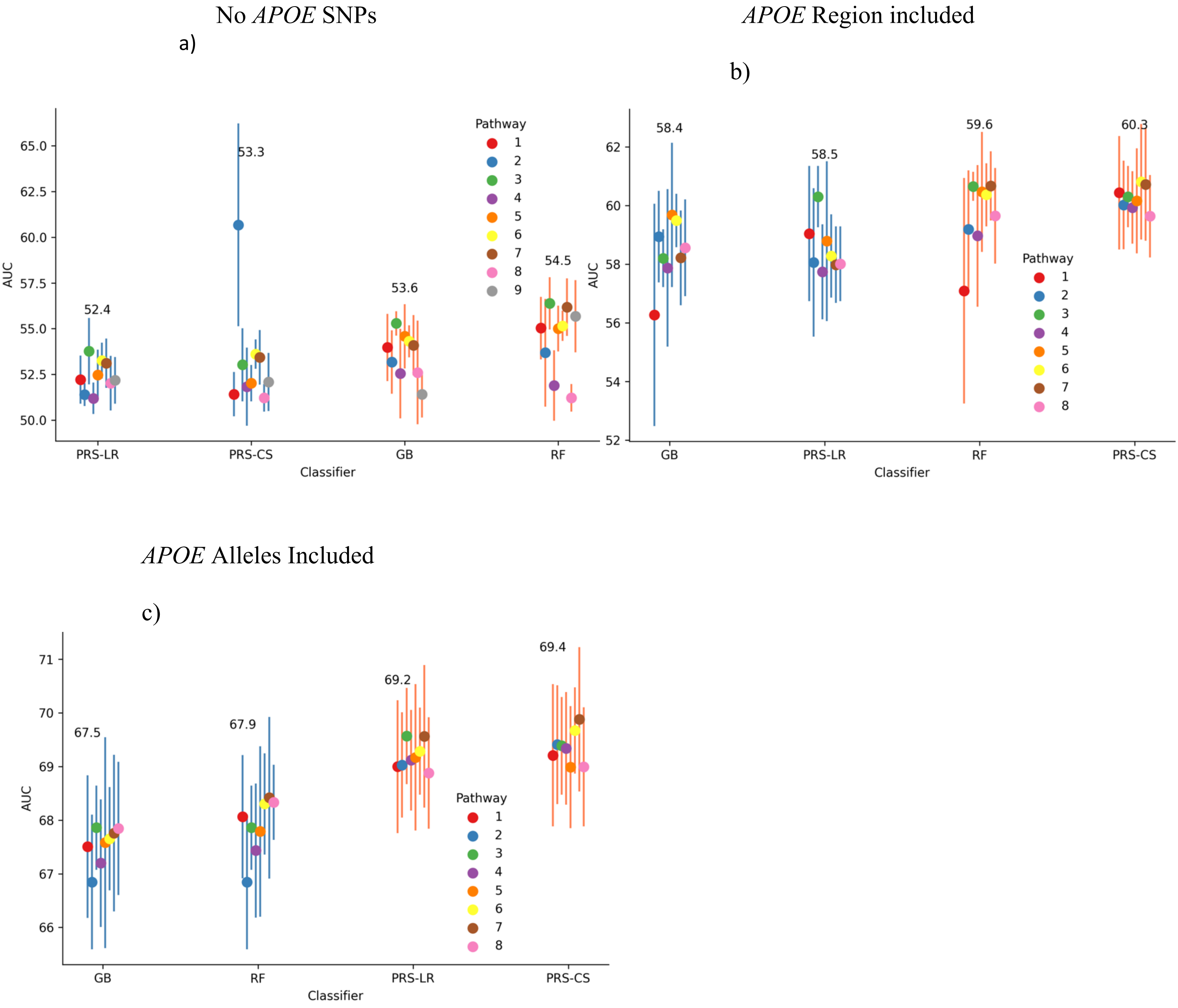
7.3.3 Individual pathway analysis

The predictive performance of both ML and PRS-LR in each pathway was assessed in this section. For all figures, prediction performance is grouped by classifier and is reported as AUC. Each coloured dot represents prediction performance for a particular pathway, with the mean value given across all 9 pathways. Calibration plots are also presented for analyses. Figures show a model's predicted output versus observed class memberships in the dataset. A loess smoother is used to outline this relationship. The left-hand plot of each figure demonstrates probabilities prior to calibration, with calibrated probabilities shown in the right-hand plot. The classifier chosen for each example is the RF, with data provided from pathway 1 for individual pathway analyses, with *APOE* SNPs included. It was decided not to plot calibration statistics for every analysis as this would result in too many figures, with similarities observed across many analyses also.

7.3.3.1 Non-imputed data

The first set of analyses represent the comparison of PRS-LR and ML techniques using non-imputed genotypes for all 9 pathways.

Figure 7.2: The Comparison of PRS-LR (P-value threshold 0.1), PRS-CS vs Selected Classifiers (RF, GB) for LD Pruned SNPs (non-Imputed Genotypes)



No *APOE* SNPs

a)

APOE Region included

b)

APOE Alleles Included

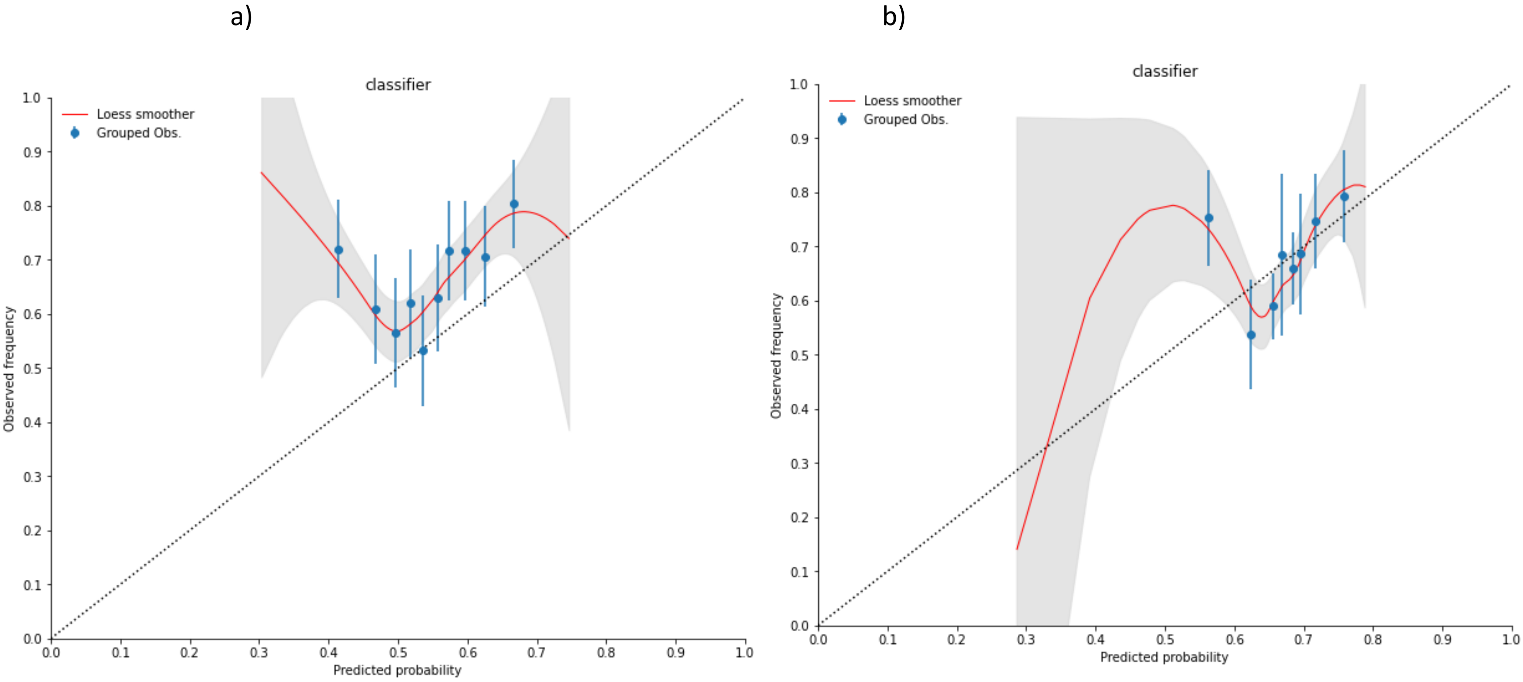
c)

Y-axis represents AUC in %; with classifiers placed on the X axis. Each dot represents the score for the prediction algorithm for all p-value thresholds, with accompanying 95% CI. The numbers placed centrally are the mean score across pathway SNP sets; GB Gradient Boosting; RF Random Forest; PRS-LR Polygenic Risk Scores Logistic Regression; AUC Area Under the Curve. Plots b and c contain 8 pathways only as pathway 9 does not include SNPs within the *APOE* region.

Figure 7.2 displays the three different sets of analyses conducted for non-imputed singular pathways. When considering analyses without *APOE* SNPs and their subsequent re-inclusion, results demonstrate the removal of SNPs led to a 5-6% reduction in AUC for all classifiers. When excluding SNPs within the *APOE* region, the best performing ML algorithm was RFs. This was evidenced by results of pairwise t-tests shown in Supplementary Table 45, in which RFs achieved better prediction than other algorithms in several pathways (1, 2, 3, 5 and 7). When including the *APOE* region, both RFs and PRS-CS achieved the highest mean AUC across all pathways. This was corroborated by the results of pairwise t-tests, as both RFs and PRS-CS performed significantly better than GB and PRS-LR across several pathways (3, 6, 7) (Supplementary Table 45). After inclusion of SNPs in the *APOE* region (Figure 7.2b), AUC for both GB and RFs was superior (1-3%) to comparative analyses in Chapters 5&6, in which SNPs were selected on a genome-wide basis.

When including *APOE* alleles for ML, PRS-CS and PRS-LR, AUC increased when compared to the two previous analyses, with an increase of 7-15% for all classifiers. This level of increase in prediction was also observed when using genome-wide variants in Chapters 5&6. Mean AUC for both PRS methods was 2-3% greater than both GB and RFs. This increase in performance was statistically significant in six pathways (1, 2, 3, 4, 6, 7) as shown in Supplementary Table 45. When comparing the performances of PRS-LR and PRS-CS-LR, PRS-CS achieved higher mean AUC across all three analyses. However, this increase in prediction performance was only significant in one pathway (6) as shown by statistics in Supplementary Table 45.

Figure 7.3: The Comparison of non-Calibrated vs Calibrated Prediction Probabilities.



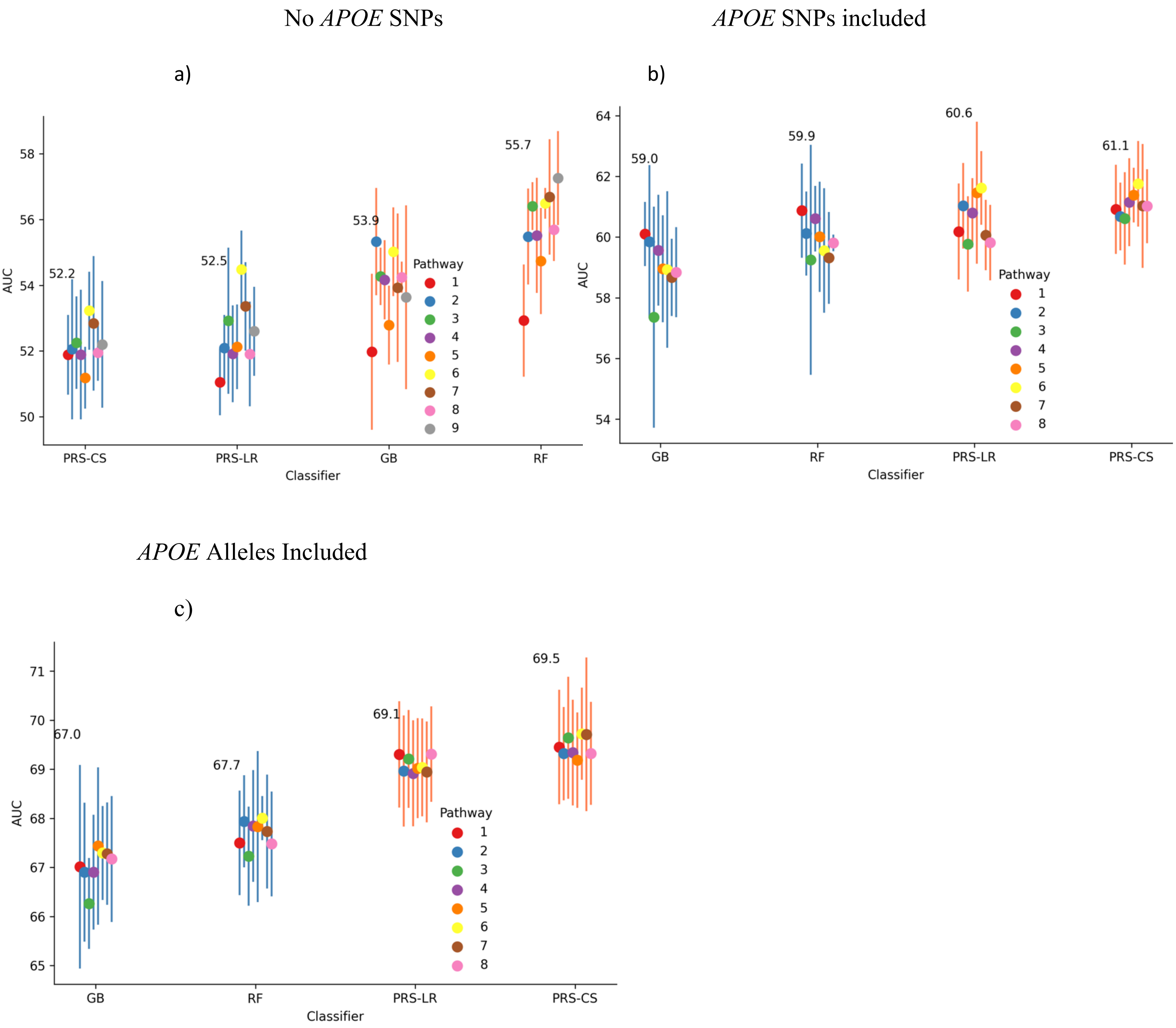
These two figures display calibration plots for the RF (Pathway one) in Figure 7.2. The left-hand plot a) displays pre-calibrated probabilities, whilst the right-hand plot b) shows post-calibration. The predicted probabilities are marked along the X-axis, whilst observed probabilities are measured on the Y-axis. Grouped observations represent the average observed prediction value for each decile of predicted probabilities, accompanied by a 95% confidence interval. The overall relationship between predicted probabilities and observed frequencies is given by the fitted loess smoother, with a 95% (grey shaded area) used.

Figures 7.3a and 7.3b demonstrate calibration statistics for a RF from Figure 7.2b. The left-hand plot demonstrated an under-estimation of risk due to the loess smoother lying above the diagonal, suggesting that the algorithm was not predicting disease risk accurately. Following Platt scaling (Platt, 1999) (sigmoidal correction) (Figure 7.3b), probabilities remained mostly underestimated and therefore the model can still be determined to be poorly calibrated. However, the loess smoother in the right-hand plot lies closer to the diagonal from 0.7-0.9 predicted probabilities.

7.3.3.2 Imputed genotypes

Analyses conducted in the previous section were repeated using imputed genotypes. SNPs were selected using the in-built GWAS method per CV. PRS-CS-LR was not used for analyses using imputed SNPs. This was due to the reference panels provided by the designers of PRS-CS not covering a large enough amount of the SNPs in imputed GERAD.

Figure 7.4: The Comparison of PRS-LR (P-value threshold 0.1) vs Selected Classifiers (RF, GB) for pathways defined by LD Pruned SNPs (Imputed Genotypes)



No *APOE* SNPs

APOE SNPs included

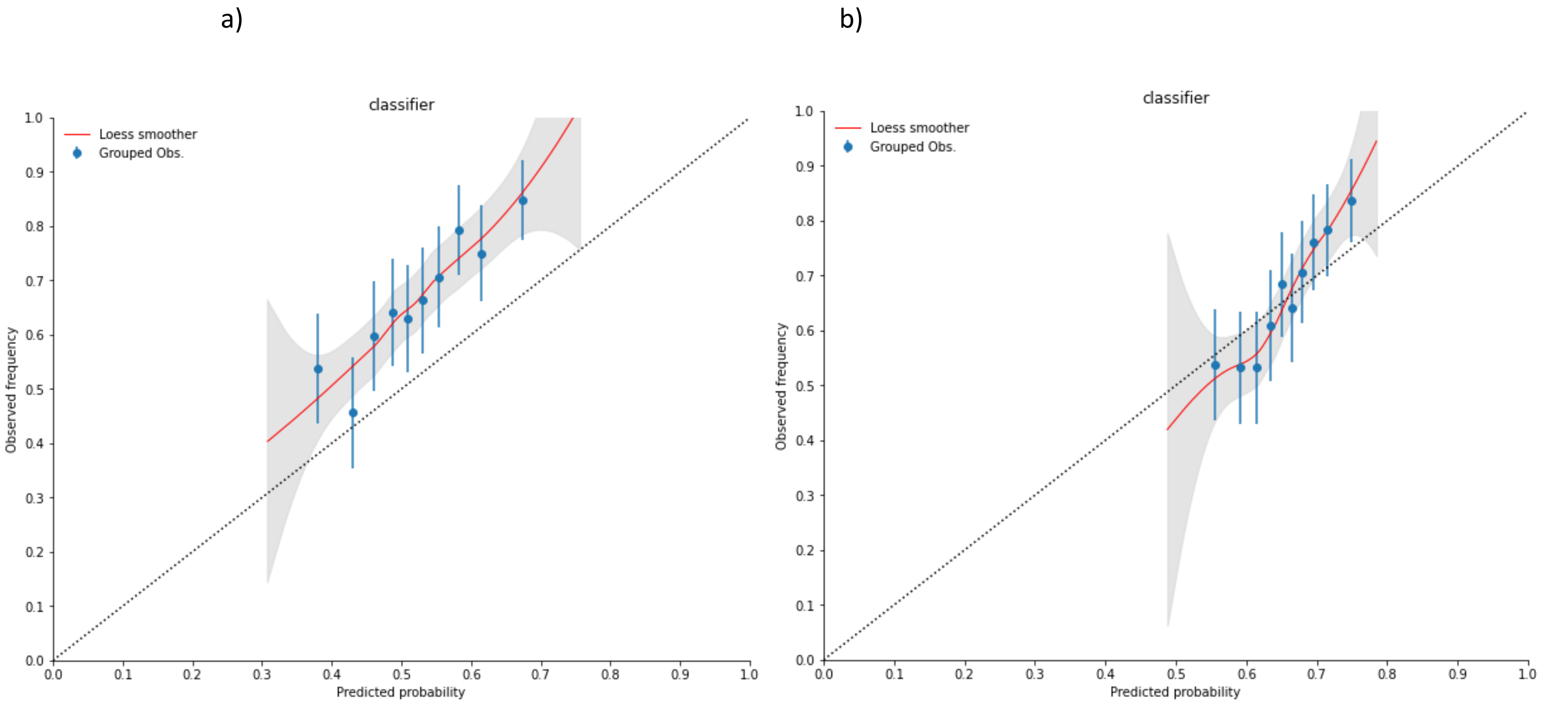
APOE Alleles Included

Y-axis represents AUC in %; with classifiers placed on the X axis. Each dot represents the score for the prediction algorithm for all p-value thresholds, with accompanying 95% CI. The numbers placed centrally are the mean score across pathway SNP sets; GB Gradient Boosting; RF Random Forest; PRS-LR Polygenic Risk Scores Logistic Regression; AUC Area Under the Curve. Plots b and c contain 8 pathways only as pathway 9 does not include SNPs within the *APOE* region.

When comparing results in Figure 7.4 with those in Figure 7.2, prediction performance for classifiers was similar. When excluding SNPs within the *APOE* region, RFs achieved a higher mean AUC than GB, PRS-CS and PRS-LR. The increased performance of RFs over both PRS-LR and PRC-CS was statistically significant in some pathways (1,3,4,5,7,8) (Supplementary Table 46), this was also true for GB when compared to PRS-LR (2, 4) (Supplementary Table 46). In comparison with Figure 7.2b, AUC for SNP-based analyses including the *APOE* region increased by 1-2% for imputed variants (Figures 7.2b and 7.4b). Prediction performance for both RFs and GB when including variants from the *APOE* region was also 2-3% greater than similar genome-wide analyses conducted in Chapters 5&6, however, AUC for PRS-LR remained similar. This provided further evidence for the observation of biologically informed SNP sets resulting in better prediction performance than variants chosen on a genome wide scale.

When comparing classifier performance for results within 7.4b, paired t-tests demonstrated that differences between algorithms were only statistically significant in two pathways (4,6). RFs achieved superior prediction performance to PRS-LR in one pathway (6), whilst the higher values of AUC for GB were also shown to be statistically significant when compared to PRS-LR in one pathway (4). The inclusion of the *APOE* alleles resulted in an increase of prediction performance as also observed in Figure 7.2. PRS-CS and PRS-LR achieved higher mean AUC when compared to GB and RFs, and this was shown to be statistically significant for several pathways (2, 3, 4, 5,6,7) (Supplementary Table 46).

Figure 7.5: Comparison of non-Calibrated vs Calibrated Prediction Probabilities.



These two figures display calibration plots for the RF (Pathway one) in Figure 7.4. The left-hand plot a) displays pre-calibrated probabilities, whilst the right-hand plot b) shows post-calibration. The predicted probabilities are marked along the X-axis, whilst observed probabilities are measured on the Y-axis. Grouped observations represent the average observed prediction value for each decile of predicted probabilities, accompanied by a 95% confidence interval. The overall relationship between predicted probabilities and observed frequencies is given by the fitted loess smoother, with a 95% (grey shaded area) used.

All predicted probabilities in Figure 7.5a were above the diagonal, representing a consistent underestimation of risk. This underestimation was partially corrected following calibration, as shown in Figure 7.5b. The loess smoother lies closer to the diagonal with some deviation.

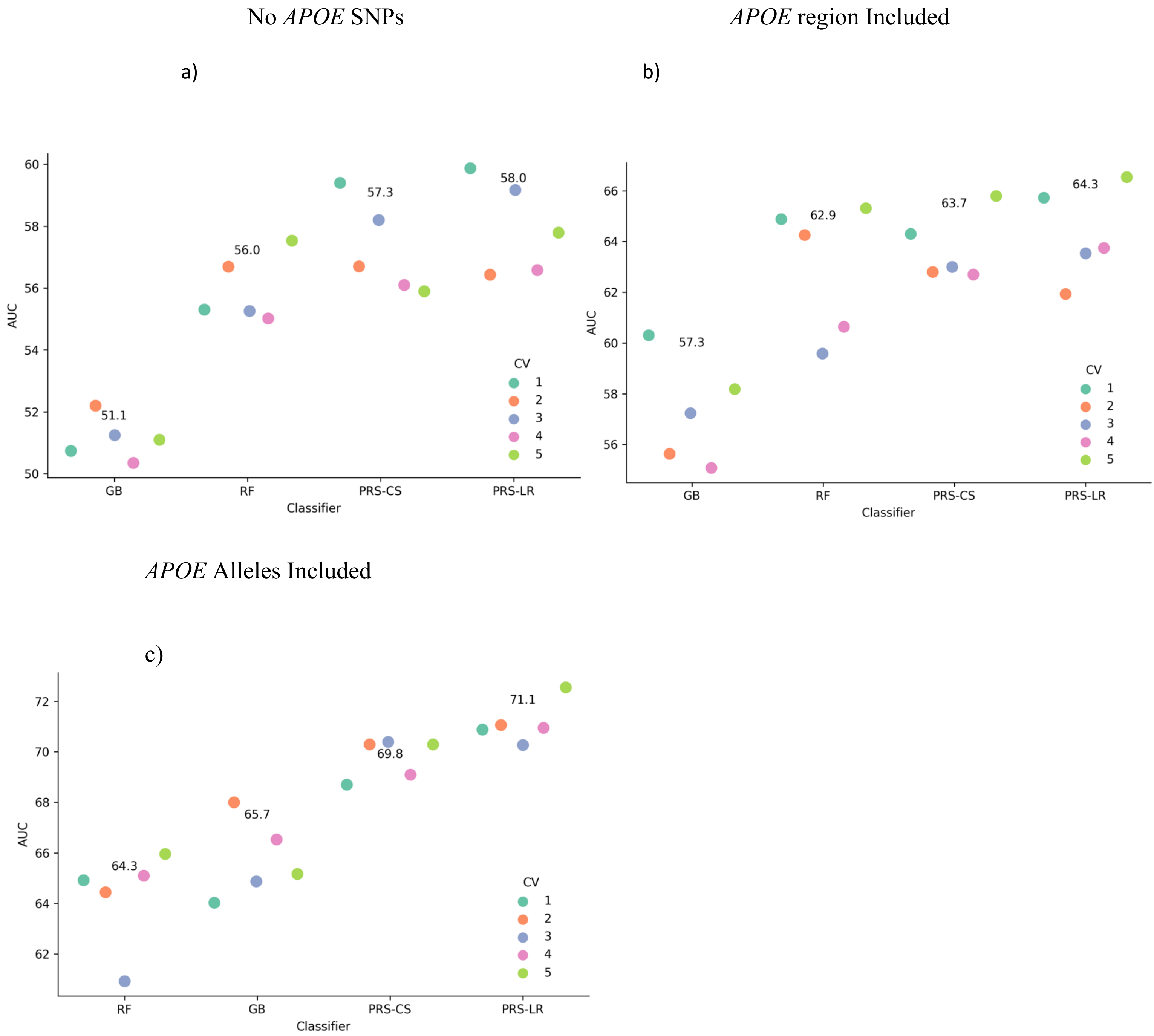
7.3.3.3 Multivariable analysis

Following analysis of singular pathways, a joint modelling approach was used for AD prediction. The nine separate imputed SNP sets were used in a multivariable model for both ML and PRS prediction. Two different types of inputs were used for ML, these were genotypes and PRSs. Calibration plots were again used to assess model prediction, on this occasion, values from one round of CV within the multivariable method were used.

7.3.3.3.1 Pathway analyses results using SNPs

Analyses described below in Figure 7.6 represent the use if a multivariable approach imputed using genotypes as inputs.

Figure 7.6: The Comparison of PRS-LR (P-value threshold 0.1) vs Selected Classifiers (RF, GB) for LD Pruned SNPs in Imputed Genotypes.



No *APOE* SNPs

APOE region Included

APOE Alleles Included

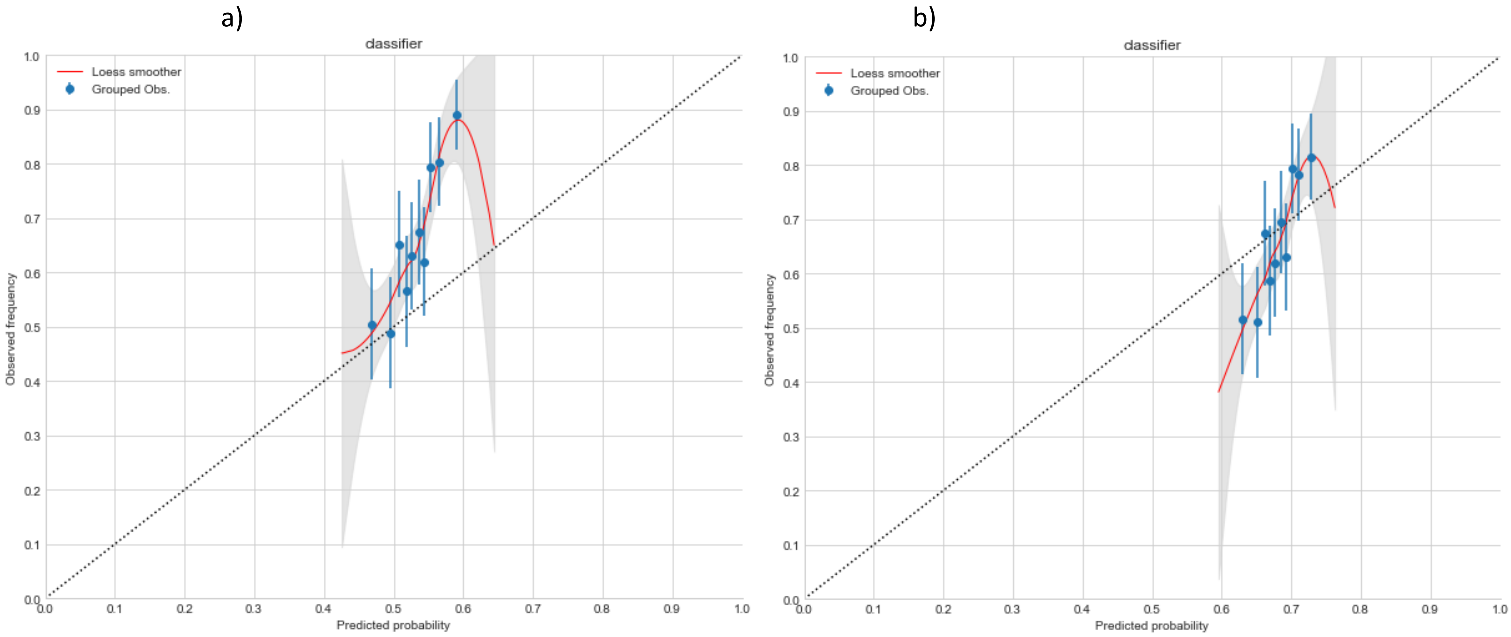
Y-axis represents AUC in %; X-axis represents each classifier's results for each round of CV; The numbers placed centrally are the mean prediction performance across 5 folds of CV; GB Gradient Boosting; RF Random Forest; PRS-LR Polygenic Risk Scores Logistic Regression; AUC Area Under the Curve.

When comparing analyses shown in Figure 7.6a with results in Figure 7.4a, prediction ML performance for the multivariable analysis did not increase relative to analyses of single pathways when excluding SNPs within the *APOE* region. When including SNPs within the *APOE* gene (comparing Figure 7.6b to Figure 7.4b), AUC for RFs increased by 3%, however prediction performance for GB reduced by 1-2%. Therefore, a definitive conclusion on whether using pathways in a multivariate method improved ML performance could not be made.

Prediction performance for both PRS-LR and PRS-CS increased by 4-6% when including SNPs within the *APOE* region, in comparison to singular pathways. This increased performance for both PRS methods resulted in significant results in Supplementary Table 47 for comparisons with mean AUC for GB (p-values = 0.003, 0.003) (no *APOE*) (p-values = 0.003, 0.003) (including *APOE*). This was also true for when comparing performance between RFs and GB, as RFs were shown to be statistically superior (p-values = 0.003 and p-value = 0.006, without and with including *APOE, respectively*). When comparing the prediction performance after inclusion of the *APOE* alleles between Figure 7.6c and Figure 7.4c, mean AUCs were similar for both GBs and RFs in the multivariable method, whilst prediction performance for PRS-LR was increased. As previously, both PRS-LR and PRS-CS outperformed RFs and GB (p-values = 0.003, 0.003), whilst GB achieved greater prediction accuracy than RFs (p-value = 0.012) when including the *APOE* alleles (Supplementary Table 47).

The reason for the increased performance of PRS-LR and not ML when comparing multivariable to single pathway analysis could be related to dimensionality issues. When combining the nine available pathways, the number of SNPs present for prediction is greater than analyses of single pathways. This increase might contribute random noise to the ML algorithms, limiting predictive accuracy. Similar observations were also seen when increasing the amount of SNPs for prediction in Chapters 5&6. PRS-LR is this instance is not limited by the same issue, as the LR has only nine inputs irrespective of the number of SNPs used, therefore dimensionality issues are not as important. Analysis was also conducted to assess the significance of each pathway in the multivariable PRS-LR analyses displayed in Figures 7.6b and 7.6c. Mean values for both p-values and beta coefficients were taken across five folds of CV and recorded in Supplementary Tables 49&50. The only variable shown to be significant (p-value < 0.05) was the summation of the *APOE* alleles multiplied by their respective effect sizes. None of the 9 pathways were shown to be individually significant after correcting for the effects of other pathways. Some pathways were significant for individual CV folds, however this altered when calculating average values across CV.

