University of Reading

A Data Mining Approach for Automated Classification of Alzheimer's Disease

Alexander Luke Spedding

A thesis presented for the degree of

Doctor of Philosophy

School of Mathematical, Physical and Computational Sciences

University of Reading

United Kingdom

July 2018

Declaration

I confirm that this is my own work and the use of all material from other sources has been properly and fully acknowledged.

Signed: Alexander Luke Spedding

Date: July 2018

Abstract

This work presents the creation of classifiers able to automatically diagnose Alzheimer's disease from structural magnetic resonance images. Measurements of regions inside the human brain and how they relate to a diagnosis of Alzheimer's disease are investigated. A genetic algorithm is used to extract a subset of brain measurements which have some of the highest predictive power when compared to the rest of the measurements. A descriptive classifier is created which uses the size of each of the hippocampal volumes of a subject and compares it to the distribution of other subjects. Based on how the individual subject's measurements compares to the distribution, we can determine whether it is an Alzheimer's disease predictive volume. This classifier achieves an accuracy equivalent to state-of-the-art approaches. A second descriptive classifier is created using a regression model to predict a healthy subject's age and applying this to Alzheimer's disease positive subjects to generate an apparent brain age. The classification accuracy of when the apparent brain age is compared to the subject's real age is comparable to the state-of-the-art methods in the literature.

List of Abbreviations

Abbreviation	Meaning	Page First Used
AD	Alzheimer's disease	1
ADNI	The Alzheimer's Disease Neuroimaging Initiative	2
APOE	apolipoprotein E	72
AUROC	area under the receiver operating characteristic curve	50
BOLD	blood-oxygen-level dependent	14
BRC	binary region classification algorithm	113
CA	cornu ammonis	10
CAD	computer-aided diagnosis	86
CRAN	Comprehensive R Archive Network	75
CSF	cerebral spinal fluid	9
dMRI	diffusion magnetic resonance image	14
DSM	Diagnostic and Statistical Manual of Mental Disorders	7
FCBF	fast correlation-based filter	53
FDG	fluorodeoxyglucose	73
fMRI	functional magnetic resonance image	14
FN	false negative	48
FP	false positive	48
FPR	false positive rate	50
GA	genetic algorithm	95

Continued on next page

Abbreviation	Meaning	Page First Used
GCA	Gaussian classifier atlas	34
GM	grey matter	9
GPU	graphics processing unit	36
HC	healthy control	2
ICV	intracranial volume	62
IXI	Information Extraction from Images	2
LDA	linear discriminant analysis	42
MCI	mild cognitive impairment	2
MMSE	mini mental-state examination	16
MRI	magnetic resonance image	1
MRMR	minimum-redundancy-maximum-relevance	54
NPV	negative predictive value	48
PCA	principal component analysis	59
PET	positron emission tomography	85
PWC	probability-weight classification	115
QA	quality assurance	157
RBF	radial basis filter	46
RF	radio frequency	13
ROC	receiver operating characteristic	49
SD	standard deviation	116
SNR	signal-to-noise ratio	59
SVM	support vector machine	45
TN	true negative	48
TP	true positive	48
TPR	true positive rate	49
WAIS	Wechsler Adult Intelligence Scale	18

Continued on next page

Abbreviation	Meaning	Page First Used
WCST	Wisconsin card sorting test	18
WM	white matter	9
WV	Weighted Voting Strategy	102

Contents

List of Figures

Chapter 1

Introduction

1.1 An Introduction to the Thesis

The work in this thesis is aimed at providing a solution for the automated diagnosis of Alzheimer's disease (AD) using structural magnetic resonance images (MRIs). AD is a chronic neurodegenerative disease which is responsible for 60% to 70% of cases of dementia. AD primarily affects the elderly as around 6% of people aged 65 and above are diagnosed with the disease. As the average age of sample groups of people increases, the percentage of the sample diagnosed with AD increases, showing that the disease becomes more prevalent with respect to age (Burns and Iliffe, 2009). The automated diagnosis would provide a tool for medical professionals to be able to have more evidence at their disposal when it comes to the diagnosis of patients with AD, potentially if the automated diagnosis achieves a sufficiently high accuracy then it would be able to replace the medical diagnosis of an expert. However, there are risks of a misdiagnosis: for a healthy subject misdiagnosed as AD these include additional medical costs incurred by the subject requiring more medical services (Hunter et al., 2015). If the misdiagnosis was an AD subject misdiagnosed as healthy then they would not be given treatment which may be able to delay the onset of the disease or reduce the symptoms. This work aims to develop algorithms which the tools for the medical

professionals could utilise.

This work takes a data-driven approach which means that the progress of the work is based on data and does not use specific domain knowledge for the AD diagnosis unless it can be verified by the data. The data used is the MRI scans of healthy control (HC) patients, patients with mild cognitive impairment (MCI) and patients with AD. This means that the work in this thesis is based solely on existing MRI scans of AD positive subjects rather than intuition. As the work in this thesis uses a data-driven approach, the methods may be able to be applied successfully to other domains, however, this has not been tested in this thesis and is discussed as potential future work in the conclusion.

There are two issues with the automated diagnosis of AD which the work aims to improve: the first aim is the classification accuracy, such that automated predictions are more likely to predict AD correctly; the second aim involves improving the descriptive ability of classifiers, such that a medical professional may understand intuitively why a decision was made by the classifier. This information could potentially aid the understanding into how AD affects the brain. The first aim is numerically measurable thus all classifiers developed in this thesis produce various statistics which enable their accuracy in predicting AD to be compared. The second aim is harder to evaluate numerically but the performance at achieving this aim will be discussed where appropriate.

1.2 Introduction to the Data Workflow

This section will discuss briefly how the MRI data is obtained, preprocessed and made ready for data mining. The first part is obtaining the data, there are various databases online with open access to structural MRI, these are discussed further in Section 7.1. The two sources used are the Alzheimer's Disease Neuroimaging Initiative (ADNI) and Information Extraction from Images (IXI) . MRIs are downloaded from these databases depending on certain criteria which is used to filter the MRIs. Once the MRIs are downloaded they are

ready for preprocessing. Due to the complexity of the preprocessing step an open source tool called Freesurfer is used, the preprocessing step and Freesurfer are discussed thoroughly in Chapter 3. Freesurfer processes the 3D MRI data into a set of measurements of various regions in the brain. After this step the data is in a numeric form rather than image form and traditional data mining techniques can be applied.

The main data mining techniques used in this work were: classification, feature selection, z-score normalisation, and regression. The techniques used will be briefly discussed here; a more in-depth discussion can be found in Chapter 4. The classification techniques used were: support vector machines, naive Bayes classifiers, and linear discriminant analysis. The feature selection techniques used were genetic algorithms, a novel t-test-based feature selection algorithm, and also feature selection based on domain knowledge where the location of the measurement in the brain determines whether it is selected. The regression technique used was LASSO regression. Many other data mining techniques have been tested through the work in this thesis but they did not provide a good enough result for the task to be included in this thesis.

1.3 Research Questions, Hypotheses and Objectives

The following research questions this thesis aims to answer are:

- Can novel predictive models be created with greater performance than current state-of-the-art models for predicting AD from a structural MRI?
- Can the performance of state-of-the-art black-box predictive models be reached using a novel descriptive predictive model for diagnosing AD?
- Does using feature selection improve the accuracy and descriptiveness of a model predicting a diagnosis of AD?

- Simultaneously with this work, can we create a software library to improve the speed at which structural MRIs can be analysed?
- Can we utilise the atrophy of a brain caused by AD to be able to generate a predicted age for AD patients which is greater than their actual age, whereas for HC patients their predicted age is similar to their actual age?

These lead to the respective hypotheses:

- We can create novel predictive models with a greater performance than current state-of-the-art models to diagnose AD from structural MRIs.
- Descriptive predictive models can be created to diagnose AD with the same or better performance than state-of-the-art black-box classifiers.
- Feature selection improves the accuracy and descriptiveness of a predictive model diagnosing AD.
- The common functionality to create predictive models can be implemented as a software library.
- A model can be created to predict the age of an MRI, and this age will be higher than the actual age for AD subjects, and similar to the actual age for HC subjects.

These hypotheses can become the respective objectives:

- To identify and use some state-of-the-art and black-box approaches as baseline performance for the accuracy, such that newly designed predictive models have a baseline to be evaluated against.
- To design a descriptive novel predictive model achieving a state-of-the-art accuracy.
- To investigate different techniques for feature selection to avoid the curse of dimensionality and to allow the predictive models to have a greater descriptive expressiveness.

- To develop a software library to support and speed up the data manipulation tasks which are a fundamental requirement of every predictive model in this work.
- To define the apparent brain age as the difference between the expected age of a brain assuming the brain is healthy and the actual age of a brain, and then use this apparent brain age to provide a predictive model with excellent descriptive information.

1.4 Contributions and Thesis Outline

There are four novel bodies of work in this thesis:

- A package for the programming language, R, to allow easier analysis of structural MRI data
- The usage of a genetic algorithm for feature selection to find the most predictive attributes of an MRI for AD
- A novel classification algorithm based on probabilities looking at a domain-specific subset of features
- The generation of a new descriptive feature of the difference between a person's actual age and their predicted age when it is assumed they have a healthy brain

The structure of thesis can be divided into two parts: the first part covers the theoretical background and a review of relevant literature; and the second part is about the work. Introductions to neuroanatomy and AD are covered in Chapter 2. Chapter 3 covers the preprocessing of the MRIs with an open source software suite and briefly covers how this software works internally. Chapter 4 introduces data mining techniques and the theory behind the machine learning algorithms used in the novel work. Chapter 5 reviews literature where data mining techniques have been used in the context of either MRI classification or predicting AD from various biomarkers.

The contributions of the thesis start at Chapter 6 which describes a custom package which was built for the R language to aid in the use of importing data from Freesurfer and the functionality to make it easy to manipulate the data once it is imported into R; as well as describing the functionality it compares the package against similar already existing packages and discusses the novel features the package implements. Chapter 7 covers preliminary analysis of the data and tests various hypotheses and provides a basis for why the novel work was performed; including how a genetic algorithm is used for feature selection of the MRI features and this builds upon previous literature where filter methods were used for feature selection. Chapter 8 introduces a novel probability-based classifier which infers a set of descriptive features to predict AD with a high predictive power based on MRI measurements of the brain. Chapter 9 introduces a method to produce a descriptive feature based on a subject's predicted age and uses this a descriptive summary of hundreds of the other features, augmented with the hippocampal volumes it can achieve a similar accuracy as when the entire feature set is used.

Chapter 2

Alzheimer's Disease and Neuroanatomy

2.1 Dementia and Alzheimer's Disease

Alzheimer's disease was named after the German psychologist Alois Alzheimer who first discovered the disease in 1906, he discovered how the brain is affected by it and the symptoms it causes (Alzheimer, 1907). While AD is the main cause of dementia there are other factors which cause it such as vascular dementia (Ott et al., 1995), genetics (Goedert and Spillantini, 2006), Lewy bodies (Hanson and Lippa, 2009) and frontotemporal dementia (Englund et al., 1994). Vascular dementia is caused by diseased blood vessels not able to supply enough blood to the brain. Genetics are another factor as inheriting the $\epsilon 4$ allele of apolipoprotein E gives an individual an increased chance of developing AD. Lewy bodies (these are circular lumps of protein which develop inside the cells of the brain) are believed to affect neurotransmitters causing less regulation of brain functions; dementia caused by Lewy bodies is closely linked to Parkinson's disease. Frontotemporal dementia is caused by damage and shrinking of the frontal and temporal lobes, and this form of dementia is linked to motor neurone disease.

The Diagnostic and Statistical Manual of Mental Disorders (DSM) (American Psychiatric

Association, 2000) is a manual published by the American Psychiatric Association and shows criteria for classification of mental disorders. DSM-IV-TR refers to a text revision (TR) of the fourth version (IV) of the DSM. Symptoms of AD based on the DSM-IV-TR criteria include: memory impairment; aphasia (language disturbance); apraxia (impaired motor functionality); agnosia (failure to recognise objects) and a disturbance in executive functioning (a set of mental skills which help a person get things done such as: time management, paying attention and multitasking) (Feldman, 2007). Early detection of AD is challenging as there have been no biological markers found to definitively diagnose it at this stage. The current diagnosis of AD involves clinical approaches and these are a set of neuropsychological tests to assess the patient (McKhann et al., 1984); since the patients are being assessed based on their symptoms AD can only be diagnosed once it has had a notable effect on the patient's lifestyle. It must be noted that while these tests are used to give a diagnosis, this diagnosis is still not definitive, a definitive diagnosis can only be given post-mortem.

While there is no cure yet for AD, symptomatic treatments exist to help patients reduce the symptoms they are suffering from as well as delay the onset. Clinical trials are still running to develop new treatments aimed to lower the chance of developing AD or delaying the onset and progression of it (Klafki et al., 2006).

Mild cognitive impairment is a condition involving diminished brain functionality beyond what is expected based on the age of the patient. It is similar to AD though the symptoms of MCI are of a lesser extent than that of AD. A person suffering with MCI will be able to continue with their daily activities to nearly the same extent as they did before developing the condition whereas AD sufferers cannot (Petersen et al., 1999). MCI is of interest in the automated detection of AD as it is frequently seen that a person suffering with MCI has a higher chance of developing AD (compared to a healthy person developing AD). This conversion rate of MCI developing into AD is 10% to 15% per year higher than the 6% chance of a healthy person developing AD in their lifetime (Grundman et al., 2004).

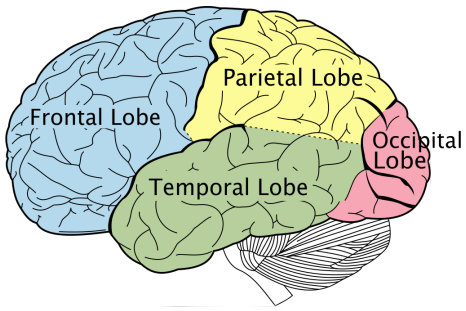

Figure 2.1: How the lobes are organised within the brain.

Source: Modified image from Gray's Anatomy, Figure 728

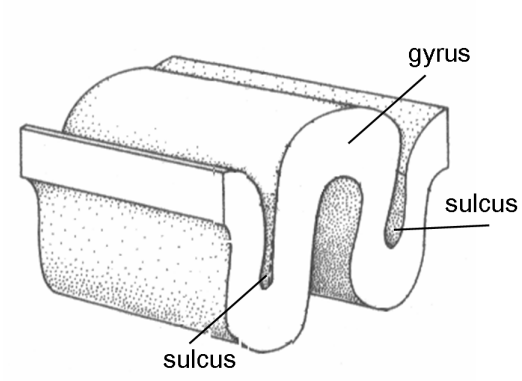

Figure 2.2: Diagram of the gyri and sulci in the cortex.

Source: Albert Kok at Dutch Wikipedia
https://commons.wikimedia.org/wiki/File:Gyrus_sulcus.png

2.2 Neuroanatomy

The brain is formed of two cerebral hemispheres, these can be divided into the left and right hemisphere. Within the brain there are four lobes: frontal, temporal, parietal and occipital; Figure 2.1 shows how these lobes are positioned within the brain, in this figure the brain is positioned such that the person would be facing the left. The outer part of the hemispheres is the cerebral cortex, the surface of which is folded into sulci and gyri. Sulci and gyri are shown in Figure 2.2 and they are respectively the depressions and grooves in the folded cortex. There are fissures throughout the surface of the cerebral cortex and this refers to the large gaps in the cortex which divide the brain into its lobes. The cortex can be divided into two parts: the neocortex and the allocortex. The neocortex is the newest part of the cerebral cortex to evolve, and is the part of the brain involving sensory perception, cognition, language, spatial reasoning and motor skills. The other part of the cerebral cortex is referred to as the allocortex which makes up the rest of the cerebral cortex.