Figure 7.7: The Comparison of non-Calibrated vs Calibrated Prediction Probabilities



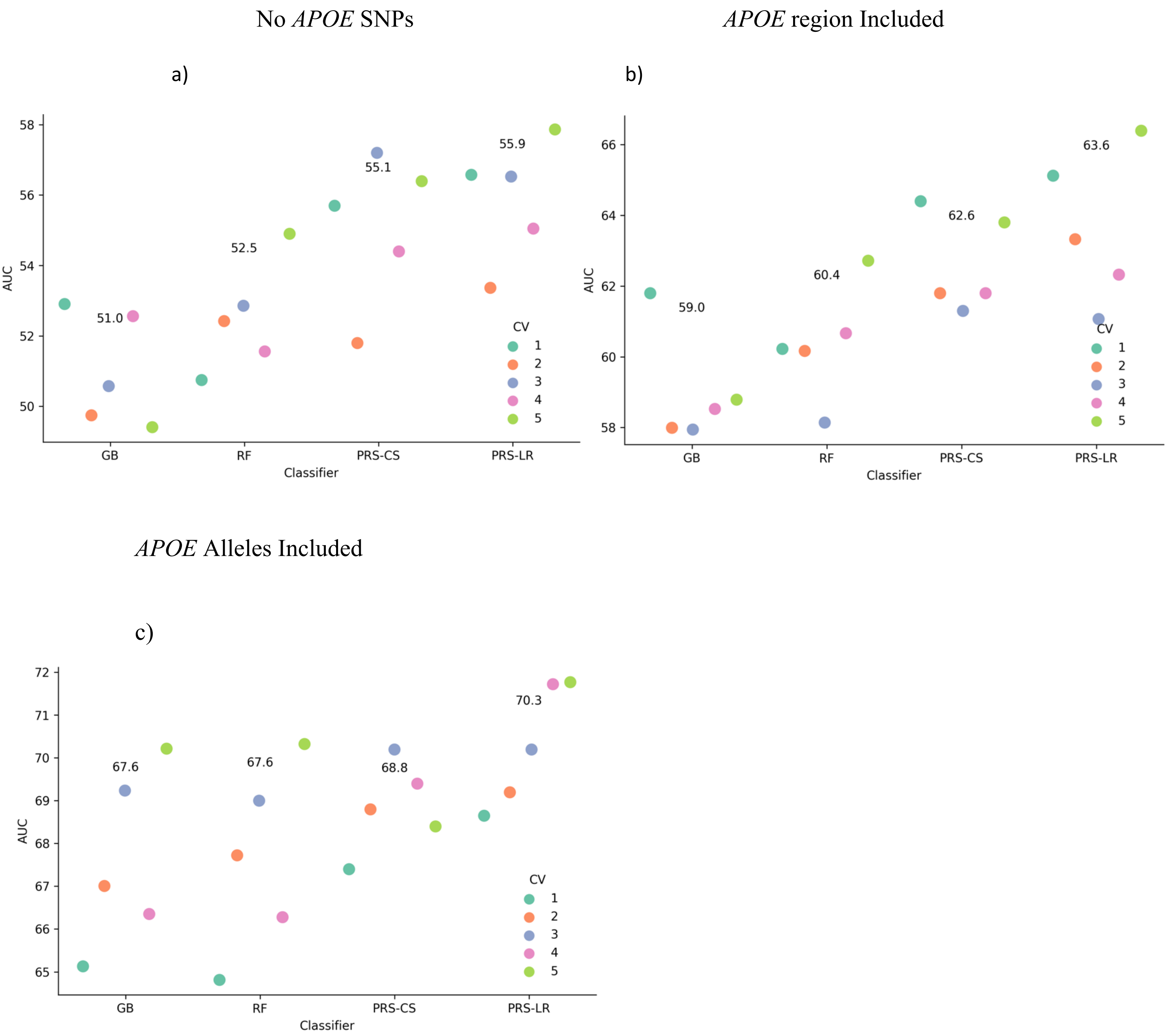
These two figures display calibration plots for the RF (Protein-lipid complex assembly) in Figure 7.6. The left-hand plot a) displays pre-calibrated probabilities, whilst the right-hand plot b) shows post-calibration. The predicted probabilities are marked along the X-axis, whilst observed probabilities are measured on the Y-axis. Grouped observations represent the average observed prediction value for each decile of predicted probabilities, accompanied by a 95% confidence interval. The overall relationship between predicted probabilities and observed frequencies is given by the fitted loess smoother, with a 95% (grey shaded area) used.

Non-calibrated probabilities in Figure 7.7a were consistently above the diagonal, therefore risk was under-estimated. Following calibration, probabilities lay beneath the diagonal, showing that calibration did not improve the estimation of risk.

7.3.3.3.2 Pathway analyses results using SNPs PRSs as inputs

Analysis in this section differed to Section 7.3.2.3.1, as ML, PRS-CS and PRS-LR algorithms were run using PRS values generated using external information from the Kunkle-noGERAD summary statistics.

Figure 7.8: The Comparison of PRS-LR (P-value threshold 0.1) vs Selected Classifiers (RF, GB) for LD Pruned SNPs, with PRS used as inputs. Variants in the *APOE* region initially excluded, included and followed by the inclusion of *APOE* alleles.



No *APOE* SNPs

a)

APOE region Included

b)

APOE Alleles Included

c)

Y-axis represents AUC in %; X-axis represents each classifier's results for each round of CV. The numbers placed centrally are the mean prediction performance across 5 folds of CV; GB Gradient Boosting; RF Random Forest; PRS-LR Polygenic Risk Scores Logistic Regression; AUC Area Under the Curve.

When comparing analysis in Figure 7.8 with results in Figure 7.6, prediction performance for ML was altered. Mean AUC for RFs fell by 3.5% when excluding the *APOE* region, there was also a reduction of 2.5% AUC for RFs when including SNPs within the *APOE* region. However, this reduction in prediction performance for RFs was not seen when including *APOE* alleles. This might suggest that the combination of SNPs into a unified dataset results in poorer performance, however the inclusion of *APOE* alleles increases AUC to levels seen in previous analyses. Therefore, the method used for non *APOE* related SNPs might not matter as alleles are the main component for prediction. Prediction performance for PRS-LR and PRS-CS were similar however when compared to results in Figure 7.6 when both including and excluding SNPs in the *APOE* region. This is surprising given PRS for analyses shown in Figure 7.6 were generated using Kunkle-noGERAD (external information) summary statistics, this should have provided greater information than the PRS generated from internal GWAS used in the analyses in Figure 7.6.

However, when comparing prediction performance between methods within Figure 7.8, both PRS-C and PRS-LR were shown to still significantly outperform RFs and GB both removing (p-values = 0.020, 0.019) and including SNPs within the *APOE* region (p-values = 0.008, 0.011) as shown in Supplementary Table 47. When using *APOE* e2/e4 alleles as an extra predictor, AUC for both GB and PRS-LR was similar to those seen in previous analyses. However, on this occasion only PRS-LR achieved superior performance to both GB when comparing AUCs using pairwise t-tests (p-values = 0.047) (Supplementary Table 47).

Figure 7.9: The Comparison of non-Calibrated vs Calibrated Prediction Probabilities

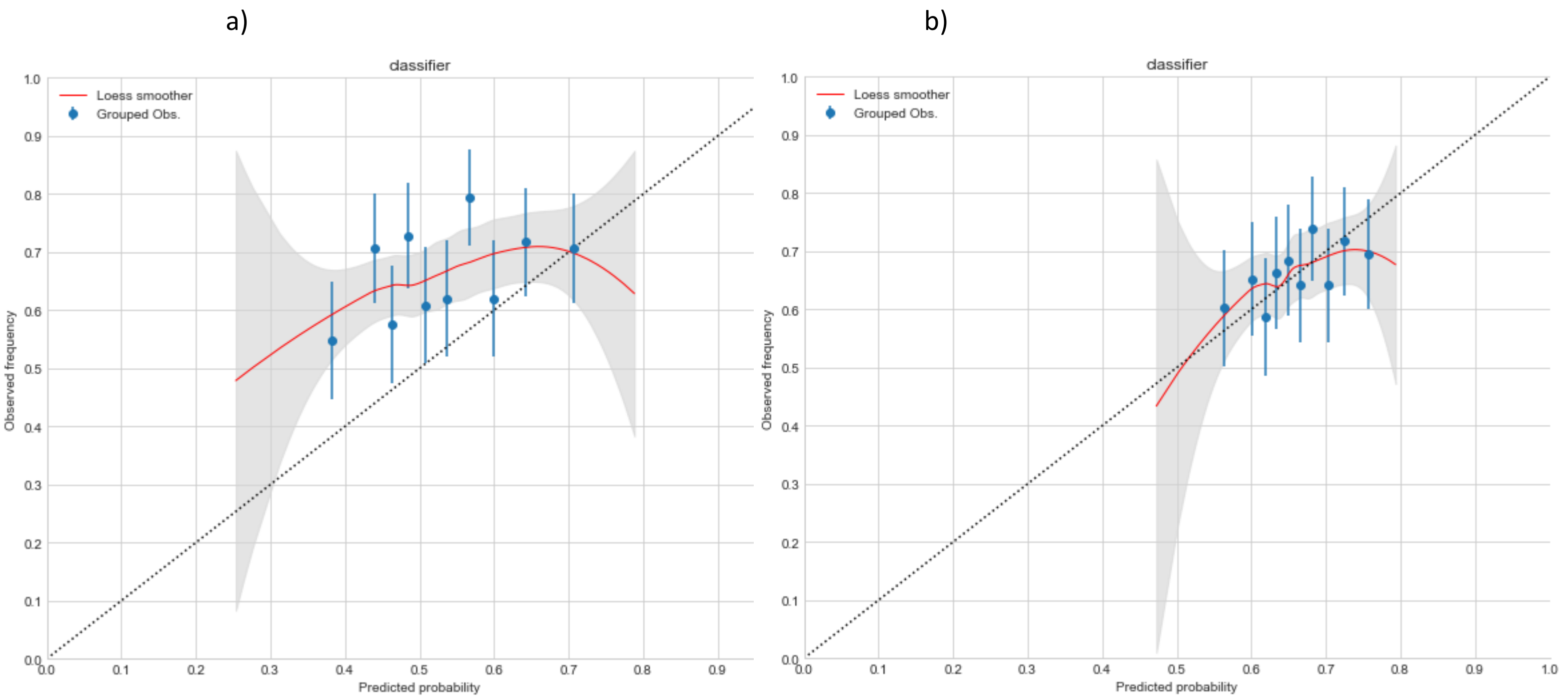


Figure 7.8. These two figures display calibration plots for the RF (Protein-lipid complex assembly) in Figure 7.8. The left-hand plot a) displays pre-calibrated probabilities, whilst the right-hand plot b) shows post-calibration. The predicted probabilities are marked along the X-axis, whilst observed probabilities are measured on the Y-axis. Grouped observations represent the average observed prediction value for each decile of predicted probabilities, accompanied by a 95% confidence interval. The overall relationship between predicted probabilities and observed frequencies is given by the fitted loess smoother, with a 95% (grey shaded area) used.

Initial probabilities depicted in Figure 7.9a displayed an under-estimation of risk as most of the loess smoother resided above the diagonal. Following calibration, probabilities were aligned more to the diagonal, suggesting a better estimation of risk.

7.3.3.4 Combined pathway dataset

Analysis in this section used a single combined dataset derived from the 9 imputed pathway SNP sets. Results are shown in Supplementary Figures 3-6. When comparing the use of genotypes for prediction, (Supplementary Figure 3) to the multivariable approach (Figures 7.6a and 7.6b), mean AUC was reduced for PRS-LR and PRS-CS by 5-9%. Therefore, it appears that modelling PRS-LR and PRS-CS in a multivariable fashion achieves better disease prediction than using a unified SNP set. AUC for GB and RFs was also generally lower when compared to results in Figures 7.6a and 7.6b. This reduction in AUC for both PRS-LR and ML classifiers was true both when including and excluding variants within the *APOE* region. When comparing classifier performance within analyses, there were no significant differences. When including *APOE* alleles, prediction performance was similar to multivariable methods.

When assessing calibration statistics for the use of genotypes, pre-calibrated probabilities shown in Supplementary Figure 4 were under-estimating risk due to all probabilities lying above the diagonal. This issue was not addressed as most probabilities resided above or below the diagonal following attempted correction.

The second half of analyses in this section used PRS values for inputs to ML algorithms, with results displayed in Supplementary Figures 5 and 6. When comparing performance with that in Supplementary Figure 3, prediction performance for RFs was lower by 4%. This was the only noticeable difference, as all other comparisons between results were within 1-2% AUC. The performances of PRS-LR and PRCS-CS were again lower than that observed under the multivariable modelling of pathway (Figure 7.6), whether SNPs within the *APOE* region were included. This reinforces the finding that, under a PRS-LR analysis, modelling pathways in a multivariable fashion gives greater prediction performance than either modelling single pathways or combining SNPs in all pathways into a single PRS. When comparing prediction performance between ML and PRS-LR within each plot, none of the pairwise t-tests conducted returned significant results. Prediction probabilities were initially underestimating risk as shown in Supplementary Figure 6. Calibration realigned probabilities closer to the diagonal, however some deviation from the diagonal was still present.

7.4 Discussion

Analyses in this chapter assessed the predictive performance of the 9 AD associated pathways reported in Kunkle 2019 (Kunkle *et al.*, 2019). Predictions from two ML algorithms (RFs and GB) were compared to both PRS-LR, and PRS-CS-LR. Genotypes used were from non-imputed and imputed variant sets. Pathways were initially assessed separately, before multivariate and joint models gauged the predictive performance of combining all SNP sets.

7.4.1 Individual pathway analysis

Comparing performance for ML vs PRS-LR/PRS-CS-LR for non-imputed genotypes, both RFs and GB achieved higher mean AUC than both PRS techniques when excluding the *APOE* region. These differences were shown to be statistically significant when comparing CV AUCs of classifiers within pathways (Supplementary Table 45). This is the first occasion in which ML has achieved higher mean AUC than PRS based models in this thesis. However, both PRS-LR and PRS-CS achieved marginally greater mean AUC when reintroducing *APOE* related variants when compared to GB and RFs. The marginal difference in mean AUC across CV resulted in only three instances in which PRS based algorithms significantly outperformed ML (Supplementary Table 45). When including *APOE* alleles instead of *APOE* SNPs for modelling purposes, PRS methods outperformed RFs and GB when using both non-imputed and imputed variants. These differences were again shown to be statistically significant when using t-tests in a number of pathways, with results shown in Supplementary Table 45.

RFs and GB continued to achieve higher mean AUC than both PRS-LR and PRS-CS when excluding *APOE* related SNPs, with differences across several pathways shown to be statistically significant (Supplementary Table 46). Similarly, to observations made when using non-imputed genotypes, PRS based methods regained superiority over both GB and RFs when reincluding variants in the *APOE* region. However, the difference between prediction methods was marginal. Results when including the *APOE* alleles were also similar to when using non-imputed SNPs, with prediction performance for PRS methods shown to statistically significant when compared to ML in some pathways (Supplementary Table 46). Analyses conducted in Chapter 5&6 involved the use of SNP sets derived on a genome wide basis. The highest achieved AUC for both GB and RFs without the inclusion of *APOE* alleles was 57-58%, with prediction performance declining as greater amounts of SNPs were introduced. Therefore, both GB and RFs achieved AUC 2-3% greater in this chapter when assessing single pathways. Analyses conducted in Chapter 4 used a select number of SNPs assessed to be genome-wide significant in Kunkle et al., 2019. Prediction performance for classifiers was again below that observed in this chapter, with only a few examples of AUC close to 60% when using an increased amount of SNPs (422 SNPs).

This suggests that using SNP sets selected due biological relevance in AD appears to be a good strategy for ML. This might be due to the increase in the proportion of disease related variants and simultaneously removing nonrelated SNPs included. The overall reduction in SNP volume might also reduce the impact of dimensionality issues.

7.4.2 Multivariable and combined pathway analysis

Following analysis of individual pathways, focus moved to assessing prediction of disease status when using all pathways simultaneously. When assessing results detailed in Figure 7.1, prediction performance for ML when using a multivariable analysis was not superior to analysing pathways individually. When using imputed variants, mean AUC across the 9 pathways was similar for individual pathway analysis to that given by the multivariable model. This was true both when including the *APOE* region and excluding variants within it. Including the *APOE* alleles also resulted in little difference between the two types of pathway analysis. This was also true for using PRS as inputs to ML in the multivariable method, as AUC for ML was similar between the singular pathway analysis and multivariable model.

The only noticeable difference in prediction performance between individual and multivariable pathway analyses was the performance of both PRS-LR and PRS-CS. AUC for the multivariable model using both internal and external information was 4-6% higher than individual pathways. This suggests that using multiple pathways can achieve higher discrimination for PRS based linear methods. Comparisons between the multivariate PRS-LR and PRS-CS with ML resulted in superior performance for the linear based models in most circumstances (Supplementary Table 47). As previously discussed in Section 7.3.2.3.1, reasoning for the increased performance of PRS-LR and PRS-CS over ML when analysing pathways in a multivariable manner might be due to dimensionality issues. Increasing the number of SNPs for PRS calculation does not elevate the number of predictors in the LR, whilst number of features for ML is directly increased. This rise in the number of inputs might limit prediction due to the introduction of random noise.

Further analysis used a combined pathway SNP set for prediction. Results for both RF and GB were marginally lower (1-2% AUC) in comparison to using the multivariate method, this was true when either using an in-built GWAS to generate summary statistics and effect sizes form Kunkle-noGERAD. Prediction performance for both PRS-LR and PRS-CS (internal and external information) fell by 5-8% AUC when compared to results for the multivariable model. The use of external information (Kunkle-noGERAD summary statistics) did not result in better prediction performance than when using an internal GWAS. This was also true for ML.

7.4.3 Predictive capability of pathways versus genome wide analyses

When comparing ML performance to analyses in previous chapters, the use of pathway related variants appeared to increase predictive performance for RFs and GB. Previous analyses which used SNPs as inputs reached AUC of 57% (Chapters 5&6). However, analyses for individual pathways in this chapter achieved AUC of 59-60% when using imputed genotypes (including *APOE* region SNPs). Prediction performance of PRS-LR and PRS-CS using a multivariate approach, including SNPs within the *APOE* region, also outperformed the whole-genome analyses. A similar result was observed for ML analyses trained on non-imputed genotypes. However, when including the *APOE* alleles, AUC for the multivariate PRS approach did not reach the level of 72% recorded for whole genome wide analyses in Chapter 5.

7.4.4 Comparing the performance of different inputs to ML

When combining the SNPs in all 9 pathways into both multivariable and amalgamated analyses, inputs to ML could take the form either of genotypes or PRS values. The latter were generated using external information. When assessing results in Figure 7.1, it can be determined that using external information to generate PRS values did not improve prediction performance over using genotypes as inputs. This is likely due to the loss of flexibility incurred by combining multiple SNPs into a single input. More surprisingly, using external information to generate the PRS for PRS-LR and PRS-CS did not result in better prediction performance than using an internal GWAS. This was true for both the multivariable and amalgamated analyses. This can be considered a surprising result as using external information are expected to provide classifiers with greater information for prediction.

7.4.5 Comparing the performance of PRS-CS and PRS-LR

Results demonstrated that PRS-CS was able to achieve increased prediction accuracy over PRS-LR when using non-imputed variants and including SNPs within the *APOE* region. However, the increase in prediction accuracy was not observed when using imputed genotypes. The most likely reason is the reference panels provided by PRS-CS. Both the 1000 genomes and UK biobank reference panels provided information for around one million SNPs, whereas the imputed GERAD dataset comprises six million variants. Therefore, only a fraction of SNPs can be used in PRS-CS modelling, many fewer than for PRS-LR. Despite this disadvantage however, PRS-CS gave similar prediction accuracy to PRS-LR for most analyses using imputed variants. This provided further evidence that PRS-CS could provide greater accuracy for AD prediction compared to the traditional approach of clumping and thresholds, provided reference panels of the future cover a larger number of genomic variants.

7.4.6 Calibration

All models were shown to be underestimating risk. Following calibration using platt scaling, calibrated probabilities produced a mixture of results. In some circumstances, risk estimation was realigned successfully, however in other scenarios, calibration had no effect on the distribution of probabilities. The underestimation of risk by models in this chapter was similar to calibration statistics also seen for the whole genome analyses previous chapters.

7.4.7 Conclusions

The central focus of this chapter was to assess prediction for both ML and PRS linear methods using the 9 pathways deemed significantly enriched for AD risk SNPs (Kunkle et al., 2019). Performance of individual pathways reached 58-60% AUC for ML when including variants within the *APOE* region. This is higher than comparable analyses run in previous chapters, in which variants from the whole genome were used. This suggests narrowing feature sets by using biologically relevant AD related SNPs instead of variants from the whole genome may improve classifier performance.

Despite the improved prediction performance for ML algorithms, AUC for both PRS-LR and PRS-CS was not higher than previous analyses using genome wide approaches. The use of a multivariable approach for ML did not result in higher AUCs, suggesting that combining pathways does not assist ML performance. However, this was not the case for both PRS-LR and PRS-CS, as the multivariable model achieved 4-5% greater AUC than single pathways.

Comparing ML versus the PRS bases approaches of PRS-LR and PRS-CS, using both non-imputed and imputed genotypes with *APOE* related SNPs removed in individual pathways, both GB and RFs achieved better performance than PRS-LR and PRS-CS. This was the first occasion throughout this thesis in which ML outperformed PRS based algorithms. This was supported through statistically significant outcomes when comparing performance by using pairwise t-tests. However, when reincluding SNPs within the *APOE* region, both PRS-CS and PRS-LR outperformed ML in most subsequent analyses.

In summary, analyses in this chapter have shown the use of pathway SNPs on an individual basis resulted in superior prediction than analyses using whole genome variants in Chapters 5&6. This suggests that using biologically informed SNP sets can aid prediction performance for AD. ML performance for singular pathways also resulted in similar performance to PRS based methods, this is not in line with results shown in Chapters 5&6, where PRS outperforms ML by a clear margin. However, PRS based methods regained superiority over ML when using a multivariate approach in this chapter. Values for AUC increased above those observed for PRS for singular pathway analyses, suggesting that greater prediction accuracy can be achieved when combining pathway SNPs.

8 Discussion

8.1 Overview

Alzheimer's Disease (AD) is a neurological condition which impairs an individual's cognitive abilities. Onset of disease is most common after the age of 65, with initial symptoms involving loss of memory and motor skills (Farfel *et al.*, 2019). As the disease progresses, care requirements increase until the individual becomes solely reliant on others and loses self-awareness. The final stage of illness is most commonly death through infection (Allen *et al.*, 2003). The burden of AD on patients and society is predicted to increase in coming decades due to increasing life expectancies and population growth (Nichols *et al.*, 2022). Therefore, the ability to understand the underlying causes of the disease for treatment purposes is crucial. Causal mutations in genes have been identified for the less common autosomal form of the disease, however no causal variant has yet been identified for the sporadic form. The genetic component of the more prevalent type of AD has been shown to be polygenic, with genetic causality the likely effect of interactions between many variants (Baker and Escott-Price, 2020).

A common approach for assessing the risk of an individual at a genome-wide scale for a certain phenotype is polygenic risk score (PRS). Its simplicity of use and predictive capability has resulted in its use in various research fields (Lewis and Vassos, 2020c). For AD prediction, PRS has achieved prediction accuracy of between 75-84% AUC in multiple publications (Baker and Escott-Price, 2020). Despite this, research has suggested that PRS is limited when predicting complex disorders. Linear prediction models such as PRS assess linear relationships between SNPs. However, it is most likely that polygenic disorders such as AD are the result of interactions between multiple genes. A method more suited to assessing non-linear relationships is machine learning (ML). However, comparisons of both PRS and ML performance for the prediction of AD are rare.

A systematic review to assess the current literature for ML prediction in AD using genetic data was conducted as a part of this thesis. Twelve articles from an initial set of 4,020 were included for review. Several insights into the current literature for AD prediction were highlighted. Metrics used for reporting results varied, with only 5 studies reporting area under the curve (AUC). The majority of the remaining 7 articles used accuracy to report prediction performance, however this method has been shown to be susceptible to imbalanced class distributions (Francisco J. Valverde-Albacete and Peláez-Moreno, 2014). Calibration statistics represent the confidence with which predictions are made by ML algorithms. Prediction tools used in medical settings are required to make well informed decisions, as mistakes can result in issues such as misdiagnosis and incorrect treatment (Van Calster et al., 2019). Only one article of the 12 reviewed reported calibration statistics. Sample size was another area highlighted. Events per variable (EPV) is a metric used to assess the balance between samples and predictors. Research has identified that values below 20 features per sample can increase the likelihood of overfitting (Austin and Steyerberg, 2017c). It was identified that most datasets used across the 12 articles were below this threshold. Finally, most articles used the ADNI dataset, which is publicly available, reducing the diversity of sources used.

Analyses in Chapter 4 were an exploratory assessment of the predictive capabilities of PRS and ML for AD prediction. Twenty-three genome-wide significant SNPs from Kunkle et al., 2019 were used for predictors, as these were likely to have strong contribution with AD risk. However, these were not present within the non-imputed GERAD dataset (Harold *et al.*, 2009), most likely due to the differences in genotype platforms used. Therefore, two sets of proxy SNPs were tested. The analyses in this chapter identified the spurious effect on results caused by the presence of the 1958 birth cohort. Therefore, these samples were removed from GERAD for analyses from Chapter 4 onwards. The predictive performance of PRS using a LR (PRS-LR) was compared to several ML algorithms, including random forests (RFs), gradient boosting (GB), naïve bayes (NB) and SVMs with several kernels. Initial analyses using the reduced dataset focused on adjusting SNPs for population stratification. Comparisons of classifier performance identified a superior performance for PRS-LR to all other ML algorithms. RFs, GB and SVMs achieved similar levels of AUC, with prediction performance generally increasing as the number of SNPs for prediction also increased from 23 to 422. This increase in AUC did not occur for NB however, most likely due to the absence of a penalisation method for Bayesian models. Adjusting genotypes with principal components, age and sex values and also using a balanced dataset, with cases and controls matched upon both sex and age did not change classification accuracy.

Analyses conducted in Chapter 5 involved the use of variants on a genome-wide scale. Datasets for prediction were generated by LD-clumping using a range of p-value thresholds (0.0001-0.5). It was hypothesised that using external summary statistics might be providing PRS-LR with greater information than ML for predictions. Therefore, in this chapter an in-built GWAS was calculated within each training fold of CV, with subsequent summary statistics used to select SNPs for prediction.

Two decision tree-based algorithms were used for ML, these were RFs and GB. The least stringent p-value thresholds of (0.2-0.5) resulted in SNP sets comprising 50,000-100,000 variants. Since ML algorithms are susceptible to an issue known as the ‘curse of dimensionality', feature selection algorithms were used to reduce the impact of this issue. When not using feature selection, prediction performance for both RFs and GB ranged from 66-57% AUC, with mean performance across all p-value thresholds residing at 60-61% AUC. This was however below the mean performance of 71.5% AUC for PRS-LR. The use of feature selection also improved average performance for ML, with the use of the RFs and ExtraTrees feature selection algorithms improving mean AUC to 65-66%. Whilst this was an improvement on not using feature selection, it was still below the mean performance of PRS-LR. This result was unaltered when exploring other methodology, such as removing the clumping phase and using a less significant r^2 when clumping.

Chapter 6 introduced the use of imputed genotypes for prediction algorithms. Imputation techniques have been used to increase the coverage by the GWAS by estimating values of SNPs not directly genotyped. Imputed genotypes were initially in the format of genotype dosages and then subsequently converted to best-guess dosages. Analyses in this chapter were conducted on both formats of imputed data. Results demonstrated that the use of imputed variants did not improve ML performance above the use of non-imputed SNPs. No discernible difference was seen when using either dosages or best-guess genotypes.

Chapter 7 used a different approach to select relevant SNPs for AD prediction. Analyses in this chapter derived SNPs from 9 biological pathways associated with AD reported by (Kunkle *et al.*, 2019). These biological processes were mainly related to lipid assembly and the regulation of tau and beta-amyloid. The non-imputed and imputed GERAD datasets were filtered using the genomic base positions of genes within each pathway. Initial analyses assessed prediction performance of each pathway on an individual basis. Results demonstrated that PRS using clumping and thresholding and PRS-CS approaches outperformed ML. AUC achieved for both RFs and GB, was 1-3% higher than analyses in previous chapters. This was true when using both non-imputed and imputed genotypes.

Following the analysis of single pathways, further analysis assessed the joint predictive capability of all pathways. This combined approach resulted in an increase in AUC for PRS based methods when compared to singular pathway analyses, whereas prediction performance for ML remained unchanged. The analysis of joint pathways also investigated the use of PRS for inputs to ML, with the intention of assessing whether this could result in better prediction than using genotypes. Despite the use of the external summary statistics to generate PRS, results for this analysis were no different to using genotypes. A further method for assessing the joint prediction of pathways was the amalgamation of all SNPs within each pathway into one large SNP set. Results for this method involved a reduction in AUC for PRS based methods when compared to the multivariable model, whilst prediction performance for ML remained similar. This was true for when using either genotype or PRS as inputs to models.

8.2 How well can machine learning predict Alzheimer's disease from genetic data?

One of the main aims of this thesis was to assess the capability of ML for predicting AD from SNPs. Analyses have shown that performance ranged from 57-62% AUC for the best performing algorithms across all chapters. This increased to 69% when including the *APOE* alleles (e2, e4) directly. Therefore, ML is yet to significantly outperform previously used linear methods. One reason for this might be the use of the clumping to selecting significant SNPs for prediction. This allows PRS to use main effects to discriminate between case/control status but may impair the ability to assess interactions (non-linear effects) between SNPs, which is one of the main advantages of ML over PRS.

8.3 How does the predictive capability of ML compare to PRS?

It has been hypothesised that ML might achieve greater prediction for polygenic disorders, due to their ability to assess complex interactions between predictors (Gola, Erdmann, Müller-Myhsok, *et al.*, 2020). The results of comparisons for other polygenic disorders have been mixed, with ML outperforming PRS in some instances. Results in this thesis demonstrated a consistent trend of superior performance for PRS based methods over ML. When using SNPs derived on a genome-wide scale, linear methods achieved 3-5% AUC better prediction performance. However, when using smaller datasets informed by biological processes, the difference between the two methods reduced, with ML performing as well in some instances.

A possible explanation for a failure of ML to outperform PRS when predicting AD is the relatively low number of samples available for analysis. For scenarios involving large amounts of features, complex ML algorithms require sufficient cohort sizes in order to overcome possible dimensionality issues (Rajput, Wang and Chen, 2023). Sample sizes can be restricted due to the cost of recruiting individuals and collecting measurements for biomarkers. Prediction performance for ML is optimised when the ratio between data instances follows a ratio of 10:1 or greater. As discussed throughout this chapter, prediction of complex diseases through genetic data usually results in an imbalance between features and samples. This imbalance reduces the likelihood of developing generalisable models. Increases in cohort sizes will reduce this, whilst also increasing the ability of the dataset to represent underlying trends in the population.

8.4 Notable achievements

One of the highlights of this thesis was the ability to analyse large amounts of SNPs. Previous studies using PRS based methods for the prediction of AD have analysed 100,000 to 250,000 SNPs. However, the use of this many variants when using ML algorithms is rare. Analyses in Chapters 5 and 6 in this thesis used up to 200,000 when comparing both ML and PRS methods. This was made possible due to the use of the *subprocess* function in *Python*, in which external processes are facilitated whilst the main program runs. However, despite the ability to run complex processes within *Python,* analysing SNP sets comprising 50,000 variants or more required greater computational resources available to a local machine. Therefore, the supercomputing system Hawk was used to run large datasets. The combination of Hawk and the use of the *subproces*s function allowed for the analysis of an uncommon number of SNPs for AD prediction.

Another highlight of this thesis was the comprehensive review of the use of ML for the prediction of AD using SNPs. This was the first investigation of its kind and highlighted several key issues in this research field. These included the consistent use of the same data source (ADNI), as well as sample sizes below the desired threshold when comparing the number of samples and predictors. Areas of concern in methodology were also found, such as inconsistent reporting of cross-validation methods and calibration statistics. This review gave an insight into requirements for analyses in subsequent chapters.

8.5 Limitations

Several limitations have been discussed throughout this thesis. Cohorts used for the prediction of AD using genetics typically comprise of more features than samples. Challenges in collecting data for analysis involve the financial cost of recruiting individuals and bureaucratic barriers to accessing previously assembled datasets (Moore, Asselbergs and Williams, 2010). Imbalances between samples sizes and the number of predictive variables are prevalent in GWAS as the human genome comprises of millions of SNPs. These imbalances increase the likelihood of non-generalisable ML algorithms due to the curse of the dimensionality (Chattopadhyay and Lu, 2019). The issue of sample size is also relevant to analyses in this thesis following the removal the 1958 birth cohort from GERAD, reducing the size of the cohort from 10687 samples to 4603. This was required due to the skewing of the age distribution from the presence of 5400 controls aged 45. The skewed nature of the dataset resulted in spurious prediction results due to relationship between age and AD. Using a smaller sample size for analyses may have limited ML performance, as SNPs with low minor allele frequency or effect sizes might not contribute to prediction.

Despite the relatively small sample size of 4603 individuals, the number of SNPs used for analyses caused computational issues. When less significant p-value thresholds such as 0.4 and 0.5, were used in the clumping process, the resulting number of SNPs was 200,000 or more. Modelling this number of variables for ML could be considered an achievement due to the computational requirements needed. However, further computational burdens were realised when optimising hyperparameters for ML algorithms. Methods for optimisation typically fit multiple models from predefined parameter distributions to find the combination which achieve the highest AUC. This requires additional amounts of time and memory, which are expanded when using larger datasets. Despite the use of the supercomputing system, the amount of resources available limited the depth of the hyperparameter search.

8.6 Future work

The main question arising from this thesis is whether larger sample sizes might further improve the capabilities of ML for AD prediction. As the number of AD cases rise in coming decades alongside improvements in resource collection, data sources will inevitably increase in size. This is already being seen in the increasing size of cohorts used for GWAS, which have led to the discovery of further GWAS significant SNPs (Bellenguez, Küçükali, Iris E. Jansen, et al., 2022). However, as shown in this thesis, larger datasets require greater computational resource to analyse. This is relevant to ML algorithms, which typically require large sample sizes and thorough hyperparameter searches for optimisation (Adadi, 2021). Despite the use of supercomputing hardware, hyperparameter tuning was limited when vastly increasing the number of predictors. Therefore, computational capabilities will also need to improve to assess the optimum ability of ML to predict polygenic diseases.