The two hemispheres are separated by the great longitudal fissure which contains the falx cerebri and the corpus callosum, the latter containing fibres linking the two hemispheres together. The brain is made up of three tissue types: white matter (WM), grey matter (GM), and cerebral spinal fluid (CSF). WM contains many myelinated axons (these are axons

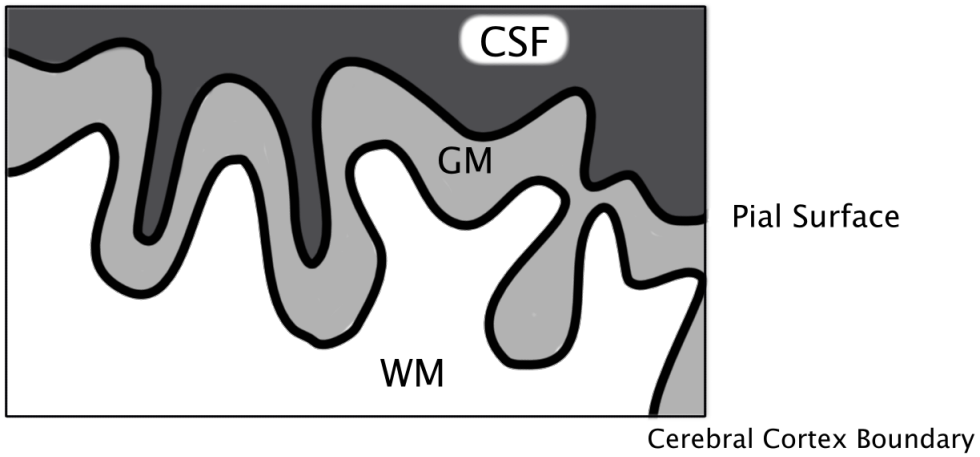

Figure 2.3: The location of the brain tissue boundaries within the brain

covered by a myelin sheath allowing them to conduct action potentials at greater velocities than unmyelinated axons) and few cell bodies, it affects how the brain learns, functions and coordinates between the different regions in the brain. GM contains many cell bodies and few myelinated axons, and it handles processing and cognition. CSF is a clear fluid found in the brain and the spine, it acts as a buffer for the cortex of the brain protecting against damage that can be caused via bacteria, viruses and physical impact. The boundary between the GM and WM is called the cerebral cortex boundary, and the boundary between the CSF and GM is called the pial surface, both of these boundaries are shown in Figure 2.3.

The subcortex is a part of the brain which is enveloped by the cerebral cortex and sits at the bottom of the cerebral cortex, it can be divided into three parts: the basal ganglia (involved in motor control and skills learning); the limbic system (consisting of the amygdala responsible for fear and the hippocampus); and the diencephalon (consisting of the thalamus, hypothalamus, epithalamus, subthalamus). The hippocampus (Figure 2.4) is a major component of the brain and is located within the temporal lobe. The hippocampus is composed of GM, and can be divided into four subfields: CA1 (cornu ammonis 1), CA2, CA3 and CA4. The location of these subfields are shown in Figure 2.5.

The ventricular system refers to the set of four connected ventricles in the brain: the left and right lateral ventricles, the third ventricle, and the fourth ventricle; the location of these

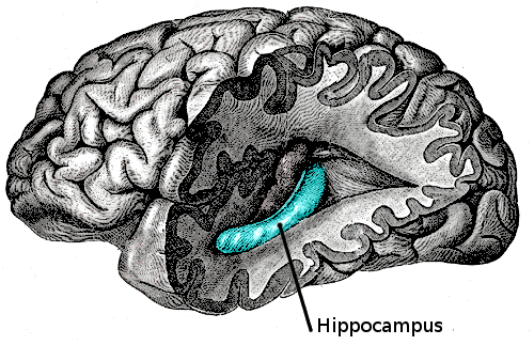

Figure 2.4: Where the hippocampus is located within the brain.

Source: Modified image from Gray's Anatomy, Figure 739

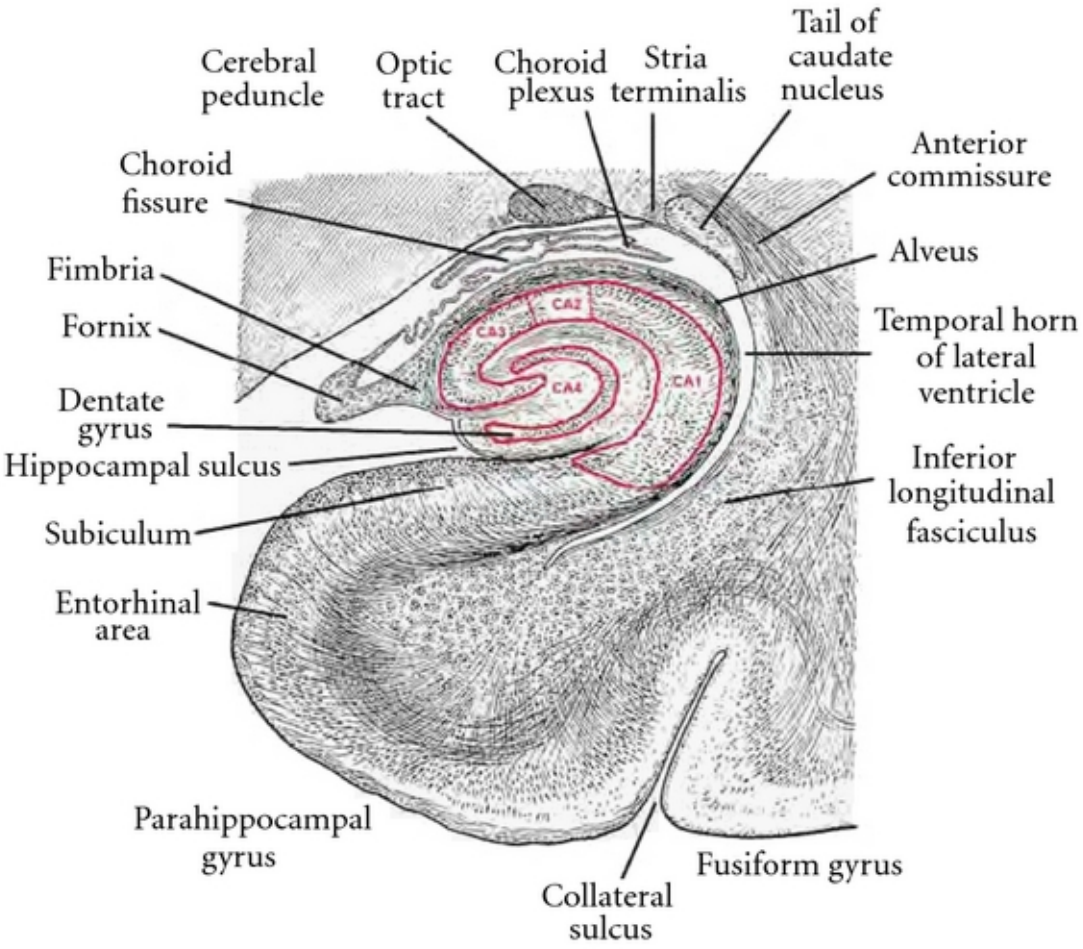

Figure 2.5: Subfields of the hippocampus and surrounding areas

Source: Modifed image by J. A. Kiernan, original from L. Edinger, The Anatomy of the Central Nervous System of Man and of Vertebrates in General (1899)

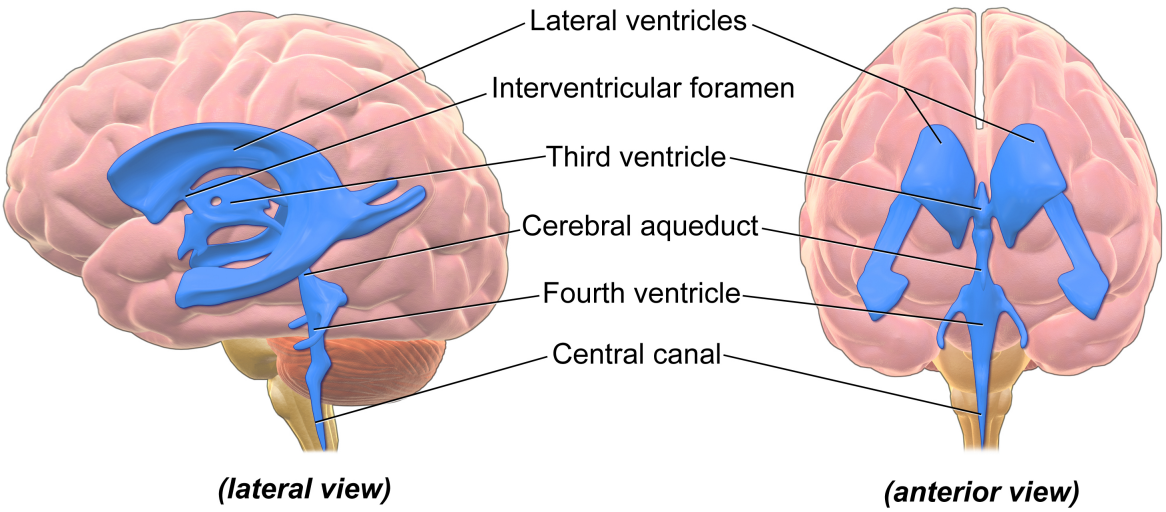

Figure 2.6: Location of the ventricular system in the brain.

Source: Bruce Blausen, Wikipedia

within the brain is shown in Figure 2.6. The ventricles are cavities in the brain containing the choroid plexus which produces CSF which then flows through the ventricular system to different parts around the brain. Both the hippocampus and the ventricles have been found to be strongly affected by AD, with the hippocampus atrophying (Jin et al., 2004; Ball et al., 1985) and the ventricles enlarging.

2.3 Neuroimaging

2.3.1 Magnetic Resonance Imaging

Magnetic resonance imaging is a medical technique to produce detailed images of internal body structures including the brain, these are known as MRIs. To create an MRI it requires a powerful magnetic field, radio frequency pulses, and a computer. There are multiple types of MRI which can be taken for the brain, the simplest of which is a structural MRI.

Structural MRI

A structural brain MRI is created using a subject positioned inside a MRI scanner, and the scanner forms a strong magnetic field around the brain. The aim of the scanner is to scan the signal from the protons in the brain tissue. Initially, an oscillating magnetic field is applied at a frequency to cause the protons to resonate. The protons are then excited by the resonant frequency and emit a radio frequency (RF) signal which the scanner measures using a receiving coil. The main magnetic field is varied using gradient coils; this causes the RF signal to encode position information. Different brain tissues have different times for the protons to return to equilibrium from the excited state.

The scan generates a 3D image of the brain using voxels, where each voxel has four values: $\{x, y, z, i\}$ where $\{x, y, z\}$ is the 3D coordinate position of the voxel and i is the intensity value of the voxel. Structural MRIs can be divided into two types of weighting: T1 and T2; where T1 refers to the rate of longitudinal relaxation, and T2 refers to the rate of transverse relaxation. Further information about the mathematics of the weighting can be found in Appendix B.

In T1 scans the intensity of a voxel corresponds to the type of brain tissue that is at the location: a low intensity means there is CSF, typically represented by a black colour in an MRI scan; a medium intensity means there is GM, typically represented by a grey colour; and a high intensity means there is WM, typically represented by a white colour. T1-weighting is useful for looking at structures in a brain which are high in fat or near CSF. In T2: CSF intensity is highest, then the GM intensity is of medium intensity with WM having the lowest intensity. T2-weighting is useful for analysis of structures containing CSF and looking for inflammation resulting in structures being filled with CSF. The difference between T1-weighted and T2-weighted MRIs is shown in Figure 2.7.

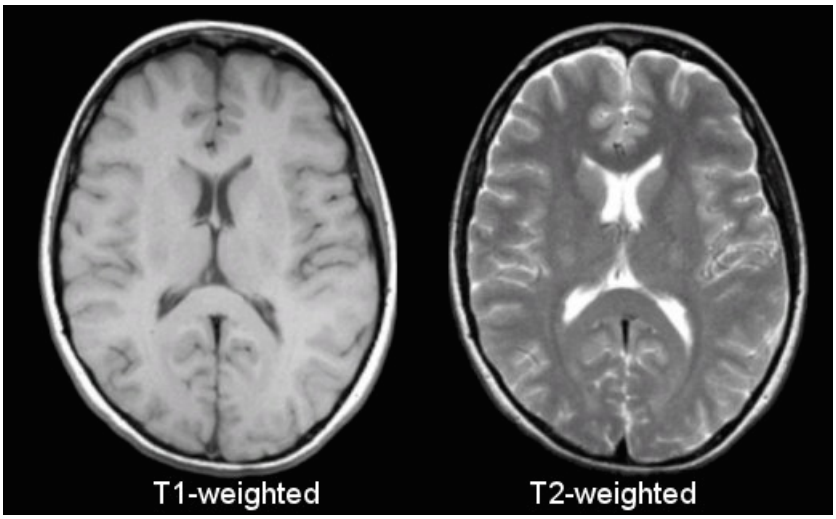

Figure 2.7: The difference between T1-weighted and T2-weighted MRIs.
Source: David C. Preston http://casemed.case.edu/clerkships/neurology/Web%20Neurorad/MRI%20Basics.htm

Functional MRI

A functional MRI (fMRI) is used to show how different regions in the brain respond to external stimuli (Logothetis, 2008). fMRIs measure the blood oxygenation level dependent (BOLD) signal; this is the signal from the interplay of blood volume, blood flow and blood oxygenation in the brain. Neuronal activity causes the brain to send blood to flow to the area where the activity took place. It sends more blood than is needed to replenish the oxygen used by the cells, this leads to a surplus of oxygen at that area which, the BOLD signal measures (Poldrack et al., 2011). The value of the BOLD signal can be mapped to a structural MRI so it can be easily viewed how the brain is responding to the stimuli.

Diffusion MRI

A diffusion MRI (dMRI) measures the diffusion process of molecules in the tissues of the brain, it is similar to a structural MRI except the intensity of a voxel measures the best estimate of the rate of water diffusion at that location (Bihan, 2003). The rate of water diffusion is discovered by measuring the fractional anisotropy of the Brownian motion of water molecules in the brain. When the water molecules are in fibre tracts they are more likely to be anisotropic (this is the property of being directionally dependent) as their movement

is restricted to moving parallel to the tract; water molecules located outside of fibre tracts will be less anisotropic as they can move in various directions. Fractional anisotropy is a value between 0 and 1 describing the degree of anisotropy of a diffusion process; thus water molecules in fibre tracts will have a higher fractional anisotropy than water molecules outside of the fibre tracts.

File Formats of MRIs

A number of file formats exist that MRIs are stored in. Larobina and Murino (2014) explain the various file formats in-depth. This subsection will discuss some of these and their benefits.

The format Analyze was used by the commercial software Analyze at the end of the 1980s and was the standard file format for MRIs for over ten years. It was designed for volumetric data and could handle MRIs in 3D and 4D (with the fourth dimension being time, such as in fMRIs). The downsides of this file format are that there is a fixed 348 byte header meaning that additional information such as the image orientation cannot be stored; and it also does not support a number of basic data types which is limiting because a user may have to scale their data to fit an allowable data type.

The Dicom format was created in 1993 by the American College of Radiology and the National Electric Manufacturers Association. The idea behind Dicom is that the metadata cannot be separated from the pixel data. It can only store pixel data as integers, but it can store a scale factor to be able to transform the pixel values to values which are in the range of what was initially recorded. Another benefit is that it supports compressed image formats such as JPEG.

The Nifti format was created by the National Institutes of Health in 2004 with the aim to address the downsides of the Analyze format while retaining its advantages. It includes: support for additional data types which Analyze did not enable; additional information in the header such as the image orientation. It has widely replaced Analyze, and has been adopted as the default format in most software packages. The work in this thesis uses data

in the Nifti format.

2.4 Medical diagnosis of Alzheimer's Disease

Currently a diagnosis of AD can only be definitive if it is diagnosed post-mortem, when the brain tissue can be analysed using a microscope. However, a diagnosis of AD can still be given to a patient, and taken directly from the DSM-IV-TR (Spitzer et al., 1980), the diagnosis criteria can be summarised as follows:

1. The patient's memory is impaired and they are suffering from one or more of the following (aphasia, apraxia, agnosia, disturbance in executive functioning).

2. The previously mentioned cognitive defects cause a large impairment in the subject's ability to function in social or occupational situations.

3. The cognitive defects are not caused by other central nervous system defects, or systemic conditions which have been known to cause dementia, or a mental disorder, or by the induction of substances.

4. There is a gradual onset of these cognitive defects and they are continuing to decline.

5. The cognitive defects do not only occur during the subject being in a state of delirium.

To measure the severity of these symptoms an examination is used.

2.4.1 Neuropsychological Tests

The Mini-mental State Examination (MMSE) is a test to determine the severity of a subject's symptoms relating to their memory and other mental abilities, which means it can be used to measure the severity of symptoms caused by MCI or AD. It is able to measure both the progression and severity of the disease. The subject is asked a series of questions which require them to use their cognitive abilities, and for every question they answer correctly they

score points. Thus at the end of the test, the amount of points they receive is an indication of their cognitive ability so a higher score means their mental ability is good whereas a lower score means their mental ability has been impaired.

There are a maximum of 30 points available, and the general time for the test to take is between 5 and 10 minutes meaning that the test is cheap and simple to perform. The questions can be grouped into seven categories with each category representing a different cognitive function (Tombaugh and McIntyre, 1992), the following is a list of these categories along with the type of questions used to evaluate them:

1. Time orientation: the subject is asked for the current year, season, date, day and month.

2. Place orientation: the subject is asked where they currently are - such as the county, the town, the hospital, and the floor they are on.

3. Attention and calculation: the subject is asked to count backwards from a number in multiples of another number, or spell a word backwards.

4. Registration of three words: the subject is told three words and asked to repeat them. If they fail this, then the recall category cannot be verified.

5. Recall of three words: the subject is told three words earlier in the test and at a certain point they are asked to remember what those three words were.

6. Visual construction: the subject is asked to copy an image.

7. Language: The subject is asked to name items such as a pencil and a wristwatch.

An example MMSE can be found in (Folstein et al., 1975). (Perneczky et al., 2006) suggests how the scores can be assigned to the severity of dementia: an MMSE score of 30 is healthy; 26-29 is questionable dementia; 21-25 is mild dementia; 11-20 is moderate dementia; and 0-10 is severe dementia.

Another test is the Wisconsin Card Sorting Test (WCST) (Puente, 1985), which is used for patients with acquired brain injury, mental illness, or a neurodegenerative disease, hence it can be used with AD sufferers. The WCST result will determine the performance of a number of frontal lobe functions. The test itself involves the participant being given a set of cards with images on, these cards will have different images and a different quantity of these images. Then the participant is asked to match the cards, but they are not told how to match them only that a given match is correct or incorrect. Using a modified version of the original WCST it was discovered that AD positive subjects make more mistakes than healthy subjects (Bondi et al., 1993).

The Wechsler Adult Intelligence Scale (WAIS) (Wechsler, 1955) is an IQ test to measure intelligence and cognitive ability; it tests verbal comprehension, perceptual reading, working memory and processing speed. It has multiple versions as it has been refined over the years since its inception and its current version is WAIS-IV (Wechsler, 2014). (Ryan and Paolo, 1989) discovered that a profile of a WAIS subtest can be associated with AD (however, it only detects AD and not other types of dementia).

2.4.2 Data-driven Classification of Alzheimer's Disease

The amount of data being collected is increasing dramatically in all fields; the term used to describe this is called big data (Murdoch and Detsky, 2013). Big data generally has five features: volume, value, velocity, variety and veracity (Kuo et al., 2014). The volume of data refers to the amount of data that needs to be processed. The value of the data is that the information it can give when it is analysed - what can be learned from the patterns in the data. The velocity of the data refers to the speed at which the data can be accessed, as data is likely to be continually growing. The variety of the data refers to the sources and types of data, for example in the healthcare industry there will be data of different types from different sources such as MRIs from MRI scanners and x-rays from x-ray machines. The veracity of data refers to the noise and bias in the data, data collection in most cases

is subject to various conditions such as noise or human error and this adds an additional challenge to the data being analysed.

The healthcare field has experienced this surge of data, as for example: the adoption of electronic health records in the US doubled in 2009 - 2011 (Charles et al., 2012); and in 2011, data from the US healthcare system reached 150 exabytes and has been increasing ever since (Cottle et al., 2013). The massive amount of data produced cannot all be analysed by humans as there is too much of it. However, data mining and machine learning techniques can be applied to discover patterns in the data. These patterns could lead to new discoveries of what correlates with diseases. One of the aims of the work in thesis was to discover these patterns in structural MRIs that link to a diagnosis of AD.

Problems of data-driven medical diagnosis are discussed in Obermeyer and Emanuel (2016). One such problem is overfitting where the algorithms can learn the data they are trained on too well and this can lead to exaggeration about the performance of the model and when it is used in the real world it may be much less accurate and diagnose incorrectly. To overcome this, the models should be tested on data sets from different populations that played no role in model development. Another problem is the computational resources required for certain machine learning algorithms. In Li et al. (2014) a convolutional neural network was used to predict HC, MCI and AD from MRIs; while they found promising results they stated that the neural networks were limited to two hidden layers due to the computational complexity required.

2.5 Biomarkers of Alzheimer's Disease

2.5.1 Discovery of a Diagnostic Biomarker for Alzheimer's Disease

One method to aid the early diagnosis of AD, is to find a diagnostic biomarker for AD. A biomarker is a biological feature which leads to a diagnosis in a subject. In general, biomarkers can be of three types: genetic, biochemical or neuroimage-based. Biomarkers of

each of these three types exist for AD. Genetic biomarkers exist for AD as heredity is the major causal factor in late-onset AD (Bergem et al., 1997). Biochemical markers exist as tests based on CSF Total-tau protein and CSF β-amyloid are a good indictor of AD (Hampel et al., 2004). Neuroimage-based biomarkers exist and will be discussed throughout the thesis.

In Ronald et al. (1998) there are recommended steps to follow to establish a biomarker for AD these are defined as:

1. "There should be at least two independent studies that specify the biomarker's sensitivity, specificity, and positive and negative predictive values."

2. "Sensitivity and specificity should be no less than 80%; positive predictive value should approach 90%."

3. "The studies should be well powered, conducted by investigators with expertise to conduct such studies, and the results published in peer-reviewed journals."

4. "The studies should specify type of control subjects, including normal subjects and those with a dementing illness but not AD."

5. "Once a marker is accepted, follow-up data should be collected and disseminated to monitor its accuracy and diagnostic value."

Some of the above steps have been fully met, one has been partially met, and others have not been met. The first criterion is met as every study in this report is based on previous studies in the literature by other authors. The second criterion can be evaluated using the results of a study as the work throughout this thesis produces a number of metrics. The third criterion can be met as some of the work has been peer reviewed. The fourth criterion can be met by including healthy subjects in the evaluation of biomarkers. The fifth criterion cannot be met in the timeframe of this thesis as the biomarkers will have to monitored after its publication. Ronald et al. (1998) also define a set of properties a diagnostic biomarker for AD should meet:

1. “Able to detect a fundamental feature of Alzheimer’s neuropathology”,
2. “Validated in neuropathologically confirmed AD cases”,
3. “Precise (able to detect AD early in its course and distinguish it from other dementias)”,
4. “Reliable”,
5. “Non-invasive”,
6. “Simple to perform”, and
7. “Inexpensive”.

The above properties can be shown to apply to a structural MRI biomarker as the discovery of a biomarker will mean that the first two criteria are met, as it will be an indictor of AD; the precision and reliability of the biomarker will be based on the performance of the biomarker when it is evaluated; and the last three criteria are automatically met due to the adoption of structural MRI.

Chapter 3

A Review of MRI Processing

This chapter will discuss the theoretical models on which MRI processing is based on and how these relate to the processing tasks performed by Freesurfer, the particular software adopted in this study.

3.1 Transforming a Magnetic Resonance Image

Affine transformations are linear transformations which can be used to modify an image. These are the simplest transformations which are used. A translation is shown in Figure 3.1; a rotation in Figure 3.2; scaling in Figure 3.3; and shearing in Figure 3.4. These transformations can be defined mathematically, and will assume they are being applied to two dimensional data.

A translation can be defined by a translation vector and is added to a position vector. The position vector is the x and y coordinates which are being translated:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} a \\ b \end{bmatrix}. \tag{3.1}$$

A rotation multiplies a position vector by a rotation matrix, where θ is the amount to rotate the image (clockwise from the origin) by:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}. \tag{3.2}$$

Scaling multiplies a position vector by a scaling matrix, where k defines the amount to scale by. If $k < 1$ the image will be shrunk; if $k = 1$ then there is no change; if $k > 1$ then the image will be enlarged. The following equation uses two versions of k, with k_x scaling along the x-axis and k_y scaling along the y-axis:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} k_x & 0 \\ 0 & k_y \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}. \tag{3.3}$$

Shearing multiplies a position vector by a shearing matrix, where k_x defines the amount to shear by on the x-axis and k_y is the shear amount on the y-axis:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} 1 & k_x \\ k_y & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}. \tag{3.4}$$

Note that these transformations are generally applied to MRIs in three dimensions using similar parameters to the 2D transformations, the 3D translation is computed by:

$$\begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} = \begin{bmatrix} x \\ y \\ z \end{bmatrix} + \begin{bmatrix} a \\ b \\ c \end{bmatrix} \tag{3.5}$$

Rotation in 3D can be applied to three axes (x, y and z), rotation on each of the three axes uses a different rotation matrix, respectively the rotation matrices for the x, y and z axes are:

$$R_x(\theta) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -\cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{bmatrix}, \tag{3.6}$$

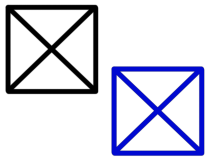

Figure 3.1: A graphical example of an affine translation, the black box is the original and the blue box is it after translation

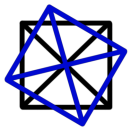

Figure 3.2: A graphical example of an affine rotation, the black box is the original and the blue box is it after rotation

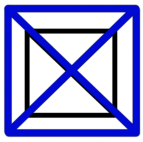

Figure 3.3: A graphical example of an affine scale, the black box is the original and the blue box is it after scaling

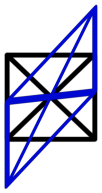

Figure 3.4: A graphical example of an affine shear, the black box is the original and the blue box is it after shearing

$$R_y(\theta) = \begin{bmatrix} \cos\theta & 0 & \sin\theta \\ 0 & 1 & 0 \\ -\sin\theta & 0 & \cos\theta \end{bmatrix}, \tag{3.7}$$

$$R_z(\theta) = \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{3.8}$$

And the rotation matrix is multiplied with the original position to generated the rotated coordinates:

$$\begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} = R_i(\theta) \begin{bmatrix} x \\ y \\ z \end{bmatrix}. \tag{3.9}$$

Scaling is computed by:

$$\begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} = \begin{bmatrix} k_x & 0 & 0 \\ 0 & k_y & 0 \\ 0 & 0 & k_z \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix}. \tag{3.10}$$

Shearing is computed by:

$$\begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} = \begin{bmatrix} 1 & sh_x^y & sh_x^z \\ sh_y^x & 1 & sh_y^z \\ sh_z^x & sh_z^y & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix}. \tag{3.11}$$

When shearing along the z-axis, the z coordinate is untouched and it is the x and y coordinates which are affected.

Affine transformations can be extended by breaking the original image down into multiple sections and different affine translations are applied on these different sections, these are called piecewise linear transformations. Because affine transformations are linear they can be quite limited; nonlinear transformations allow for greater flexibility. Nonlinear transformations can be defined as basis functions such as the polynomial expansion:

$$x' = a_1 + a_2x + a_3y + a_4x^2 + a_5y^2 \tag{3.12}$$

$$y' = b_1 + b_2x + b_3y + b_4x^2 + b_5y^2 \tag{3.13}$$

3.1.1 Cost Function

A cost function must be defined which produces a value based on how similar two images are. A cost function should return a low value when images are similar, and a high value when they are different. If we define A_v and B_v as being the intensity of the v^{th} pixel in images A and B respectively we can define some common cost functions. Such as:

Using the residual sum of squares:

$$C(A,B) = \sum_{v=1}^{n} (A_v - B_v)^2 \tag{3.14}$$

Using the normalised correlation:

$$C(A,B) = 1 - \frac{\sum_{v=1}^{n} A_v B_v}{\left(\sum_{v=1}^{n} A_v^2\right)^{0.5} \left(\sum_{v=1}^{n} B_v^2\right)^{0.5}} \tag{3.15}$$

3.1.2 Estimating the Transformation

To align two images, the set of transformations must be found which is consistent with the lowest resulting cost value when one transformed image is compared to another. Optimisation methods are used to determine the set of transformations to use as it is infeasible to perform a brute-force approach where every possible set of transformations are applied. Typically for these problems a gradient descent is performed, however, the downside to this method is that they can easily get stuck in a local optimum with no way to improve. The two methods used instead are regularisation and multiscale optimisation. Regularisation involves shrinking any coefficients produced by a model to as close to zero as possible in order to reduce the complexity of the model. In the context of MRI transformation, it will apply a penalty to more complex transformations to encourage simpler, smoother transformations. The idea behind multiscale optimisation is to start estimating the transformations at a low resolution (these are transformations which affect large regions of the brain and thus have a large effect). Once these larger features are aligned, it will then start looking at and trying to align the smaller features.

3.1.3 Intensity Interpolation

Once the transformation parameters have been estimated, they are then applied to the original MRI to create the transformed MRI. To create the transformed MRI, for each transformation it is necessary to take the original voxels and their intensities and transform them to

create a set of new voxels. However, it is likely that the transformed voxel will reside in more than one position meaning you cannot just map the intensity of that voxel to another voxel. The target voxel may contain multiple voxels such as $\frac{1}{5}$ of voxel A, $\frac{2}{5}$ of voxel B, and $\frac{2}{5}$ of voxel C. Thus interpolation methods must be used to determine the intensity value of this new voxel. Using nearest neighbour interpolation, the new voxel takes the intensity of the nearest transformed voxel. This is the simplest method but it has downsides such as making the new image look blocky (as there tends to be big changes from the intensity of one voxel to the intensity of its neighbour).

Tri-linear interpolation is applied in three dimensions and takes a weighted average of the adjacent voxels, this is a fairly fast method but has the downside of blurring the image. Tri-linear interpolation on a unit cube will be defined. If we denote the eight closest neighbours of the voxel we wish to interpolate as $(0,0,0),(0,0,1),...,(1,1,1)$, see Figure 3.5 for a visual example. Let U_{ijk} denote the value of the intensity of a voxel at (i,j,k). Assume we have a set of basis functions $\lambda_{ijk}(x,y,z)$, such that $\lambda_{ijk}(i,j,k) = 1$ and for all others 0 is returned. The set of basis functions are defined below:

$$\begin{aligned}
\lambda_{000} =& (1-x)(1-y)(1-z) \\
\lambda_{001} =& (1-x)(1-y)z \\
\lambda_{010} =& (1-x)(1-z)y \\
\lambda_{011} =& (1-x)yz \\
&\vdots \\
\lambda_{111} =& xyz
\end{aligned} \tag{3.16}$$

Then the formula for tri-linear interpolation is:

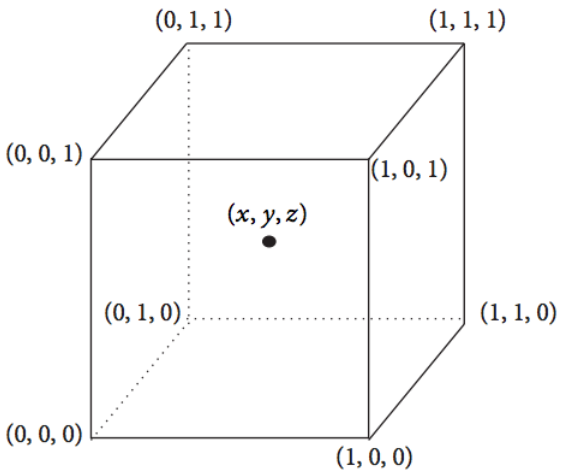

Figure 3.5: A visual representation of tri-linear interpolation.