Analyses within thesis used SNPs for prediction only. Future work might involve the use of other forms of genetic variants, such as CNVs (copy number variants) and rare SNVs (single nucleotide variants) derived from whole genome sequencing. The use of this form of data has increased in recent years, as issues initially existed concerning both the financial and computational cost of mapping genomes. However, as advancements in technology have continued, the burdens of retrieving such data have reduced but not disappeared (Muir *et al.*, 2016). Despite being in its relative infancy, comparisons have been drawn between the use of sequencing and genotyping for disease prediction. The result of which was substantial differences in the risk predictions for various polygenic disorders (Morgan, Chen and Butte, 2012).

The presence of correlation (LD) between variants can result in inaccurately assigned effect sizes and masking of the true signal between causal variants and disease risk (Grady, Torstenson and Ritchie, 2011). Methods such as clumping and pruning have been introduced to limit the impact of such correlations, however the effect cannot be completely removed. An avenue for future analyses might involve comparisons of how LD between causal and proxy variants effects disease predictions for ML and PRS on an individual basis.

A further avenue not explored in this thesis is the use of deep learning. Traditional ML algorithms such as RFs, GB and SVMs are shallow learners. Deep learning algorithms known as neural networks (NN) use layers of interconnected nodes to learn complex patterns between features and targets. This interconnection of nodes allows for the transformation of data inputs in a non-linear fashion, allowing the model to assess complex non-linear relationships. Deep learning models have achieved greater prediction accuracies than less complex ML algorithms for a range of applications (Najafabadi *et al.*, 2015). Therefore, it is hoped that NNs can improve upon the performance of ML for complex disorders. However, as discussed in the previous paragraph, ML algorithms require large datasets to achieve desired outcomes. This is more of an issue for NNs, as deep learning requires richer data sources than shallow learning to assess more complex patterns. The interconnected architecture of the model and the large datasets analysed also result in significant computational burden (Lippmann, 1987). Therefore, large amounts of computational resources shall be required to analyse larger GWAS.

8.7 Clinical Applications

Research into the ability of ML algorithms to predict AD is conducted with the aim of assessing its use in a clinical setting. The results of this thesis suggest that further development is required as model accuracies were below clinically accepted thresholds. Prediction tools used by clinicians are required to achieve high levels of success due to the severity of false predictions. Errors could result in incorrect treatment packages and potential harm to patients (Verma *et al.*, 2021).

8.8 Conclusion

The results of analyses across chapters demonstrate that ML can only at best compete with PRS, with the linear method outperforming ML in most cases. The difference between two methods varied by chapter, with PRS methods achieving the highest difference in prediction accuracy when selecting SNPs on a genome wide level. The gap between both methodologies reduced when selecting variants using biological information (pathways), with AUC for ML algorithms improving when compared to genome wide analyses. This suggests that selecting variants using prior AD related information might enrich SNP sets for better prediction.

Reasons for the consistent superior performance of PRS over ML methodologies lie in the structure of datasets. Situations in which features outnumber samples increase the likelihood of overfitting, whereby algorithms perform poorly on unseen samples (Ying, 2019). This scenario is relevant in the field of disease prediction using genetics as economic factors often limit the number of available samples, whilst the human genome comprises millions of SNPs

(Manthena *et al.*, 2022). Results in this thesis demonstrated that prediction accuracy reduced as the number of variants increased, reinforcing the issue of dimensionality burden for ML. However, increasing the number of variants for PRS does not result in an increase for LR as only a singular score variable is used. Therefore, PRS related methods do not suffer from the same dimensionality related issues, with this pitfall overriding any ability of ML to assess non-linear patterns.

Results of this thesis suggest that ML is yet to improve upon PRS based methods for AD prediction. It is hypothesised that this improvement might not occur until sample sizes for GWAS increase. Therefore, the recommendation based on results is to continue with use of both PRS and ML related methodologies in the near future. However, as available cohort sizes increase and advances in computational technologies increase, research into ML techniques should continue. The use of increased amounts of features in this thesis prevented the use of SVMs despite adequate performance in Chapter 4. As barriers to computation decrease, further investigation into the use of these algorithms should be explored. Increased cohort sizes could also aid the use of neural networks, which have been shown to be powerful prediction algorithms in some circumstances. Results also suggest that disease prediction also benefits from datasets generated using prior biological information, therefore efforts should be made to continue the development of these.

8.9 Supplementary Information

The code used to develop all analyses in this thesis has not been published. The dataset used in this thesis is also not publicly available.

9 Supplementary Tables

Supplementary Tables for Chapter 4:

Supplementary Table 1. Results of comparison between PRS-LR and ML, in which genotypes and PRS were adjusted by three PCs, age and sex.

SNP Type[a]	PRS[b]	Random Forest[c]	PRS2[d]	Gradient Boosting[e]	PRS3[f]	Naïve Bayes[g]	PRS4[h]	SVM Linear[i]	PRS5[j]	SVM RBF[k]
P-values (23-SNPS)	56.90	89.30	54.10	90.50	56.90	55.10	60.30	59.70	60.50	58.70
R-squared (23-SNPS)	55.90	90.20	53.80	90.80	55.60	55.20	59.80	59.50	59.30	59.10
No-apoe-pvalues(21-SNPS)	54.13	89.40	54.80	69.80	54.00	51.90	57.50	56.80	55.90	56.50
No -apoe-rsquared (21 -SNPS)	54.00	89.90	54.70	71.20	53.90	53.20	56.80	55.90	57.10	55.20
Increased SNP's (422-SNPS)	57.50	90.70	57.60	92.10	57.70	52.80	61.20	60.40	60.50	61.70

a – Type of SNP data set explained in method section; b – First PRS % score compared to machine learning method. c – Random forest algorithm % compared to column b; d – Second PRS % score compared to machine learning method; e – Gradient Boosting algorithm % compared to column d; f - Third PRS % score compared to machine learning method; g – Naïve Bayes algorithm % compared to column f; h – Fourth PRS % score compared to machine learning method; i – SVM-Linear algorithm % compared to column h; j - Fifth PRS % score compared to machine learning method; k – SVM-RBF algorithm % compared to column j.

Supplementary Table 2. Comparison between PRS-LR and ML, in which both PRS and genotypes were adjusted by PCs, with age/sex added as separate variables.

SNP Type[a]	PRS[b]	Random Forest[c]	PRS2[d]	Gradient Boosting[e]	PRS3[f]	Naïve Bayes[g]	PRS4[h]	SVM Linear[i]	PRS5[j]	SVM RBF[k]
P-values (23-SNPS)	90.80	91.20	90.90	91.40	90.90	90.90	90.00	90.30	90.70	90.50
R-squared (23-SNPS)	90.80	91.60	90.70	91.70	90.70	90.50	90.60	90.60	90.20	90.40
No-apoe-pvalues(21-SNPS)	90.70	91.10	90.60	91.00	90.60	90.60	90.80	89.80	89.90	90.10
No -apoe-rsquared (21 -SNPS)	90.40	91.30	90.40	91.30	90.30	89.80	89.90	90.20	90.10	90.60
Increased SNP's (422-SNPS)	91.30	91.20	91.30	91.70	91.30	89.50	91.10	91.20	91.50	91.00

a – Type of SNP data set explained in method section; b – First PRS % score compared to machine learning method. c – Random forest algorithm % compared to column b; d – Second PRS % score compared to machine learning method; e – Gradient Boosting algorithm % compared to column d; f - Third PRS % score compared to machine learning method; g – Naïve Bayes algorithm % compared to column f; h – Fourth PRS % score compared to machine learning method; i – SVM-Linear algorithm % compared to column h; j - Fifth PRS % score compared to machine learning method; k – SVM-RBF algorithm % compared to column j.

Supplementary Table 3. Results of PRS-LR versus all ML algorithms with the 1958 birth cohort removed from GERAD, genotypes and PRS have been PC adjusted.

SNP Type[a]	PRS[b] %	Random Forest[c] %	Gradient Boosting[d] %	Naïve Bayes[e] %	SVM% Linear[f]	SVM RBF[g]
P-values (23-SNPS)	60.00	65.00	65.50	56.70	58.00	56.90
R-squared (23-SNPS)	59.80	67.30	67.40	58.10	58.10	57.70
No-apoe-pvalues(21-SNPS)	57.00	64.40	63.40	54.10	54.90	54.50
No -apoe-rsquared (21 -SNPS)	58.30	67.00	65.40	56.00	56.50	54.20
Increased SNP's (422-SNPS)	61.50	66.00	66.60	55.60	59.10	59.50

a- Type of SNP data set explained in method section; b – PRS % score compared to machine learning methods; c - Random Forest algorithm % compared to column b; d - Gradient Boosting algorithm % compared to column b; e - Gradient Boosting algorithm % compared to column b; d - Gradient Boosting algorithm % compared to column c; e NB algorithm % compared to column b; f SVM- Linear algorithm % compared to column b; g SVM-RBF algorithm compared to column b,

Supplementary Table 4. Results of PRS-LR versus all ML algorithms with the 1958 birth cohort removed and CV used, genotypes and PRS have been PC adjusted only.

SNP Type[a]	PRS[b] %	Random Forest[c] %	Gradient Boosting[d] %	Naïve Bayes[e] %	SVM % Linear[f]	SVM% RBF[g]
P-values (23-SNPS)	60.10	57.80	55.70	57.90	57.50	56.70
R-squared (23-SNPS)	60.60	57.80	55.90	58.10	57.30	57.30
No-apoe-pvalues(21-SNPS)	57.40	54.30	53.00	55.30	54.10	55.40
No -apoe-rsquared (21 -SNPS)	57.70	54.60	53.00	55.90	53.80	54.00
Increased SNP’s (422-SNPS)	61.60	59.60	58.40	55.80	59.00	59.60

a- Type of SNP data set explained in method section; b – PRS % score compared to machine learning methods; c - Random Forest algorithm % compared to column b; d - Gradient Boosting algorithm % compared to column b; e - Gradient Boosting algorithm % compared to column b; d - Gradient Boosting algorithm % compared to column c; e NB algorithm % compared to column b; f SVM- Linear algorithm % compared to column b; g SVM-RBF algorithm compared to column b,

Supplementary Table 5. Results of PRS-LR versus all ML algorithms with the 1958 birth cohort removed, with CV used and genotypes and PRS adjusted by PCs, age and sex.

SNP Type[a]	PRS[b] %	Random Forest[c] %	Gradient Boosting[d] %	Naïve Bayes[e] %	SVM % Linear[f]	SVM % RBF[g]
P-values (23-SNPS)	60.70	58.40	57.10	58.30	57.80	57.20
R-squared (23-SNPS)	59.10	56.60	56.00	58.20	56.00	57.00
No-apoe-pvalues(21 -SNPS)	59.30	54.50	52.00	55.60	54.20	55.10
No -apoe-rsquared (21 - SNPS)	57.70	54.80	53.80	55.80	53.50	53.80
Increased SNP's (422-SNPS)	64.60	59.40	56.10	56.20	59.40	60.20

a- Type of SNP data set explained in method section; b – PRS % score compared to machine learning methods; c - Random Forest algorithm % compared to column b; d - Gradient Boosting algorithm % compared to column b; e - Gradient Boosting algorithm % compared to column b; d - Gradient Boosting algorithm % compared to column c; e NB algorithm % compared to column b; f SVM- Linear algorithm % compared to column b; g SVM-RBF algorithm compared to column b,

Supplementary Table 6. Results of PRS-LR versus all ML algorithms with the 1958 birth cohort removed and balanced dataset, CV used, and genotypes and PRS have been PC adjusted.

SNP Type[a]	PRS[b] %	Random Forest[c] %	Gradient Boosting[d] %	Naïve Bayes[e] %	SVM[f] % Linear	SVM RBF
P-values (23-SNPS)	62.50	58.60	55.80	58.80	59.30	58.70
R-squared (23-SNPS)	60.40	56.10	55.80	57.90	57.60	57.30
No-apoe-pvalues(21 -SNPS)	59.70	55.80	53.20	56.10	55.80	54.30
No -apoe-rsquared (21 - SNPS)	61.80	57.80	55.70	59.30	57.90	56.80
Increased SNP's (422-SNPS)	59.80	57.30	55.30	53.80	58.80	58.80

a- Type of SNP data set explained in method section; b – PRS % score compared to machine learning methods; c - Random Forest algorithm % compared to column b; d - Gradient Boosting algorithm % compared to column b; e - Gradient Boosting algorithm % compared to column b; d - Gradient Boosting algorithm % compared to column c; e NB algorithm % compared to column b; f SVM- Linear algorithm % compared to column b; g SVM-RBF algorithm compared to column b,

Supplementary Table 7. Comparison of classifier performance using pairwise t-tests for analyses with the 1958 birth cohort removed, the use of CV and both PRS/genotypes adjusted by PCs only.

Classifier Comparison[a]	No-apoe-pvalues[b] (21-SNPS)	No -apoe-rsquared[c] (21 - SNPS)	P-values[d] (23-SNPS)	R-squared[e] (N SNPs = 23)	Increased SNP's[f] (422- SNPS)
PRS vs RF	Statistic = 8.31 p-value = 0.031	Statistic = 1.65 p-value = 0.284	Statistic = 2.18 p-value = 0.198	Statistic = 2.35 p-value = 0.173	Statistic = 0.625 p-value = 0.663
PRS vs GB	Statistic = 5.793 p-value = 0.031	Statistic = 3.00 p-value = 0.120	Statistic = 5.69 p-value = 0.044	Statistic = 3.86 p-value = 0.066	Statistic = 2.03 p-value = 0.228
PRS vs NB	Statistic = 5.15 p-value = 0.125	Statistic = 2.56 p-value = 0.162	Statistic = 3.84 p-value = 0.066	Statistic = 3.93 p-value = 0.066	Statistic = 5.19 p-value = 0.050
PRS vs SVM-Lin	Statistic = 4.74 p-value = 0.062	Statistic = 2.56 p-value = 0.162	Statistic = 2.48 p-value = 0.165	Statistic = 8.06 p-value = 0.031	Statistic = 7.56 p-value = 0.031
PRS vs SVM-RBF	Statistic = 2.38 p-value = 0.173	Statistic = 4.13 p-value = 0.066	Statistic = 3.72 p-value = 0.070	Statistic = 8.34 p-value = 0.031	Statistic = 4.31 p-value = 0.070
RF vs GB	Statistic = -0.794 p-value = 0.589	Statistic = 1.60 p-value = 0.295	Statistic = 3.52 p-value = 0.080	Statistic = 1.15 p-value = 0.443	Statistic = 4.423 p-value = 0.066
RF vs NB	Statistic = -2.53 p-value = 0.162	Statistic = -0.445 p-value = 0.760	Statistic = -0.470 p-value = 0.754	Statistic = 1.15 p-value = 0.443	Statistic = 1.84 p-value = 0.261
RF vs SVM-Lin	Statistic = -0.152 p-value = 0.923	Statistic = -1.242 p-value = 0.415	Statistic = -1.21 p-value = 0.423	Statistic = -1.67 p-value = 0.284	Statistic = 0.719 p-value = 0.620
RF vs SVM-RBF	Statistic = -1.74 p-value = 0.272	Statistic = 0.914 p-value = 0.533	Statistic = 0.637 p-value = 0.663	Statistic = -0.032 p-value = 0.989	Statistic = 0.377 p-value = 0.777
GB vs NB	Statistic = -1.10 p-value = 0.459	Statistic = -1.66 p-value = 0	Statistic = -2.21 p-value = 0.092	Statistic = -1.95 p-value = 0.243	Statistic = 1.09 p-value = 0.337
GB vs SVM-Lin	Statistic = 0.432 p-value = 0.770	Statistic = -2.60 p-value = 0.162	Statistic = -5.80 p-value = 0.044	Statistic = -3.00 p-value = 0.120	Statistic = -0.404 p-value = 0.768
GB vs SVM-RBF	Statistic = -0.844 p-value = 0.567	Statistic = -0.103 p-value = 0.948	Statistic = -1.81 p-value = 0.263	Statistic = -0.953 p-value = 0.519	Statistic = -0.962 p-value = 0.519
NB vs SVM-Lin	Statistic = -1.85 p-value = 0.261	Statistic = 1.51 p-value = 0.319	Statistic = 1.79 p-value = 0.147	Statistic = 2.38 p-value = 0.173	Statistic = 3.95 p-value = 0.066
NB vs SVM-RBF	Statistic = 0.009 p-value = 0.993	Statistic = -1.32 p-value = 0.386	Statistic = -1.33 p-value = 0.386	Statistic = -0.582 p-value = 0.682	Statistic = 4.21 p-value = 0.066
SVM-RBF vs SVM-Lin	Statistic = 0.763 p-value = 0.600	Statistic = -4.00 p-value = 0.066	Statistic = -4.53 p-value = 0.066	Statistic = -5.69 p-value = 0.044	Statistic = 2.67 p-value = 0.161

a- The comparison of classifiers; b – SNPs chosen by the p-value method and *APOE* SNPs removed; c - SNPs chosen by the r-squared method and *APOE* SNPs removed; d - SNPs chosen by the p-value method; e - SNPs chosen by the r-squared method; ; f – SNPs chosen by the clumping method

Supplementary Table 8. Comparison of classifier performance using pairwise t-tests for analyses with the 1958 birth cohort removed, the use of CV and both PRS/genotypes adjusted by PCs, age and sex.

Classifier Comparison[a]	No-apoe-pvalues(21-SNPS)[b]	No -apoe-rsquared (21 - SNPS)[c]	P-values (23-SNPS)[d]	R-squared (23-SNPS)[e]	Increased SNP's (422- SNPS)[f]
PRS vs RF	Statistic = 4.57 p-value = 0.033	Statistic = 3.61 p-value = 0.054	Statistic = 2.97 p-value = 0.082	Statistic = 2.33 p-value = 0.139	Statistic = 7.57 p-value = 0.011
PRS vs GB	Statistic = 7.27 p-value = 0.011	Statistic = 4.067 p-value = 0.004	Statistic = 6.90 p-value = 0.012	Statistic = 4.36 p-value = 0.036	Statistic = 9.28 p-value = 0.006
PRS vs NB	Statistic = 9.97 p-value = 0.006	Statistic = 1.88 p-value = 0.195	Statistic = 4.72 p-value = 0.031	Statistic = 1.11 p-value = 0.425	Statistic = 9.52 p-value = 0.006
PRS vs SVM-Lin	Statistic = 11.6 p-value = 0.005	Statistic = 0.990 p-value = 0.473	Statistic = 6.86 p-value = 0011	Statistic = -0.629 p-value = 0.650	Statistic = 12.2 p-value = 0.005
PRS vs SVM-RBF	Statistic = 7.17 p-value = 0.011	Statistic = 5.79 p-value = 0.019	Statistic = 5.33 p-value = 0.023	Statistic = 4.10 p-value = 0.038	Statistic = 14.2 p-value = 0.005
RF vs GB	Statistic = 1.81 p-value = 0.145	Statistic = 0.096 p-value = 0.954	Statistic = 4.25 p-value = 0.036	Statistic = 0.794 p-value = 0.561	Statistic = 3.15 p-value = 0.072
RF vs NB	Statistic = -0.585 p-value = 0.671	Statistic = -3.29 p-value = 0.067	Statistic = 0.769 p-value = 0.568	Statistic = -1.96 p-value = 0.181	Statistic = 3.12 p-value = 0.072
RF vs SVM-Lin	Statistic = -2.03 p-value = 0.171	Statistic = -5.03 p-value = 0.026	Statistic = 1.86 p-value = 0.197	Statistic = -4.22 p-value = 0.036	Statistic = 0.223 p-value = 0.869
RF vs SVM-RBF	Statistic = -0.542 p-value = 0.690	Statistic = -0.970 p-value = 0.476	Statistic = 2.27 p-value = 0.140	Statistic = -0.512 p-value = 0.700	Statistic = -1.19 p-value = 0.394
GB vs NB	Statistic = -4.29 p-value = 0.036	Statistic = -8.37 p-value = 0.008	Statistic = -0.891 p-value = 0.512	Statistic = -2.51 p-value = 0.118	Statistic = -0.023 p-value = 0.983
GB vs SVM-Lin	Statistic = -5.78 p-value = 0.019	Statistic = 12.4 p-value = 0.005	Statistic = -0.307 p-value = 0.818	Statistic = -4.48 p-value = 0.034	Statistic = -2.91 p-value = 0.082
GB vs SVM-RBF	Statistic = -3.02 p-value = 0.077	Statistic = -1.67 p-value = 0.236	Statistic = -0.066 p-value = 0.964	Statistic = -1.51 p-value = 0.279	Statistic = -3.46 p-value = 0.061
NB vs SVM-Lin	Statistic = 3.17 p-value = 0.034	Statistic = 2.10 p-value = 0.162	Statistic = -1.23 p-value = 0.384	Statistic = 2.31 p-value = 0.139	Statistic = 3.81 p-value = 0.019
NB vs SVM-RBF	Statistic = -0.338 p-value = 0.806	Statistic = -2.75 p-value = 0.094	Statistic = -0.994 p-value = 0.473	Statistic = -2.15 p-value = 0.155	Statistic = 5.05 p-value = 0.026
SVM-RBF vs SVM-Lin	Statistic = -2.30 p-value = 0.083	Statistic = -6.77 p-value = 0.012	Statistic = -0.415 p-value = 0.760	Statistic = -26.8 p-value = 1.6e-05	Statistic = 3.32 p-value = 0.067

a- The comparison of classifiers; b – SNPs chosen by the p-value method and *APOE* SNPs removed; c - SNPs chosen by the r-squared method and *APOE* SNPs removed; d - SNPs chosen by the p-value method; e - SNPs chosen by the r-squared method; ; f – SNPs chosen by the clumping method

Supplementary Table 9. Comparison of classifier performance using pairwise t-tests for analyses with the 1958 birth cohort removed, the use of a balanced dataset by age and sex, with PRS/genotypes adjusted by PCs only.

Classifier Comparison[a]	No-apoe-pvalues(21-SNPS)[b]	No -apoe-rsquared (21 - SNPS)[c]	P-values (23-SNPS)[d]	R-squared (23-SNPS)[e]	Increased SNP's (422- SNPS)[f]
PRS vs RF	Statistic = 1.64 p-value = 0.307	Statistic = 2.66 p-value = 0.137	Statistic = 7.44 p-value = 0.043	Statistic = 3.25 p-value = 0.103	Statistic = 4.78 p-value = 0.067
PRS vs GB	Statistic = 2.084 p-value = 0.208	Statistic = 4.53 p-value = 0.011	Statistic = 4.97 p-value = 0.067	Statistic = 6.20 p-value = 0.043	Statistic = 3.35 p-value = 0.103
PRS vs NB	Statistic = 2.35 p-value = 0.080	Statistic = 2.67 p-value = 0.137	Statistic = 6.22 p-value = 0.043	Statistic = 4.61 p-value = 0.066	Statistic = 3.72 p-value = 0.103
PRS vs SVM-Lin	Statistic = 1.89 p-value = 0.243	Statistic = 4.71 p-value = 0.066	Statistic = 3.15 p-value = 0.107	Statistic = 7.02 p-value = 0.043	Statistic = 0.890 p-value = 0.539
PRS vs SVM-RBF	Statistic = 2.19 p-value = 0.196	Statistic = 4.011 p-value = 0.092	Statistic = 3.25 p-value = 0.103	Statistic = 8.09 p-value = 0.043	Statistic = 0.990 p-value = 0.498
RF vs GB	Statistic = 0.798 p-value = 0.578	Statistic = 2.83 p-value = 0.132	Statistic = 1.54 p-value = 0.311	Statistic = 3.53 p-value = 0.103	Statistic = 0.100 p-value = 0.957
RF vs NB	Statistic = -0.071 p-value = 0.957	Statistic = -1.54 p-value = 0.311	Statistic = -6.53 p-value = 0.043	Statistic = -0.621 p-value = 0.657	Statistic = 1.94 p-value = 0.238
RF vs SVM-Lin	Statistic = 0.079 p-value = 0.957	Statistic = -0.994 p-value = 0.498	Statistic = -5.00 p-value = 0.066	Statistic = 0.368 p-value = 0.807	Statistic = -1.38 p-value = 0.355
RF vs SVM-RBF	Statistic = 1.62 p-value = 0.307	Statistic = -0.538 p-value = 0.703	Statistic = -2.46 p-value = 0.158	Statistic = 0.09 p-value = 0.957	Statistic = -1.35 p-value = 0.358
GB vs NB	Statistic = -0.954 p-value = 0.510	Statistic = -2.78 p-value = 0.132	Statistic = -3.42 p-value = 0.103	Statistic = -3.65 p-value = 0.103	Statistic = 2.66 p-value = 0.137
GB vs SVM-Lin	Statistic = -0.620 p-value = 0.657	Statistic = -2.20 p-value = 0.196	Statistic = -2.325 p-value = 0.049	Statistic = -1.56 p-value = 0.311	Statistic = -1.02 p-value = 0.498
GB vs SVM-RBF	Statistic = 1.88 p-value = 0.243	Statistic = -1.40 p-value = 0.350	Statistic = -3.63 p-value = 0.103	Statistic = -2.16 p-value = 0.197	Statistic = -1.14 p-value = 0.450
NB vs SVM-Lin	Statistic = -0.293 p-value = 0.852	Statistic = -3.28 p-value = 0.103	Statistic = 0.782 p-value = 0.578	Statistic = -1.54 p-value = 0.311	Statistic = 2.50 p-value = 0.157
NB vs SVM-RBF	Statistic = -1.70 p-value = 0.292	Statistic = -3.11 p-value = 0.107	Statistic = -0.056 p-value = 0.957	Statistic = -0.661 p-value = 0.649	Statistic = 2.94 p-value = 0.122
SVM-RBF vs SVM-Lin	Statistic = -1.52 p-value = 0.311	Statistic = -0.830 p-value = 0.566	Statistic = 0.782 p-value = 0.578	Statistic = 0.464 p-value = 0.746	Statistic = -0.268 p-value = 0.860

a- The comparison of classifiers; b – SNPs chosen by the p-value method and *APOE* SNPs removed; c - SNPs chosen by the r-squared method and *APOE* SNPs removed; d - SNPs chosen by the p-value method; e - SNPs chosen by the r-squared method; ; f – SNPs chosen by the clumping method

Supplementary Tables for Chapter 5:

Supplementary Table 10. Comparison between ML and PRS with no Feature Selection and no *APOE* Alleles.

P-value Threshold[a]	Random Forest[b]	Gradient Boosting[c]	PRS[d]	Mean Number of SNPs[e]
0.0001	56.7	57.7	54.5	20
0.001	55.3	55.9	51.3	187
0.01	53.5	53.5	57.9	1981
0.05	53.8	54.6	60.6	8909
0.1	53.9	55.6	60.5	16409
0.2	51.6	54.5	60.1	29210
0.3	52.1	56.0	59.8	39984
0.4	53.8	54.5	59.3	49268
0.5	55.8	54.5	59.0	57225

a- p-value thresholds used for analyses; b – AUC for RFs across five folds of CV; c – AUC for GB across five folds of CV; d – AUC for PRS across five folds of CV; e – Mean number of SNPs used across five folds of CV.

Supplementary table 11. Comparison between ML and PRS without Feature Selection

P-value Threshold[a]	Random Forest[b]	Gradient Boosting[c]	PRS[d]	Mean Number of SNPs[e]
0.0001	66.3	65.9	68.7	17
0.001	65.7	65.5	70.0	191
0.01	63.0	65.8	71.8	1976
0.05	59.1	60.2	72.6	8869
0.1	59.4	60.5	72.6	16403
0.2	58.3	57.4	72.6	29232
0.3	56.7	59.3	72.3	40012
0.4	58.1	58.8	72.1	49301
0.5	57.4	59.0	71.8	57260

a- p-value thresholds used for analyses; b – AUC for RFs across five folds of CV; c – AUC for GB across five folds of CV; d – AUC for PRS across five folds of CV; e – Mean number of SNPs used across five folds of CV.

Supplementary table 12. Comparison between ML and PRS with RF as Feature Selection

P-value Threshold[a]	Random Forest[b]	Gradient Boosting[c]	PRS[d]	Average Number of SNPs[e]	Mean Number of SNPs after FS[f]	Overlap in SNPs[g]
0.0001	66.0	65.0	69.5	17	18	N/A
0.001	65.8	66.2	70.0	191	1	1
0.01	65.2	66.3	71.8	1975	21	1
0.05	66.8	67.7	72.5	8920	3	1
0.1	66.9	65.7	72.4	16435	55	2
0.2	67.1	66.7	72.4	29192	20	2
0.3	66.4	65.1	72.4	39980	124	4
0.4	67.0	66.8	72.1	49245	96	3
0.5	67.7	65.1	71.8	57239	1715	20

a- p-value thresholds used for analyses; b – AUC for RFs across five folds of CV; c – AUC for GB across five folds of CV; d – AUC for PRS across five folds of CV; e – Mean number of SNPs used across five folds of CV; f –Mean number of SNPs across five folds of CV; - g – Mean number of shared SNPs between the current and previous p-value threshold over 5 folds of CV.

Supplementary table 13. Comparison between ML and PRS with ExtraTree as Feature Selection

P-value Threshold[a]	Random Forest[b]	Gradient Boosting[c]	PRS[d]	Average Number of SNPs[e]	Mean Number of SNPs after FS[f]	Overlap in SNPs[g]
0.0001	65.9	64.7	69.0	18	20	N/A
0.001	65.6	65.8	70.3	190	1	1
0.01	65.7	65.9	71.7	1976	1	1
0.05	66.6	65.7	72.7	8893	3	1
0.1	67.0	66.4	72.4	16411	4	1
0.2	66.6	60.6	72.5	29225	579	2
0.3	66.8	62.2	72.3	40006	496	8
0.4	66.2	61.5	72.1	49301	1266	10
0.5	67.7	62.5	72.0	57211	489	8

a- p-value thresholds used for analyses; b – AUC for RFs across five folds of CV; c – AUC for GB across five folds of CV; d – AUC for PRS across five folds of CV; e – Mean number of SNPs used across five folds of CV; f –Mean number of SNPs across five folds of CV; - g – Mean number of shared SNPs between the current and previous p-value threshold over 5 folds of CV.

Supplementary table 14. Comparison between ML and PRS with LASSO as Feature Selection

P-Value Threshold[a]	Random Forest[b]	Gradient Boosting[c]	PRS[d]	Average Number of SNPs[e]	Mean Number of SNPs after FS[f]	Overlap in SNPs[g]
0.0001	65.9	64.3	69.1	18	18	N/A
0.001	65.4	62.2	70.3	189	190	12
0.01	64.2	61.7	71.8	1965	1433	60
0.05	62.4	59.5	72.7	8890	2506	255
0.1	63.1	62.1	72.3	16404	2653	204
0.2	62.2	62.1	72.3	29223	2813	188
0.3	62.0	59.3	72.4	40019	2851	142
0.4	62.4	61.0	72.3	49304	2997	100
0.5	62.9	59.5	72.0	57236	2921	84

a- p-value thresholds used for analyses; b – AUC for RFs across five folds of CV; c – AUC for GB across five folds of CV; d – AUC for PRS across five folds of CV; e – Mean number of SNPs used across five folds of CV; f –Mean number of SNPs across five folds of CV; - g – Mean number of shared SNPs between the current and previous p-value threshold over 5 folds of CV.

Supplementary table 15. Comparison between ML and PRS with Elastic Net as Feature Selection

P-value Threshold[a]	Random Forest[b]	Gradient Boosting[c]	PRS[d]	Average Number of SNPs[e]	Mean Number of SNPs after FS[f]	Overlap of SNPs[g]
0.0001	66.4	64.6	69.3	17	19	N/A
0.001	65.6	62.7	69.8	187	193	8
0.01	62.4	60.9	71.6	1967	1796	86
0.05	62.7	60.1	72.3	8883	3558	343
0.1	62.0	60.5	72.5	16409	3670	337
0.2	62.5	60.8	72.3	29189	3702	244
0.3	59.0	60.4	72.2	40012	3941	265
0.4	57.8	59.5	72.1	49304	3800	274
0.5	59.0	59.7	72.1	57233	4036	301

a- p-value thresholds used for analyses; b – AUC for RFs across five folds of CV; c – AUC for GB across five folds of CV; d – AUC for PRS across five folds of CV; e – Mean number of SNPs used across five folds of CV; f –Mean number of SNPs across five folds of CV; - g – Mean number of shared SNPs between the current and previous p-value threshold over 5 folds of CV.

Supplementary Table 16. Comparison between ML and PRS with no Feature Selection, with SNPs not Clumped.

P-value Thresholds[a]	Random Forest[b]	Gradient Boosting[c]	PRS[d]	Mean Number of SNPs[e]
0.0001	65.2	64.8	69.3	24
0.001	65.0	61.8	69.7	269
0.01	59.8	60.4	71.1	3356
0.05	58.1	56.7	72.5	18562
0.1	56.6	58.7	72.4	38806
0.2	55.9	58.2	72.2	80289
0.3	55.2	57.9	69.4	122264
0.4	55.0	57.0	69.4	164508
0.5	55.1	57.5	69.4	207078

a- p-value thresholds used for analyses; b – AUC for RFs across five folds of CV; c – AUC for GB across five folds of CV; d – AUC for PRS across five folds of CV; e – Mean number of SNPs used across five folds of CV.

Supplementary Table 17. Comparison between ML and PRS with RF as Feature Selection, with SNPs not Clumped.

P-value Thresholds[a]	Random Forest[b]	Gradient Boosting[c]	PRS[d]	Average Number of SNPs[e]	Mean Number of SNPs after FS[f]	Overlap of SNPs[g]
0.0001	65.9	65.3	69.6	22	24	N/A
0.001	66.4	66.4	70.0	269	1	1
0.01	65.3	65.1	71.3	3372	1	1
0.05	65.6	66.5	72.3	18612	1	1
0.1	65.9	66.1	72.5	38856	1	1
0.2	66.6	65.3	72.2	80319	1562	10
0.3	67.5	66.3	69.6	122252	186	3
0.4	65.5	66.4	72.2	164803	215	5
0.5	68.1	66.4	69.4	207037	295	9

a- p-value thresholds used for analyses; b – AUC for RFs across five folds of CV; c – AUC for GB across five folds of CV; d – AUC for PRS across five folds of CV; e – Mean number of SNPs used across five folds of CV; f –Mean number of SNPs across five folds of CV; - g – Mean number of shared SNPs between the current and previous p-value threshold over 5 folds of CV.

Supplementary Table 18 Comparison between ML and PRS with ExtraTree as Feature Selection, with SNPs not Clumped.

P-value Thresholds[a]	Random Forest[b]	Gradient Boosting[c]	PRS[d]	Average Number of SNPs[e]	Mean Number of SNPs after FS[f]	Overlap of SNPs[g]
0.0001	65.6	64.0	69.3	22	21	N/A
0.001	66.0	66.3	69.7	272	1	1
0.01	65.4	64.8	71.0	3376	1	1
0.05	67.5	67.4	72.5	18538	2	1
0.1	66.3	63.0	72.5	38761	297	6
0.2	67.4	63.1	72.3	80226	54	1
0.3	67.2	64.4	69.7	122219	105	2
0.4	67.8	61.7	69.3	164639	84	2
0.5	68.2	64.4	69.4	206987	128	4

a- p-value thresholds used for analyses; b – AUC for RFs across five folds of CV; c – AUC for GB across five folds of CV; d – AUC for PRS across five folds of CV; e – Mean number of SNPs used across five folds of CV; f –Mean number of SNPs across five folds of CV; - g – Mean number of shared SNPs between the current and previous p-value threshold over 5 folds of CV.

Supplementary table 19. Comparison between ML and PRS with no Feature Selection, with a more Lenient r^2.

P-value Thresholds[a]	Random Forest[b]	Gradient Boosting[c]	PRS[d]	Mean Number of SNPs[e]
0.0001	65.5	63.5	69.0	19
0.001	65.7	61.0	70.0	208
0.01	62.6	62.0	71.7	2395
0.05	58.3	61.1	70.0	12376
0.1	58.6	58.7	70.2	24924
0.2	57.0	58.1	69.8	49715
0.3	55.2	58.9	69.5	73997
0.4	57.4	56.4	69.5	97693
0.5	56.3	57.2	69.2	120579

a- p-value thresholds used for analyses; b – AUC for RFs across five folds of CV; c – AUC for GB across five folds of CV; d – AUC for PRS across five folds of CV; e – Mean number of SNPs used across five folds of CV.

Supplementary table 20. Comparison between ML and PRS with RF as Feature Selection, with a Less Lenient r^2.

P-value Thresholds[a]	Random Forest[b]	Gradient Boosting[c]	PRS[d]	Average Number of SNPs[e]	Mean Number of SNPs after FS[f]	Overlap of SNPs[g]
0.0001	66.3	65.0	69.2	25	21	N/A
0.001	64.9	65.5	69.7	210	1	1
0.01	66.1	65.3	71.7	2386	1	1
0.05	66.1	65.7	70.1	12365	1	1
0.1	67.9	66.8	69.8	24916	26	1
0.2	67.6	66.1	69.8	49727	10	1
0.3	67.2	66.9	69.6	73964	66	2
0.4	66.7	66.0	69.5	97628	55	1
0.5	66.7	66.4	69.3	120541	38	1

a- p-value thresholds used for analyses; b – AUC for RFs across five folds of CV; c – AUC for GB across five folds of CV; d – AUC for PRS across five folds of CV; e – Mean number of SNPs used across five folds of CV; f –Mean number of SNPs across five folds of CV; - g – Mean number of shared SNPs between the current and previous p-value threshold over 5 folds of CV.

Supplementary Table 21. Comparison between ML and PRS with the ExtraTree algorithm as Feature Selection, with a less Lenient r^2.

P-value Thresholds[a]	Random Forest[b]	Gradient Boosting[c]	PRS[d]	Average Number of SNPs[e]	Mean Number of SNPs after FS[f]	Overlap of SNPs[g]
0.0001	66.3	66.2	69.2	23	20	N/A
0.001	65.0	66.5	69.9	188	1	1
0.01	65.7	66.6	71.7	1982	1	1
0.05	66.3	65.6	69.9	12378	2	1
0.1	65.8	67.2	70.0	24925	3	1
0.2	67.2	64.3	69.9	49680	163	8
0.3	68.2	63.0	69.6	73905	137	8
0.4	67.1	62.4	69.4	97639	123	12
0.5	68.1	63.3	69.5	120571	128	5

a- p-value thresholds used for analyses; b – AUC for RFs across five folds of CV; c – AUC for GB across five folds of CV; d – AUC for PRS across five folds of CV; e – Mean number of SNPs used across five folds of CV; f –Mean number of SNPs across five folds of CV; - g – Mean number of shared SNPs between the current and previous p-value threshold over 5 folds of CV.

Supplementary Table 22. Comparison between ML and PRS with Microglia related SNPs

P-value Thresholds[a]	Random Forest[b]	Gradient Boosting[c]	PRS[d]	Mean Number of SNPs[e]
0.0001	67.9	66.5	68.9	2
0.001	67.0	63.8	69.3	12
0.01	67.0	64.0	68.5	124
0.05	66.1	59.1	69..4	638
0.1	66.1	61.7	69.5	1298
0.2	65.2	59.2	69.2	2593
0.3	64.0	59.1	69.2	3786
0.4	63.1	60.9	69.6	4983
0.5	63.3	60.2	69.4	6096

a- p-value thresholds used for analyses; b – AUC for RFs across five folds of CV; c – AUC for GB across five folds of CV; d – AUC for PRS across five folds of CV; e – Mean number of SNPs used across five folds of CV.

Supplementary Table 23. Comparison between ML and PRS with Synapse related SNPs

P-value Thresholds[a]	Random Forest[b]	Gradient Boosting[c]	PRS[d]	Mean Number of SNPs[e]
0.0001	67.6	65.7	69.0	3
0.001	66.4	64.5	68.8	17
0.01	66.9	61.7	69.7	183
0.05	66.7	59.8	69.5	839
0.1	65.9	61.6	69.3	1557
0.2	65.3	60.5	69.4	2753
0.3	61.8	66.9	69.6	3768
0.4	63.2	61.8	69.2	4682
0.5	62.8	60.2	69.0	5452

a- p-value thresholds used for analyses; b – AUC for RFs across five folds of CV; c – AUC for GB across five folds of CV; d – AUC for PRS across five folds of CV; e – Mean number of SNPs used across five folds of CV.

Supplementary Table 24: Paired T-test Statistics without the use of *Apoe* Alleles and Feature Selection.

P-value Threshold[a]	Comparison Statistics[b]
0.0001	PRS had superior performance to GB. PRS vs GB = (-4.48, 0.033)
0.01	PRS had superior performance to GB. PRS vs GB = (6.39, 0.011) PRS vs RFs = (4.11, 0.034)
0.1	PRS had superior performance to both GB and RFs PRS vs GB = (10.3, 0.004) PRS vs RFs = (7.75, 0.007)
0.3	PRS had superior performance to RFs. PRS vs RFs = (9.87, 0.001) PRS vs GB = (3.35, 0.050)
0.5	PRS had superior performance to RFs. PRS vs RFs = (3.44, 0.050) PRS vs GB = (4.04, 0.034)

a- p-value thresholds used for analyses; b – Reported statistics for pairwise t-tests between classifiers, statistics were only reported if significant (<0.05)

Supplementary Table 25: Paired T-test Statistics with the use of *Apoe* Alleles and Feature Selection on LD Pruned SNPs

P-value Threshold[a]	No Feature Selection[b]	Random Forest used for Feature Selection[c]	ExtraTrees used for Feature Selection[d]
0.0001	PRS had superior performance to GB. PRS vs GB = (5.12, 0.011)	PRS had superior performance to both GB and RFs PRS vs GB = (3.40, 0.030) PRS vs RFs = (3.63, 0.022)	PRS had superior performance to both GB and RFs PRS vs GB = (3.26, 0.042) PRS vs RFs = (5.30, 0.011)
0.01	PRS had superior performance to both GB and RFs. Whilst GB outperformed RFs. PRS vs RFs = (9.12,0.002) PRS vs GB = (7.75,0.002) RFs vs GB = (-4.07, 0.023)	PRS had superior performance to both GB and RFs PRS vs GB = (4.76, 0.032) PRS vs RFs = (3.93, 0.012)	PRS had superior performance to both GB and RFs PRS vs GB = (4.03, 0.016) PRS vs RFs = (6.37, 0.007)
0.1	PRS had superior performance to both GB and RFs PRS vs RFs = (9.64, 0.001) PRS vs GB = (9.41, 0.002)	PRS had superior performance to both GB and RFs PRS vs GB = (4.64, 0.010) PRS vs RFs = (6.59, 0.003)	PRS had superior performance to both GB and RFs PRS vs GB = (8.31, 0.003) PRS vs RFs = (10.1, 0.003)
0.3	PRS had superior performance to both GB and RFs PRS vs RFs = (19.5,0.001) PRS vs GB = (9.74, 0.002)	PRS had superior performance to both GB and RFs PRS vs GB = (14.6, 0.008) PRS vs RFs = (7.18, 0.029)	PRS had superior performance to both GB and RFs. PRS vs GB = (6.83,0.036) PRS vs RFs = (10.1,0.008) RFs vs GB = (3.01, 0.050)
0.5	PRS had superior performance to both GB and RFs PRS vs RFs = (24.4, 0.001) PRS vs GB = (8.70, 0.002	PRS had superior performance to both GB and RFs PRS vs GB = (10.9, 0.003) PRS vs RFs = (6.78, 0.008)	PRS had superior performance to both GB and RFs. Whilst GB outperformed RFs. PRS vs GB = (8.77,0.003) PRS vs RFs = (9.80, 0.003) RFs vs GB = (4.27, 0.021)

a- p-value thresholds used for analyses; b – Reported statistics for pairwise t-tests between classifiers, statistics were only reported if significant (<0.05); c – Statistics for analyses using RFs for feature selection; d - Statistics for analyses using ExtraTrees for feature selection.

Supplementary Table 26: Paired T-test Statistics with the use of *Apoe* Alleles and Feature Selection Without the use of Clumping

P-value Threshold[a]	No Feature Selection[b]	Random Forest used for Feature Selection[c]	ExtraTrees used for Feature Selection[d]
0.0001	PRS had superior performance to both GB and RFs PRS vs RFs = (4.48, 0.018) PRS vs GB = (4.25, 0.012)	PRS had superior performance to both GB and RFs PRS vs RFs = (5.76, 0.011) PRS vs GB = (6.11, 0.011)	There were no significant results in this pathway. PRS vs RFs = (7.15, 0.007) PRS vs GB = (3.91, 0.003)
0.01	PRS had superior performance to both GB and RFs PRS vs RFs = (9.20, 0.002) PRS vs GB = (7.91, 0.003)	PRS had superior performance to both GB and RFs PRS vs RFs = (9.14, 0.004) PRS vs GB = (8.30, 0.004)	PRS had superior performance to both GB and RFs PRS vs RFs = (6.78, 0.007) PRS vs GB = (7.05, 0.007)
0.1	PRS had superior performance to both GB and RFs PRS vs RFs = (12.6, 0.001) PRS vs GB = (16.3, 0.001)	PRS had superior performance to both GB and RFs PRS vs RFs = (4.02, 0.016) PRS vs GB = (3.30, 0.047)	PRS had superior performance to both GB and RFs, whilst RFs also outperformed GB. PRS vs RFs = (11.2,0.003) PRS vs GB = (3.30, 0.003) RFs vs GB = (5.31, 0.011)
0.3	PRS had superior performance to both GB and RFs PRS vs RFs = (28.7,0.0001) PRS vs GB = (8.21, 0.003)	PRS had superior performance to both GB and RFs PRS vs RFs = (4.02, 0.034) PRS vs GB = (9.13, 0.001)	PRS had superior performance to both GB and RFs, whilst RFs also outperformed GB. PRS vs RFs = (5.34, 0.011) PRS vs GB = (5.97, 0.010) RFs vs GB = (3.00, 0.050)
0.5	PRS had superior performance to both GB and RFs PRS vs RFs = (26.2, 0.0002) PRS vs GB = (11.2, 0.006)	PRS had superior performance to both GB and RFs PRS vs RFs = (3.344, 0.050) PRS vs GB = (3.25, 0.050)	PRS had superior performance to both GB and RFs PRS vs RFs = (3.33, 0.043) PRS vs GB = (3.19, 0.050)

a- p-value thresholds used for analyses; b – Reported statistics for pairwise t-tests between classifiers, statistics were only reported if significant (<0.05); c – Statistics for analyses using RFs for feature selection; d - Statistics for analyses using ExtraTrees for feature selection.

Supplementary Table 27: Paired T-test Statistics with the use of *Apoe* Alleles and Feature Selection and with the use of a More Lenient Value of r^2

P-value Threshold[a]	No Feature Selection[b]	Random Forest used for Feature Selection[c]	ExtraTrees used for Feature Selection[d]
0.0001	PRS had superior performance to both GB and RFs PRS vs GB = (7.15, 0.002) PRS vs RFs = (3.91, 0.030)	There were no significant results in this pathway.	There were no significant results in this pathway.
0.01	PRS had superior performance to both GB and RFs PRS vs GB = (6.78, 0.007) PRS vs RFs = (7.05, 0.007)	PRS had superior performance to GB. PRS vs GB = (11.8,0.004)	PRS had superior performance to both GB and RFs PRS vs GB = (11.8, 0.004) PRS vs RFs = (5.85,0.021)
0.1	PRS had superior performance to both GB and RFs. RFs also outperformed GB. PRS vs GB = (12.4, 0.002) PRS vs RFs = (11.2, 0.002 RFs vs GB = (5.30, 0.006)	There were no significant results in this pathway.	There were no significant results in this pathway.
0.3	PRS had superior performance to both GB and RFs. RFs also outperformed GB. PRS vs GB = (5.97, 0.010) PRS vs RFs = (5.34, 0.011) RFs vs GB = (3.00, 0.050)	PRS had superior performance to RFs. PRS vs RFs = (7.21,0.030)	PRS had superior performance to RFs. PRS vs RF = (7.22,0.015)
0.5	PRS had superior performance to both GB and RFs PRS vs GB = (3.33, 0.043) PRS vs RFs = (3.19, 0.050)	There were no significant results in this pathway.	There were no significant results in this pathway.

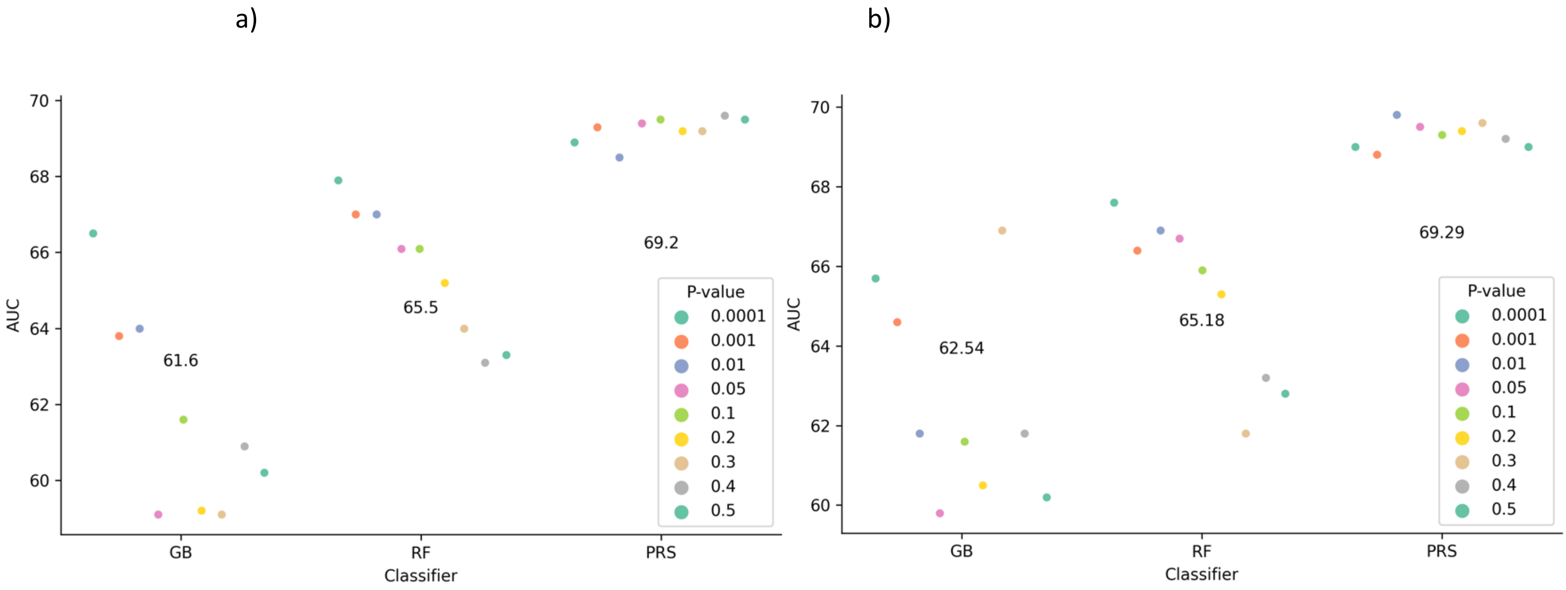
a- p-value thresholds used for analyses; b – Reported statistics for pairwise t-tests between classifiers, statistics were only reported if significant (<0.05); c – Statistics for analyses using RFs for feature selection; d - Statistics for analyses using ExtraTrees for feature selection.

Supplementary Table 28: Paired T-test Statistics with the use of *Apoe* Alleles and SNPs Selected using Biological Information

P-value Threshold[a]	Microglia[b]	Synapse[c]
0.0001	There were no significant results in this pathway.	PRS had superior performance to GB. RFs also outperformed GB. PRS vs GB = (5.40, 0.010) RFs vs GB = (9.24, 0.006)
0.01	PRS had superior performance to GB. Whilst RFs outperformed GB. PRS vs GB = (9.09, 0.003) RFs vs GB = (11.8, 0.002)	PRS had superior performance to both GB and RFs. RFs also outperformed GB. PRS vs GB = (5.27, 0.010) PRS vs RFs = (4.63, 0.015) RFs vs GB = (3.19, 0.040)
0.1	PRS had superior performance to GB. Whilst RFs outperformed GB. PRS vs GB = (12.8, 0.002) RFs vs GB = (4.28, 0.024)	PRS had superior performance to both GB and RFs. PRS vs GB = (5.45, 0.010) PRS vs RFs = (18.3, 0.001)
0.3	PRS had superior performance to both GB and RFs. RFs also outperformed GB. PRS vs GB = (9.99, 0.003) PRS vs RFs = (8.96, 0.003) RFs vs GB = (4.00, 0.016)	PRS had superior performance to both GB and RFs. RFs also outperformed GB. PRS vs GB = (5.86, 0.010) PRS vs RFs = (5.47, 0.010) RFs vs GB = (-3.92, 0.022)
0.5	PRS had superior performance to both GB and RFs PRS vs GB = (6.29, 0.003) PRS vs RFs = (7.99, 0.001)	PRS had superior performance to both GB and RFs. RFs also outperformed GB. PRS vs GB = (7.58, 0.008) PRS vs RFs = (6.76, 0.009) RFs vs GB = (4.16, 0.019)

a-p-value thresholds used for analyses; b – Reported statistics for pairwise t-tests between classifiers, statistics were only reported if significant (<0.05); c – Statistics for analyses using RFs for feature selection; d - Statistics for analyses using ExtraTrees for feature selection.

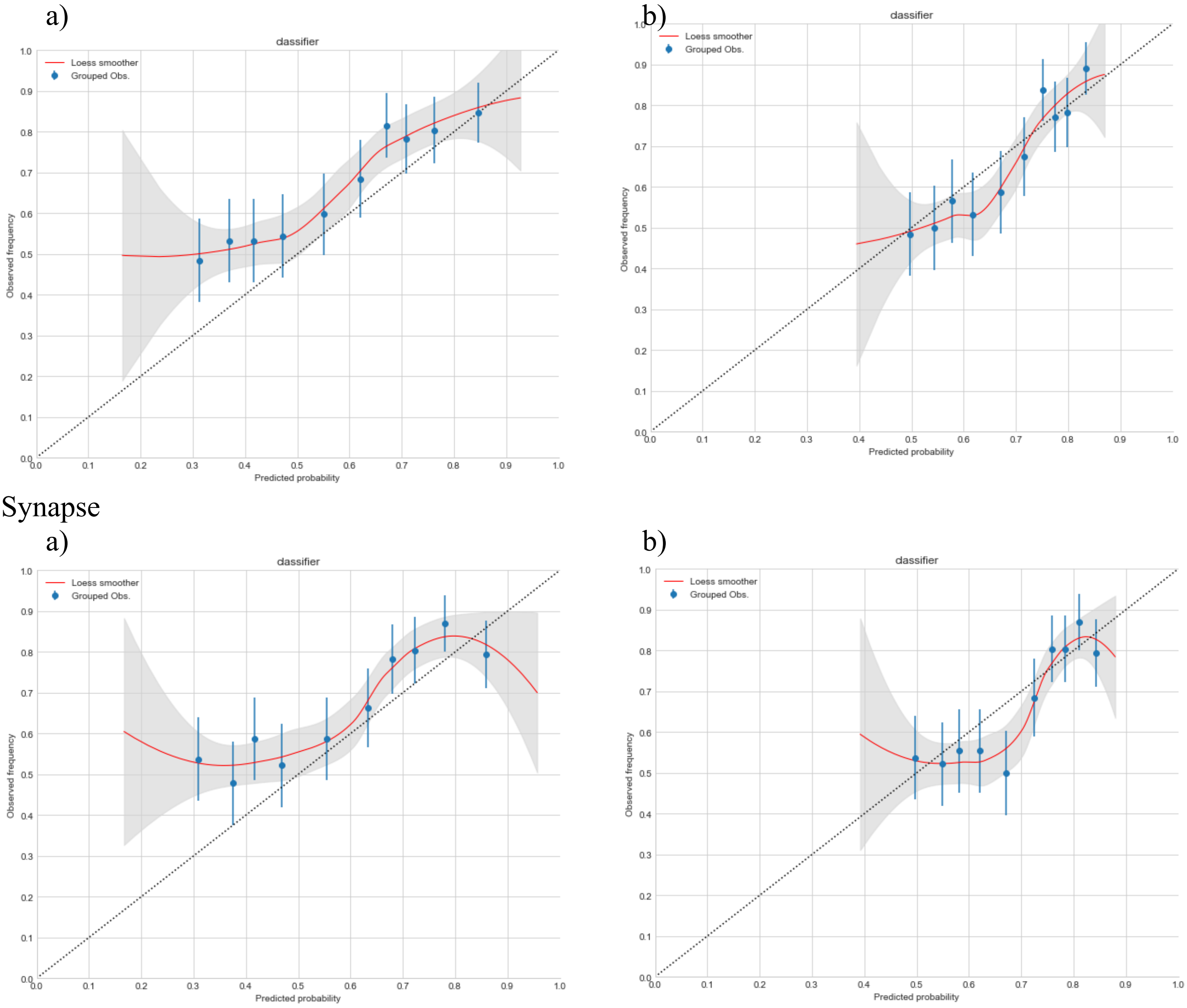
Supplementary Figure 1: The Comparison of PRS vs Chosen Classifiers (RF, GB) for LD Pruned ($r^2 = 0.1$) SNPs using Microglia and Synapse related SNPs.



Y-axis represents AUC in %; with classifiers placed on the X axis. Each dot represents the score for the prediction algorithm for all p-value thresholds. The numbers placed centrally are the mean score across p-value threshold values; GB Gradient Boosting; RF Random Forest; PRS-LR Polygenic Risk Scores Logistic Regression; AUC Area Under the Curve.

Supplementary Figure 2: The Comparison of non-Calibrated vs Calibrated Prediction Probabilities for RFs when using Biological Information

Microglia



a)

b)

Synapse

a)

b)

These figures represent pre a) and post b) calibration plots for the related RF algorithm (Figure 5.9) (p-value 0.0001). The x-axis represents the prediction output of the classifier in terms of the probability of being a case. With the y-axis denoting observed class frequencies. Perfect calibration in which predicted probabilities match observed accuracies is denoted by the diagonal dotted line. The blue dots represent the mean probability values within each quantile and are accompanied by a 95% confidence interval (blue bar). The overall relationship between predicted probabilities and observed frequencies (calibration curve) is given by the fitted loess smoother (red line), with a 95% (grey shaded area) used.

Supplementary Tables for Chapter 6:

Supplementary Table 29. Comparison between ML and PRS with No Feature Selection using Genotypes

P-value threshold[a]	Random Forest[b]	Gradient Boosting[c]	Polygenic Risk Score[d]	Mean Number of SNPs[e]
0.0001	65.5	63.0	67.7	100
0.1	54.7	61.5	68.1	44037
0.5	53.8	58.0	68.4	128575

a-p-value thresholds used for analyses; b – Mean value of AUC across 5 p-value thresholds for the RF; c Mean value of AUC across 5 p-value thresholds for the GB; – d - Mean value of AUC across 5 p-value thresholds for PRS; - e Mean number of SNPs across p-value thresholds

Supplementary Table 30. Comparison between ML and PRS with no Feature Selection using Dosages

P-value threshold[a]	Random Forest[b]	Gradient Boosting[c]	Polygenic Risk Score[d]	Mean Number of SNPs[e]
0.0001	65.6	62.8	68.2	130
0.1	56.8	58.7	69.6	64374
0.5	56.6	58.9	69.5	204392

a-p-value thresholds used for analyses; b – Mean value of AUC across 5 p-value thresholds for the RF; c Mean value of AUC across 5 p-value thresholds for the GB; – d - Mean value of AUC across 5 p-value thresholds for PRS; - e Mean number of SNPs across p-value thresholds

Supplementary Table 31. Comparison between ML and PRS with RF Feature Selection using Genotypes

P-value threshold[a]	Random Forest[b]	Gradient Boosting[c]	Polygenic Risk Score[d]	Mean Number of SNPs[e]	Mean Number of SNPs for FS[f]	Mean Number of Overlapping SNPs[g]
0.0001	67.7	67.3	68.1	101	1	N/A
0.1	66.5	68.1	68.3	44063	16	1
0.5	65.0	65.0	68.3	128613	1267	1

a-p-value thresholds used for analyses; b – Mean value of AUC across 5 p-value thresholds for the RF; c Mean value of AUC across 5 p-value thresholds for the GB; – d - Mean value of AUC across 5 p-value thresholds for PRS; - e Mean number of SNPs across p-value thresholds; -f Mean number of SNPs across p-value thresholds for feature selection; -g The overlap of SNPs used for feature selection between the current and previous p-value threshold.

Supplementary Table 32. Comparison between ML and PRS with RF Feature Selection using Dosages

P-value threshold[a]	Random Forest[b]	Gradient Boosting[c]	Polygenic Risk Score[d]	Mean Number of SNPs[e]	Mean Number of SNPs for FS[f]	Mean Number of Overlapping SNPs[g]
0.0001	65.8	65.0	68.6	131	13	N/A
0.1	66.2	67.1	69.3	66403	1	1
0.5	65.4	66.1	69.1	204479	5	1

a-p-value thresholds used for analyses; b – Mean value of AUC across 5 p-value thresholds for the RF; c Mean value of AUC across 5 p-value thresholds for the GB; – d - Mean value of AUC across 5 p-value thresholds for PRS; - e Mean number of SNPs across p-value thresholds; -f Mean number of SNPs across p-value thresholds for feature selection; -g The overlap of SNPs used for feature selection between the current and previous p-value threshold.

Supplementary Table 33. Comparison between ML and PRS with ExtraTrees as Feature Selection using Genotypes

P-value Threshold[a]	Random Forest[b]	Gradient Boosting[c]	Polygenic Risk Score[d]	Mean Number of SNPs[e]	Mean Number of SNPs for FS[f]	Mean Number of Overlapping SNPs[g]
0.0001	66.2	65.7	67.8	100	1	N/A
0.1	66.5	65.3	68.3	44057	4	1
0.5	68.6	66.4	67.0	128582	175	1

a-p-value thresholds used for analyses; b – Mean value of AUC across 5 p-value thresholds for the RF; c Mean value of AUC across 5 p-value thresholds for the GB; – d - Mean value of AUC across 5 p-value thresholds for PRS; - e Mean number of SNPs across p-value thresholds; -f Mean number of SNPs across p-value thresholds for feature selection; -g The overlap of SNPs used for feature selection between the current and previous p-value threshold.

Supplementary Table 34. Comparison between ML and PRS with ExtraTrees Feature Selection using Dosages

P-value Threshold[a]	Random Forest[b]	Gradient Boosting[c]	Polygenic Risk Score[d]	Mean Number of SNPs[e]	Mean Number of SNPs for FS[f]	Mean Number of Overlapped SNPs[g]
0.0001	66.8	65.8	68.5	132	1	N/A
0.1	67.9	67.3	69.0	64414	3	1
0.5	66.4	66.6	69.6	204481	1	1

a-p-value thresholds used for analyses; b – Mean value of AUC across 5 p-value thresholds for the RF; c Mean value of AUC across 5 p-value thresholds for the GB; – d - Mean value of AUC across 5 p-value thresholds for PRS; - e Mean number of SNPs across p-value thresholds; -f Mean number of SNPs across p-value thresholds for feature selection; -g The overlap of SNPs used for feature selection between the current and previous p-value threshold.

Supplementary Table 35. Pairwise t-test for Classifier Comparisons when using Genotypes and no Feature Selection.

P-value Threshold[a]	Comparison Statistics[b]
0.0001	PRS had superior performance to GB. PRS vs GB = (5.26, 0.011)
0.1	PRS had superior performance to both GB and RFs PRS vs GB = (7.19, 0.004) PRS vs RFs = (17.3, 0.000)
0.5	PRS had superior performance to both GB and RFs PRS vs GB = (8.85, 0.003) PRS vs RFs = (55.7, 5.61e-06)

a – P-value threshold used; b – Results of classifier comparisons using t-tests when including all SNPs, only those tests which returned significant results (p-value <0.05) were are detailed.

Supplementary Table 36. Pairwise t-test for Classifier Comparisons when using Dosages and no Feature Selection.

P-value Threshold[a]	Comparison Statistics[b]
0.0001	PRS had superior performance to GB. PRS vs GB = (5.26, 0.011)
0.1	PRS had superior performance to both GB and RFs PRS vs GB = (7.19, 0.002) PRS vs RFs = (17.3, 6.55e-05)
0.5	PRS had superior performance to both GB and RFs PRS vs GB = (8.85, 0.003) PRS vs RFs = (55.7, 6.23e-07)

a – P-value threshold used; b – Results of classifier comparisons using t-tests when including all SNPs, only those tests which returned significant results (p-value <0.05) were are detailed.

Supplementary Table 37. Pairwise t-test Statistics for Classifier Comparisons when using Genotypes and RFs for Feature Selection.

P-value Threshold[a]	Comparison Statistics[b]
0.0001	There were no differences between methods.
0.1	PRS had superior performance to RFs. PRS vs RFs = (4.48, 0.050)
0.5	PRS had superior performance to GB. PRS vs RFs = (5.15, 0.050)

a – P-value threshold used; b – Results of classifier comparisons using t-tests when including all SNPs, only those tests which returned significant results (p-value <0.05) were are detailed.

Supplementary Table 38. Pairwise t-test Statistics for Classifier Comparisons when using Dosages and RFs for Feature Selection.

P-value Threshold[a]	Comparison Statistics[b]
0.0001	There were no differences between methods.
0.1	PRS had superior performance to both GB and RFs PRS vs RFs = (9.60, 0.003) PRS vs GB = (7.56, 0.005)
0.5	PRS had superior performance to both GB and RFs PRS vs RFs = (9.95, 0.003) PRS vs GB = (4.44, 0.030)

a – P-value threshold used; b – Results of classifier comparisons using t-tests when including all SNPs, only those tests which returned significant results (p-value <0.05) are detailed.

Supplementary Table 39. Pairwise t-test for Classifier Comparisons when using Genotypes and the ExtraTree Algorithm for Feature Selection.

P-value Threshold[a]	Comparison Statistics[b]
0.0001	There were no differences between methods.
0.1	There were no differences between methods.
0.5	Both PRS and RFs had superior performance to both GB PRS vs GB = (6.03, 0.020) RFs vs GB = (6.58, 0.020)

a – P-value threshold used; b – Results of classifier comparisons using t-tests when including all SNPs, only those tests which returned significant results (p-value <0.05) are detailed.