Source: National Institutes of Health Center for Information Technology, Rockville, MD, USA.

$$\hat{u}(x,y,z) = \sum_{i,j,k \in \{0,1\}} U_{ijk} \lambda_{ijk}(x,y,z) \tag{3.17}$$

As this was defined for a unit cube it guarantees that there will be a voxel at the position we wish the interpolate, for example if $(x,y,z) = (0,1,0)$ then only λ_{010} is 1 and: $\hat{u}(0,1,0) = U_{010}$. In practice, tri-linear interpolation will not be applied to a unit cube, thus the above formulae must be adjusted to handle a non-unit cube, one way to accomplish this is to rewrite the basis functions such that the integer portion of (x,y,z) is removed and it will lie in the unit cube. First we define $x', y',$ and z' as being the input coordinates mapped to the unit cube:

$$\begin{aligned} x' &= xn_x - \lfloor xn_x \rfloor \\ y' &= yn_x - \lfloor yn_y \rfloor \\ z' &= zn_x - \lfloor zn_z \rfloor \end{aligned} \tag{3.18}$$

And then we define the basis functions using the mapped coordinates:

$$
\begin{aligned}
\lambda_{000} =& (1-x')(1-y')(1-z') \\
\lambda_{001} =& (1-x')(1-y')z' \\
\lambda_{010} =& (1-x')(1-z')y' \\
\lambda_{011} =& (1-x')y'z' \\
&\vdots \\
\lambda_{111} =& x'y'z'
\end{aligned}
\tag{3.19}
$$

Higher-order interpolation refers to a group of interpolation methods which operate on a broader area than tri-linear interpolation. In the transformation of MRIs, the Whittaker-Shannon interpolation (or sinc interpolation) (Whittaker, 1935) is the most commonly used. The sinc function is defined by:

$$
\operatorname{sinc}(x) = \frac{\sin(\pi x)}{\pi x} \tag{3.20}
$$

And the unbounded sinc interpolation function is:

$$
\hat{s}(x,y,z) = \sum_{x'=-\infty}^{\infty} \sum_{y'=-\infty}^{\infty} \sum_{z'=-\infty}^{\infty} \left(i(x',y',z') \operatorname{sinc}(x-x') \operatorname{sinc}(y-y') \operatorname{sinc}(z-z') \right) \tag{3.21}
$$

Where $\hat{s}(x,y,z)$ is the interpolated voxel intensity of the transformed MRI at location (x,y,z) and $i(x',y',z')$ is the uninterpolated transformed image intensity at (x',y',z'). Theoretically, the sinc interpolation is applied to every voxel in the image, however, in practice due to the size of the MRI, it is applied with a window to be computationally faster.

3.2 Motion Correction

While in an MRI scanner, a subject may move their head and even a minor movement will have a notable effect on the resulting MRI. Since MRI slices are recorded one after the other, then the movement of a head will cause the image to move between slices; this is called bulk motion. To account for this bulk motion, MRI preprocessing typically applies motion correction techniques to realign the images between the MRI slices. When bulk motion is detected between slices, the techniques described in Section 3.1 are such that the misaligned image is aligned with the previous image. It is recommended to use an image from the centre of the slices as a base and align all other images to that one as on average this image is the most likely to be the closest to all other images in the slices. Cost functions are used in this process to evaluate the alignment.

3.3 Noise Removal

Noise in an MRI can be defined as interference in the MRI signal causing an irregular granular pattern across the MRI, this is represented by abnormal voxel intensities at various points on the MRI. Noise removal can be implemented by transforming the MRI into a frequency domain where the noise components are separate from the actual MRI data.

A Fourier transform transforms data in the spatial (or temporal) domain into the frequency domain, in this case, the space domain will be the MRI voxels. The frequency domain will contain components which each represent a particular frequency in the spatial domain, Figure 3.6 shows an image in the spatial domain, and the same image once transformed to the frequency domain. Once in the frequency domain, various filters can be applied with the aim to remove noise in the spatial domain. A low-pass filter can be used to remove low frequency signals, while a high-pass filter can be used to remove high frequency signals. This means that generally a low-pass filter will smooth out noise while high-pass filter will amplify it. After a filter has been applied, the image can be transformed back to the spatial domain

Figure 3.6: A spectrum image computed with the Fourier transform. The left image is the spatial domain, the right image is the spectrum image.

via an inverse Fourier transform.

The Fourier transform of a function $f(t)$ and the inverse Fourier transform are denoted respectively by:

$$\mathscr{F}\Big(f(t)\Big) = F(\omega) = \int_{-\infty}^{\infty} f(t) e^{-2\pi i \omega t} dt \tag{3.22}$$

$$\mathscr{F}^{-1}\Big(F(\omega)\Big) = f(t) = \int_{-\infty}^{\infty} F(\omega) e^{2\pi i \omega t} d\omega \tag{3.23}$$

The above notation is for the continuous Fourier transform. The discrete Fourier transform is the Fourier transform generalised to a discrete function, the discrete Fourier transform and its inverse are denoted respectively by:

$$F_n = \sum_{k=0}^{N-1} f_k e^{-i\frac{2\pi}{N}kn} \tag{3.24}$$

$$f_k = \frac{1}{N} \sum_{n=0}^{N-1} F_n e^{i\frac{2\pi}{N}kn} \tag{3.25}$$

3.4 Normalisation to Atlases

To compare individuals' MRIs they must first be aligned to a template so the locations of the regions within their brain are roughly in the same location. A 3D proportional grid was proposed by Talairach (1967) and this was based on anatomical landmarks which create a bounding box showing the extent of the size of the brain in each dimension, this bounding box is often referred to as the Talairach bounding box; while the anatomical landmarks are known as the Talairach landmarks and the coordinate space is referred to as the Talairach space.

An atlas is a map providing the locations of important anatomical features in a coordinate space; in the context of MRIs, an atlas points out the location of anatomical features in the previously defined coordinate space. The most common atlas is the one which was defined by Talairach and updated (Talairach and Tournoux, 1988). Once data has been normalised to the Talairach space then it is simple to apply the atlas and find the location of various anatomical features, however the normalisation itself is not a simple process.

A template is a representative image of the atlas and is used as a target to which individual MRIs can be mapped. The most commonly used templates are the Montreal Neurological Institute (MNI) templates. These templates were developed with the aim to be used in automated segmentation of the MRI. The first commonly used MNI template is the MNI305 which was created by aligning 305 MRIs to the Talairach atlas based on anatomical features (Evans et al., 1993). Once the 305 images were aligned, a mean image was created by averaging the 305 aligned MRIs, and then the 305 initial images were aligned to the mean image.

3.5 Non-brain Tissue Removal

When an MRI is created, it does not just image the brain of the subject; their skull, eyes and other non-brain tissue are scanned too. Thus to analyse the brain, the skull needs to be

removed from the MRI. Typically a watershed algorithm is used to achieve this; a watershed algorithm is an image processing technique used for image segmentation. There are various implementations of the watershed algorithm, one of the most common is Meyer's flooding algorithm (Beucher and Meyer, 1992). The watershed algorithm treats an image as if it were a topological surface with high intensity areas represented as hills and low intensity areas as valleys. Each valley is then filled with water and as the water level rises, valleys will merge together connected by the same pool of water. Where the water merges together, barriers are created, and this process is repeated until all the hills are submerged in water. The created barriers give the segmentation result.

3.6 Tissue Segmentation

Another key part of MRI preprocessing is to separate the different tissues (WM, GM and CSF). It is not accurate enough to determine threshold intensity values to determine which voxels are WM, GM or CSF; this is due to noise varying the intensities at voxels, overlapping intensities of WM and GM, voxels containing multiple types of tissue. Clarke et al. (1995) review a number of techniques for tissue segmentation, and typical methods use a probabilistic atlas to determine the probability of a certain voxel being a certain type of brain tissue.

3.7 Cortical Reconstruction by Freesurfer

In this section, the reconstruction process of Freesurfer will be summarised, the process can be broken into two stages: the surface-based pipeline (more information can be found in Dale et al. (1999) and Fischl et al. (1999)), and the volume-based pipeline (further details in Fischl et al. (2002b) and Fischl et al. (2004)).

The first step in the reconstruction process is for the input MRI to be converted from the format it is currently in (mostly likely the native scanner format) into the COR format which

is used by Freesurfer (the format is designed to store high-resolution structural data[1]). If multiple scans exist then they can be averaged together with the aim to correct any motion of the subject in the scanner. MRIs are then aligned to a common template to make comparisons possible. The next step aims to correct the MRI where the signal intensity varies smoothly across an image. This tends to be caused by factors of the scanner such as poor RF field uniformity or magnetic field gradients switching and creating eddy currents. It uses a method entitled Non-parametric Non-uniform Intensity Normalisation (Boyes et al., 2008) and this method works independently of any knowledge about the MRI scan itself, it solely works with the image data.

Then the set of transformations required to transform the original MRI onto the MNI305 atlas are computed (Evans et al., 1994), so after this step the MRI can be mapped to a common atlas (refer to Section 3.4 for more information). Next the intensity of the voxels are normalised such that the average intensity of a WM voxel is 110, this attempts to correct for fluctuations in intensity and makes intensity-based segmentation easier. Then Freesurfer attempts to remove the skull from the MRI, so afterwards only the brain remains. A watershed algorithm is used and it enables parameters to be adjusted so that is it more or less likely to remove parts of the brain. This means that if the skull stripping has removed part of the brain, then you can adjust parameters to make this less likely to happen on the next run.

The next step finds the set of linear transformations which map the MRI to the default gaussian classifier atlas (GCA) (Fischl et al., 2002a), this is a proprietary format comprised of three volumes: the mean intensity of the most likely label at each voxel, the indices of the most likely label at each voxel, and the probability of the label occurring at each voxel. Next canonical normalisation is performed based on the GCA, and then the set of nonlinear transformations to map the MRI onto the GCA are found. The neck is then removed from the MRI and the linear transformation to map to the GCA is found. Then the subcortical

[1]https://surfer.nmr.mgh.harvard.edu/fswiki/FsTutorial/CorFormat

structures are labelled based on the GCA model, as well as statistics based on the segmented cortical structures.

Next another intensity normalisation is performed, this time it is applied to the brain after the skull stripping and thus performs better. After this the WM segmentation is performed where Freesurfer attempts to extract the WM from everything else in the MRI. In the next step based on the extracted WM, the mid brain is cut from the cerebral cortex and the hemispheres are cut from each other, this creates the subcortical mass which is then tessellated to produce a triangulation of the subcortical surface containing vertices (based on the triangulation), this surface will be referred to as the original surface. The problem is that because the tesselation is based on the voxels, then the triangles will be at right angles to each other, so an algorithm is run to smooth the triangulation. The surface is then inflated with the aim to preserve distances and areas. The inflated surface is then checked for topology defects such as holes which are then fixed on the original surface. Based on the original surface; the white surface, pial surface and thickness and curvature data are created. Finally a binary volume (where voxels are 0 or 1) is made of the cortical ribbon with voxels being a 1 if they belong to the ribbon.

Spherical inflation is applied to the original surface so it can be applied to both an ipsilateral hemisphere and contralateral hemisphere atlas. The average curvature of the atlases for the subject are resampled, this allows for brain activity to be displayed on a folded surface. Next cortical parcellation takes place where an anatomical label is given to each cortical structure as well as statistics about the parcellation.

3.8 Freeview - Visualisation of Reconstruction by Freesurfer

Freeview is software provided by Freesurfer to enable the visualisation of the brain and the regions into which it has been divided by Freesurfer. It can be used to give each region of the MRI a unique colour, and then when that region is selected (via a mouse click) the

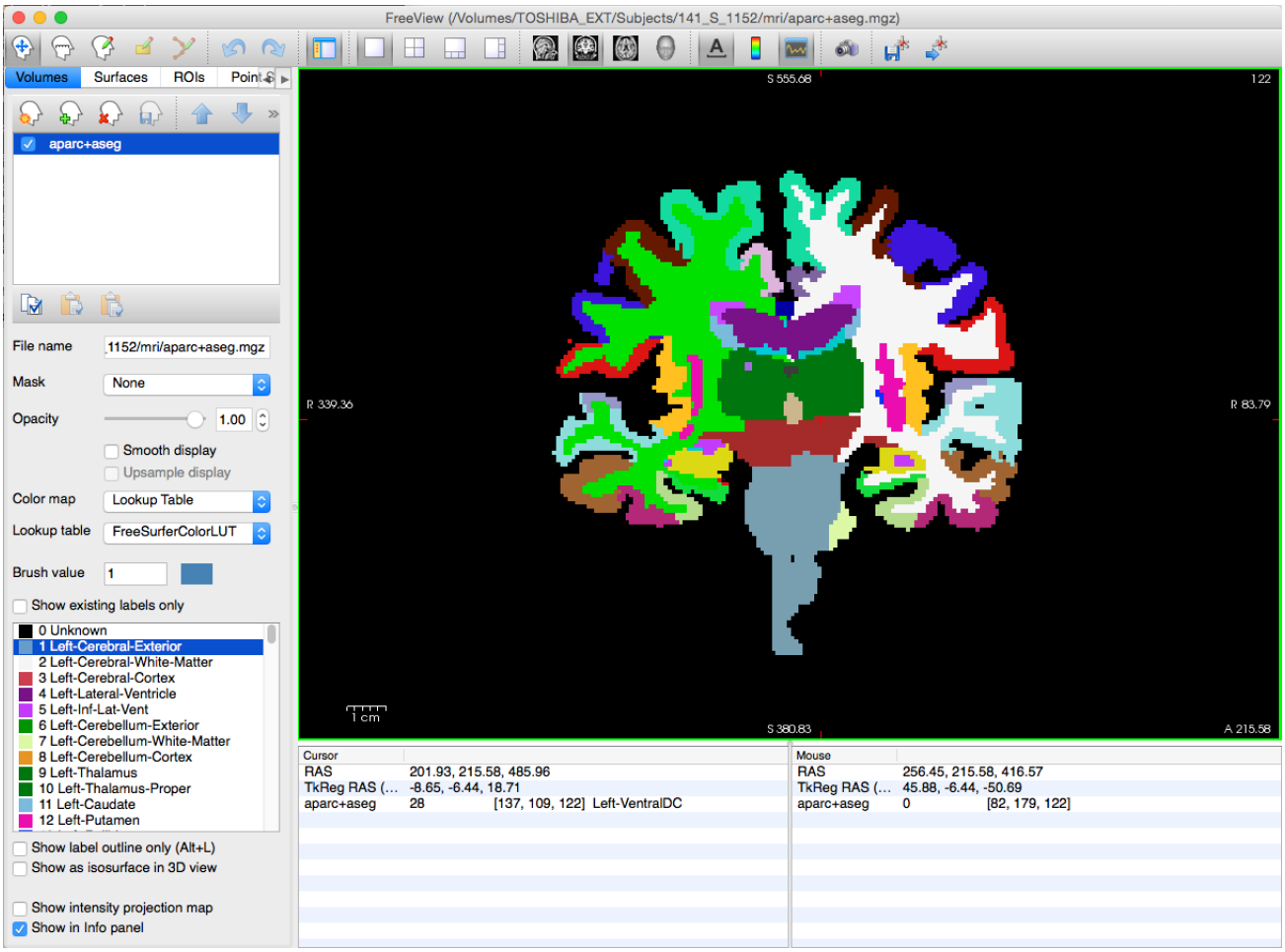

Figure 3.7: Freeview being used to view the volumes segmented by Freesurfer

system will tell you which that region is. An image of Freeview being used to visualise the segmentations of a brain is shown in Figure 3.7 and a screenshot of Freeview being used to view the extracted pial surface is shown in Figure 3.8.