Supplementary Table 40. Pairwise t-test for Classifier Comparisons when using Dosages and the ExtraTree Algorithm for Feature Selection.

P-value Threshold[a]	Comparison Statistics[b]
0.0001	There were no differences between methods.
0.1	There were no differences between methods.
0.5	PRS had superior performance GB only. PRS vs GB = (6.85, 0.021)

a – P-value threshold used; b – Results of classifier comparisons using t-tests when including all SNPs, only those tests which returned significant results (p-value <0.05) are detailed .

Supplementary Tables for Chapter 7:

Supplementary Table 41. Results of Individual Pathway Analyses using non-imputed Genotypes

Pathway[a]	No *APOE*[b]	*APOE* Included[c]	*APOE* Alleles[d]

1	RF = 55.0
	GB = 54.0
	PRS-LR = 52.2
	PRS-CS = 51.4
		RF = 57.1
		GB = 56.3
		PRS-LR = 59.0
		PRS-CS = 60.4
			RF = 68.1
			GB = 67.5
			PRS-LR = 69.0
			PRS-CS = 69.2

2	RF = 53.7
	GB = 53.2
	PRS-LR = 51.4
	PRS-CS = 60.7
		RF = 59.2
		GB = 58.9
		PRS-LR = 58.1
		PRS-CS = 60.0
			RF = 67.5
			GB = 66.8
			PRS-LR = 69.0
			PRS-CS = 69.4

3	RF = 56.4
	GB = 55.3
	PRS-LR = 53.8
	PRS-CS = 53.0
		RF = 60.6
		GB = 58.2
		PRS-LR = 58.1
		PRS-CS = 60.3
			RF = 67.9
			GB = 67.4
			PRS-LR = 69.6
			PRS-CS = 69.4

4	RF = 51.9
	GB = 52.5
	PRS-LR = 51.2
	PRS-CS = 51.8
		RF = 59.0
		GB = 57.9
		PRS-LR = 57.7
		PRS-CS = 60.0
			RF = 67.4
			GB = 67.2
			PRS-LR = 69.1
			PRS-CS = 69.3

5	RF = 55.1
	GB = 54.0
	PRS-LR = 52.3
	PRS-CS = 51.6
		RF = 60.4
		GB = 59.7
		PRS-LR = 58.8
		PRS-CS = 60.2
			RF = 67.8
			GB = 67.6
			PRS-LR = 69.2
			PRS-CS = 69.0

6	RF = 55.2
	GB = 54.6
	PRS-LR = 53.3
	PRS-CS = 53.6
		RF = 60.4
		GB = 59.5
		PRS-LR = 58.3
		PRS-CS = 60.8
			RF = 68.3
			GB = 67.7
			PRS-LR = 69.3
			PRS-CS = 69.7

7	RF = 56.2
	GB = 54.1
	PRS-LR = 52.5
	PRS-CS = 52.1
		RF = 60.7
		GB = 58.2
		PRS-LR = 58.0
		PRS-CS = 60.7
			RF = 68.4
			GB = 67.8
			PRS-LR = 69.6
			PRS-CS = 69.9

8	RF = 53.6
	GB = 52.6
	PRS-LR = 52.0
	PRS-CS = 51.2
		RF = 59.6
		GB = 58.6
		PRS-LR = 58.0
		PRS-CS = 58.6
			RF = 68.3
			GB = 67.8
			PRS-LR = 68.9
			PRS-CS = 69.0

9	RF = 55.7
	GB = 51.4
	PRS-LR = 52.2
	PRS-CS = 52.1		

a – The pathway set used for analysis; b – Results of analysis when all *APOE* related SNPs were removed from the SNP set. c – Results of analysis when all *APOE* related SNPs were included the SNP set; d – Results of analysis where APOE related SNPs were removed, and alleles included.

Supplementary Table 42. Results of Individual Pathway Analyses using Imputed Genotypes

Pathway[a]	No *APOE*[b]	*APOE* Included[c]	*APOE* Alleles[d]

1	RF = 52.3
	GB = 51.9
	PRS-LR = 51.1
	PRS-CS = 51.9
		RF = 60.1
		GB = 60.0
		PRS-LR = 60.1
		PRS-CS = 60.9
			RF = 67.5
			GB = 67.0
			PRS-LR = 69.3
			PRS-CS = 69.5

2	RF = 55.5
	GB = 55.3
	PRS-LR = 52.1
	PRS-CS = 52.1
		RF = 60.1
		GB = 59.8
		PRS-LR = 61.0
		PRS-CS = 60.7
			RF = 67.9
			GB = 66.8
			PRS-LR = 68.9
			PRS-CS =

3	RF = 56.4
	GB = 54.3
	PRS-LR = 52.9
	PRS-CS = 52.3
		RF = 59.8
		GB = 60.6
		PRS-LR = 59.8
		PRS-CS = 60.6
			RF = 67.9
			GB = 66.9
			PRS-LR = 68.9
			PRS-CS = 69.3

4	RF = 55.5
	GB = 54.2
	PRS-LR = 51.9
	PRS-CS = 51.9
		RF = 60.6
		GB = 59.6
		PRS-LR = 60.8
		PRS-CS = 61.2
			RF = 67.8
			GB = 66.9
			PRS-LR = 68.9
			PRS-CS = 69.3

5	RF = 54.7
	GB = 52.8
	PRS-LR = 52.1
	PRS-CS = 51.2
		RF = 60.0
		GB = 58.9
		PRS-LR = 61.5
		PRS-CS = 61.4
			RF = 67.8
			GB = 67.4
			PRS-LR = 69.0
			PRS-CS = 69.2

6	RF = 56.5
	GB = 55.0
	PRS-LR = 54.5
	PRS-CS = 53.2
		RF = 59.3
		GB = 58.0
		PRS-LR = 61.4
		PRS-CS = 61.9
			RF = 68.0
			GB = 67.3
			PRS-LR = 69.0
			PRS-CS = 69.7

7	RF = 56.7
	GB = 53.9
	PRS-LR = 53.4
	PRS-CS = 52.8
		RF = 59.3
		GB = 58.7
		PRS-LR = 60.0
		PRS-CS =61.0
			RF = 67.7
			GB = 67.3
			PRS-LR = 68.9
			PRS-CS = 69.7

8	RF = 55.7
	GB = 54.2
	PRS-LR = 51.9
	PRS-CS = 51.9
		RF = 59.8
		GB = 58.8
		PRS-LR = 59.8
		PRS-CS = 61.0
			RF = 67.5
			GB = 67.2
			PRS-LR = 69.3
			PRS-CS = 69.3

9	RF = 57.3
	GB = 53.6
	PRS-LR = 52.6
	PRS-CS = 52.2		

a – The pathway set used for analysis; b – Results of analysis when all *APOE* related SNPs were removed from the SNP set. c – Results of analysis when all *APOE* related SNPs were included the SNP set; d – Results of analysis where APOE related SNPs were removed, and alleles included.

Supplementary Table 43. Results of Multivariable Analyses using Imputed Genotypes.

Type of Input[a]	Without *APOE*[b]	*APOE* Included[c]	With *APOE* Alleles[d]

Genotypes	RF = 56.0
	GB = 51.1
	PRS-LR = 58.0
	PRS-CS = 57.3
		RF = 62.9
		GB = 57.3
		PRS-LR = 64.3
		PRS-CS = 63.7
			RF = 64.3
			GB = 65.7
			PRS-LR = 71.1
			PRS-CS = 69.8

PRS Values	RF = 52.5
	GB = 51.0
	PRS-LR = 55.9
	PRS-CS = 55.1
		RF = 60.4
		GB = 59.0
		PRS-LR = 63.6
		PRS-CS = 62.6
			RF = 67.6
			GB = 67.6
			PRS-LR = 70.3
			PRS-CS = 68.8

a – Type of data input to classifiers; b – Results of analysis when including all SNPs. c – Results of analysis when all *APOE* related SNPs were removed, and alleles included.

Supplementary Table 44. Results of Analyses using the Unified SNP set.

Type of Input[a]	Without *APOE*[b]	*APOE* Included[c]	With Alleles[d]

Genotypes	RF = 55.8
	GB = 53.0
	PRS-LR = 52.7
	PRS-CS = 51.7
		RF = 57.8
		GB = 56.0
		PRS-LR = 55.5
		PRS-CS = 5.5
			RF = 68.0
			GB = 66.3
			PRS-LR = 68.3
			PRS-CS = 67.9

PRS Values	RF = 51.0
	GB = 50.8
	PRS-LR = 52.1
	PRS-CS = 51.6
		RF = 55.3
		GB = 56.5
		PRS-LR = 57.5
		PRS-CS = 56.9
			RF = 67.6
			GB = 66.7
			PRS-LR = 67.5
			PRS-CS = 66.7

a – Type of data input to classifiers; b – Results of analysis when including all SNPs. c – Results of analysis when all *APOE* related SNPs were removed, and alleles included.

Supplementary Table 45. Results of Classifier Comparisons for Individual Pathway Analysis (non-imputed SNPs) using the t-test.

Pathway[a]	No *APOE*[b]	*APOE* Included[c]	*APOE* Alleles[d]
1	Both ML methods outperformed both PRS methods. Whilst RFs also outperformed GB RF vs PRS-LR = (-13.0, 0.025) GB vs PRS-LR = (-6.34, 0.036) RFs vs GB = (4.52, 0.048) RFs vs PRS-CS = (-6.03, 0.036)	There were no differences between methods.	PRS-LR outperformed RFs. Whilst PRS_CS outperformed RFs and GB. PRS-LR vs RFs = (4.57, 0.048) PRS_CS vs RFs = (6.14, 0.036) PRS_CS vs GB = (5.44, 0.040)
2	RFs achieved greater prediction to PRS-LR RFs vs PRS-LR = (4.91, 0.043)	There were no differences between methods.	Both PRS-LR and PRS_CS outperformed RFs and GB. PRS-LR vs RFs = (4.93, 0.043) PRS-LR vs GB = (7.69, 0.029) PRS_CS vs RFs = (7.09, 0.035) PRS_CS vs GB = (9.83, 0.025)
3	RFs outperformed both PRS methods. RF vs PRS-LR = (-6.18, 0.003) RFs vs PRS_CS = (-8.23, 0.025)	RFs outperformed both PRS-LR and GB. Whilst PRS_CS also outperformed GB. RF vs PRS-LR = (-4.66, 0.048) RFs vs GB = (5.15, 0.043) PRS_CS vs GB = (4.52, 0.048)	Both PRS-LR and PRS_CS outperformed RFs. PRS-LR vs RFs = (6.71, 0.035) PRS_CS vs RFs = (8.33, 0.025)
4	There were no differences between methods.	There were no differences between methods.	PRS-LR outperformed RFs. Whilst PRS_CS outperformed RFs and GB. PRS-LR vs RFs = (5.03, 0.043) PRS-LR vs GB = (8.98, 0.025) PRS_CS vs RFs = (9.77, 0.025) PRS_CS vs GB = (5.07, 0.042)
5	RFs achieved greater prediction to PRS_CS RFs vs PRS_CS = (-5.78, 0.037)	There were no differences between methods.	There were no differences between methods.
6	There were no differences between methods.	PRS_CS outperformed PRS only. PRS_CS vs PRS-LR = (5.77, 0.037)	PRS-CS outperformed both PRS_LR and GB. PRS_CS vs PRS_LR = (6.18, 0.036) PRS_CS vs GB = (4.95, 0.043)
7	RFs achieved greater prediction to PRS-LR RF vs PRS-LR = (-4.59, 0.048)	RFs outperformed both PRS-LR and GB. Whilst PRS_CS also outperformed PRS-LR. RF vs PRS-LR = (-5.63, 0.039) RFs vs GB = (6.82, 0.035) PRS_CS vs PRS-LR = (5.06, 0.043)	Both PRS-LR and PRS_CS outperformed RFs. PRS-LR vs RFs = (5.43, 0.040) PRS_CS vs RFs = (8.41, 0.025)
8	There were no differences between methods.	There were no differences between methods.	There were no differences between methods.
9	There were no differences between methods.	There were no differences between methods.	There were no differences between methods.

a – The pathway set used for analysis; b – Results of classifier comparisons using t-tests when all *APOE* related SNPs were removed from the SNP set. c – Results of classifier comparisons using t-tests when all APOE related SNPs were included the SNP set; d – Results of classifier comparisons using t-tests where *APOE* related SNPs were removed, and alleles included.

Supplementary Table 46. Results of Classifier Comparisons for Individual Pathway Analysis (Imputed SNPs) using the t-test.

Pathway[a]	No *APOE*[b]	*APOE* Included[c]	*APOE* Alleles[d]
1	There were no differences between methods.	There were no differences between methods.	PRS-CS outperformed RFs only. PRS-CS vs RFs = (7.59, 0.033)
2	Both RFs and GB outperformed PRS-LR RF vs PRS-LR = (-5.47, 0.045) GB vs PRS-LR = (-5.85, 0.044)	There were no differences between methods.	PRS-LR outperformed GB only. PRS-LR vs GB = (7.91,0.044)
3	RFs outperformed PRS-CS only. RF vs PRS-CS = (-5.27, 0.041)	There were no differences between methods.	PRS-CS outperformed GB only. PRS-CS vs GB = (5.43, 0.044)
4	Both RFs and GB achieved greater prediction than PRS-LR RFs vs PRS-LR = (-7.45,0.044) GB vs PRS-LR = (-6.44, 0.044)	PRS-LR outperformed GB only. PRS-LR vs GB = (-5.85, 0.044)	Both PRS-LR and PRS-CS outperformed RFs and GB. PRS-CS vs GB = (6.58, 0.033) PRS-LR vs GB = (5.30, 0.041) PRS-LR vs RFs = (5.39 0.041)
5	There were no differences between methods.	There were no differences between methods.	PRS-LR outperformed GB only. PRS-LR vs GB = (4.87, 0.045)
6	There were no differences between methods.	PRS-LR outperformed RFs only. PRS-LR vs RFs = (4.90, 0.045)	Both PRS-LR and PRS-CS outperformed GB. PRS-LR vs GB = (5.61, 0.041) PRS-CS vs GB = (8.50, 0.033)
7	RFs outperformed both PRS-LR and PRS-CS. RFs vs PRS-LR = (-5.35, 0.045) RFs vs PRS-CS = (-6.73, 0.033)	There were no differences between methods.	PRS-CS outperformed RFs only. PRS-CS vs RFs = (6.06, 0.039)
8	RFs outperformed PRS-CS only. RF vs PRS-CS = (-6.45, 0.033)	There were no differences between methods.	There were no differences between methods.
9	There were no differences between methods.		

a – The pathway set used for analysis; b – Results of classifier comparisons using t-tests when all *APOE* related SNPs were removed from the SNP set. c – Results of classifier comparisons using t-tests when all APOE related SNPs were included the SNP set; d – Results of classifier comparisons using t-tests where *APOE* related SNPs were removed, and alleles included.

Supplementary Table 47. Results of Classifier Comparisons for Multivariable Analyses using the t-test.

Type of Input[a]	Without *APOE*[b]	*APOE* Included[c]	With Alleles[d]
Genotypes	PRS-LR, PRS-CS and RFs outperformed GB. PRS-LR vs GB = (8.29,0.003) PRS-CS vs GB = (11.0, 0.003) RFs vs GB = (11.7,0.003)	PRS-LR outperformed both RFs and GB, whilst PRS-CS outperformed GB. PRS-LR vs GB = (11.0, 0.003) PRS-CS vs GB = (7.84, 0.007) RFs vs GB = (5.24, 0.006)	PRS-LR and PRS-CS outperformed GB and RFs, PRS-LR vs RFs = (9.33,0.003) PRS-LR vs GB = (9.35,0.003) PRS-CS vs RF = (5.31,0.012) PRS-CS vs GB = (5.09, 0.020)
PRS Values	Both PRS-LR and PRS-CS outperformed GB and RFs. PRS-LR vs GB = (4.32, 0.023) PRS-LR vs RF = (4.32,0.022) PRS-CS vs RF = (4.74, 0.020) PRS-CS vs GB = (4.79, 0.020)	PRS-LR and PRS-CS outperformed GB and RFs. PRS-LR vs RFs = (4.41, 0.023) PRS-LR vs GB = (5.05, 0.020) PRS-CS vs RF = (10.6, 0.003) PRS-CS vs GB = (7.11, 0.008)	PRS-LR outperformed GB only. PRS-LR vs GB = (3.45,0.047)

a – Type of data input to classifiers; b – Results of classifier comparisons using t-tests when including all SNPs. c – Results of analysis when all *APOE* related SNPs were removed, and alleles included.

Supplementary Table 48. Results of Classifier Comparisons for Unified SNP set Analyses using the t-test.

Type of Input[a]	No *APOE*[b]	*APOE* Included[c]	With Alleles[d]
Genotypes	There were no differences between methods.	There were no differences between methods.	There were no differences between methods.
PRS Values	There were no differences between methods.	There were no differences between methods.	There were no differences between methods.

a – Type of data input to classifiers; b – Results of classifier comparisons using t-tests when including all SNPs. c – Results of analysis when all APOE related SNPs were removed, and alleles included.

Supplementary Table 49. Results of Pathway significance in a multivariable PRS-LR when including all SNPs (including *APOE* SNPs)

Pathway[a]	Coefficients[b]	Standard Error[c]	Z Score[d]	P-value[e]
1	0.1797	0.269	0.669	0.504
2	0.4097	0.187	2.174	0.302
3	-0.0322	0.110	-0.294	0.769
4	-0.3283	0.198	-1.656	0.198
5	0.3160	0.116	2.728	0.455
6	0.4445	0.155	2.873	0.203
7	-0.1388	0.175	-0.795	0.427
8	-0.2322	0.222	-1.044	0.297
9	0.7270	0.073	1.287	0.198

a – Type of data input to classifiers; b – Results of classifier comparisons using t-tests when including all SNPs. c – Results of analysis when all APOE related SNPs were removed, and alleles included.

Supplementary Table 50. Results of Pathway in a Multivariable PRS-LR when excluding *APOE* related SNPs and including *APOE* alleles

Pathway[a]	Coefficients[b]	Standard Error[c]	Z Score[d]	P-value[e]
1	0.1554	0.201	0.773	0.439
2	0.3446	0.134	2.571	0.120
3	-0.0772	0.093	-0.831	0.406
4	-0.3629	0.133	-2.722	0.306
5	0.0178	0.087	0.204	0.838
6	0.1352	0.119	1.137	0.255
7	0.1194	0.153	0.779	0.436
8	-0.3092	0.167	-1.856	0.163
9	0.0787	0.076	1.040	0.298
10	0.2317	0.001	8.798	0.001

a – Type of data input to classifiers; b – Results of classifier comparisons using t-tests when including all SNPs. c – Results of analysis when all APOE related SNPs were removed, and alleles included.

Supplementary Table 51. Results of Pathway in a Multivariable PRS-LR when excluding *APOE* related SNPs and including *APOE* alleles

Type of Analysis	No *APOE* Region	*APOE* Region Included	*APOE* Alleles Included

Individual Pathways	Non-Imputed SNPs:
	PRS-LR = 52.4
	PRS-CS = 53.3
	RFs = 54.5
	GB = 53.6
	Imputed SNPs:
	PRS-LR = 52.5
	PRS-CS = 52.2
	RFs = 55.5
	GB = 54.0
		Non-Imputed SNPs:
		PRS-LR = 58.2
		PRS-CS = 60.0
		RFs = 59.7
		GB = 59.0
		Imputed SNPs:
		PRS-LR = 60.6
		PRS-CS = 61.1
		RFs = 59.9
		GB = 59.0
			Non-Imputed SNPs:
			PRS-LR = 69.2
			PRS-CS = 69.4
			RFs = 67.9
			GB = 67.5
			Imputed SNPs:
			PRS-LR = 69.1
			PRS-CS = 69.5
			RFs = 67.7
			GB = 67.0

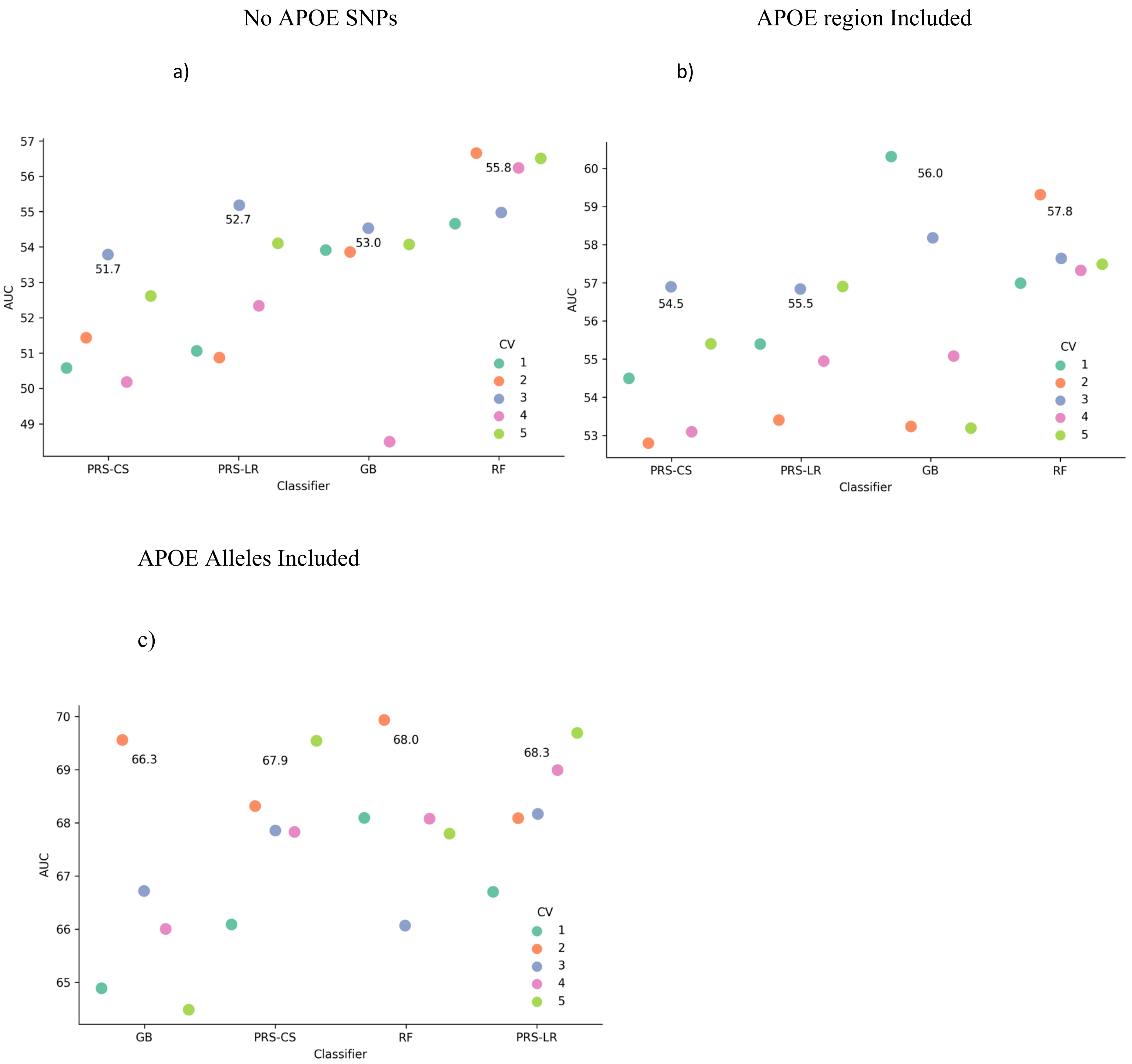
Multivariable Analysis	Imputed SNPs:
	Genotypes (internal information):
	PRS-LR = 58.0
	PRS-CS = 57.3
		Imputed SNPs:
		Genotypes (internal information):
		PRS-LR = 64.3
		PRS-CS = 63.7
			Imputed SNPs:
			Genotypes (internal information):
			PRS-LR = 71.1
			PRS-CS = 69.8

	RFs = 56.0
	GB = 51.1
	PRS (external information):
	PRS-LR = 55.9
	PRS-CS = 55.1
	RFs = 52.5
	GB = 51.0
		RFs = 62.9
		GB = 57.3
		PRS (external information):
		PRS-LR = 63.7
		PRS-CS = 62.6
		RFs = 60.1
		GB = 58.6
			RFs = 64.3
			GB = 65.7
			PRS (external information):
			PRS-LR = 70.1
			PRS-CS = 68.8
			RFs = 68.8
			GB = 68.4

Combined Pathway Analysis	Imputed SNPs:
	Genotypes (internal information):
	PRS-LR = 52.7
	PRS-CS = 51.7
	RFs = 55.8
	GB = 53.0
	PRS (external information):
	PRS-LR = 52.1
	PRS-CS = 51.6
	RFs = 51.0
	GB = 50.8
		Imputed SNPs:
		Genotypes (internal information):
		PRS-LR = 55.5
		PRS-CS = 54.5
		RFs = 57.8
		GB = 56.0
		PRS (external information):
		PRS-LR = 57.5
		PRS-CS = 56.9
		RFs = 55.3
		GB =56.5
			Imputed SNPs:
			Genotypes (internal information):
			PRS-LR = 68.3
			PRS-CS = 67.9
			RFs = 68.0
			GB = 66.3
			PRS (external information):
			PRS-LR = 67.5
			PRS-CS = 66.7
			RFs = 67.6
			GB = 66.7

Analyses are split into the three sections investigated in Chapter 7, with further splitting into the three different ways *APOE* was modelled.

Supplementary Figure 3: Comparison of PRS vs Selected Classifiers (RF, GB) for LD Pruned SNPs in Imputed Genotypes. Variants in the *APOE* region initially excluded, included and followed by the inclusion of *APOE* alleles.



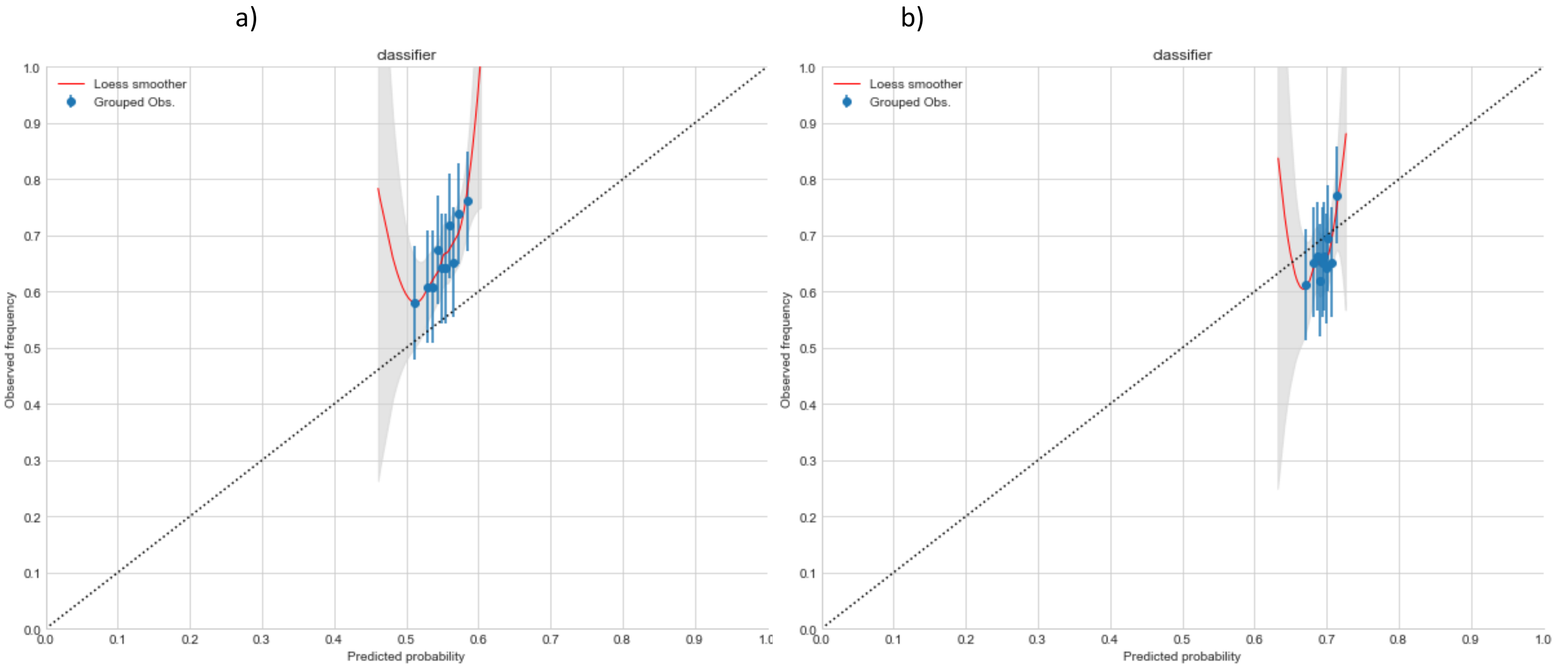
No APOE SNPs

APOE region Included

APOE Alleles Included

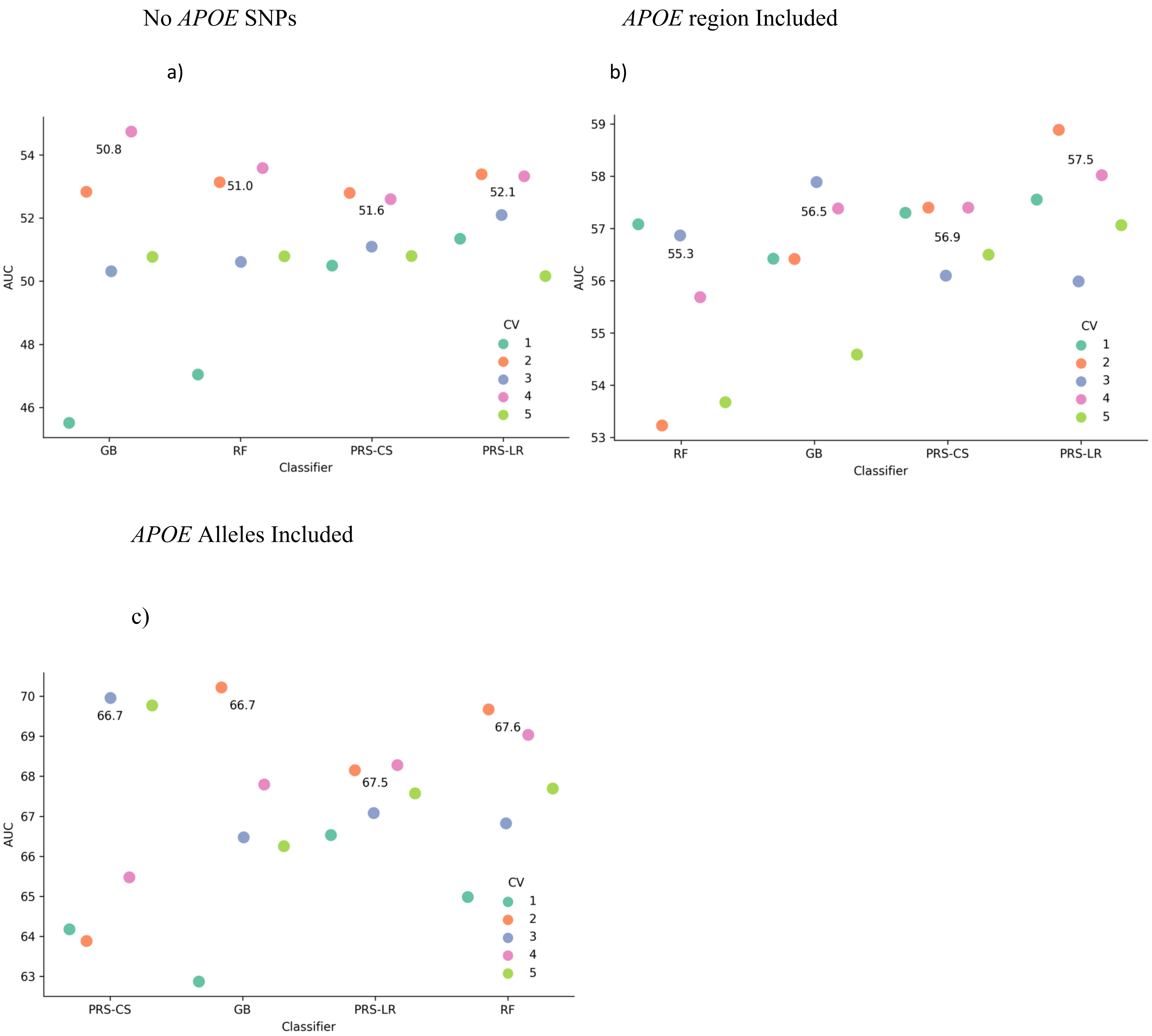
Y-axis represents AUC in %; X-axis represents each classifier's results for each round of CV. The numbers placed centrally are the mean prediction performance across 5 folds of CV; GB Gradient Boosting; RF Random Forest; PRS-LR Polygenic Risk Scores Logistic Regression; PRS-CS.AUC Area Under the Curve.

Supplementary Figure 4: The Comparison of non-Calibrated vs Calibrated Prediction Probabilities for the RF from Figure 7.9.



These two figures display calibration plots for the RF (Protein-lipid complex assembly) in Figure 7.9. The left-hand plot a) displays pre-calibrated probabilities, whilst the right-hand plot b) shows post-calibration. The predicted probabilities are marked along the X-axis, whilst observed probabilities are measured on the Y-axis. Grouped observations represent the average observed prediction value for each decile of predicted probabilities, accompanied by a 95% confidence interval. The overall relationship between predicted probabilities and observed frequencies is given by the fitted loess smoother, with a 95% (grey shaded area) used.

Supplementary Figure 5: The Comparison of PRS-LR vs Selected Classifiers (RF, GB) for LD Pruned SNPs. with PRSs used as inputs. Variants in the *APOE* region initially excluded, included and followed by the inclusion of APOE alleles.



No *APOE* SNPs

a)

APOE region Included

b)

APOE Alleles Included

c)

Y-axis represents AUC in %; X-axis represents each classifier's results per CV fold. The numbers placed centrally are the mean prediction performance across 5 folds of CV; GB Gradient Boosting; RF Random Forest; PRS-LR Polygenic Risk Scores Logistic Regression; AUC Area Under the Curve.

Supplementary Figure 6: The Comparison of non-Calibrated vs Calibrated Prediction Probabilities

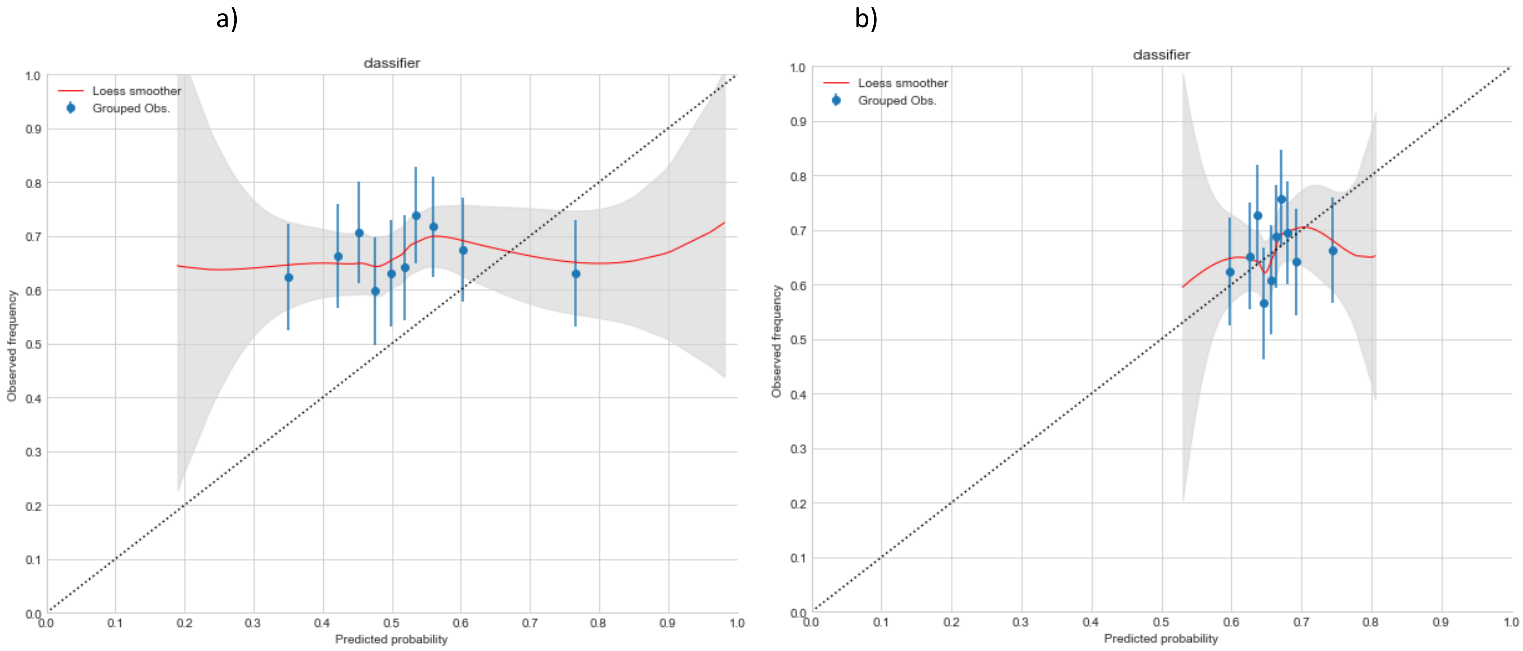


Figure 7.12. These two figures display calibration plots for the RF (Protein-lipid complex assembly) in Figure 7.11. The left-hand plot a) displays pre-calibrated probabilities, whilst the right-hand plot b) shows post-calibration. The predicted probabilities are marked along the X-axis, whilst observed probabilities are measured on the Y-axis. Grouped observations represent the average observed prediction value for each decile of predicted probabilities, accompanied by a 95% confidence interval. The overall relationship between predicted probabilities and observed frequencies is given by the fitted loess smoother, with a 95% (grey shaded area) used.

10 References

‘<scp>2021</scp> Alzheimer’s disease facts and figures’ (2021) *Alzheimer’s & Dementia*, 17(3), pp. 327–406. Available at: https://doi.org/10.1002/alz.12328.

‘2020 Alzheimer’s disease facts and figures’ (2020) *Alzheimer’s & Dementia*, 16(3), pp. 391–460. Available at: https://doi.org/10.1002/alz.12068.

Abd El Hamid, M.M., Mabrouk, M.S. and Omar, Y.M.K. (2019) ‘DEVELOPING AN EARLY PREDICTIVE SYSTEM FOR IDENTIFYING GENETIC BIOMARKERS ASSOCIATED TO ALZHEIMER’S DISEASE USING MACHINE LEARNING TECHNIQUES’, *Biomedical Engineering: Applications, Basis and Communications*, 31(05), p. 1950040. Available at: https://doi.org/10.4015/S1016237219500406.

Adadi, A. (2021) ‘A survey on data-efficient algorithms in big data era’, *Journal of Big Data*, 8(1), p. 24. Available at: https://doi.org/10.1186/s40537-021-00419-9.

Aljohani, N.R., Fayoumi, A. and Hassan, S.-U. (2021) ‘A novel focal-loss and class-weight-aware convolutional neural network for the classification of in-text citations’, *Journal of Information Science*, p. 016555152199102. Available at: https://doi.org/10.1177/0165551521991022.

Allen, R. *et al.* (2003) ‘End-of-Life Issues in the Context of Alzheimer’s Disease’, *Alzheimer’s Care Quarterly*, 4(4), pp. 312–330.

Altman, N. and Krzywinski, M. (2017) ‘Ensemble methods: bagging and random forests’, *Nature Methods*, 14(10), pp. 933–934. Available at: https://doi.org/10.1038/nmeth.4438.

An, L. *et al.* (2017) ‘A Hierarchical Feature and Sample Selection Framework and Its Application for Alzheimer’s Disease Diagnosis’, *Scientific Reports*, 7(1), p. 45269. Available at: https://doi.org/10.1038/srep45269.

Andrade, C. (2019) ‘The P Value and Statistical Significance: Misunderstandings, Explanations, Challenges, and Alternatives.’, *Indian journal of psychological medicine*, 41(3), pp. 210–215. Available at: https://doi.org/10.4103/IJPSYM.IJPSYM_193_19.

Arevalo-Rodriguez, I. *et al.* (2015) ‘Mini-Mental State Examination (MMSE) for the detection of Alzheimer’s disease and other dementias in people with mild cognitive impairment (MCI).’, *The Cochrane database of systematic reviews*, 2015(3), p. CD010783. Available at: https://doi.org/10.1002/14651858.CD010783.pub2.

Ashburner, M. *et al.* (2000) ‘Gene Ontology: tool for the unification of biology’, *Nature Genetics*, 25(1), pp. 25–29. Available at: https://doi.org/10.1038/75556.

Attaran, M. and Deb, P. (2018a) ‘Machine Learning: The New “Big Thing” for Competitive Advantage’, *International Journal of Knowledge Engineering and Data Mining*, 5(1), p. 1. Available at: https://doi.org/10.1504/IJKEDM.2018.10015621.

Attaran, M. and Deb, P. (2018b) ‘Machine learning: the new “big thing” for competitive advantage’, *International Journal of Knowledge Engineering and Data Mining*, 5(4), p. 277. Available at: https://doi.org/10.1504/IJKEDM.2018.095523.

Auria, L. and Moro, R.A. (2008) ‘Support Vector Machines (SVM) as a Technique for Solvency Analysis’, *SSRN Electronic Journal* [Preprint]. Available at: https://doi.org/10.2139/ssrn.1424949.

Austin, P.C. and Steyerberg, E.W. (2017a) ‘Events per variable (EPV) and the relative performance of different strategies for estimating the out-of-sample validity of logistic regression models.’, *Statistical methods in medical research*, 26(2), pp. 796–808. Available at: https://doi.org/10.1177/0962280214558972.

Austin, P.C. and Steyerberg, E.W. (2017b) ‘Events per variable (EPV) and the relative performance of different strategies for estimating the out-of-sample validity of logistic regression models’, *Statistical Methods in Medical Research*, 26(2), pp. 796–808. Available at: https://doi.org/10.1177/0962280214558972.

Austin, P.C. and Steyerberg, E.W. (2017c) ‘Events per variable (EPV) and the relative performance of different strategies for estimating the out-of-sample validity of logistic regression models’, *Statistical Methods in Medical Research*, 26(2), pp. 796–808. Available at: https://doi.org/10.1177/0962280214558972.

Awad, M. and Khanna, R. (2015) *Efficient Learning Machines*. Berkeley, CA: Apress. Available at: https://doi.org/10.1007/978-1-4302-5990-9.

Awada, A.A. (2015) ‘Early and late-onset Alzheimer’s disease: What are the differences?’, *Journal of neurosciences in rural practice*, 6(3), pp. 455–6. Available at: https://doi.org/10.4103/0976-3147.154581.

Bachiller, S. *et al.* (2018) ‘Microglia in Neurological Diseases: A Road Map to Brain-Disease Dependent-Inflammatory Response.’, *Frontiers in cellular neuroscience*, 12, p. 488. Available at: https://doi.org/10.3389/fncel.2018.00488.

Baker, E. and Escott-Price, V. (2020a) ‘Polygenic Risk Scores in Alzheimer’s Disease: Current Applications and Future Directions’, *Frontiers in Digital Health*, 2. Available at: https://doi.org/10.3389/fdgth.2020.00014.

Baker, E. and Escott-Price, V. (2020b) ‘Polygenic Risk Scores in Alzheimer’s Disease: Current Applications and Future Directions’, *Frontiers in Digital Health*, 2. Available at: https://doi.org/10.3389/fdgth.2020.00014.

Balakrishnan, K., Dhanalakshmi, R. and Khaire, U. (2022) ‘Analysing stable feature selection through an augmented marine predator algorithm based on <scp>opposition-based</scp> learning’, *Expert Systems*, 39(1). Available at: https://doi.org/10.1111/exsy.12816.

Barnes, L.L. (2022) ‘Alzheimer disease in African American individuals: increased incidence or not enough data?’, *Nature reviews. Neurology*, 18(1), pp. 56–62. Available at: https://doi.org/10.1038/s41582-021-00589-3.

Bauermeister, S. *et al.* (2020) 'The Dementias Platform UK (DPUK) Data Portal.', *European journal of epidemiology*, 35(6), pp. 601–611. Available at: https://doi.org/10.1007/s10654-020-00633-4.

Bekris, L.M. *et al.* (2010a) 'Genetics of Alzheimer disease.', *Journal of geriatric psychiatry and neurology*, 23(4), pp. 213–27. Available at: https://doi.org/10.1177/0891988710383571.

Bekris, L.M. *et al.* (2010b) 'Genetics of Alzheimer disease.', *Journal of geriatric psychiatry and neurology*, 23(4), pp. 213–27. Available at: https://doi.org/10.1177/0891988710383571.

Bellenguez, C., Küçükali, F., Jansen, Iris E, *et al.* (2022) 'New insights into the genetic etiology of Alzheimer's disease and related dementias.', *Nature genetics*, 54(4), pp. 412–436. Available at: https://doi.org/10.1038/s41588-022-01024-z.

Bellenguez, C., Küçükali, F., Jansen, Iris E., *et al.* (2022a) 'New insights into the genetic etiology of Alzheimer's disease and related dementias', *Nature Genetics*, 54(4), pp. 412–436. Available at: https://doi.org/10.1038/s41588-022-01024-z.

Bellenguez, C., Küçükali, F., Jansen, Iris E., *et al.* (2022b) 'New insights into the genetic etiology of Alzheimer's disease and related dementias', *Nature Genetics*, 54(4), pp. 412–436. Available at: https://doi.org/10.1038/s41588-022-01024-z.

Ben-Hur, A. *et al.* (2008) 'Support Vector Machines and Kernels for Computational Biology', *PLoS Computational Biology*, 4(10), p. e1000173. Available at: https://doi.org/10.1371/journal.pcbi.1000173.

Berrar, D. (2019) 'Cross-Validation', in *Encyclopedia of Bioinformatics and Computational Biology*. Elsevier, pp. 542–545. Available at: https://doi.org/10.1016/B978-0-12-809633-8.20349-X.

Bey, R. *et al.* (2020) 'Fold-stratified cross-validation for unbiased and privacy-preserving federated learning', *Journal of the American Medical Informatics Association*, 27(8), pp. 1244–1251. Available at: https://doi.org/10.1093/jamia/ocaa096.

Bi, X. *et al.* (2019) 'Effective Diagnosis of Alzheimer's Disease via Multimodal Fusion Analysis Framework', *Frontiers in Genetics*, 10. Available at: https://doi.org/10.3389/fgene.2019.00976.

Biau, G. (2010) 'Analysis of a Random Forests Model'.
Biessels, G.J., Kappelle, L.J. and Utrecht Diabetic Encephalopathy Study Group (2005) 'Increased risk of Alzheimer's disease in Type II diabetes: insulin resistance of the brain or insulin-induced amyloid pathology?', *Biochemical Society transactions*, 33(Pt 5), pp. 1041–4. Available at: https://doi.org/10.1042/BST0331041.

Blagus, R. and Lusa, L. (2013) 'SMOTE for high-dimensional class-imbalanced data', *BMC Bioinformatics*, 14(1), p. 106. Available at: https://doi.org/10.1186/1471-2105-14-106.

Bom, P.R.D. and Rachinger, H. (2020) 'A <scp>generalized-weights</scp> solution to sample overlap in <scp>meta-analysis</scp>', *Research Synthesis Methods*, 11(6), pp. 812–832. Available at: https://doi.org/10.1002/jrsm.1441.

Bracher-Smith, M. *et al.* (2022) ‘Machine learning for prediction of schizophrenia using genetic and demographic factors in the UK biobank’, *Schizophrenia Research*, 246, pp. 156–164. Available at: https://doi.org/10.1016/j.schres.2022.06.006.

Bracher-Smith, M., Crawford, K. and Escott-Price, V. (2021) ‘Machine learning for genetic prediction of psychiatric disorders: a systematic review’, *Molecular Psychiatry*, 26(1), pp. 70–79. Available at: https://doi.org/10.1038/s41380-020-0825-2.

Breitner, J.C.S. *et al.* (1993) ‘Use of twin cohorts for research in Alzheimer’s disease’, *Neurology*, 43(2), pp. 261–261. Available at: https://doi.org/10.1212/WNL.43.2.261.

Brodeur, Z.P., Herman, J.D. and Steinschneider, S. (2020) ‘Bootstrap Aggregation and Cross-Validation Methods to Reduce Overfitting in Reservoir Control Policy Search’, *Water Resources Research*, 56(8). Available at: https://doi.org/10.1029/2020WR027184.

Brothers, H.M., Gosztyla, M.L. and Robinson, S.R. (2018) ‘The Physiological Roles of Amyloid-β Peptide Hint at New Ways to Treat Alzheimer’s Disease.’, *Frontiers in aging neuroscience*, 10, p. 118. Available at: https://doi.org/10.3389/fnagi.2018.00118.

Burns, M.E. and Augustine, G.J. (1995) ‘Synaptic structure and function: dynamic organization yields architectural precision.’, *Cell*, 83(2), pp. 187–94. Available at: https://doi.org/10.1016/0092-8674(95)90160-4.

Cacace, R., Sleegers, K. and Van Broeckhoven, C. (2016) ‘Molecular genetics of early-onset Alzheimer’s disease revisited.’, *Alzheimer’s & dementia : the journal of the Alzheimer’s Association*, 12(6), pp. 733–48. Available at: https://doi.org/10.1016/j.jalz.2016.01.012.

Calabrò, M. *et al.* (2021) ‘The biological pathways of Alzheimer disease: a review.’, *AIMS neuroscience*, 8(1), pp. 86–132. Available at: https://doi.org/10.3934/Neuroscience.2021005.

Calafato, M.S. *et al.* (2018) ‘Use of schizophrenia and bipolar disorder polygenic risk scores to identify psychotic disorders.’, *The British journal of psychiatry : the journal of mental science*, 213(3), pp. 535–541. Available at: https://doi.org/10.1192/bjp.2018.89.

Van Calster, B. *et al.* (2019a) ‘Calibration: the Achilles heel of predictive analytics’, *BMC Medicine*, 17(1), p. 230. Available at: https://doi.org/10.1186/s12916-019-1466-7.

Van Calster, B. *et al.* (2019b) ‘Calibration: the Achilles heel of predictive analytics’, *BMC Medicine*, 17(1), p. 230. Available at: https://doi.org/10.1186/s12916-019-1466-7.

Calus, M.P.L. and Vandenplas, J. (2018) ‘SNPrune: an efficient algorithm to prune large SNP array and sequence datasets based on high linkage disequilibrium’, *Genetics Selection Evolution*, 50(1), p. 34. Available at: https://doi.org/10.1186/s12711-018-0404-z.

Camacho, D.M. *et al.* (2018) ‘Next-Generation Machine Learning for Biological Networks’, *Cell*, 173(7), pp. 1581–1592. Available at: https://doi.org/10.1016/j.cell.2018.05.015.

Castro, D.M. *et al.* (2010) 'The economic cost of Alzheimer's disease: Family or public health burden?', *Dementia & neuropsychologia*, 4(4), pp. 262–267. Available at: https://doi.org/10.1590/S1980-57642010DN40400003.

Cervantes, J. *et al.* (2020) 'A comprehensive survey on support vector machine classification: Applications, challenges and trends', *Neurocomputing*, 408, pp. 189–215. Available at: https://doi.org/10.1016/j.neucom.2019.10.118.

Chanda, P. *et al.* (2012) 'Comprehensive evaluation of imputation performance in African Americans', *Journal of Human Genetics*, 57(7), pp. 411–421. Available at: https://doi.org/10.1038/jhg.2012.43.

Chang, Y.-C. *et al.* (2020) 'GenEpi: gene-based epistasis discovery using machine learning', *BMC Bioinformatics*, 21(1), p. 68. Available at: https://doi.org/10.1186/s12859-020-3368-2.
Chattopadhyay, A. and Lu, T.-P. (2019) 'Gene-gene interaction: the curse of dimensionality', *Annals of Translational Medicine*, 7(24), pp. 813–813. Available at: https://doi.org/10.21037/atm.2019.12.87.

Chaudhury, S. *et al.* (2019) 'Alzheimer's disease polygenic risk score as a predictor of conversion from mild-cognitive impairment.', *Translational psychiatry*, 9(1), p. 154. Available at: https://doi.org/10.1038/s41398-019-0485-7.

Chen, H. *et al.* (2021) 'Improved naive Bayes classification algorithm for traffic risk management', *EURASIP Journal on Advances in Signal Processing*, 2021(1), p. 30. Available at: https://doi.org/10.1186/s13634-021-00742-6.

Chen, S.-F. *et al.* (2020) 'Genotype imputation and variability in polygenic risk score estimation', *Genome Medicine*, 12(1), p. 100. Available at: https://doi.org/10.1186/s13073-020-00801-x.

Chen, X. and Ishwaran, H. (2012) 'Random forests for genomic data analysis', *Genomics*, 99(6), pp. 323–329. Available at: https://doi.org/10.1016/j.ygeno.2012.04.003.

Chen, Z.-L. *et al.* (2019) 'A high-speed search engine pLink 2 with systematic evaluation for proteome-scale identification of cross-linked peptides', *Nature Communications*, 10(1), p. 3404. Available at: https://doi.org/10.1038/s41467-019-11337-z.

Chengsheng, T., Huacheng, L. and Bing, X. (2017) 'AdaBoost typical Algorithm and its application research', *MATEC Web of Conferences*, 139, p. 00222. Available at: https://doi.org/10.1051/matecconf/201713900222.

Ching, T. *et al.* (2018) 'Opportunities and obstacles for deep learning in biology and medicine', *Journal of The Royal Society Interface*, 15(141), p. 20170387. Available at: https://doi.org/10.1098/rsif.2017.0387.

Cho, G. *et al.* (2019) 'Review of Machine Learning Algorithms for Diagnosing Mental Illness', *Psychiatry Investigation*, 16(4), pp. 262–269. Available at: https://doi.org/10.30773/pi.2018.12.21.2.

Choi, S.W., Mak, T.S.-H. and O'Reilly, Paul F (2020a) 'Tutorial: a guide to performing polygenic risk score analyses.', *Nature protocols*, 15(9), pp. 2759–2772. Available at: https://doi.org/10.1038/s41596-020-0353-1.

Choi, S.W., Mak, T.S.-H. and O'Reilly, Paul F (2020b) 'Tutorial: a guide to performing polygenic risk score analyses.', *Nature protocols*, 15(9), pp. 2759–2772. Available at: https://doi.org/10.1038/s41596-020-0353-1.

Choi, S.W., Mak, T.S.-H. and O'Reilly, Paul F. (2020) 'Tutorial: a guide to performing polygenic risk score analyses', *Nature Protocols*, 15(9), pp. 2759–2772. Available at: https://doi.org/10.1038/s41596-020-0353-1.

Coleman, J.R.I. *et al.* (2016) 'Quality control, imputation and analysis of genome-wide genotyping data from the Illumina HumanCoreExome microarray.', *Briefings in functional genomics*, 15(4), pp. 298–304. Available at: https://doi.org/10.1093/bfgp/elv037.

Collister, J.A., Liu, X. and Clifton, L. (2022a) 'Calculating Polygenic Risk Scores (PRS) in UK Biobank: A Practical Guide for Epidemiologists', *Frontiers in Genetics*, 13. Available at: https://doi.org/10.3389/fgene.2022.818574.

Collister, J.A., Liu, X. and Clifton, L. (2022b) 'Calculating Polygenic Risk Scores (PRS) in UK Biobank: A Practical Guide for Epidemiologists', *Frontiers in Genetics*, 13. Available at: https://doi.org/10.3389/fgene.2022.818574.

Collobert, R. and Bengio, S. (2004) 'Links between perceptrons, MLPs and SVMs', in *Twenty-first international conference on Machine learning - ICML '04*. New York, New York, USA: ACM Press, p. 23. Available at: https://doi.org/10.1145/1015330.1015415.

Contreras, P. *et al.* (2021) 'Influence of Random Forest Hyperparameterization on Short-Term Runoff Forecasting in an Andean Mountain Catchment', *Atmosphere*, 12(2), p. 238. Available at: https://doi.org/10.3390/atmos12020238.

Correa, D.D. *et al.* (2014) 'APOE polymorphisms and cognitive functions in patients with brain tumors.', *Neurology*, 83(4), pp. 320–7. Available at: https://doi.org/10.1212/WNL.0000000000000617.

Cortes, C. and Vapnik, V. (1995) 'Support-vector networks', *Machine Learning*, 20(3), pp. 273–297. Available at: https://doi.org/10.1007/BF00994018.

Crawford, K. *et al.* (2022) 'Golgi apparatus, endoplasmic reticulum and mitochondrial function implicated in Alzheimer's disease through polygenic risk and RNA sequencing', *Molecular Psychiatry* [Preprint]. Available at: https://doi.org/10.1038/s41380-022-01926-8.

Crisci, C., Ghattas, B. and Perera, G. (2012) 'A review of supervised machine learning algorithms and their applications to ecological data', *Ecological Modelling*, 240, pp. 113–122. Available at: https://doi.org/10.1016/j.ecolmodel.2012.03.001.

Cristianini, N. and Shawe-Taylor, J. (2000) *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press. Available at: https://doi.org/10.1017/CBO9780511801389.

Cui, P. *et al.* (2018) ‘Shared Biological Pathways Between Alzheimer’s Disease and Ischemic Stroke.’, *Frontiers in neuroscience*, 12, p. 605. Available at: https://doi.org/10.3389/fnins.2018.00605.

Das, D., Nayak, M. and Pani, S.K. (2019) ‘Missing Value Imputation-A Review’, *International Journal of Computer Sciences and Engineering*, 7(4), pp. 548–558. Available at: https://doi.org/10.26438/ijcse/v7i4.548558.

Das, S. *et al.* (2016) ‘Next-generation genotype imputation service and methods’, *Nature Genetics*, 48(10), pp. 1284–1287. Available at: https://doi.org/10.1038/ng.3656.

Daunt, P. *et al.* (2021) ‘Polygenic Risk Scoring is an Effective Approach to Predict Those Individuals Most Likely to Decline Cognitively Due to Alzheimer’s Disease.’, *The journal of prevention of Alzheimer’s disease*, 8(1), pp. 78–83. Available at: https://doi.org/10.14283/jpad.2020.64.

Davis-Turak, J. *et al.* (2017) ‘Genomics pipelines and data integration: challenges and opportunities in the research setting’, *Expert Review of Molecular Diagnostics*, 17(3), pp. 225–237. Available at: https://doi.org/10.1080/14737159.2017.1282822.

Debie, E. and Shafi, K. (2019) ‘Implications of the curse of dimensionality for supervised learning classifier systems: theoretical and empirical analyses’, *Pattern Analysis and Applications*, 22(2), pp. 519–536. Available at: https://doi.org/10.1007/s10044-017-0649-0.

Debray, T.P. *et al.* (2019) ‘A framework for meta-analysis of prediction model studies with binary and time-to-event outcomes’, *Statistical Methods in Medical Research*, 28(9), pp. 2768–2786. Available at: https://doi.org/10.1177/0962280218785504.

Delaneau, O., Zagury, J.-F. and Marchini, J. (2013) ‘Improved whole-chromosome phasing for disease and population genetic studies.’, *Nature methods*, 10(1), pp. 5–6. Available at: https://doi.org/10.1038/nmeth.2307.

Deng, H. *et al.* (2021) ‘Ensemble learning for the early prediction of neonatal jaundice with genetic features’, *BMC Medical Informatics and Decision Making*, 21(1), p. 338. Available at: https://doi.org/10.1186/s12911-021-01701-9.

DeTure, M.A. and Dickson, D.W. (2019) ‘The neuropathological diagnosis of Alzheimer’s disease.’, *Molecular neurodegeneration*, 14(1), p. 32. Available at: https://doi.org/10.1186/s13024-019-0333-5.

Dukart, J. *et al.* (2011) ‘Age correction in dementia--matching to a healthy brain.’, *PloS one*, 6(7), p. e22193. Available at: https://doi.org/10.1371/journal.pone.0022193.

Duong, S., Patel, T. and Chang, F. (2017) ‘Dementia: What pharmacists need to know.’, *Canadian pharmacists journal : CPJ = Revue des pharmaciens du Canada : RPC*, 150(2), pp. 118–129. Available at: https://doi.org/10.1177/1715163517690745.

Edler, M.K., Mhatre-Winters, I. and Richardson, J.R. (2021) ‘Microglia in Aging and Alzheimer’s Disease: A Comparative Species Review’, *Cells*, 10(5), p. 1138. Available at: https://doi.org/10.3390/cells10051138.

Elgart, M. *et al.* (2022) ‘Non-linear machine learning models incorporating SNPs and PRS improve polygenic prediction in diverse human populations’, *Communications Biology*, 5(1), p. 856. Available at: https://doi.org/10.1038/s42003-022-03812-z.

Elgeldawi, E. *et al.* (2021) ‘Hyperparameter Tuning for Machine Learning Algorithms Used for Arabic Sentiment Analysis’, *Informatics*, 8(4), p. 79. Available at: https://doi.org/10.3390/informatics8040079.

Elgendy, N. and Elragal, A. (2014) ‘Big Data Analytics: A Literature Review Paper’, in, pp. 214–227. Available at: https://doi.org/10.1007/978-3-319-08976-8_16.

Ertekin-Taner, N. (2007) ‘Genetics of Alzheimer’s disease: a centennial review.’, *Neurologic clinics*, 25(3), pp. 611–67, v. Available at: https://doi.org/10.1016/j.ncl.2007.03.009.

Escott-Price, V., Sims, R., Bannister, C., Harold, D., Vronskaya, M., Majounie, E., Badarinarayan, N., Morgan, K., *et al.* (2015) ‘Common polygenic variation enhances risk prediction for Alzheimer’s disease’, *Brain*, 138(12), pp. 3673–3684. Available at: https://doi.org/10.1093/brain/awv268.

Escott-Price, V., Sims, R., Bannister, C., Harold, D., Vronskaya, M., Majounie, E., Badarinarayan, N., GERAD/PERADES, *et al.* (2015a) ‘Common polygenic variation enhances risk prediction for Alzheimer’s disease.’, *Brain : a journal of neurology*, 138(Pt 12), pp. 3673–84. Available at: https://doi.org/10.1093/brain/awv268.

Escott-Price, V., Sims, R., Bannister, C., Harold, D., Vronskaya, M., Majounie, E., Badarinarayan, N., GERAD/PERADES, *et al.* (2015b) ‘Common polygenic variation enhances risk prediction for Alzheimer’s disease.’, *Brain : a journal of neurology*, 138(Pt 12), pp. 3673–84. Available at: https://doi.org/10.1093/brain/awv268.

Escott-Price, V. *et al.* (2017) ‘Polygenic score prediction captures nearly all common genetic risk for Alzheimer’s disease.’, *Neurobiology of aging*, 49, pp. 214.e7-214.e11. Available at: https://doi.org/10.1016/j.neurobiolaging.2016.07.018.

Falahati, F. *et al.* (2016) ‘The Effect of Age Correction on Multivariate Classification in Alzheimer’s Disease, with a Focus on the Characteristics of Incorrectly and Correctly Classified Subjects.’, *Brain topography*, 29(2), pp. 296–307. Available at: https://doi.org/10.1007/s10548-015-0455-1.

Fan, J., Han, F. and Liu, H. (2014) ‘Challenges of Big Data analysis’, *National Science Review*, 1(2), pp. 293–314. Available at: https://doi.org/10.1093/nsr/nwt032.

Farfel, J.M. *et al.* (2019) ‘Alzheimer’s disease frequency peaks in the tenth decade and is lower afterwards’, *Acta Neuropathologica Communications*, 7(1), p. 104. Available at: https://doi.org/10.1186/s40478-019-0752-0.

Fawagreh, K., Gaber, M.M. and Elyan, E. (2014) 'Random forests: from early developments to recent advancements', *Systems Science & Control Engineering*, 2(1), pp. 602–609. Available at: https://doi.org/10.1080/21642583.2014.956265.

Fernandez, C.G. *et al.* (2019) 'The Role of APOE4 in Disrupting the Homeostatic Functions of Astrocytes and Microglia in Aging and Alzheimer's Disease.', *Frontiers in aging neuroscience*, 11, p. 14. Available at: https://doi.org/10.3389/fnagi.2019.00014.

Ferreira, L.K. and Busatto, G.F. (2011) 'Neuroimaging in Alzheimer's disease: current role in clinical practice and potential future applications.', *Clinics (Sao Paulo, Brazil)*, 66 Suppl 1(Suppl 1), pp. 19–24. Available at: https://doi.org/10.1590/s1807-59322011001300003.

Fillenbaum, G.G. *et al.* (2008) 'Consortium to Establish a Registry for Alzheimer's Disease (CERAD): the first twenty years.', *Alzheimer's & dementia : the journal of the Alzheimer's Association*, 4(2), pp. 96–109. Available at: https://doi.org/10.1016/j.jalz.2007.08.005.

Findeis, M.A. (2007) 'The role of amyloid beta peptide 42 in Alzheimer's disease.', *Pharmacology & therapeutics*, 116(2), pp. 266–86. Available at: https://doi.org/10.1016/j.pharmthera.2007.06.006.

Flach, P. (2019) 'Performance Evaluation in Machine Learning: The Good, the Bad, the Ugly, and the Way Forward', *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01), pp. 9808–9814. Available at: https://doi.org/10.1609/aaai.v33i01.33019808.

Fonseca, P.G. and Lopes, H.D. (2017) 'Calibration of Machine Learning Classifiers for Probability of Default Modelling'.

Forlenza, O. V. *et al.* (2013) 'Mild cognitive impairment (part 1): clinical characteristics and predictors of dementia', *Revista Brasileira de Psiquiatria*, 35(2), pp. 178–185. Available at: https://doi.org/10.1590/1516-4446-2012-3503.

Frederick, T., O'Connor, C. and Koziarski, J. (2018) 'Police Interactions with People Perceived to have a Mental Health Problem: A Critical Review of Frames, Terminology, and Definitions', *Victims & Offenders*, 13(8), pp. 1037–1054. Available at: https://doi.org/10.1080/15564886.2018.1512024.

Freijeiro-González, L., Febrero-Bande, M. and González-Manteiga, W. (2022) 'A Critical Review of LASSO and Its Derivatives for Variable Selection Under Dependence Among Covariates', *International Statistical Review*, 90(1), pp. 118–145. Available at: https://doi.org/10.1111/insr.12469.