3.9 Processing the Subject Data

The MRI data used in this thesis are all processed by Freesurfer, and since Freesurfer is not multi-threaded, nor does it take advantage of a graphics processing unit (GPU), to process the large number of subjects, it has been used in conjunction with GNU Parallel (Tange, 2011), to enable an instance of Freesurfer to run on each core of the central processing unit. All of the MRI data was processed using the same version of Freesurfer and the same version

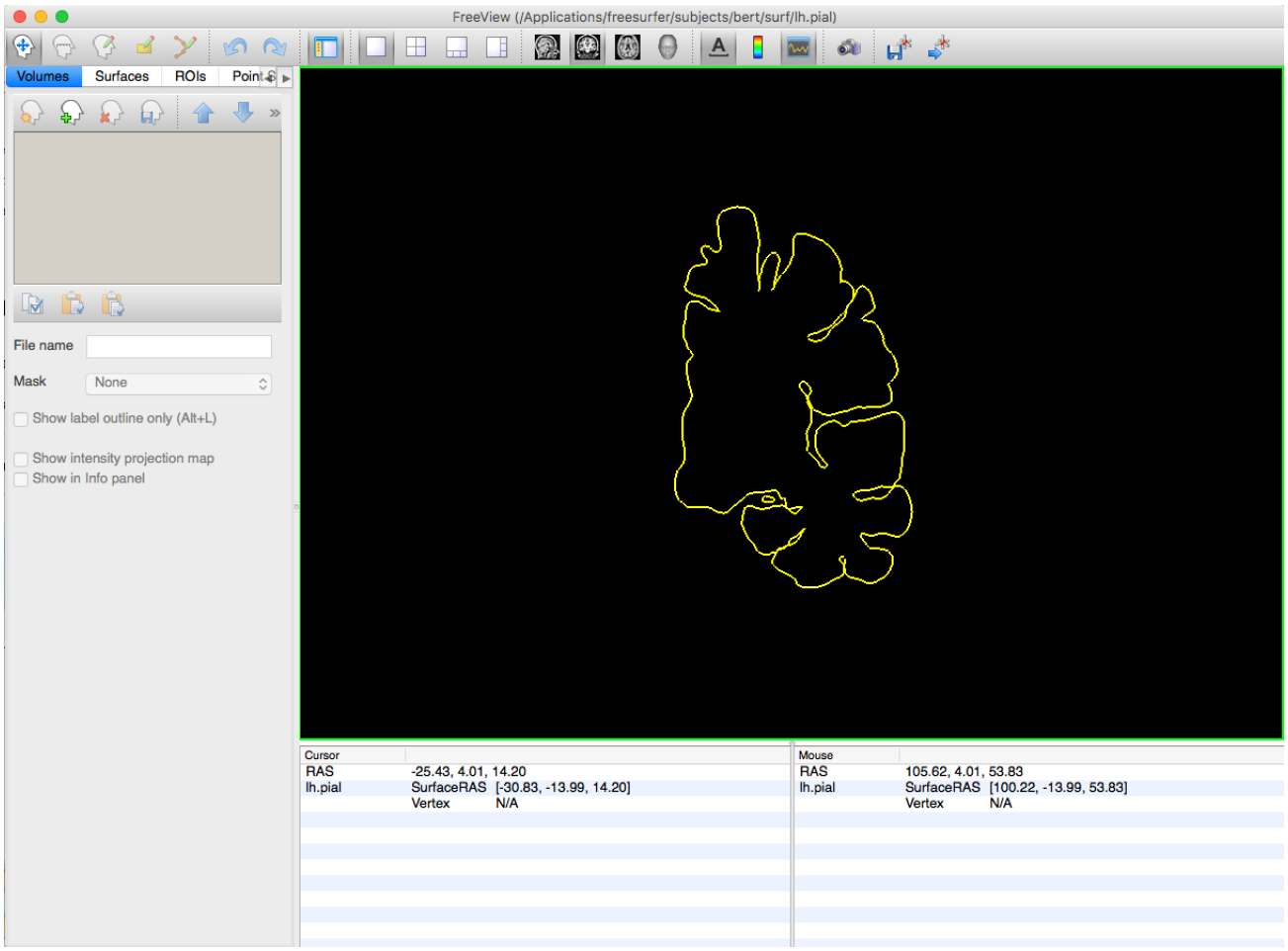

Figure 3.8: Freeview being used to view the pial surface generated by Freesurfer

of the operating system (as according to Gronenschild et al. (2012) and Chepkoech et al. (2016), differences in these two factors can affect the output). The specific command used was:

```
recon-all -i /inputdata/mri*.nii.gz -subjid mri* -sd /outputdata -all -hippo-
    subfields
```

Chapter 4

Literature Review of Machine Learning Techniques

4.1 Data Mining

Data mining is the process of discovering patterns in large sets of data with the goal of extracting information from the data transforming it into a format understandable for users. It is one of the five steps of the knowledge discovery in databases (KDD) process (Fayyad et al., 1996):

1. Selection,
2. Pre-processing,
3. Transformation,
4. Data Mining, and
5. Evaluation.

The selection step comprises deciding which data will be used: in this thesis it involves the selection of structural MRIs from various databases. The pre-processing step is the conversion

of the MRI images to numerical data using Freesurfer; the elimination of subjects whose MRIs were not processed correctly; elimination of outlying subjects and the normalisation of the data. The data mining stage typically involves one of six tasks (Fayyad et al., 1996):

1. Anomaly Detection,
2. Association Rule Learning,
3. Clustering,
4. Classification,
5. Regression, and
6. Summarisation.

Anomaly detection is detecting unusual MRI measurements, perhaps a subject has a volume which is two times greater than any other subject, this could be caused by an error in the preprocessing step or they may just have an unusually sized brain; anomaly detection has been implemented in Chapter 9. Association rule learning generate rules between variables based on their relationships to each other. Clustering finds groups or structures in the data without relying on class labels. Neither clustering nor rule detection techniques have been implemented in this work, but are discussed in Section 10.1.1 as a way to extend the work in this thesis. Classification is a model learning the class labels of one set of data, and then the model is applied to a set of data with unknown class labels to predict them; classification is implemented throughout this thesis in Chapters 7, 8 and 9. Regression is similar to classification, except rather than predicting a nominal class label, it predicts a continuous value; it was implemented in Chapter 9. Summarisation involves generating a representation of the data which is easier to understand than the raw data itself such as report generation or visualisation; visualisations are produced in Chapters 7, 8, 9 to explain the data or the development of the methods in this thesis.

4.2 Machine Learning

Machine learning is a sub-field of artificial intelligence where computers are able to learn how to complete a given task rather than being programmed to complete the task with step-by-step instructions. Machine learning algorithms are able to learn from input values and a model is built enabling decisions or predictions to be made, these learned rules can then be applied to new inputs which have not been seen by the model before and a classification or decision is made based on the new data. Machine learning has many applications with the common ones including filtering spam emails, computer vision, optical character recognition and medical diagnosis (the latter being the element that this thesis focuses on). Machine learning problems can be categorised into two groups: supervised learning and unsupervised learning. Supervised learning is where the model is trained using a set of data which already have class labels, an example would be classification of AD from a structural MRI where in order to learn the model is given structural MRIs as input and it is told the diagnosis of each structural MRI, it can then learn which features are associated with AD and so be applied to new data and predict the class labels. Unsupervised learning is where the model must find patterns based on input data without class labels. An example would be to group unlabelled MRIs by whether they have a brain tumour or not - brain tumour diagnosis requires an invasive process to diagnose it.

4.3 Classification Algorithms

Classification algorithms are a set of algorithms which build a model using supervised learning in order to be able to classify new instances of data as accurately as possible. A more formal definition would be, given an n-dimensional input vector $\overline{x} = [\, x_1 \cdots x_n \,]$ to predict its output class y where $y \in C$ and C is the set of all possible classes. Each input instance of data the classifier is trained on is in the form $(\overline{x}, y)$ and the classifier builds a model to predict the class label of new data which could be represented by: $\hat{y} = f(\overline{x})$ where $\hat{y}$ is the predicted

class of the data. The aforementioned notation covers a single data observation, the input and classes of multiple data observations can be denoted by $(X, \overline{y})$ where X is a matrix of all data observations and $\overline{y}$ is the vector with all of the classes of the observations.

4.3.1 Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is a binary classification technique (Fisher, 1936) which involves finding a combination of features for the data which gives both the maximum separation between the means of each class when projected to this new dimension, and minimises the variances for each class when projected to the new dimension. This transformation can be expressed by:

$$f(\overline{x}) = w_0 + \sum_{i=1}^{n} w_i x_i^T \tag{4.1}$$

Where $\overline{w} = [\, w_0 \cdots w_n \,]$ denotes the decision boundary the LDA algorithm will determine. A feature vector, $\overline{x}$, is classified as C_0 or C_1, dependent on which side of the decision boundary it falls: if $f(\overline{x}) \geq 0$, then it is classified as C_0; if $f(\overline{x}) < 0$, then it is classified as C_1. An example of a decision boundary found using LDA is shown in Figure 4.1, the decision boundary is denoted by the solid line and is perpendicular to the feature weight vector which is denoted by the dashed line.

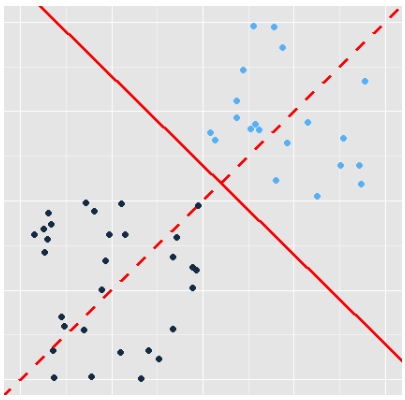

Figure 4.1: An example of the decision boundary created using LDA where the solid line is the decision boundary and the dashed line is the normal position of the decision boundary

To find the decision boundary, $\overline{w}$ needs to be found such that the two above criteria are optimised (maximise class separation, minimise class variance). The first step is to find the scatter matrices (unnormalised covariance matrices) for the between-class and the within-class. If we define the mean vector of both classes as $\overline{\mu}_0$ and $\overline{\mu}_1$, the overall mean (of both classes) as $\overline{\mu}$, and the number of observations in a class as N_i, then the between-class scatter matrix can be defined as:

$$S_b = \sum_{i=0}^{1} N_i(\overline{\mu}_i - \overline{\mu})(\overline{\mu}_i - \overline{\mu})^T. \tag{4.2}$$

Next we define the covariances of both classes as Σ_0 and Σ_1. The within-class scatter matrix can be defined as:

$$S_w = \sum_{i=0}^{1} \sum_{j=1}^{N_i} (\overline{x}_j - \overline{\mu}_i)(\overline{x}_j - \overline{\mu}_i)^T. \tag{4.3}$$

Since the aim of LDA is to minimise S_w and maximise S_b, the optimisation problem LDA solves is to find $\overline{w}$ which optimises:

$$\overline{w}_{LDA} = \max_{\overline{w}} \frac{\overline{w}^T S_b \overline{w}}{\overline{w}^T S_w \overline{w}}. \tag{4.4}$$

The magnitude of $\overline{w}$ is irrelevant and thus the optimisation can be simplified to:

$$\overline{w}_{LDA} = \max_{\overline{w}} \overline{w} S_b \overline{w}^T \text{ subject to } \overline{w} S_w \overline{w}^T = 1. \tag{4.5}$$

4.3.2 Naive Bayes Classifier

The Naive Bayes model assumes a set of independent features as input, $\overline{x} = x_1, ..., x_n$. The input variables are then conditioned on an output variable which is the class of $\overline{x}$ denoted by y where $y \in C$. We can then define the likelihood function which returns the probability that an observation $\overline{x}$ belongs to class c:

$$\text{likelihood} = P(\overline{x} \mid c). \tag{4.6}$$

We can calculate the prior probability, which is the probability a randomly selected observation belongs to a class c where $c \in C$ as:

$$\text{prior} = P(c) = \frac{\text{Number of observations of class } c}{\text{Number of observations}}. \tag{4.7}$$

And we can calculate the evidence as:

$$\text{evidence} = P(\overline{x}) = \frac{\text{Numbers of observations that are } \overline{x}}{\text{Number of observations}}. \tag{4.8}$$

Bayes' Rule is shown in Equation 4.9, and is used to obtain the posterior probability. The posterior probability in this case is $P(c|\overline{x})$: the probability of a given observation being class c.

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} \qquad \text{posterior} = \frac{prior \times likelihood}{evidence}. \tag{4.9}$$

If we apply Bayes' rule to the observations we get:

$$P(c \mid \overline{x}) = \frac{P(\overline{x} \mid c)P(c)}{P(\overline{x})}. \tag{4.10}$$

Since the evidence does not depend on c, and the values of $\overline{x}$ are given, it is a constant thus we can state:

$$P(c \mid \overline{x}) \propto P(\overline{x} \mid c)P(c). \tag{4.11}$$

$P(\overline{x} \mid c)P(c)$ is equivalent to the joint probability model, and can be rewritten using the chain rule for repeated applications of conditional probability:

$$P(c \mid \overline{x}) \propto P(c) \prod_{i=1}^{n} P(\overline{x}_i \mid c). \tag{4.12}$$

Now we have the feature model, the Naive Bayes' classifier adds a decision rule to this model by selecting the class with the highest probability:

$$\hat{y} = \max_{c \in C} P(c) \prod_{i=1}^{n} P(\overline{x}_i \mid c). \tag{4.13}$$

4.3.3 Support Vector Machine

The design behind the support vector machine (SVM) (Vapnik, 1995) was to create a classifier which does not overfit the data by controlling the complexity of the model by only picking important data instances when building the separating hyperplane to distinguish between the classes. An SVM can be defined by the following optimisation problem:

$$\min_{\overline{w},b,\overline{\xi}} \frac{1}{2} \overline{w} \overline{w}^T + C \sum_{i=1}^{n} \overline{\xi}_i \tag{4.14}$$

$$\text{subject to: } y_i\big(\overline{w}\phi(\overline{x}_i)^T + b\big) \geq 1 - \xi_i \text{ and } \xi_i \geq 0$$

Where $\overline{x}_i \in R^m$, $\overline{y} \in \{1, -1\}^l$, n is the number of training data instances, and ϕ is a function mapping $\overline{x}_i$ to a higher dimensional feature space (potentially infinite dimensions). C is the cost hyperparameter and is the penalty parameter of the error term, it is always greater than zero. A low value of C can be used to create soft-margin SVMs while a high value creates hard-margin SVMs. A soft-margin SVM is one which allows for misclassifications; whereas a hard-margin SVM penalises miscalculations greatly and thus can lead to overfitting. ξ_i refer to slack variables which allow the SVM to learn training data and be able to misclassify some data. $\overline{w}$ is a vector of coefficients, and b is a single coefficient representing the hyperplane. Solving this optimisation problem will return a linearly separating hyperplane for the data; even if the data is non-linearly separable, an SVM is able to transform it to a higher dimension

where it may be linearly separable and return the hyperplane in that dimension.

A kernel function is applied to the inputs $\overline{x}$ to map them to a different feature space. A kernel function is denoted as $k(\overline{x}_i, \overline{x}_j)$, and this maps $\overline{x}$ to $\phi(\overline{x})$. Examples of commonly used kernels include:

$$\begin{aligned} \text{Linear: } & k(\overline{x}_i, \overline{x}_j) = \overline{x}_i \cdot \overline{x}_j \\ \text{Polynomial: } & k(\overline{x}_i, \overline{x}_j) = (\overline{x}_i \cdot \overline{x}_j + a)^b \\ \text{Sigmoidal: } & k(\overline{x}_i, \overline{x}_j) = \tanh(a\overline{x}_i \cdot \overline{x}_j - b) \\ \text{Radial Basis Filter: } & k(\overline{x}_i, \overline{x}_j) = \exp(-\gamma ||\overline{x}_i - \overline{x}_j||^2) \end{aligned} \tag{4.15}$$

Note, that while kernels can allow the SVM to handle more complex patterns in the data, some of the kernel functions require additional parameters such as the radial basis filter (RBF) kernel requiring γ meaning that the use of a kernel function may mean there is extra tuning required to optimise the parameters.