Fullerton, J.M. and Nurnberger, J.I. (2019) 'Polygenic risk scores in psychiatry: Will they be useful for clinicians?', *F1000Research*, 8. Available at: https://doi.org/10.12688/f1000research.18491.1.

Gatz, M. *et al.* (2006) 'Role of genes and environments for explaining Alzheimer disease.', *Archives of general psychiatry*, 63(2), pp. 168–74. Available at: https://doi.org/10.1001/archpsyc.63.2.168.

Gaugler, J.E. *et al.* (2013) ‘Characteristics of patients misdiagnosed with Alzheimer’s disease and their medication use: an analysis of the NACC-UDS database’, *BMC Geriatrics*, 13(1), p. 137. Available at: https://doi.org/10.1186/1471-2318-13-137.

Ge, T. *et al.* (2019) ‘Polygenic prediction via Bayesian regression and continuous shrinkage priors.’, *Nature communications*, 10(1), p. 1776. Available at: https://doi.org/10.1038/s41467-019-09718-5.

Ghaffari, H.R. (2021) ‘Speeding up the testing and training time for the support vector machines with minimal effect on the performance’, *The Journal of Supercomputing*, 77(10), pp. 11390–11409. Available at: https://doi.org/10.1007/s11227-021-03729-0.

Ghojogh, B. and Crowley, M. (2019a) ‘The Theory Behind Overfitting, Cross Validation, Regularization, Bagging, and Boosting: Tutorial’.

Ghojogh, B. and Crowley, M. (2019b) ‘The Theory Behind Overfitting, Cross Validation, Regularization, Bagging, and Boosting: Tutorial’.

Ghojogh, B. and Crowley, M. (2019c) ‘The Theory Behind Overfitting, Cross Validation, Regularization, Bagging, and Boosting: Tutorial’.

Gola, D., Erdmann, J., Müller-Myhsok, B., *et al.* (2020) ‘Polygenic risk scores outperform machine learning methods in predicting coronary artery disease status.’, *Genetic epidemiology*, 44(2), pp. 125–138. Available at: https://doi.org/10.1002/gepi.22279.

Gola, D., Erdmann, J., Müller-Myhsok, B., *et al.* (2020) ‘Polygenic risk scores outperform machine learning methods in predicting coronary artery disease status’, *Genetic Epidemiology*, 44(2), pp. 125–138. Available at: https://doi.org/10.1002/gepi.22279.

Gosselin, D. *et al.* (2017) ‘An environment-dependent transcriptional network specifies human microglia identity.’, *Science (New York, N.Y.)*, 356(6344). Available at: https://doi.org/10.1126/science.aal3222.

de Graaf, M.A. *et al.* (2011) ‘Matching, an Appealing Method to Avoid Confounding?’, *Nephron Clinical Practice*, 118(4), pp. c315–c318. Available at: https://doi.org/10.1159/000323136.

Grady, B.J., Torstenson, E.S. and Ritchie, M.D. (2011) ‘The effects of linkage disequilibrium in large scale SNP datasets for MDR’, *BioData Mining*, 4(1), p. 11. Available at: https://doi.org/10.1186/1756-0381-4-11.

Graffelman, J. and Weir, B.S. (2016) ‘Testing for Hardy-Weinberg equilibrium at biallelic genetic markers on the X chromosome.’, *Heredity*, 116(6), pp. 558–68. Available at: https://doi.org/10.1038/hdy.2016.20.

Gross, A.L. *et al.* (2016) ‘Alzheimer’s disease severity, objectively determined and measured’, *Alzheimer’s & Dementia: Diagnosis, Assessment & Disease Monitoring*, 4(1), pp. 159–168. Available at: https://doi.org/10.1016/j.dadm.2016.08.005.

Grueso, S. and Viejo-Sobera, R. (2021) 'Machine learning methods for predicting progression from mild cognitive impairment to Alzheimer's disease dementia: a systematic review.', *Alzheimer's research & therapy*, 13(1), p. 162. Available at: https://doi.org/10.1186/s13195-021-00900-w.

Grupe, A. *et al.* (2007) 'Evidence for novel susceptibility genes for late-onset Alzheimer's disease from a genome-wide association study of putative functional variants.', *Human molecular genetics*, 16(8), pp. 865–73. Available at: https://doi.org/10.1093/hmg/ddm031.

Le Guen, Y. *et al.* (2021) 'A novel age-informed approach for genetic association analysis in Alzheimer's disease', *Alzheimer's Research & Therapy*, 13(1), p. 72. Available at: https://doi.org/10.1186/s13195-021-00808-5.

Guo, C. *et al.* (2017) 'On Calibration of Modern Neural Networks'.
Guo, X. *et al.* (2008) 'On the Class Imbalance Problem', in *2008 Fourth International Conference on Natural Computation*. IEEE, pp. 192–201. Available at: https://doi.org/10.1109/ICNC.2008.871.

Gupta, D. and Rani, R. (2019) 'A study of big data evolution and research challenges', *Journal of Information Science*, 45(3), pp. 322–340. Available at: https://doi.org/10.1177/0165551518789880.

GUZE, S.B. (1995) 'Diagnostic and Statistical Manual of Mental Disorders, 4th ed. (DSM-IV)', *American Journal of Psychiatry*, 152(8), pp. 1228–1228. Available at: https://doi.org/10.1176/ajp.152.8.1228.

Haga, S.B. (2010) 'Impact of limited population diversity of genome-wide association studies', *Genetics in Medicine*, 12(2), pp. 81–84. Available at: https://doi.org/10.1097/GIM.0b013e3181ca2bbf.

Hajian-Tilaki, K. (2013) 'Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation.', *Caspian journal of internal medicine*, 4(2), pp. 627–35.

Halperin, E. and Stephan, D.A. (2009) 'SNP imputation in association studies.', *Nature biotechnology*, 27(4), pp. 349–51. Available at: https://doi.org/10.1038/nbt0409-349.

Hancock, D.B. *et al.* (2012) 'Assessment of Genotype Imputation Performance Using 1000 Genomes in African American Studies', *PLoS ONE*, 7(11), p. e50610. Available at: https://doi.org/10.1371/journal.pone.0050610.
Hand, D.J. (2006) 'Classifier Technology and the Illusion of Progress', *Statistical Science*, 21(1). Available at: https://doi.org/10.1214/088342306000000060.

Hanks, S.C. *et al.* (2022) 'Extent to which array genotyping and imputation with large reference panels approximate deep whole-genome sequencing', *The American Journal of Human Genetics*, 109(9), pp. 1653–1666. Available at: https://doi.org/10.1016/j.ajhg.2022.07.012.

Hannun, A., Guo, C. and van der Maaten, L. (2021) 'Measuring Data Leakage in Machine-Learning Models with Fisher Information'.

Hao, X. *et al.* (2016) 'Identifying Multimodal Intermediate Phenotypes Between Genetic Risk Factors and Disease Status in Alzheimer's Disease', *Neuroinformatics*, 14(4), pp. 439–452. Available at: https://doi.org/10.1007/s12021-016-9307-8.

Hardy, J. and Escott-Price, V. (2019) 'Genes, pathways and risk prediction in Alzheimer's disease', *Human Molecular Genetics* [Preprint]. Available at: https://doi.org/10.1093/hmg/ddz163.

Harold, D. *et al.* (2009) 'Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease.', *Nature genetics*, 41(10), pp. 1088–93. Available at: https://doi.org/10.1038/ng.440.

Hasan, Md.A.M. *et al.* (2016) 'Feature Selection for Intrusion Detection Using Random Forest', *Journal of Information Security*, 07(03), pp. 129–140. Available at: https://doi.org/10.4236/jis.2016.73009.

Hayes, G., Hudspith, R. and Francis, P. (2012) 'P3-104: The Brains for Dementia Research cohort: Demographics of people agreeing to regular assessment and brain donation for research', *Alzheimer's & Dementia*, 8(4S_Part_13). Available at: https://doi.org/10.1016/j.jalz.2012.05.1324.

Hillel, T. *et al.* (2021) 'A systematic review of machine learning classification methodologies for modelling passenger mode choice', *Journal of Choice Modelling*, 38, p. 100221. Available at: https://doi.org/10.1016/j.jocm.2020.100221.

Hira, Z.M. and Gillies, D.F. (2015) 'A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data.', *Advances in bioinformatics*, 2015, p. 198363. Available at: https://doi.org/10.1155/2015/198363.

Ho, D.S.W. *et al.* (2019a) 'Machine Learning SNP Based Prediction for Precision Medicine', *Frontiers in Genetics*, 10. Available at: https://doi.org/10.3389/fgene.2019.00267.

Ho, D.S.W. *et al.* (2019b) 'Machine Learning SNP Based Prediction for Precision Medicine', *Frontiers in Genetics*, 10. Available at: https://doi.org/10.3389/fgene.2019.00267.

Hosseini, M. *et al.* (2020) 'I tried a bunch of things: The dangers of unexpected overfitting in classification of brain data', *Neuroscience & Biobehavioral Reviews*, 119, pp. 456–467. Available at: https://doi.org/10.1016/j.neubiorev.2020.09.036.

Husain, M.A., Laurent, B. and Plourde, M. (2021) 'APOE and Alzheimer's Disease: From Lipid Transport to Physiopathology and Therapeutics.', *Frontiers in neuroscience*, 15, p. 630502. Available at: https://doi.org/10.3389/fnins.2021.630502.

Ibrahim, A.M. and Bennett, B. (2014) 'The Assessment of Machine Learning Model Performance for Predicting Alluvial Deposits Distribution', *Procedia Computer Science*, 36, pp. 637–642. Available at: https://doi.org/10.1016/j.procs.2014.09.067.

Iddi, S. *et al.* (2019) 'Predicting the course of Alzheimer's progression', *Brain Informatics*, 6(1), p. 6. Available at: https://doi.org/10.1186/s40708-019-0099-0.

Iwata, H. and Jannink, J. (2010) ‘Marker Genotype Imputation in a Low-Marker-Density Panel with a High-Marker-Density Reference Panel: Accuracy Evaluation in Barley Breeding Lines’, *Crop Science*, 50(4), pp. 1269–1278. Available at: https://doi.org/10.2135/cropsci2009.08.0434.

Jack, C.R. *et al.* (2011) ‘Introduction to the recommendations from the National Institute on Aging-Alzheimer’s Association workgroups on diagnostic guidelines for Alzheimer’s disease.’, *Alzheimer’s & dementia : the journal of the Alzheimer’s Association*, 7(3), pp. 257–62. Available at: https://doi.org/10.1016/j.jalz.2011.03.004.

James, G. *et al.* (2013) *An Introduction to Statistical Learning*. New York, NY: Springer New York. Available at: https://doi.org/10.1007/978-1-4614-7138-7.

Jansen, I.E. *et al.* (2019) ‘Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer’s disease risk.’, *Nature genetics*, 51(3), pp. 404–413. Available at: https://doi.org/10.1038/s41588-018-0311-9.

Jekel, K. *et al.* (2015) ‘Mild cognitive impairment and deficits in instrumental activities of daily living: a systematic review’, *Alzheimer’s Research & Therapy*, 7(1), p. 17. Available at: https://doi.org/10.1186/s13195-015-0099-0.

Jiang, Z. *et al.* (2021) ‘A New Oversampling Method Based on the Classification Contribution Degree’, *Symmetry*, 13(2), p. 194. Available at: https://doi.org/10.3390/sym13020194.

Johnson, A.D. (2009) ‘Single-nucleotide polymorphism bioinformatics: a comprehensive review of resources.’, *Circulation. Cardiovascular genetics*, 2(5), pp. 530–6. Available at: https://doi.org/10.1161/CIRCGENETICS.109.872010.

Johnson, J.M. and Khoshgoftaar, T.M. (2019) ‘Survey on deep learning with class imbalance’, *Journal of Big Data*, 6(1), p. 27. Available at: https://doi.org/10.1186/s40537-019-0192-5.

Jones, L., Holmans, Peter A., *et al.* (2010) ‘Genetic Evidence Implicates the Immune System and Cholesterol Metabolism in the Aetiology of Alzheimer’s Disease’, *PLoS ONE*, 5(11), p. e13950. Available at: https://doi.org/10.1371/journal.pone.0013950.

Jones, L., Holmans, Peter A, *et al.* (2010) ‘Genetic evidence implicates the immune system and cholesterol metabolism in the aetiology of Alzheimer’s disease.’, *PloS one*, 5(11), p. e13950. Available at: https://doi.org/10.1371/journal.pone.0013950.

Jovic, A., Brkic, K. and Bogunovic, N. (2015) ‘A review of feature selection methods with applications’, in *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. IEEE, pp. 1200–1205. Available at: https://doi.org/10.1109/MIPRO.2015.7160458.

Kalina, J. (2014) ‘Classification methods for high-dimensional genetic data’, *Biocybernetics and Biomedical Engineering*, 34(1), pp. 10–18. Available at: https://doi.org/10.1016/j.bbe.2013.09.007.

Kang, H. (2013) 'The prevention and handling of the missing data.', *Korean journal of anesthesiology*, 64(5), pp. 402–6. Available at: https://doi.org/10.4097/kjae.2013.64.5.402.

Kappen, T.H. *et al.* (2018) 'Evaluating the impact of prediction models: lessons learned, challenges, and recommendations', *Diagnostic and Prognostic Research*, 2(1), p. 11. Available at: https://doi.org/10.1186/s41512-018-0033-6.

Karamizadeh, S. *et al.* (2013) 'An Overview of Principal Component Analysis', *Journal of Signal and Information Processing*, 04(03), pp. 173–175. Available at: https://doi.org/10.4236/jsip.2013.43B031.

Karanicolas, P.J., Farrokhyar, F. and Bhandari, M. (2010) 'Practical tips for surgical research: blinding: who, what, when, why, how?', *Canadian journal of surgery. Journal canadien de chirurgie*, 53(5), pp. 345–8.

Kassraian-Fard, P. *et al.* (2016) 'Promises, Pitfalls, and Basic Guidelines for Applying Machine Learning Classifiers to Psychiatric Imaging Data, with Autism as an Example', *Frontiers in Psychiatry*, 7. Available at: https://doi.org/10.3389/fpsyt.2016.00177.

Kelly, C.J. *et al.* (2019) 'Key challenges for delivering clinical impact with artificial intelligence', *BMC Medicine*, 17(1), p. 195. Available at: https://doi.org/10.1186/s12916-019-1426-2.

Khan, A.U.M. (2015) 'Early diagnosis of Alzheimer.s disease using machine learning techniques: A review paper', *IEEE* [Preprint].

Khan, S.I. and Hoque, A.S.M.L. (2020) 'SICE: an improved missing data imputation technique', *Journal of Big Data*, 7(1), p. 37. Available at: https://doi.org/10.1186/s40537-020-00313-w.

Kirasich, K., Smith, T. and Sadler, B. (2018) 'Random Forest vs Logistic Regression: Binary Classification for Heterogeneous Datasets', *Data Science Review*, 1(3).

Klatka, L.A. *et al.* (1996) 'Incorrect diagnosis of Alzheimer's disease. A clinicopathologic study.', *Archives of neurology*, 53(1), pp. 35–42. Available at: https://doi.org/10.1001/archneur.1996.00550010045015.

Knopman, D.S. and Petersen, R.C. (2014) 'Mild cognitive impairment and mild dementia: a clinical perspective.', *Mayo Clinic proceedings*, 89(10), pp. 1452–9. Available at: https://doi.org/10.1016/j.mayocp.2014.06.019.
Koedam, E.L.G.E. *et al.* (2010) 'Early-versus late-onset Alzheimer's disease: more than age alone.', *Journal of Alzheimer's disease : JAD*, 19(4), pp. 1401–8. Available at: https://doi.org/10.3233/JAD-2010-1337.

Koopmans, F. *et al.* (2019) 'SynGO: An Evidence-Based, Expert-Curated Knowledge Base for the Synapse.', *Neuron*, 103(2), pp. 217-234.e4. Available at: https://doi.org/10.1016/j.neuron.2019.05.002.

Korte, A. and Farlow, A. (2013) 'The advantages and limitations of trait analysis with GWAS: a review', *Plant Methods*, 9(1), p. 29. Available at: https://doi.org/10.1186/1746-4811-9-29.

K.P., M.N. and P., T. (2022) 'Feature selection using efficient fusion of Fisher Score and greedy searching for Alzheimer's classification', *Journal of King Saud University - Computer and Information Sciences*, 34(8), pp. 4993–5006. Available at: https://doi.org/10.1016/j.jksuci.2020.12.009.

Krithika, S. *et al.* (2012) 'Evaluation of the imputation performance of the program IMPUTE in an admixed sample from Mexico City using several model designs.', *BMC medical genomics*, 5, p. 12. Available at: https://doi.org/10.1186/1755-8794-5-12.

Kumar, Er.P.S.Er.P. (2014) 'Artificial Neural Networks-A Atudy', *International Journal of Emerging Engineering Research and Technology*, 2(2), pp. 143–148.

Kumar, R. and Indrayan, A. (2011) 'Receiver operating characteristic (ROC) curve for medical researchers.', *Indian pediatrics*, 48(4), pp. 277–87. Available at: https://doi.org/10.1007/s13312-011-0055-4.

Kumar, Y. *et al.* (2022) 'Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda.', *Journal of ambient intelligence and humanized computing*, pp. 1–28. Available at: https://doi.org/10.1007/s12652-021-03612-z.

Kumari, K. and Yadav, S. (2018) 'Linear regression analysis study', *Journal of the Practice of Cardiovascular Sciences*, 4(1), p. 33. Available at: https://doi.org/10.4103/jpcs.jpcs_8_18.

Kunkle, B.W. *et al.* (2019) 'Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing', *Nature Genetics*, 51(3), pp. 414–430. Available at: https://doi.org/10.1038/s41588-019-0358-2.

Lambert, J.-C., Heath, S., Even, G., Campion, D., Sleegers, K., Amouyel, P., *et al.* (2009a) 'Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease.', *Nature genetics*, 41(10), pp. 1094–9. Available at: https://doi.org/10.1038/ng.439.

Lambert, J.-C., Heath, S., Even, G., Campion, D., Sleegers, K., Amouyel, P., *et al.* (2009b) 'Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease.', *Nature genetics*, 41(10), pp. 1094–9. Available at: https://doi.org/10.1038/ng.439.

Lambert, J.-C., Heath, S., Even, G., Campion, D., Sleegers, K., Hiltunen, M., *et al.* (2009) 'Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease', *Nature Genetics*, 41(10), pp. 1094–1099. Available at: https://doi.org/10.1038/ng.439.

Lambert, J.C. *et al.* (2013) 'Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease.', *Nature genetics*, 45(12), pp. 1452–8. Available at: https://doi.org/10.1038/ng.2802.

Langfelder, P. and Horvath, S. (2008) 'WGCNA: an R package for weighted correlation network analysis', *BMC Bioinformatics*, 9(1), p. 559. Available at: https://doi.org/10.1186/1471-2105-9-559.

Langley, P. and Sage, S. (2013) 'Induction of Selective Bayesian Classifiers'.
Latha, P.H. and Mohanasundaram, R. (2019) 'A New Hybrid Strategy for Malware Detection Classification with Multiple Feature Selection Methods and Ensemble Learning Methods', *International Journal of Engineering and Advanced Technology*, 9(2), pp. 4013–4018. Available at: https://doi.org/10.35940/ijeat.B4666.129219.

Lee, C.A. *et al.* (2011) 'Recent Developments in High Performance Computing for Remote Sensing: A Review', *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 4(3), pp. 508–527. Available at: https://doi.org/10.1109/JSTARS.2011.2162643.

Lee, J.H. (2008) 'Analyses of the National Institute on Aging Late-Onset Alzheimer's Disease Family Study', *Archives of Neurology*, 65(11), p. 1518. Available at: https://doi.org/10.1001/archneur.65.11.1518.

Leonenko, G., Shoai, M., *et al.* (2019) 'Genetic risk for alzheimer disease is distinct from genetic risk for amyloid deposition.', *Annals of neurology*, 86(3), pp. 427–435. Available at: https://doi.org/10.1002/ana.25530.

Leonenko, G., Sims, R., *et al.* (2019) 'Polygenic risk and hazard scores for Alzheimer's disease prediction', *Annals of Clinical and Translational Neurology*, 6(3), pp. 456–465. Available at: https://doi.org/10.1002/acn3.716.

Leonenko, G. *et al.* (2021) 'Identifying individuals with high risk of Alzheimer's disease using polygenic risk scores', *Nature Communications*, 12(1), p. 4506. Available at: https://doi.org/10.1038/s41467-021-24082-z.

Lewis, C.M. and Vassos, E. (2020a) 'Polygenic risk scores: from research tools to clinical instruments', *Genome Medicine*, 12(1), p. 44. Available at: https://doi.org/10.1186/s13073-020-00742-5.

Lewis, C.M. and Vassos, E. (2020b) 'Polygenic risk scores: from research tools to clinical instruments', *Genome Medicine*, 12(1), p. 44. Available at: https://doi.org/10.1186/s13073-020-00742-5.

Lewis, C.M. and Vassos, E. (2020c) 'Polygenic risk scores: from research tools to clinical instruments', *Genome Medicine*, 12(1), p. 44. Available at: https://doi.org/10.1186/s13073-020-00742-5.

Li, X.-L. *et al.* (2014) 'Behavioral and Psychological Symptoms in Alzheimer's Disease', *BioMed Research International*, 2014, pp. 1–9. Available at: https://doi.org/10.1155/2014/927804.

Li, Y. *et al.* (2009a) 'Genotype imputation.', *Annual review of genomics and human genetics*, 10, pp. 387–406. Available at: https://doi.org/10.1146/annurev.genom.9.081307.164242.

Li, Y. *et al.* (2009b) ‘Genotype imputation.’, *Annual review of genomics and human genetics*, 10, pp. 387–406. Available at: https://doi.org/10.1146/annurev.genom.9.081307.164242.

Liberati, A. *et al.* (2009) ‘The PRISMA Statement for Reporting Systematic Reviews and Meta-Analyses of Studies That Evaluate Health Care Interventions: Explanation and Elaboration’, *PLoS Medicine*, 6(7), p. e1000100. Available at: https://doi.org/10.1371/journal.pmed.1000100.

Lippmann, R. (1987) ‘An introduction to computing with neural nets’, *IEEE ASSP Magazine*, 4(2), pp. 4–22. Available at: https://doi.org/10.1109/MASSP.1987.1165576.

Liu, Q. and Wu, Y. (2012) ‘Supervised Learning’, in *Encyclopedia of the Sciences of Learning*. Boston, MA: Springer US, pp. 3243–3245. Available at: https://doi.org/10.1007/978-1-4419-1428-6_451.

Lvovs, D., Favorova, O.O. and Favorov, A. V (2012) ‘A Polygenic Approach to the Study of Polygenic Diseases.’, *Acta naturae*, 4(3), pp. 59–71.

M, H. and M.N, S. (2015) ‘A Review on Evaluation Metrics for Data Classification Evaluations’, *International Journal of Data Mining & Knowledge Management Process*, 5(2), pp. 01–11. Available at: https://doi.org/10.5121/ijdkp.2015.5201.

Ma, H. and Qin, M. (2009) ‘Research Method of Customer Churn Crisis Based on Decision Tree’, in *2009 International Conference on Management and Service Science*. IEEE, pp. 1–4. Available at: https://doi.org/10.1109/ICMSS.2009.5305403.

Majnik, M. and Bosnić, Z. (2013) ‘ROC analysis of classifiers in machine learning: A survey’, *Intelligent Data Analysis*, 17(3), pp. 531–558. Available at: https://doi.org/10.3233/IDA-130592.

Maleki, F. *et al.* (2020a) ‘Machine Learning Algorithm Validation’, *Neuroimaging Clinics of North America*, 30(4), pp. 433–445. Available at: https://doi.org/10.1016/j.nic.2020.08.004.

Maleki, F. *et al.* (2020b) ‘Machine Learning Algorithm Validation’, *Neuroimaging Clinics of North America*, 30(4), pp. 433–445. Available at: https://doi.org/10.1016/j.nic.2020.08.004.

Maleki, F. *et al.* (2020c) ‘Machine Learning Algorithm Validation’, *Neuroimaging Clinics of North America*, 30(4), pp. 433–445. Available at: https://doi.org/10.1016/j.nic.2020.08.004.

Manabe, T. *et al.* (2019) ‘Pneumonia-associated death in patients with dementia: A systematic review and meta-analysis.’, *PloS one*, 14(3), p. e0213825. Available at: https://doi.org/10.1371/journal.pone.0213825.

Manthena, V. *et al.* (2022) ‘Evaluating dimensionality reduction for genomic prediction’, *Frontiers in Genetics*, 13. Available at: https://doi.org/10.3389/fgene.2022.958780.

Marcus, C., Mena, E. and Subramaniam, R.M. (2014) ‘Brain PET in the diagnosis of Alzheimer’s disease.’, *Clinical nuclear medicine*, 39(10), pp. e413-22; quiz e423-6. Available at: https://doi.org/10.1097/RLU.0000000000000547.

Marioni, R.E. *et al.* (2018) ‘GWAS on family history of Alzheimer’s disease.’, *Translational psychiatry*, 8(1), p. 99. Available at: https://doi.org/10.1038/s41398-018-0150-6.

Martin, G.P. *et al.* (2017) ‘Clinical prediction in defined populations: a simulation study investigating when and how to aggregate existing models’, *BMC Medical Research Methodology*, 17(1), p. 1. Available at: https://doi.org/10.1186/s12874-016-0277-1.

Marttinen, P. *et al.* (2013) ‘Genome-wide association studies with high-dimensional phenotypes’, *Statistical Applications in Genetics and Molecular Biology*, 12(4). Available at: https://doi.org/10.1515/sagmb-2012-0032.

McCarroll, S.A. and Hyman, S.E. (2013) ‘Progress in the genetics of polygenic brain disorders: significant new challenges for neurobiology.’, *Neuron*, 80(3), pp. 578–87. Available at: https://doi.org/10.1016/j.neuron.2013.10.046.

McKhann, G.M. *et al.* (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), pp. 263–269. Available at: https://doi.org/10.1016/j.jalz.2011.03.005.

McNeish, D.M. (2015) ‘Using Lasso for Predictor Selection and to Assuage Overfitting: A Method Long Overlooked in Behavioral Sciences’, *Multivariate Behavioral Research*, 50(5), pp. 471–484. Available at: https://doi.org/10.1080/00273171.2015.1036965.

Medina-Gomez, C. *et al.* (2015) ‘Challenges in conducting genome-wide association studies in highly admixed multi-ethnic populations: the Generation R Study.’, *European journal of epidemiology*, 30(4), pp. 317–30. Available at: https://doi.org/10.1007/s10654-015-9998-4.

Mena Mamani, N. (2020) ‘Machine Learning techniques and Polygenic Risk Score application to prediction genetic diseases’, *ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal*, 9(1), pp. 5–14. Available at: https://doi.org/10.14201/ADCAIJ202091514.

Mendez, M.F. (2017) ‘Early-Onset Alzheimer Disease.’, *Neurologic clinics*, 35(2), pp. 263–281. Available at: https://doi.org/10.1016/j.ncl.2017.01.005.

Mendez, M.F. (2019) ‘Early-onset Alzheimer Disease and Its Variants.’, *Continuum (Minneapolis, Minn.)*, 25(1), pp. 34–51. Available at: https://doi.org/10.1212/CON.0000000000000687.

Mishra, R. and Li, B. (2020) ‘The Application of Artificial Intelligence in the Genetic Study of Alzheimer’s Disease’, *Aging and disease*, 11(6), p. 1567. Available at: https://doi.org/10.14336/AD.2020.0312.

Misra, S. and Li, H. (2020) ‘Noninvasive fracture characterization based on the classification of sonic wave travel times’, in *Machine Learning for Subsurface Characterization*. Elsevier, pp. 243–287. Available at: https://doi.org/10.1016/B978-0-12-817736-5.00009-0.

Mitt, M. *et al.* (2017) 'Improved imputation accuracy of rare and low-frequency variants using population-specific high-coverage WGS-based imputation reference panel', *European Journal of Human Genetics*, 25(7), pp. 869–876. Available at: https://doi.org/10.1038/ejhg.2017.51.

Moloney, C.M., Lowe, V.J. and Murray, M.E. (2021) 'Visualization of neurofibrillary tangle maturity in Alzheimer's disease: A clinicopathologic perspective for biomarker research.', *Alzheimer's & dementia : the journal of the Alzheimer's Association*, 17(9), pp. 1554–1574. Available at: https://doi.org/10.1002/alz.12321.

Moons, K.G.M. *et al.* (2014) 'Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies: The CHARMS Checklist', *PLoS Medicine*, 11(10), p. e1001744. Available at: https://doi.org/10.1371/journal.pmed.1001744.

Moore, J.H., Asselbergs, F.W. and Williams, S.M. (2010) 'Bioinformatics challenges for genome-wide association studies.', *Bioinformatics (Oxford, England)*, 26(4), pp. 445–55. Available at: https://doi.org/10.1093/bioinformatics/btp713.

Morgan, A.A., Chen, R. and Butte, A.J. (2012) 'Clinical utility of sequence-based genotype compared with that derivable from genotyping arrays.', *Journal of the American Medical Informatics Association : JAMIA*, 19(e1), pp. e21-7. Available at: https://doi.org/10.1136/amiajnl-2011-000737.

Moskvina, V. and Schmidt, K.M. (2008) 'On multiple-testing correction in genome-wide association studies.', *Genetic epidemiology*, 32(6), pp. 567–73. Available at: https://doi.org/10.1002/gepi.20331.

Mostafa Abd El Hamid, M., Omar, Y.M.K. and Mabrouk, M.S. (2016) 'Identifying genetic biomarkers associated to Alzheimer's disease using Support Vector Machine', in *2016 8th Cairo International Biomedical Engineering Conference (CIBEC)*. IEEE, pp. 5–9. Available at: https://doi.org/10.1109/CIBEC.2016.7836087.

Muir, P. *et al.* (2016) 'The real cost of sequencing: scaling computation to keep pace with data generation', *Genome Biology*, 17(1), p. 53. Available at: https://doi.org/10.1186/s13059-016-0917-0.

Musoro, J.Z. *et al.* (2014) 'Validation of prediction models based on lasso regression with multiply imputed data', *BMC Medical Research Methodology*, 14(1), p. 116. Available at: https://doi.org/10.1186/1471-2288-14-116.

Nahm, F.S. (2022) 'Receiver operating characteristic curve: overview and practical use for clinicians.', *Korean journal of anesthesiology*, 75(1), pp. 25–36. Available at: https://doi.org/10.4097/kja.21209.

Naj, A.C. *et al.* (2011) 'Common variants at MS4A4/MS4A6E, CD2AP, CD33 and EPHA1 are associated with late-onset Alzheimer's disease.', *Nature genetics*, 43(5), pp. 436–41. Available at: https://doi.org/10.1038/ng.801.

Najafabadi, M.M. *et al.* (2015) 'Deep learning applications and challenges in big data analytics', *Journal of Big Data*, 2(1), p. 1. Available at: https://doi.org/10.1186/s40537-014-0007-7.

Namipashaki, A.R.-M.Z.A.-P.N. (2015) 'The Essentiality of Reporting Hardy-Weinberg Equilibrium Calculations in Population-Based Genetic Association Studies', *Cell Journal*, 17(2), pp. 187–192.

Natekin, A. and Knoll, A. (2013) 'Gradient boosting machines, a tutorial.', *Frontiers in neurorobotics*, 7, p. 21. Available at: https://doi.org/10.3389/fnbot.2013.00021.

Neal, B. *et al.* (2018) 'A Modern Take on the Bias-Variance Tradeoff in Neural Networks'.

Nebel, R.A. *et al.* (2018) 'Understanding the impact of sex and gender in Alzheimer's disease: A call to action.', *Alzheimer's & dementia : the journal of the Alzheimer's Association*, 14(9), pp. 1171–1183. Available at: https://doi.org/10.1016/j.jalz.2018.04.008.

Network and Pathway Analysis Subgroup of Psychiatric Genomics Consortium (2015) 'Psychiatric genome-wide association study analyses implicate neuronal, immune and histone pathways.', *Nature neuroscience*, 18(2), pp. 199–209. Available at: https://doi.org/10.1038/nn.3922.

Nichols, E. *et al.* (2022) 'Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: an analysis for the Global Burden of Disease Study 2019', *The Lancet Public Health*, 7(2), pp. e105–e125. Available at: https://doi.org/10.1016/S2468-2667(21)00249-8.

Nixon, J. *et al.* (2019) 'Measuring Calibration in Deep Learning'.

Oshiro, T.M., Perez, P.S. and Baranauskas, J.A. (2012) 'How Many Trees in a Random Forest?', in, pp. 154–168. Available at: https://doi.org/10.1007/978-3-642-31537-4_13.

Osipowicz, M. *et al.* (2021) 'Careful feature selection is key in classification of Alzheimer's disease patients based on whole-genome sequencing data.', *NAR genomics and bioinformatics*, 3(3), p. lqab069. Available at: https://doi.org/10.1093/nargab/lqab069.

Palmqvist, S. *et al.* (2012) 'Comparison of brief cognitive tests and CSF biomarkers in predicting Alzheimer's disease in mild cognitive impairment: six-year follow-up study.', *PloS one*, 7(6), p. e38639. Available at: https://doi.org/10.1371/journal.pone.0038639.

Park, Y. and Ho, J.C. (2020) 'CaliForest', in *Proceedings of the ACM Conference on Health, Inference, and Learning*. New York, NY, USA: ACM, pp. 40–50. Available at: https://doi.org/10.1145/3368555.3384461.

Parvandeh, S. *et al.* (2020) 'Consensus features nested cross-validation', *Bioinformatics*, 36(10), pp. 3093–3098. Available at: https://doi.org/10.1093/bioinformatics/btaa046.

Pavlou, M. *et al.* (2015a) 'How to develop a more accurate risk prediction model when there are few events', *BMJ*, p. h3868. Available at: https://doi.org/10.1136/bmj.h3868.

Pavlou, M. *et al.* (2015b) 'How to develop a more accurate risk prediction model when there are few events', *BMJ*, p. h3868. Available at: https://doi.org/10.1136/bmj.h3868.

Pe'er, I. *et al.* (2008) 'Estimation of the multiple testing burden for genomewide association studies of nearly all common variants.', *Genetic epidemiology*, 32(4), pp. 381–5. Available at: https://doi.org/10.1002/gepi.20303.

Perl, D.P. (2010) 'Neuropathology of Alzheimer's disease.', *The Mount Sinai journal of medicine, New York*, 77(1), pp. 32–42. Available at: https://doi.org/10.1002/msj.20157.

Petersen, R. C. *et al.* (2010a) 'Alzheimer's Disease Neuroimaging Initiative (ADNI): Clinical characterization', *Neurology*, 74(3), pp. 201–209. Available at: https://doi.org/10.1212/WNL.0b013e3181cb3e25.

Petersen, R C *et al.* (2010) 'Alzheimer's Disease Neuroimaging Initiative (ADNI): clinical characterization.', *Neurology*, 74(3), pp. 201–9. Available at: https://doi.org/10.1212/WNL.0b013e3181cb3e25.

Petersen, R. C. *et al.* (2010b) 'Alzheimer's Disease Neuroimaging Initiative (ADNI): Clinical characterization', *Neurology*, 74(3), pp. 201–209. Available at: https://doi.org/10.1212/WNL.0b013e3181cb3e25.

Platt, J. (1999) 'Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods'.

van der Ploeg, T., Austin, P.C. and Steyerberg, E.W. (2014) 'Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints', *BMC Medical Research Methodology*, 14(1), p. 137. Available at: https://doi.org/10.1186/1471-2288-14-137.

Pretorius, A., Bierman, S. and Steel, S.J. (2016) 'A meta-analysis of research in random forests for classification', in *2016 Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech)*. IEEE, pp. 1–6. Available at: https://doi.org/10.1109/RoboMech.2016.7813171.

Price, A.L. *et al.* (2006) 'Principal components analysis corrects for stratification in genome-wide association studies', *Nature Genetics*, 38(8), pp. 904–909. Available at: https://doi.org/10.1038/ng1847.

Privé, F. *et al.* (2019) 'Making the Most of Clumping and Thresholding for Polygenic Scores.', *American journal of human genetics*, 105(6), pp. 1213–1221. Available at: https://doi.org/10.1016/j.ajhg.2019.11.001.

Probst, P., Bischl, B. and Boulesteix, A.-L. (2018a) 'Tunability: Importance of Hyperparameters of Machine Learning Algorithms'.

Probst, P., Bischl, B. and Boulesteix, A.-L. (2018b) 'Tunability: Importance of Hyperparameters of Machine Learning Algorithms'.

Probst, P., Wright, M. and Boulesteix, A.-L. (2018) ‘Hyperparameters and Tuning Strategies for Random Forest’. Available at: https://doi.org/10.1002/widm.1301.

Prokopenko, D. *et al.* (2021) ‘Whole-genome sequencing reveals new Alzheimer’s disease–associated rare variants in loci related to synaptic function and neuronal development’, *Alzheimer’s & Dementia*, 17(9), pp. 1509–1527. Available at: https://doi.org/10.1002/alz.12319.

Prosvirin, A., Duong, B.P. and Kim, J.-M. (2019) ‘SVM Hyperparameter Optimization Using a Genetic Algorithm for Rub-Impact Fault Diagnosis’, in, pp. 155–165. Available at: https://doi.org/10.1007/978-981-13-6861-5_14.

Prusty, S., Patnaik, S. and Dash, S.K. (2022) ‘SKCV: Stratified K-fold cross-validation on ML classifiers for predicting cervical cancer’, *Frontiers in Nanotechnology*, 4. Available at: https://doi.org/10.3389/fnano.2022.972421.

Psaty, B.M. *et al.* (2009) ‘Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium: Design of prospective meta-analyses of genome-wide association studies from 5 cohorts.’, *Circulation. Cardiovascular genetics*, 2(1), pp. 73–80. Available at: https://doi.org/10.1161/CIRCGENETICS.108.829747.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, Manuel A R, *et al.* (2007) ‘PLINK: a tool set for whole-genome association and population-based linkage analyses.’, *American journal of human genetics*, 81(3), pp. 559–75. Available at: https://doi.org/10.1086/519795.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, Manuel A.R., *et al.* (2007) ‘PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses’, *The American Journal of Human Genetics*, 81(3), pp. 559–575. Available at: https://doi.org/10.1086/519795.

Qiu, C., Kivipelto, M. and von Strauss, E. (2009) ‘Epidemiology of Alzheimer’s disease: occurrence, determinants, and strategies toward intervention.’, *Dialogues in clinical neuroscience*, 11(2), pp. 111–28. Available at: https://doi.org/10.31887/DCNS.2009.11.2/cqiu.

Radovic, M. *et al.* (2017) ‘Minimum redundancy maximum relevance feature selection approach for temporal gene expression data’, *BMC Bioinformatics*, 18(1), p. 9. Available at: https://doi.org/10.1186/s12859-016-1423-9.

Raiha, I. *et al.* (1996) ‘Alzheimer’s disease in Finnish twins.’, *Lancet (London, England)*, 347(9001), pp. 573–8. Available at: https://doi.org/10.1016/s0140-6736(96)91272-6.

Raileanu, L.E. and Stoffel, K. (2004) ‘Theoretical Comparison between the Gini Index and Information Gain Criteria’, *Annals of Mathematics and Artificial Intelligence*, 41(1), pp. 77–93. Available at: https://doi.org/10.1023/B:AMAI.0000018580.96245.c6.

Rait, G. *et al.* (2010) ‘Survival of people with clinical diagnosis of dementia in primary care: cohort study.’, *BMJ (Clinical research ed.)*, 341, p. c3584. Available at: https://doi.org/10.1136/bmj.c3584.

Rajput, D., Wang, W.-J. and Chen, C.-C. (2023) ‘Evaluation of a decided sample size in machine learning applications’, *BMC Bioinformatics*, 24(1), p. 48. Available at: https://doi.org/10.1186/s12859-023-05156-9.

Ramspek, C.L. *et al.* (2021) ‘External validation of prognostic models: what, why, how, when and where?’, *Clinical Kidney Journal*, 14(1), pp. 49–58. Available at: https://doi.org/10.1093/ckj/sfaa188.

Rasmussen, J. and Langerman, H. (2019) ‘Alzheimer’s Disease - Why We Need Early Diagnosis.’, *Degenerative neurological and neuromuscular disease*, 9, pp. 123–130. Available at: https://doi.org/10.2147/DNND.S228939.

Reiman, E.M. *et al.* (2007) ‘GAB2 Alleles Modify Alzheimer’s Risk in APOE ε4 Carriers’, *Neuron*, 54(5), pp. 713–720. Available at: https://doi.org/10.1016/j.neuron.2007.05.022.

Ricciarelli, R. and Fedele, E. (2017) ‘The Amyloid Cascade Hypothesis in Alzheimer’s Disease: It’s Time to Change Our Mind.’, *Current neuropharmacology*, 15(6), pp. 926–935. Available at: https://doi.org/10.2174/1570159X15666170116143743.

Ridge, Perry G. *et al.* (2013) ‘Alzheimer’s Disease: Analyzing the Missing Heritability’, *PLoS ONE*, 8(11), p. e79771. Available at: https://doi.org/10.1371/journal.pone.0079771.

Ridge, Perry G *et al.* (2013) ‘Alzheimer’s disease: analyzing the missing heritability.’, *PloS one*, 8(11), p. e79771. Available at: https://doi.org/10.1371/journal.pone.0079771.

Robbins, M., Clayton, E. and Kaminski Schierle, G.S. (2021) ‘Synaptic tau: A pathological or physiological phenomenon?’, *Acta neuropathologica communications*, 9(1), p. 149. Available at: https://doi.org/10.1186/s40478-021-01246-y.

Romero-Rosales, B.-L. *et al.* (2020a) ‘Improving predictive models for Alzheimer’s disease using GWAS data by incorporating misclassified samples modeling’, *PLOS ONE*, 15(4), p. e0232103. Available at: https://doi.org/10.1371/journal.pone.0232103.

Romero-Rosales, B.-L. *et al.* (2020b) ‘Improving predictive models for Alzheimer’s disease using GWAS data by incorporating misclassified samples modeling’, *PLOS ONE*, 15(4), p. e0232103. Available at: https://doi.org/10.1371/journal.pone.0232103.

Roshyara, N.R. *et al.* (2016) ‘Comparing performance of modern genotype imputation methods in different ethnicities’, *Scientific Reports*, 6(1), p. 34386. Available at: https://doi.org/10.1038/srep34386.

Roth, T.C., Krochmal, A.R. and Németh, Z. (2015) ‘Thinking about Change: An Integrative Approach for Examining Cognition in a Changing World’, *Integrative and Comparative Biology*, 55(3), pp. 347–353. Available at: https://doi.org/10.1093/icb/icv068.

Rountree, S.D. *et al.* (2012) ‘Factors that influence survival in a probable Alzheimer disease cohort.’, *Alzheimer’s research & therapy*, 4(3), p. 16. Available at: https://doi.org/10.1186/alzrt119.

Rowe, T.W. *et al.* (2021a) ‘Machine learning for the life-time risk prediction of Alzheimer’s disease: a systematic review’, *Brain Communications*, 3(4). Available at: https://doi.org/10.1093/braincomms/fcab246.

Rowe, T.W. *et al.* (2021b) ‘Machine learning for the life-time risk prediction of Alzheimer’s disease: a systematic review’, *Brain Communications*, 3(4). Available at: https://doi.org/10.1093/braincomms/fcab246.

Ruder, S. (2016) ‘An overview of gradient descent optimization algorithms’.

Ryo, M. and Rillig, M.C. (2017) ‘Statistically reinforced machine learning for nonlinear patterns and variable interactions’, *Ecosphere*, 8(11), p. e01976. Available at: https://doi.org/10.1002/ecs2.1976.

Safieh, M., Korczyn, A.D. and Michaelson, D.M. (2019) ‘ApoE4: an emerging therapeutic target for Alzheimer’s disease.’, *BMC medicine*, 17(1), p. 64. Available at: https://doi.org/10.1186/s12916-019-1299-4.

Sahu, M. and Prasuna, J.G. (2016) ‘Twin Studies: A Unique Epidemiological Tool.’, *Indian journal of community medicine : official publication of Indian Association of Preventive & Social Medicine*, 41(3), pp. 177–82. Available at: https://doi.org/10.4103/0970-0218.183593.

Salman, S. and Liu, X. (2019) ‘Overfitting Mechanism and Avoidance in Deep Neural Networks’.

Samala, R.K. *et al.* (2021) ‘Risks of feature leakage and sample size dependencies in deep feature extraction for breast mass classification’, *Medical Physics*, 48(6), pp. 2827–2837. Available at: https://doi.org/10.1002/mp.14678.

Sánchez García, J. and Cruz Rambaud, S. (2022) ‘Machine Learning Regularization Methods in High-Dimensional Monetary and Financial VARs’, *Mathematics*, 10(6), p. 877. Available at: https://doi.org/10.3390/math10060877.

Sánchez-Maroño, N., Alonso-Betanzos, A. and Tombilla-Sanromán, M. (no date) ‘Filter Methods for Feature Selection – A Comparative Study’, in *Intelligent Data Engineering and Automated Learning - IDEAL 2007*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 178–187. Available at: https://doi.org/10.1007/978-3-540-77226-2_19.

Sarica, A., Cerasa, A. and Quattrone, A. (2017) ‘Random Forest Algorithm for the Classification of Neuroimaging Data in Alzheimer’s Disease: A Systematic Review’, *Frontiers in Aging Neuroscience*, 9. Available at: https://doi.org/10.3389/fnagi.2017.00329.

Sarker, I.H. (2021) ‘Machine Learning: Algorithms, Real-World Applications and Research Directions’, *SN Computer Science*, 2(3), p. 160. Available at: https://doi.org/10.1007/s42979-021-00592-x.

Savas, C. and Dovis, F. (2019) ‘The Impact of Different Kernel Functions on the Performance of Scintillation Detection Based on Support Vector Machines.’, *Sensors (Basel, Switzerland)*, 19(23). Available at: https://doi.org/10.3390/s19235219.

Schurz, H. *et al.* (2019) ‘Evaluating the Accuracy of Imputation Methods in a Five-Way Admixed Population’, *Frontiers in Genetics*, 10. Available at: https://doi.org/10.3389/fgene.2019.00034.

Sham, P.C. and Purcell, S.M. (2014) ‘Statistical power and significance testing in large-scale genetic studies’, *Nature Reviews Genetics*, 15(5), pp. 335–346. Available at: https://doi.org/10.1038/nrg3706.

Shariff, N.S.Md. and Ferdaos, N.A. (2017) ‘An application of robust ridge regression model in the presence of outliers to real data problem’, *Journal of Physics: Conference Series*, 890, p. 012150. Available at: https://doi.org/10.1088/1742-6596/890/1/012150.

Shekhar, S., Bansode, A. and Salim, A. (2022) ‘A Comparative study of Hyper-Parameter Optimization Tools’.

Sherif, F.F., Zayed, N. and Fakhr, M. (2015) ‘Discovering Alzheimer Genetic Biomarkers Using Bayesian Networks’, *Advances in Bioinformatics*, 2015, pp. 1–8. Available at: https://doi.org/10.1155/2015/639367.

Shin, D.M. *et al.* (2020) ‘GEN2VCF: a converter for human genome imputation output format to VCF format’, *Genes & Genomics*, 42(10), pp. 1163–1168. Available at: https://doi.org/10.1007/s13258-020-00982-0.

Sierksma, A., Escott-Price, V. and De Strooper, B. (2020) ‘Translating genetic risk of Alzheimer’s disease into mechanistic insight and drug targets’, *Science*, 370(6512), pp. 61–66. Available at: https://doi.org/10.1126/science.abb8575.

Silberstein, M. *et al.* (2021) ‘Pathway analysis for genome-wide genetic variation data: Analytic principles, latest developments, and new opportunities.’, *Journal of genetics and genomics = Yi chuan xue bao*, 48(3), pp. 173–183. Available at: https://doi.org/10.1016/j.jgg.2021.01.007.

Silva, M.V.F. *et al.* (2019) ‘Alzheimer’s disease: risk factors and potentially protective measures.’, *Journal of biomedical science*, 26(1), p. 33. Available at: https://doi.org/10.1186/s12929-019-0524-y.

Silver, M. *et al.* (2012) ‘Identification of gene pathways implicated in Alzheimer’s disease using longitudinal imaging phenotypes with sparse regression.’, *NeuroImage*, 63(3), pp. 1681–94. Available at: https://doi.org/10.1016/j.neuroimage.2012.08.002.

Sivarajah, U. *et al.* (2017) ‘Critical analysis of Big Data challenges and analytical methods’, *Journal of Business Research*, 70, pp. 263–286. Available at: https://doi.org/10.1016/j.jbusres.2016.08.001.

Song, Y.-Y. and Lu, Y. (2015) ‘Decision tree methods: applications for classification and prediction.’, *Shanghai archives of psychiatry*, 27(2), pp. 130–5. Available at: https://doi.org/10.11919/j.issn.1002-0829.215044.

Stahl, K., Gola, D. and König, I.R. (2021) ‘Assessment of Imputation Quality: Comparison of Phasing and Imputation Algorithms in Real Data’, *Frontiers in Genetics*, 12. Available at: https://doi.org/10.3389/fgene.2021.724037.

Steyerberg, E.W. *et al.* (2010) ‘Assessing the Performance of Prediction Models’, *Epidemiology*, 21(1), pp. 128–138. Available at: https://doi.org/10.1097/EDE.0b013e3181c30fb2.

Strandén, I. and Christensen, O.F. (2011) ‘Allele coding in genomic evaluation’, *Genetics Selection Evolution*, 43(1), p. 25. Available at: https://doi.org/10.1186/1297-9686-43-25.

Südhof, T.C. (2018) ‘Towards an Understanding of Synapse Formation.’, *Neuron*, 100(2), pp. 276–293. Available at: https://doi.org/10.1016/j.neuron.2018.09.040.

Sun, S. *et al.* (2019a) ‘A Survey of Optimization Methods from a Machine Learning Perspective’.

Sun, S. *et al.* (2019b) ‘A Survey of Optimization Methods from a Machine Learning Perspective’.

SUN, Y., WONG, A.K.C. and KAMEL, M.S. (2009) ‘CLASSIFICATION OF IMBALANCED DATA: A REVIEW’, *International Journal of Pattern Recognition and Artificial Intelligence*, 23(04), pp. 687–719. Available at: https://doi.org/10.1142/S0218001409007326.

Susan, S. and Kumar, A. (2021) ‘The balancing trick: Optimized sampling of imbalanced <scp>datasets—A</scp> brief survey of the recent State of the Art’, *Engineering Reports*, 3(4). Available at: https://doi.org/10.1002/eng2.12298.

Sylvester, E.V.A. *et al.* (2018) ‘Applications of random forest feature selection for fine-scale genetic population assignment’, *Evolutionary Applications*, 11(2), pp. 153–165. Available at: https://doi.org/10.1111/eva.12524.

Talavera, L. (2005) ‘An Evaluation of Filter and Wrapper Methods for Feature Selection in Categorical Clustering’, in, pp. 440–451. Available at: https://doi.org/10.1007/11552253_40.

Taliun, D. *et al.* (2021) ‘Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program’, *Nature*, 590(7845), pp. 290–299. Available at: https://doi.org/10.1038/s41586-021-03205-y.
Tam, V. *et al.* (2019) ‘Benefits and limitations of genome-wide association studies’, *Nature Reviews Genetics*, 20(8), pp. 467–484. Available at: https://doi.org/10.1038/s41576-019-0127-1.
Tan, J. *et al.* (2021) ‘A critical look at the current train/test split in machine learning’.

Tangirala, S. (2020) ‘Evaluating the Impact of GINI Index and Information Gain on Classification using Decision Tree Classifier Algorithm*’, *International Journal of Advanced Computer Science and Applications*, 11(2). Available at: https://doi.org/10.14569/IJACSA.2020.0110277.

Tansey, K.E., Cameron, D. and Hill, M.J. (2018) ‘Genetic risk for Alzheimer’s disease is concentrated in specific macrophage and microglial transcriptional networks.’, *Genome medicine*, 10(1), p. 14. Available at: https://doi.org/10.1186/s13073-018-0523-8.

Tanveer, M. *et al.* (2020) ‘Machine Learning Techniques for the Diagnosis of Alzheimer’s Disease’, *ACM Transactions on Multimedia Computing, Communications, and Applications*, 16(1s), pp. 1–35. Available at: https://doi.org/10.1145/3344998.

Tarawneh, R. (2020) ‘Biomarkers: Our Path Towards a Cure for Alzheimer Disease.’, *Biomarker insights*, 15, p. 1177271920976367. Available at: https://doi.org/10.1177/1177271920976367.

Tibshirani, R. (1996) ‘Regression Shrinkage and Selection Via the Lasso’, *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), pp. 267–288. Available at: https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.

Touzani, S., Granderson, J. and Fernandes, S. (2018) ‘Gradient boosting machine for modeling the energy consumption of commercial buildings’, *Energy and Buildings*, 158, pp. 1533–1543. Available at: https://doi.org/10.1016/j.enbuild.2017.11.039.

Toyota, Y. *et al.* (2007) ‘Comparison of behavioral and psychological symptoms in early-onset and late-onset Alzheimer’s disease.’, *International journal of geriatric psychiatry*, 22(9), pp. 896–901. Available at: https://doi.org/10.1002/gps.1760.

Tsai, C.-W. *et al.* (2015) ‘Big data analytics: a survey’, *Journal of Big Data*, 2(1), p. 21. Available at: https://doi.org/10.1186/s40537-015-0030-3.

Tsikliras, A.C. and Froese, R. (2019) ‘Maximum Sustainable Yield’, in *Encyclopedia of Ecology*. Elsevier, pp. 108–115. Available at: https://doi.org/10.1016/B978-0-12-409548-9.10601-3.

Uddin, M.F. (2019) ‘Addressing Accuracy Paradox Using Enhanched Weighted Performance Metric in Machine Learning’, in *2019 Sixth HCT Information Technology Trends (ITT)*. IEEE, pp. 319–324. Available at: https://doi.org/10.1109/ITT48889.2019.9075071.

Vabalas, A. *et al.* (2019) ‘Machine learning algorithm validation with a limited sample size’, *PLOS ONE*, 14(11), p. e0224365. Available at: https://doi.org/10.1371/journal.pone.0224365.

Vaghela, V.B., Ganatra, A. and Thakkar, A. (2009) ‘Boost a Weak Learner to a Strong Learner Using Ensemble System Approach’, in *2009 IEEE International Advance Computing Conference*. IEEE, pp. 1432–1436. Available at: https://doi.org/10.1109/IADCC.2009.4809227.

Valverde-Albacete, Francisco J and Peláez-Moreno, C. (2014) ‘100% classification accuracy considered harmful: the normalized information transfer factor explains the accuracy paradox.’, *PloS one*, 9(1), p. e84217. Available at: https://doi.org/10.1371/journal.pone.0084217.

Valverde-Albacete, Francisco J. and Peláez-Moreno, C. (2014) ‘100% Classification Accuracy Considered Harmful: The Normalized Information Transfer Factor Explains the Accuracy Paradox’, *PLoS ONE*, 9(1), p. e84217. Available at: https://doi.org/10.1371/journal.pone.0084217.

Vardarajan, B.N. *et al.* (2014) ‘Age-Specific Incidence Rates for Dementia and Alzheimer Disease in NIA-LOAD/NCRAD and EFIGA Families’, *JAMA Neurology*, 71(3), p. 315. Available at: https://doi.org/10.1001/jamaneurol.2013.5570.

Varma, A.R. *et al.* (1999) ‘Evaluation of the NINCDS-ADRDA criteria in the differentiation of Alzheimer’s disease and frontotemporal dementia’, *Journal of Neurology, Neurosurgery & Psychiatry*, 66(2), pp. 184–188. Available at: https://doi.org/10.1136/jnnp.66.2.184.

Varma, S. and Simon, R. (2006) ‘Bias in error estimation when using cross-validation for model selection’, *BMC Bioinformatics*, 7(1), p. 91. Available at: https://doi.org/10.1186/1471-2105-7-91.

De Velasco Oriol, J. *et al.* (2019) ‘Benchmarking machine learning models for late-onset alzheimer’s disease prediction from genomic data’, *BMC Bioinformatics*, 20(1), p. 709. Available at: https://doi.org/10.1186/s12859-019-3158-x.

de Velasco Oriol, J.E.V.E.E.K.T.A.D.N.I. (2019) ‘Predicting late-onset Alzheimer’s disease from genetic data using deep neural networks’, *bioRxiv* [Preprint].

Velliangiri, S., Alagumuthukrishnan, S. and Thankumar joseph, S.I. (2019) ‘A Review of Dimensionality Reduction Techniques for Efficient Computation’, *Procedia Computer Science*, 165, pp. 104–111. Available at: https://doi.org/10.1016/j.procs.2020.01.079.

Verleysen, M. and François, D. (2005a) ‘The Curse of Dimensionality in Data Mining and Time Series Prediction’, in, pp. 758–770. Available at: https://doi.org/10.1007/11494669_93.

Verleysen, M. and François, D. (2005b) ‘The Curse of Dimensionality in Data Mining and Time Series Prediction’, in, pp. 758–770. Available at: https://doi.org/10.1007/11494669_93.

Verma, A.A. *et al.* (2021) ‘Implementing machine learning in medicine.’, *CMAJ : Canadian Medical Association journal = journal de l’Association medicale canadienne*, 193(34), pp. E1351–E1357. Available at: https://doi.org/10.1503/cmaj.202434.

Verma, S.S. *et al.* (2014) ‘Imputation and quality control steps for combining multiple genome-wide datasets.’, *Frontiers in genetics*, 5, p. 370. Available at: https://doi.org/10.3389/fgene.2014.00370.

van der Walt, S., Colbert, S.C. and Varoquaux, G. (2011) ‘The NumPy Array: A Structure for Efficient Numerical Computation’, *Computing in Science & Engineering*, 13(2), pp. 22–30. Available at: https://doi.org/10.1109/MCSE.2011.37.

Wang, H. *et al.* (2021) ‘Genome-wide epistasis analysis for Alzheimer’s disease and implications for genetic risk prediction’, *Alzheimer’s Research & Therapy*, 13(1), p. 55. Available at: https://doi.org/10.1186/s13195-021-00794-8.

Wang, M. and Xu, S. (2019) 'Statistical power in genome-wide association studies and quantitative trait locus mapping.', *Heredity*, 123(3), pp. 287–306. Available at: https://doi.org/10.1038/s41437-019-0205-3.

Wang, Q. *et al.* (2022) 'A Comprehensive Survey of Loss Functions in Machine Learning', *Annals of Data Science*, 9(2), pp. 187–212. Available at: https://doi.org/10.1007/s40745-020-00253-5.

Weerts, H.J.P., Mueller, A.C. and Vanschoren, J. (2020) 'Importance of Tuning Hyperparameters of Machine Learning Algorithms'.

Wei, W., Visweswaran, S. and Cooper, G.F. (2011a) 'The application of naive Bayes model averaging to predict Alzheimer's disease from genome-wide data', *Journal of the American Medical Informatics Association*, 18(4), pp. 370–375. Available at: https://doi.org/10.1136/amiajnl-2011-000101.

Wei, W., Visweswaran, S. and Cooper, G.F. (2011b) 'The application of naive Bayes model averaging to predict Alzheimer's disease from genome-wide data.', *Journal of the American Medical Informatics Association : JAMIA*, 18(4), pp. 370–5. Available at: https://doi.org/10.1136/amiajnl-2011-000101.

Wei, Z. *et al.* (2013) 'Large sample size, wide variant spectrum, and advanced machine-learning technique boost risk prediction for inflammatory bowel disease.', *American journal of human genetics*, 92(6), pp. 1008–12. Available at: https://doi.org/10.1016/j.ajhg.2013.05.002.

Weller, J. and Budson, A. (2018) 'Current understanding of Alzheimer's disease diagnosis and treatment.', *F1000Research*, 7. Available at: https://doi.org/10.12688/f1000research.14506.1.

Wickramasinghe, I. (2020) 'Classification of All-Rounders in the Game of ODI Cricket: Machine Learning Approach', *ATHENS JOURNAL OF SPORTS*, 7(1), pp. 21–34. Available at: https://doi.org/10.30958/ajspo.7-1-2.

Wightman, D.P. *et al.* (2021a) 'A genome-wide association study with 1,126,563 individuals identifies new risk loci for Alzheimer's disease', *Nature Genetics*, 53(9), pp. 1276–1282. Available at: https://doi.org/10.1038/s41588-021-00921-z.

Wightman, D.P. *et al.* (2021b) 'A genome-wide association study with 1,126,563 individuals identifies new risk loci for Alzheimer's disease', *Nature Genetics*, 53(9), pp. 1276–1282. Available at: https://doi.org/10.1038/s41588-021-00921-z.

Wolff, R.F. *et al.* (2019) 'PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies', *Annals of Internal Medicine*, 170(1), p. 51. Available at: https://doi.org/10.7326/M18-1376.

Wray, N.R. *et al.* (2010) 'The Genetic Interpretation of Area under the ROC Curve in Genomic Profiling', *PLoS Genetics*, 6(2), p. e1000864. Available at: https://doi.org/10.1371/journal.pgen.1000864.

Wyner, A.J. *et al.* (2015) 'Explaining the Success of AdaBoost and Random Forests as Interpolating Classifiers'.

Xia, B. *et al.* (2015) 'PETs: A Stable and Accurate Predictor of Protein-Protein Interacting Sites Based on Extremely-Randomized Trees', *IEEE Transactions on NanoBioscience*, 14(8), pp. 882–893. Available at: https://doi.org/10.1109/TNB.2015.2491303.

Xie, H., Li, J. and Xue, H. (2017) 'A survey of dimensionality reduction techniques based on random projection'.

Yadav, S. and Bhole, G.P. (2020) 'Handling Imbalanced Dataset Classification in Machine Learning', in *2020 IEEE Pune Section International Conference (PuneCon)*. IEEE, pp. 38–43. Available at: https://doi.org/10.1109/PuneCon50868.2020.9362471.

Yadav, S. and Shukla, S. (2016) 'Analysis of k-Fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality Classification', in *2016 IEEE 6th International Conference on Advanced Computing (IACC)*. IEEE, pp. 78–83. Available at: https://doi.org/10.1109/IACC.2016.25.

Yang, H. *et al.* (2010) 'A Hybrid Machine Learning Method for Fusing fMRI and Genetic Data: Combining both Improves Classification of Schizophrenia.', *Frontiers in human neuroscience*, 4, p. 192. Available at: https://doi.org/10.3389/fnhum.2010.00192.

Yang, H.D. *et al.* (2016a) 'History of Alzheimer's Disease.', *Dementia and neurocognitive disorders*, 15(4), pp. 115–121. Available at: https://doi.org/10.12779/dnd.2016.15.4.115.

Yang, H.D. *et al.* (2016b) 'History of Alzheimer's Disease.', *Dementia and neurocognitive disorders*, 15(4), pp. 115–121. Available at: https://doi.org/10.12779/dnd.2016.15.4.115.

Yiannopoulou, K.G. and Papageorgiou, S.G. (2013) 'Current and future treatments for Alzheimer's disease.', *Therapeutic advances in neurological disorders*, 6(1), pp. 19–33. Available at: https://doi.org/10.1177/1756285612461679.

Yiannopoulou, K.G. and Papageorgiou, S.G. (2020) 'Current and Future Treatments in Alzheimer Disease: An Update.', *Journal of central nervous system disease*, 12, p. 1179573520907397. Available at: https://doi.org/10.1177/1179573520907397.

Ying, X. (2019) 'An Overview of Overfitting and its Solutions', *Journal of Physics: Conference Series*, 1168, p. 022022. Available at: https://doi.org/10.1088/1742-6596/1168/2/022022.

Yousef, W.A. (2019) 'Machine Learning Assessment: implications to cybersecurity'. Available at: https://doi.org/10.1007/978-3-031-16237-4_3.

Yu, T. and Zhu, H. (2020) 'Hyper-Parameter Optimization: A Review of Algorithms and Applications'.

Zhang, Q. *et al.* (2020) 'Risk prediction of late-onset Alzheimer's disease implies an oligogenic architecture', *Nature Communications*, 11(1), p. 4799. Available at: https://doi.org/10.1038/s41467-020-18534-1.

Zhang, T. and Yu, B. (2005) ‘Boosting with early stopping: Convergence and consistency’, *The Annals of Statistics*, 33(4). Available at: https://doi.org/10.1214/009053605000000255.

Zhao, H. *et al.* (2018) ‘A practical approach to adjusting for population stratification in genome-wide association studies: principal components and propensity scores (PCAPS).’, *Statistical applications in genetics and molecular biology*, 17(6). Available at: https://doi.org/10.1515/sagmb-2017-0054.

Zheng, H.-F. *et al.* (2015) ‘Performance of genotype imputation for low frequency and rare variants from the 1000 genomes.’, *PloS one*, 10(1), p. e0116487. Available at: https://doi.org/10.1371/journal.pone.0116487.

Zhong, Y., Chalise, P. and He, J. (2020) ‘Nested cross-validation with ensemble feature selection and classification model for high-dimensional biological data’, *Communications in Statistics - Simulation and Computation*, pp. 1–18. Available at: https://doi.org/10.1080/03610918.2020.1850790.

Zhou, T., Thung, K., *et al.* (2019) ‘Effective feature learning and fusion of multimodality data using stage-wise deep neural network for dementia diagnosis’, *Human Brain Mapping*, 40(3), pp. 1001–1016. Available at: https://doi.org/10.1002/hbm.24428.

Zhou, T., Liu, M., *et al.* (2019) ‘Latent Representation Learning for Alzheimer’s Disease Diagnosis With Incomplete Multi-Modality Neuroimaging and Genetic Data’, *IEEE Transactions on Medical Imaging*, 38(10), pp. 2411–2422. Available at: https://doi.org/10.1109/TMI.2019.2913158.

Zhou, X. *et al.* (2021) ‘Polygenic Score Models for Alzheimer’s Disease: From Research to Clinical Applications’, *Frontiers in Neuroscience*, 15. Available at: https://doi.org/10.3389/fnins.2021.650220.

Zolnierek, A. and Rubacha, B. (no date) ‘The Empirical Study of the Naive Bayes Classifier in the Case of Markov Chain Recognition Task’, in, pp. 329–336. Available at: https://doi.org/10.1007/3-540-32390-2_38.

Zou, H. and Hastie, T. (2005) ‘Addendum: Regularization and variable selection via the elastic net’, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(5), pp. 768–768. Available at: https://doi.org/10.1111/j.1467-9868.2005.00527.x.