Examples of the linear SVM (both hard and soft margin) are shown in Figures 4.2 and 4.3; and examples of an SVM with the RBF kernel are shown in Figures 4.4 and 4.5.

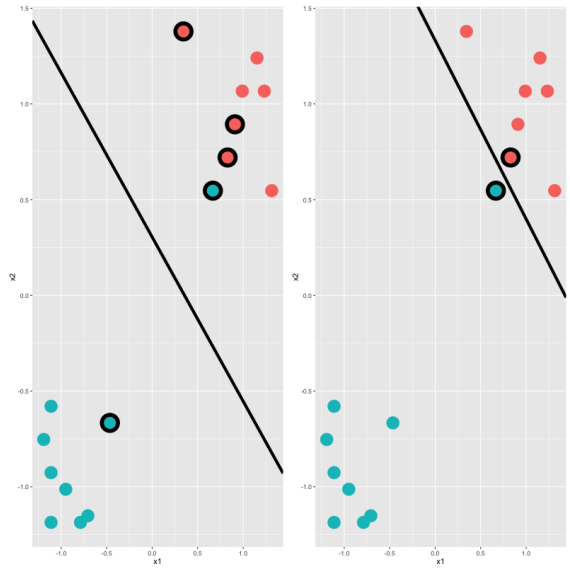

Figure 4.2: Linear SVMs solving a linearly separable problem

Left: Soft-margin SVM, $C = 1$
Right: Hard-margin SVM, $C = 10^{10}$

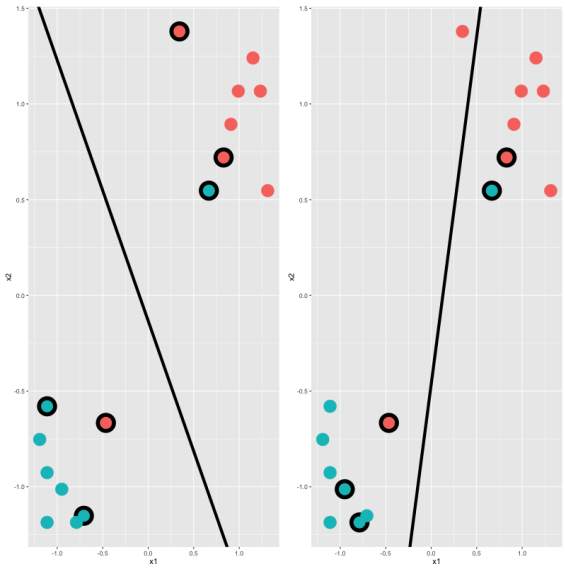

Figure 4.3: Linear SVMs solving a non-linearly separable problem

Left: Soft-margin SVM, $C = 1$
Right: Hard-margin SVM, $C = 10^{10}$

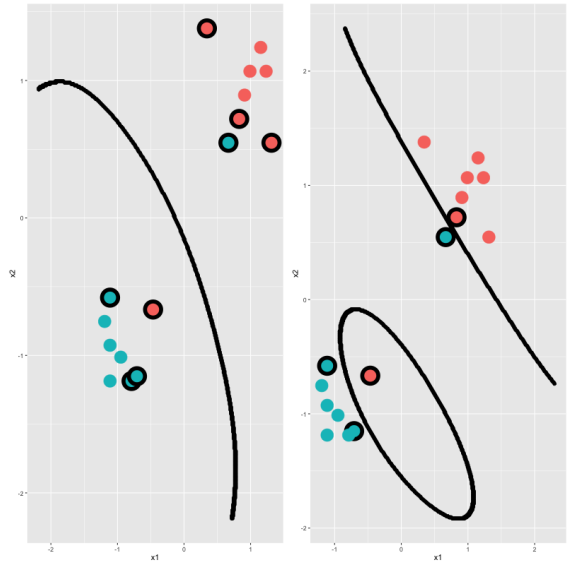

Figure 4.4: RBF SVMs solving a complex non-linearly separable problem

Left: Soft-margin SVM, $C = 1$
Right: Hard-margin SVM, $C = 10^{10}$

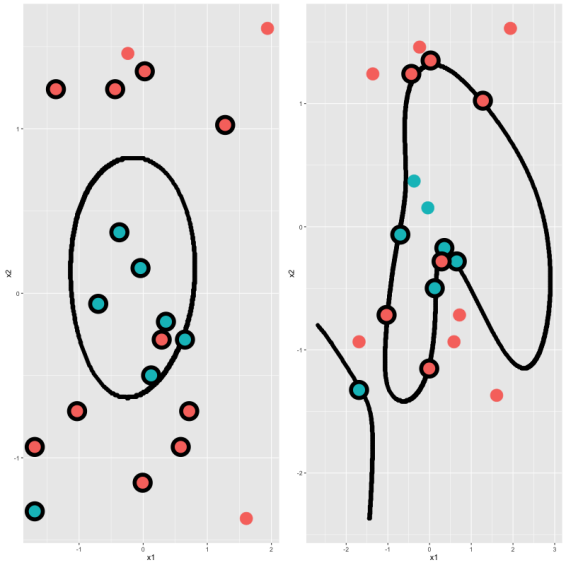

Figure 4.5: RBF SVMs solving a non-linearly separable problem

Left: Soft-margin SVM, $C = 1$
Right: Hard-margin SVM, $C = 10^{10}$

4.4 Classifier Evaluation

Multiple methods exist to evaluate the performance of a classifier. The aim of a classifier is to have high predictability on new instances of data. The notation used in classifier evaluation metrics are: true positives (TP), which are the number of positive classes predicted correctly; false positives (FP), the number of positive classes predicted incorrectly; true negatives (TN), the number of negative classes predicted correctly; and false negatives (FN) the number of negative classes predicted incorrectly. Five commonly used metrics can be computed from the above four values.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.16}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{4.17}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{4.18}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4.19}$$

$$\text{Negative Predictive Value} = \frac{TN}{TN + FN} \tag{4.20}$$

In the classification case of predicting a diagnosis of AD or that a subject is healthy; accuracy refers to the percentage of subjects (both HC and AD) who were predicted correctly; sensitivity measures the proportion of AD subjects that are correctly predicted as suffering from the disease; specificity measures the proportion of HC subjects that are correctly predicted as being healthy; precision measures how likely an AD prediction is to be correct; and the negative prediction value (NPV) measures how likely an HC prediction is to be correct.

Sensitivity is also called the true positive rate (TPR).

Even if the five above metrics are optimal (all at 100%), the classifier can still be useless when applied to a new set of data and thus useless to aid medical professionals. When a model is trained on a set of data instances, it can then be used to classify the same instances and may achieve 100% on each of the aforementioned metrics; yet when another set of new data instances (data instances not used for training) are input to it to classify, the majority of the new data instances may be predicted incorrectly. This is known as the problem of overfitting. Overfitting occurs when a classification model is too complex and ends up describing random errors and the noise of the data such that it is completely biased towards the training data. To determine whether overfitting has occurred the model must be evaluated with different data than it was trained with; one such technique is using cross validation to evaluate the performance of the model.

4.4.1 Cross Validation

Cross validation is a model evaluation technique to measure how well a model performs on a data set that it has not seen before. k-fold cross validation (Refaeilzadeh et al., 2009) works by splitting the data set into k folds where each fold contains an equal (or near equal) number of samples. Next, k iterations of training and testing are performed where $k-1$ folds are used for training and the leftover fold is used for testing. Eventually all folds will have been tested (using a different $k-1$ folds for training at each iteration). An accuracy can then be computed between the actual class of all of the samples of the data and the predicted class for all the data samples. The 10-fold cross validation is repeated ten times for all classifiers in order to try and eliminate any bias towards the data set being tested.

4.4.2 Area Under the Receiver Operating Characteristic Curve

The Receiver Operating Characteristic (ROC) curve is a graph showing the performance of a binary classifier as its discrimination threshold is varied. It is created by plotting the TPR

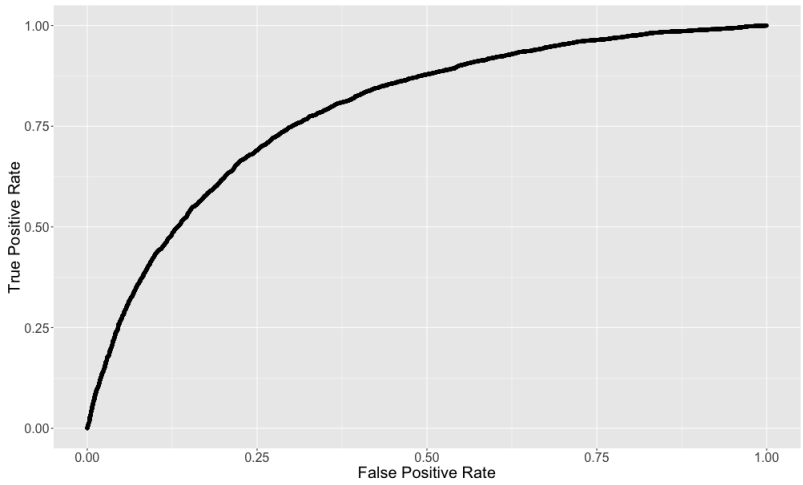

Figure 4.6: An example of a ROC curve

against the false positive rate (FPR) (defined as $FPR = 1 - \text{specificity}$). An example ROC curve is shown in Figure 4.6 which shows the TPR and FPR of a logistic regression classifier on simulated data. The ROC curve can be summarised using its area; the area under the ROC curve (AUROC). The AUROC ranges between 0.5 and 1; with an AUROC of 0.5 the curve is linear and the model has no discrimination ability, while an AUROC of 1 means perfect discrimination.

4.5 Regression Algorithms

Regression is a similar method to classification, however, instead of a discrete class value being learned and predicted, a continuous numerical value is learned and predicted. An example of the difference between the two is that a classification algorithm could predict whether it would be "cold" or "hot" on a certain day of the year; but a regression algorithm could predict the temperature in degrees celsius on a given day of the year. A more formal definition would be given an n-dimensional input vector $\overline{x} = [\, x_1 \cdots x_n \,]$, predict its output value y where $y \in \mathbb{R}$. Each input instance of data the regression is trained on is in the form $(\overline{x}, y)$ and the regression model will be built using this data and be able to predict new data which could be represented by: $\hat{y} = f(\overline{x})$ where $\hat{y}$ is the predicted value of the data and f

is the regression model. Multiple data observations can be denoted by $(X, \overline{y})$ where X is a matrix of all data observations and $\overline{y}$ is the vector with all of the values of the observations.

4.5.1 Linear Regression

Linear regression is one of the simplest regression models. The notation for a linear regression model is shown in Equation 4.21, where $\overline{x}$ is a single data observation with n features, and $\overline{w} = [\, w_0 \cdots w_n \,]$ are the unknown coefficients of the model to be learned in an attempt to make the model as accurate as possible.

$$\hat{y} = f(\overline{x}) = w_0 + \sum_{i=1}^{n} w_i x_i^T \tag{4.21}$$

4.5.2 LASSO Regression

LASSO regression is a regression model which aims to handle the downsides to a linear regression model, as such it provides implicit feature selection and shrinkage methods. The notation for a linear regression model is shown in Equation 4.21, where X is a single data observation with n features, and $W = [\, w_0 \cdots w_n \,]$ are the unknown coefficients of the model to be learned in an attempt to make the model as accurate as possible.

$$f(X) = w_0 + \sum_{i=1}^{n} w_i x_i^T \tag{4.22}$$

One of the main problems with the standard linear regression model is that when there are many correlated variables, the corresponding coefficient can exhibit high variance. Assume there are two variables in the data which correlate highly with one another, if one of the variables has a large positive coefficient and the other has a large negative coefficient then they will cancel each other out and provide no information to the model. Another problem is regarding feature selection, typical search techniques (i.e. backward, forward and stepwise searches) can be applied before the model is created, however these search techniques only

provide discrete decisions: features are either retained or discarded. To solve this problem a shrinkage method can be used, this is an additional constraint applied to the regression model which stipulates that the sum of the coefficients must be less than or equal to a user defined threshold value, t. The model for computing the coefficients for Lasso regression is shown in Equation 4.23 where m is the number of data observations and y_i is the output value of the i^{th} class. Another way of defining the model is shown in Equation 4.24, where the relationship between λ and t is: $\lambda = \frac{1}{t}$.

$$\begin{aligned} \hat{W}^{lasso} = & \min_{W} \sum_{j=1}^{m} \left(y_j - w_0 - \sum_{i=1}^{n} w_i x_{ij}^T\right)^2 \\ & \text{subject to } \sum_{i=1}^{n} |w_j| \leq t \end{aligned} \tag{4.23}$$

$$\hat{W}^{lasso} = \min_{W} \left(\frac{1}{2} \sum_{j=1}^{m} \left(y_j - w_0 - \sum_{i=1}^{n} w_i x_{ij}^T\right)^2 + \lambda \sum_{i=1}^{n} |w_j| \right) \tag{4.24}$$

4.6 Feature Selection and Dimensionality Reduction

Feature selection is a technique to reduce the dimensionality of the data, the aim is to select the smallest subset of dimensions such that the classification accuracy does not decrease and that the distribution of classes is similar to the class distribution before feature selection techniques are applied. Given a n-dimensional data set, a brute force feature selection approach where every subset of dimensions is tested would be too computationally expensive as it would require 2^n classifier models to be built and evaluated. This is feasible when n is small but it quickly becomes computationally expensive. Thus feature selection algorithms are used to reduce the computation time for finding a good (but not the best) subset of features to use for the model. These algorithms can be divided into three groups: filter models, wrapper models and embedded models.

Dimensionality reduction is a technique to reduce the dimensionality by projecting the

data to a lower number of dimensions. This transformation will change the meaning of the features such that the projected features are not equal to the original features.

4.6.1 Filter Models

Filter models evaluate features using methods that are completely independent of the classification algorithm being used.

Fast Correlation-Based Filter

Fast Correlation-Based Filter (FCBF) (Yu and Liu, 2003) first selects a subset of features S' which are highly correlated to the class label. To compute the correlation between a feature and a class label the concept of entropy is used. Entropy is a measure of how unpredictable information content is, therefore if some information has low entropy then the outcome is easier to predict than if it had high entropy; entropy is computed by the following equations:

$$H(\overline{x}) = -\sum_{i=1}^{N} p(\overline{x}_i) \log_2(p(\overline{x}_i)) \tag{4.25}$$

$$H(\overline{x} \mid \overline{y}) = -\sum_{j=1}^{M} p(\overline{y}_j) \sum_{i=1}^{N} p(\overline{x}_i \mid \overline{y}_j) \log_2(p(\overline{x}_i \mid \overline{y}_j)) \tag{4.26}$$

Information Gain is computed by the following equation and it represents a measure of dependence between two inputs:

$$IG(\overline{x}, \overline{y}) = H(\overline{x}) - H(\overline{x} \mid \overline{y}) \tag{4.27}$$

If Information Gain is computed between a feature and the class label, then a high information gain means that the feature is relevant. Features for S' are highly correlated to the class label if $SU(s, c_k) \geq \delta$ where δ is a user-defined threshold and SU is the Symmetrical Uncertainty and computed by:

$$SU(\overline{x}, \overline{y}) = 2\frac{IG(\overline{x}, \overline{y})}{H(\overline{x}) + H(\overline{y})} \tag{4.28}$$

In FCBF a feature f_i is predominant if $SU(f_i, c_k) \geq \delta$ and there exists no other feature f_j such that the following equation is true:

$$SU(f_j, f_i) \geq SU(f_i, f_j) \forall f_j \in S' \text{ where } i \neq j \tag{4.29}$$

f_j is an irrelevant feature to f_i if $SU(f_j, f_i) \geq SU(i, c_k)$. These redundant features are then placed in a set S_{P_i} which is further divided into $S^+_{P_i}$ and $S^-_{P_i}$, where $S^+_{P_i}$ contains features such that $SU(f_j, c_k) > SU(f_i, c_k)$; and $S^-_{P_i}$ contains features such that $SU(f_j, c_k) \leq SU(f_i, c_k)$. FCBF finally applies three heuristics on S_{P_i}, $S^+_{P_i}$ and $S^-_{P_i}$ where redundant features are removed and features that are most relevant to the class are left.

Minimum-Redundancy-Maximum-Relevance

The aim of Minimum-Redundancy-Maximum-Relevance (MRMR) (Peng et al., 2005) is to select features that have a large distance from the other selected features yet still have a high correlation to the class variable. MRMR is an approximation to maximising the dependency between the joint distribution of selected features and the class variable. To minimise redundancy for discrete features, the following is optimised:

$$\text{find } \min_{W_I} \text{ where } W_I = \frac{1}{|S^2|} \sum_{i,j \in S} IG(i, j) \tag{4.30}$$

And for continuous features, the following is optimised:

$$\text{find } \min_{W_C} \text{ where } W_C = \frac{1}{|S^2|} \sum_{i,j \in S} |C(i, j)| \tag{4.31}$$

Where S is the set of features, $IG(i, j)$ is the mutual information between features i and j and is computed by Equation 4.27. $C(i, j)$ is a correlation between features i and j and

can be computed by Pearson's correlation (see Equation 4.34). Maximisation of relevance for discrete features is shown in the following:

$$\text{find } \min_{V_I} \text{ where } V_I = \frac{1}{|S^2|} \sum_{i \in S} IG(c, i) \tag{4.32}$$

For continuous features:

$$\text{find } \min_{V_C} \text{ where } V_C = \frac{1}{|S^2|} \sum_{i \in S} F(i, c) \tag{4.33}$$

Where c is the target class, and $F(i, j)$ is the F-Statistic between features i and j computed by: $F = \frac{\sigma_i^2}{\sigma_j^2}$ where σ_i^2 and σ_j^2 are the variance of feature i and j respectively.

$$\rho_{X,Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y} \tag{4.34}$$

Random Forest Importance

When building a random forest an importance value for any tree can be assigned to each feature, this is known as the GINI importance, I_G, which is calculated for each feature - the higher the GINI importance, the more relevant the feature is to the classification process. The GINI importance can be described as the ability of a feature to reduce the error rate in the classification across every node, τ in every tree, T of the forest, F.

When building a decision tree, one technique which can be used to determine which feature will be used for the decision tree is that the chosen feature causes the maximum reduction in GINI impurity when compared to the reduction in GINI impurity caused by the other features. Thus throughout the forest, features which cause the highest reduction in error rate are more likely to be chosen.

To compute the GINI importance of a feature, θ, throughout the forest, F, the following computes the importance:

$$I_G(\theta, F) = \sum_{T}^{F} \sum_{\tau}^{T} \Delta i_\theta(\tau, T) \tag{4.35}$$

I_G is computed as the sum of the decrease in GINI impurity for which θ is the feature used to split the data. The decrease of impurity at a node τ is computed by:

$$\Delta i_\theta(\tau) = I(f(\tau) = \theta)\Big(i(\tau) - p_L i(\tau_L) - p_R i(\tau_R)\Big) \text{ where } p_K = \frac{n_K}{n} \tag{4.36}$$

Where $f(\tau)$ is the feature of node τ used to split the data. Since I is an indicator function, $f(\tau)$ must be θ for this impurity decrease from this node to be non-zero, thus $\Delta i_\theta(\tau) = 0$ if $f(\tau) \neq \theta$. If $|K| = 2$ (where K is the set of all possible classes) the GINI impurity for a single node is computed by:

$$i(\tau) = 1 - P(k = 0)^2 - P(k = 1)^2 \text{ where } k \in \{0, 1\} \tag{4.37}$$

And if $|K| > 2$ then this is used:

$$i(\tau) = 1 - \sum_{i}^{K} P(k = i)^2 \tag{4.38}$$

For a regression tree (a decision tree used to predict a continuous value rather than a discrete value), the GINI impurity is replaced. Rather than the GINI impurity being the measure, for regression trees the measure of residual sum of squares (RSS) can be used:

$$RSS = \sum_{i=1}^{n} \big(y_i - f(x_i)\big)^2 \tag{4.39}$$

In the RSS formula, n is the number of observations, y_i is the actual value of i and $f(x_i)$ is the predicted value of i. Thus $\Delta i_\theta(\tau)$ becomes computed by:

$$\Delta i_\theta(\tau) = I(f(\tau) = \theta)\Big(RSS(\tau) - p_L RSS(\tau_L) - p_R RSS(\tau_R)\Big) \text{ where } p_K = \frac{n_K}{n} \tag{4.40}$$

4.6.2 Wrapper Models

Unlike filter models, wrapper models actively involve the classification algorithm as they work by selecting a subset of features and then evaluate that subset using the desired classification algorithm. They will repeatedly choose subsets to evaluate until a given termination criteria is met.

Stepwise Selection

There are three forms of stepwise selection which can be used: forward selection, backward elimination, and bidirectional elimination (Hocking, 1976). Forward selection starts by being given a set of features, initially this set of features is the empty set, then it adds the feature to the set which improves the performance the best - if no other feature exists which improves the performance then the selection is terminated and the set of features is returned. Backward elimination where the entire set of features are considered first, and it removes the feature which gives the best increase in performance, this is repeated until removing any feature always makes the performance worse. Bidirectional elimination combines forward selection and backward elimination by allowing features to be added or removed at each stage in the selection process.

Simulated Annealing

Simulated Annealing (Brooks and Morgan, 1995; Rutenbar, 1989) is a search method which makes minor random changes to an initial solution and if the performance of the new solution is better than the initial solution then the new solution replaces the initial solution. If the new solution is worse, an acceptance probability is created which factors in the performance difference between the two solutions and the current iteration the search is in. If the acceptance probability is met, then the new solution replaces the initial solution even though the new solution performs worse. A basic formula to create the acceptance probability is: $AP = e^{\frac{C-N}{C \times I}}$ where C is the score of the current solution, N is the score of the new solution

and I is the current iteration. Formulae to compute the acceptance probability generally follow two rules: AP should decrease as the difference between C and N increases; AP should decrease as I increases.

4.6.3 Dimensionality Reduction

Spectral Feature Selection

Spectral Feature Selection (Zhao and Liu, 2007) is a dimensionality reduction method which works by first constructing a similarity matrix, S, using the RBF function to compute similarities:

$$S_{i,j} = exp\big(- \frac{||x_i - x_j||^2}{2\sigma^2}\big) \tag{4.41}$$

Where x_i and x_j are vectors containing the data for subject i and subject j and σ is a hyperparameter. From the similarity matrix, a graph, G and adjacency matrix W are created. Next a degree matrix, $\bar{D}$ is created from W, a degree matrix is a diagonal matrix with information about the number of edges attached to each vertex. Using W and $\bar{D}$, a Laplacian matrix, L, and normalised Laplacian matrix $\mathscr{L}$ is created. A Laplacian matrix stores information about a graph given an adjacency matrix and degree matrix and it is computed by $L = \bar{D} - W$; the normalised Laplacian matrix is computed by $\mathscr{L} = \bar{D}^{-\frac{1}{2}} W \bar{D}^{-\frac{1}{2}}$. Using the Laplacian matrices, the features are weighted according to how they relate to the graph structure such that features that behave similarly across multiple samples are ranked highly as they will help to form clusters in the data. The output from the spectral feature selection is a list of all the features assigned a weight between 0 and 1 inclusive, with a higher ranking meaning that the features are more likely to form clusters in the data.

Principal Component Analysis

Principal Component Analysis (PCA) is a non-parametric method of reducing the number of features in data sets. It provides a way to reduce a complex data set into a lower dimension. Assume we have m features and n observations, and we want to reduce the complexity of the data set. We want to keep the features that do not depend on other features and we also want the features which have a high variance. To accomplish this, the representation of the data must be changed such that the most important features can easily be extracted. A linear transformation is applied to the data to change its representation:

$$X' = XP \tag{4.42}$$

The aim of this new representation is to decrease the noise and redundancy in the data. A common measure for noise is the signal-to-noise ratio (SNR), which is computed by dividing the variance of the signal by the variance of the noise:

$$SNR = \frac{\sigma^2_{signal}}{\sigma^2_{noise}} \tag{4.43}$$

A high SNR (much greater than 1) indicates data with a high precision whereas a low SNR indicates data with a lot of noise. We must find the rotation of X which maximises the SNR. First we compute the covariance matrix of X:

$$S_X = \frac{1}{n-1} XX^T \tag{4.44}$$

The covariance matrix quantifies the correlations between all the pairs of measurements. A large covariance leads to a high SNR. Next we need to diagonalise S_X as the signal is determined by the diagonal and the redundancy determined by the non-diagonal values. Thus we need to find P such that the covariance of X', $S_{X'}$ is diagonalised.

$$S_{X'} = \frac{1}{n-1} X' X'^T = \frac{1}{n-1} PAP^T \tag{4.45}$$

Where $A = XX^T$ and A is symmetric. A can be written as $A = EDE^T$ where D is a diagonal matrix and E is a matrix of eigenvectors of A arranged as columns. We select P to be a matrix where each row is an eigenvector of XX^T therefore $P = E^T$ and $A = P^T DP$. Using this, and $P^T = P^{-1}$:

$$S_{X'} = \frac{1}{n-1} D \tag{4.46}$$

Therefore $S_{X'}$ is diagonal showing that the choice of P as the matrix where each row is an eigenvector of XX^T gives us a transformation where the covariance matrix of X' is a diagonal matrix. The transformed matrix X' is the matrix of the principal components of X, and the eigenvalues of XX^T can be used as a way to rank the importance of these principal components with a higher associated eigenvalue meaning that the principal component explains a larger portion of the variance in the data.

Chapter 5

Literature Review of Predictive Data Mining for Alzheimer's Disease

5.1 An Overview of Techniques for Data Mining Alzheimer's Disease

The techniques reviewed in this chapter will be divided into three categories, as each category operates on a different format of data requiring different analysis. The first technique is termed "Morphometry Analysis" and refers to image mining approaches operating on the raw MRI data. Image mining is a technique where raw image data is analysed with the aim of detecting patterns, it is an interdisciplinary technique utilising computer vision, image processing, image retrieval, data mining, machine learning and artificial intelligence (Hsu et al., 2002). Image mining relates to brain MRIs as they are 3D image data thus the algorithms must process multiple images for patterns. A simple example of this technique would be to take a single slice of an MRI and look for patterns in the image that correspond to an HC or an AD subject; more successful techniques will look at the entire MRI data, thus it will treat the MRI as a three-dimensional set of voxels.

The second technique focuses on literature where the data is processed MRIs such that

various measurements are extracted from the raw MRI data, thus the data will be in a numerical format rather than an image-based format that the first technique uses. These techniques will be grouped as "Data Mining of MRI Measurements". The third technique will focus on literature where the data being analysed is not from MRIs, instead it is from neuropsychological tests. This will potentially uncover methods showing how these neuropsychological tests could be augmented with MRIs to improve the ability of the classifier.

Different types of intracranial volume (ICV) normalisations are discussed, these are techniques to adjust individual volumes inside a brain based on the total volume of the brain. These techniques are commonly used in the literature of classification of AD from MRIs to improve the classification accuracy thus they are an important technique. Finally the methods used by participants of two challenges to predict AD are analysed. These challenges are: The Computer-Aided Diagnosis of Dementia Challenge and The Alzheimer's Disease DREAM Challenge.

5.2 Morphometry Analysis/Image Mining

This section covers a selection of the literature which uses techniques applied to raw MRIs. In Long and Holder (2012) a graph-based approach is used to classify individuals based on their level of cognitive impairment due to AD, education and gender. The graph-based approach represents the shape of the brain, the shape of the ventricular system and shape relative to the skull. It was discovered that AD was found to affect the shape of the ventricles of the brain to a large extent. A classification accuracy of 90.9% was achieved between HC and AD subjects. This body of work showed that evaluation of the structure of the brain led to smaller regions in the brain being discovered which could be linked with the function of the brain or life events. A minor drawback of this work was that while the method could automatically detect regions of the brain that correlated with AD, it would still require further work or a medical professional to manually look at these regions and determine what they are.

An SVM in Klöppel et al. (2008) is used to classify MRIs as healthy or AD, the voxels of the MRI are treated as points in a high dimension (where the number of voxels is equal to the number of dimensions); dimensionality is then reduced via subspace clustering to identify similar images within the data, then the remaining data is classified by the SVM and achieves a classification sensitivity of 95.0% and specificity of 95.0% using a leave-one-out method; however a drawback of this work is that they used 34 AD subjects and 34 HC subjects for their confirmed-AD versus HC classification which is a very small sample when compared to other work in the literature. This work showed that multiple datasets (MRIs from different scanners and protocols) could be combined to achieve a very high accuracy, but this would need a much larger sample to fully test whether the combination of cohorts improves the results.

Duchesne et al. (2008) use an SVM to classify HC and AD patients based on their structural MRIs using Jacobian determinants derived from dense deformation fields and scaled intensity from selected ROIs on the medial temporal lobe. It reaches a 92% accuracy when classifying between 75 HC and 75 AD patients, this is a high accuracy that is reached but it does not provide much insight to why the decision was made due to the use of the SVM and the data transformation process which takes place prior to the classification, as well as using a fairly small set of samples. This work also demonstrated how data from varying cohorts can be combined into the same dataset for a classification problem to achieve a larger dataset than could be achieved by looking at a single cohort, but again this suffers from the usage of a fairly small sample set even once they have all been combined. This work shows how the MRI processing step can be modified to affect the output data; since the work in this thesis uses Freesurfer for MRI processing this step will automate the MRI processing step but it must be kept in mind that this step could potentially be adjusted to improve the data used for the classification problem.

Plant et al. (2010) use various classifiers (Bayesian, SVM, and Voting Feature Intervals) to distinguish between HC and AD based on the clusters of voxel data of structural MRI

scans; an accuracy of 92% is achieved by the classifier between HC and AD, and it also had a 75% success rate in predicting the conversion of MCI to AD. This work uses data from 32 AD subjects, 24 MCI subjects and 18 HC subjects; this makes it a very small data set the work is using. This work demonstrated that density-based clustering for MRI voxels can generate a feature capable of distinguishing between HC and AD thus showing the connectivity of brains belonging to an HC have a vastly different structure in areas than an AD positive subject. As well as this, it also was capable of predicting MCI to AD conversions at a high rate showing that there is a physical relationship in the brain between MCI and AD.

5.3 Data Mining of MRI Measurements

This section covers a selection of literature based on using data mining techniques applied to measurements of a brain which can be inferred from an MRI. Cuingnet et al. (2011) tested ten methods to classify HC, MCI and AD based on MRI data; some of these methods included training classifiers on data extracted by Freesurfer. The most interesting method was that they used a Parzen Window on the hippocampal volumes, achieving a sensitivity of 73% and a specificity of 74% as this showed the ability of the small subset of the hippocampal volumes to achieve a high accuracy showing that the hippocampal volumes alone have a high predictive ability. They concluded that feature selection increased the sensitivity and accuracy, the highest sensitivity and specificity they achieved was 81% and 95% respectively. In this work 162 HC subjects, 210 MCI subjects and 137 AD subjects were used which is a fairly large number of subjects. A slight drawback is that all the data was from the same cohort, so it does not determine whether combining cohorts are a benefit or not.

Wan et al. (2012) propose a sparse Bayesian multi-task learning algorithm on MRI data extracted by Freesurfer to deduce relationships between neuroimaging measures and cognitive scores to show how changes in the structure of a brain can affect the cognitive condition of it. The authors looked at 222 HCs and 171 AD subjects from the same cohort, this

work discovered biomarkers capable of predicting multiple cognitive scores successfully and explained that there could be potential for these biomarkers to determine cognitive functions and define the progression of AD in a brain. The results of the work here achieved an accuracy of 73.5% at predicting the cognitive scores of the subjects. This could lead to biomarkers discovered from the work in this thesis not just being capable of predicting AD, but determining the progression of AD.

Iftikhar and Idris (2016) aim to improve classification using features computed using the cortical surface measurements of a structural MRI. The problem with cortical surface measurements is that there is large spatial variance of the cortical thicknesses and the high dimensionality of structural MRIs mean that the features defined on vertices (these are features defined from positions on a surface) are sensitive to noise. The structural MRIs are preprocessed with Freesurfer and 32 neocortical and 2 non-neocortical ROIs are used. Volumes and thickness are extracted from each of the ROIs as well as hybrid features combining both the thicknesses and volumes. Then the F-Score method is used to reduce the number of features. The ensemble classifiers used here are multiple SVMs with different kernels (linear, RBF and sigmoid) and vote mechanisms used to determine the final classification is using the decision which is the most common throughout the classifiers. The highest accuracy achieved is 98.8% with a sensitivity of 100.0% and specificity of 96.0% on 60 HC and 60 AD subjects. There is no control experiment to show how a single SVM performs on the data, thus it is impossible to determine whether using this ensemble classifier proves to be an advantage over the use of a single classifier. The ensemble would perform worse if the SVM with a certain kernel is highly accurate at classifying the data, and the other kernel SVMs in the ensemble are poor at classifying the data - these low performing SVMs would cause the ensemble to create incorrect classifications regardless of the highly accurate SVM.

dMRIs have been used to classify AD (Schouten et al., 2017) where the best performing method involves clustering the voxel-wise diffusion tensor measures, these measures are extracted from the dMRIs using tract based spatial statistics (Smith et al., 2006) where the

fractional anisotropy of the water molecules in the brain are projected onto a mean WM skeleton representing the centre of the WM tracts, this then produces a feature vector of 113282 dimensions for each subject. Next independent component analysis using MELODIC (Beckmann, 2012) is applied to the feature vector which returns 28 components. Using elastic net classification on these 28 components, the best result achieved is an accuracy of 85.1%, a sensitivity of 86.8% and a specificity of 84.4%. While the work in this thesis is based on structural MRIs and not dMRIs, this body of work demonstrated the power of a component analysis algorithm to reduce high dimensionality data and still retain a high predictive ability.

5.4 Data Mining of Neuropsychological Tests

This section covers a selection of literature based on using data mining techniques applied to the data from the results of neuropsychological tests a patient has taken. Johnson et al. (2014) used a Genetic Algorithm (GA) for feature selection of a variety of neuropsychological tests, which were then input to a logistic regressor to predict conversion from HC to MCI or AD, and conversion from MCI to AD. They also showed a GA performed better than Stepwise Variable Selection, a commonly used feature selection method and reached an AUC accuracy of 0.91 when predicting HC to MCI conversions; and 0.87 when predicting MCI to AD conversions. Work which uses Bayesian modelling in application to AD diagnosis includes the work of Prince (1996) which uses Bayes' Theorem to calculate likelihood ratios for multiple tests which are used to diagnose a patient with AD. These likelihood ratios are then combined using conditional independence where the likelihood ratios can be multiplied to generate a single probability of whether the subject has AD.

Lee et al. (2016) develop a prediction model which outputs the probability a subject is suffering from AD based on the features on a subject's MRI and their neuropsychological scores. The neuropsychological scores are based on the overall MMSE as well as scores from sections within the MMSE as well as their WSCT scores and subtests scores of WAIS-III.

The image features are volume and shapes computed from the MRI such as: area, perimeter, compactness, elongation, rectangularity, symmetry and minimum thickness of the cerebral ventricles. The prediction model is created using a probabilistic neural network and learning vector quantisation to overcome the computational cost of the probabilistic neural network when large training data sets are used. The results show that the probability correlates with the diagnosis of the subjects, and since this work looks at baseline, 1 year and 2 years scans: it shows that as time progresses, the probability for suffering from AD increases. This is valid because AD susceptibility increases with age.

5.5 Intracranial Volume Normalisation

This section covers a selection of literature regarding the usage of ICV normalisation applied to MRIs. ICV normalisation (Whitwell et al., 2001) is a technique for modifying measurements of ROIs in the brain to account for differing brain sizes (the basic concept is that the larger a person's brain, the larger the ROIs of the brain will be as $ICV \propto ROI_{vol}$ however the reality is not that simple). Many techniques have been investigated to solve this problem, however the majority of them assume a linear correlation between the volumes of the ROIs inside the brain and the ICV of the brain however this may not be the case. This technique is often used in the literature related to classifying dementia from MRIs (Chao et al., 2010; Fan et al., 2008).

The simplest ICV normalisation method is the proportional method (Nordenskjöld et al., 2013) where the aim is to express a volume as the proportion of the brain it occupies and it is computed by $v' = \frac{v}{ICV}$ where v is the unnormalised volume of the patient, ICV is the intra cranial volume of the patient and v' is the normalised volume of the patient. This method is completely independent of any other subjects.

The residual method (Nordenskjöld et al., 2013) assumes a linear relation between the volumes and the subject's ICV and estimates the relationship such that: $v' = v - w_1(ICV - \overline{ICV})$

where $\overline{ICV}$ is the average intra cranial volume across all subjects being considered, and w_1 is a coefficient computed by solving the linear regression problem of $v' = w_1 ICV + w_0$. The residual method can be further refined by assuming a different relationship between male and female subjects' volumes and their ICVs such that if the subject is male:

$$v'_M = v - w_{1,M}(ICV - \overline{ICV_M}) \tag{5.1}$$

Where $w_{1,M}$ is found by solving the linear regression problem (only considering male subjects) of:

$$v' = w_{1,M} ICV + w_{0,M} \tag{5.2}$$

And a similar problem for female patients where in the above example, males are replaced by females:

$$v'_F = v - w_{1,F}(ICV - \overline{ICV_F}) \tag{5.3}$$

Where $w_{1,F}$ is found by solving the linear regression problem (only considering female subjects) of:

$$v' = w_{1,F} ICV + w_{0,F} \tag{5.4}$$

The covariate method (Nordenskjöld et al., 2013) is similar to the residual method but it considers gender ($G \in \{0, 1\}$) at the linear relationship level such that gender is another variable in the linear regression. The normalised volume is computed via:

$$v' = v - w_1(ICV - \overline{ICV}) - w_2 G \tag{5.5}$$

Where w_1 and w_2 are found by solving the linear regression problem:

$$v' = w_0 + w_1 ICV + w_2 G \tag{5.6}$$

The power proportion method (Liu et al., 2014) computes the normalised volume as $v' = \frac{v}{ICV^{w_1}}$ where w_1 is computed by the power regression problem: $v' = w_0 ICV^{w_1}$. Note that this power regression problem can be converted to a linear regression problem by taking the natural logarithm of both sides to create $\ln(v') = w_1 \ln(ICV) + \ln(w_0)$.

The Freesurfer documentation[1] states that it computes an estimate of the ICV per brain rather than the more accurate count of non-zero voxels. The method Freesurfer uses is to estimate the ICV based on the linear transformation of the brain computed from the amount of scaling on the linear transform into MNI305 atlas space (Buckner et al., 2004). The reason the estimated ICV is computed is because in T1 MRIs the skull and CSF have a similar intensity and thus it is hard to distinguish between the two and this affects the value of the estimated ICV.

In this documentation page the reasoning behind the normalisation of volumes is discussed. It is stated that while volumes, and to a lesser extent, some areas benefit from normalisation; thicknesses do not. This is due to volume scaling with the total head size, and surface area scaling with the total head size to some extent, but since the thicknesses are a linear measure, they do not scale with the head size. On a similar note, it is only beneficial to scale those volumes and areas which correlate with the total head size.

5.6 CAD Dementia Challenge

The CAD Dementia Challenge is a challenge with the aim to improve the clinical use of computer-aided diagnosis for AD by performing a large validation of methods. The challenge works by providing a data set of T1-weighted structural MRIs from multiple medical centres of subjects classed as HC, MCI or AD. Of the data set, only 30 subjects are provided with their diagnosis labels thus the organisers encourage participants to use other data sources such as ADNI. The challenge ranks the participants on how their method classifies the 354 MRI scans which do not have the diagnosis labels provided for them. The ranking metrics

[1] http://www.freesurfer.net/fswiki/eTIV

are accuracy, true positive fraction (for HC), true positive fraction (for MCI), true positive fraction (for AD), AUROC (for all), AUROC (for HC) and AUROC (for MCI). A selection of the works submitted to this challenge will be discussed.

Sørensen et al. (2014) used a variety of image-based biomarkers: extracted Freesurfer features from the structural MRIs; using two hippocampal shape scores (one for the left and one for the right); and a hippocampal texture score. The hippocampal shape was computed by first aligning each hippocampus surface to a template hippocampus using the iterative closest point (an algorithm minimising the difference between two clouds of points), then landmarks were mapped from the template to the hippocampus. Next the hippocampi were aligned using generalised Procrustes alignment (this is a method used to compare the shapes of objects). The PCA was then applied to find the components which explained 90% of the variance. The hippocampal texture score was computed using the texture descriptor defined (Sørensen et al., 2012) but excluding the Gaussian filter. The biomarkers were then combined as input to a regularised LDA classifier and the results using 10-fold cross validation were 73.3% on the CAD Dementia training data and 62.2% on their ADNI and AIBL (Australian Imaging, Biomarker & Lifestyle Flagship Study of Ageing) data. On the CAD Dementia test data, they achieved the highest accuracy in the challenge of 63.0% (as well as second place with an accuracy of 59.9%).

In third place for the challenge was Wachinger et al. (2014). This method augments 109 features extracted using Freesurfer with shape information. This shape information is a representation of the anatomy of the brain containing information of the shapes of multiple cortical and subcortical structures. Feature selection was stepwise model selection in a generalised linear model using Akaike information criterion (this is a measure of the quality of statistical models for a given set of data). Classification is performed using a generalised linear model with age correction by linear regression, the best accuracy result was achieved when no ICV normalisation was used and the feature set was 30 thickness, 11 volume and 19 shape features, this accuracy was 80% on the challenge validation data. They

noted that Freesurfer failed to process three subjects whom they marked as AD assuming a relation between the atrophy caused by AD and processing difficulties. It also is not specified in the paper whether cross-validation or holdout was used for the accuracy evaluation, cross-validation was however used for parameter selection of the generalised linear model, it is assumed that a holdout method was used where the validation data was held back. Their challenge result was an accuracy of 59.0%.

Dolph et al. (2014) measure five features per slice of 46 MRI slices: the ratio between the amount of WM and amount of CSF in the slice; the ratio between the amount of GM and CSF in the slice; the number of CSF pixels in the slice; the number of WM pixels in the slice; the number of GM pixels in the slice. Using these features, an SVM with an RBF kernel is used to classify the data and attains an accuracy of 80% when classifying HC, MCI and AD; however only 30 samples are used to test the data using a holdout method.

Sarica et al. (2014b) investigate feature selection methods (recursive feature elimination, correlation filters, random forest filter) for SVMs applied to Freesurfer processed MRIs. One-versus-one SVMs are combined with a voting mechanism to create a three-class classifier capable of classifying between HC, MCI and AD. ICV normalisation is also tested on the subjects to determine whether it is beneficial or not. The best accuracy is achieved by the SVM which uses no ICV normalisation and the random forest filter and was 95.2%; however only 30 samples are used to test the data using a holdout method.

5.7 Alzheimer's Disease DREAM Challenge

The Alzheimer's Disease DREAM Challenge (Allen et al., 2016) was a challenge to assess capabilities for estimating cognitive ability (via predicting an MMSE score) and predicting cognitive decline (change in MMSE score over time). The challenge consisted of three sub-challenges, which participants were invited to answer one or more of them. The three sub-challenges were: predicting cognitive decline based on genetic data; prediction of resilience

to cognitive decline in subjects with abnormal β-amyloid peptide production and deposition; the estimation of cognitive state based on structural MRIs - predicting the MMSE given a structural MRI. A selection of the works submitted to the sub-challenges will be discussed.

The first sub-challenge was to predict the MMSE change in two years after the initial assessment. The inclusion of this in the challenge was because accomplishing this will enable the prediction of cognitive decline and potentially provide new methods for early AD diagnosis. With the results of sub-challenge 1, it was discovered that aside from the apolipoprotein E (APOE) haplotype, there were no other notable contribution of genetics to predict the cognitive decline. The APOE haplotype is the apolipoprotein E gene which produces a lipoprotein. A lipoprotein is a soluble protein that combines with and transports lipids in the blood plasma. A lipid is a class of organic compounds that are fatty acids that are insoluble in water but soluble in organic solvents. Blood plasma is a component of the blood which stores the blood cells and is used to transport nutrients, hormones and proteins around the body. One of the three most common APOE variants is APOE-ϵ4, and having this gene has been found to greatly increase the risk of AD (Corder et al., 1993). The performance of the entries were measured on the correlation (Pearson and Spearman) between the observed change in MMSE from 24 months to baseline and predicted change in MMSE using clinical covariates and both with and without genetic data. The top performing entry achieved the Pearson correlations of: 0.382 and 0.382 without and with genetic data respectively; and Spearman correlations of: 0.433 and 0.433 with and without genetic data respectively.

The second sub-challenge was to predict the set of cognitively normal subjects whose biomarkers indicate abnormal β-amyloid peptide production/deposition as abnormal β-amyloid peptide production and deposition, has been linked to AD (Selkoe, 1994). Discovering this link will show how certain subjects can maintain normal cognitive function while β-amyloid production exists in their brain, this could lead to the development of method to prevent AD. For this sub-challenge, no participants were able to achieve a better than random performance which indicates that information regarding cognitive resilience is hard to discover

from gene analysis.

The third sub-challenge was to estimate the cognitive state of the subjects using structural brain data by predicting a subject's MMSE score from a structural MRI. For the training data, 628 subjects were used from ADNI; and for the unseen test data, 182 subjects from the AddNeuroMed database were used. Of the thirteen teams who submitted for this challenge, three teams performed considerably higher than the others but their performances were statistically indistinguishable from each other. From the submissions, the most common features used to predict the MMSE scores were the hippocampal volume and entorhinal thickness; these features being AD predictors is evidenced by previous work (Haight et al., 2012). The performance for this sub-challenge was measured by both the Pearson correlation between the observed and predicted MMSE and the "concordance correlation coefficient for agreement on a continuous measure" between observed and predicted MMSE (Lawrence and Lin, 1989). The best entry achieved a Pearson correlation of 0.573 and a concordance correlation coefficient of 0.569.

Conclusions drawn from this challenge were that the predictive performance of the submissions was modest and the submitted methods performed similarly. There are three explanations given as to why this could be the case: the first is that the results are due to a failure in the modelling methods used by the teams. The second explanation is that the data used is inadequate, potentially there needs to be more data for the models to learn, or it could be due to noise in the data as the participants used and the MRIs. Errors could be introduced in the MRI capturing process, and also during the step where the MRIs are segmented with Freesurfer. The third explanation is that there is an imperfect relationship between brain structure and function. Since the structural MRI only contains information about the brain structure and not its function, the models are unable to map the structure of a brain to its expected function accurately, which means that it may be preferable to use newer imaging methods such as fluorodeoxyglucose (FDG) positron emission topography (PET) or tau PET to assess cognitive behaviour in AD. FDG-PET is a nuclear imaging technique showing func-

tional processes in the body, and FDG is used as a tracer. Tau PET uses a tracer which is sensitive to tau molecules.

Chapter 6

An MRI Data Importing and Manipulation Toolset in R

The work in this thesis was written in the statistical programming language, R (R Core Team, 2013). R allows the usage of packages, which are collections of R functions, data and compiled code. Many packages are included within the default installation of R, and there are thousands of packages written by users on the package repository Comprehensive R Archive Network (CRAN) . While the work was being written, many functions were being developed to interact with the data generated by the MRI segmentation software Freesurfer, perform ICV normalisation and various other tasks. Thus a package, rsurfer, containing both the functionality of a wrapper around various Freesurfer commands and functionality to manipulate the imported data was accepted onto CRAN[1].

The rsurfer package provides functionality to import the data generated by Freesurfer, specifically the data generated by the cortical restruction function of Freesurfer (recon-all). Once this data is imported, rsurfer provides functions to easily manipulate the data; and also provides commonly used ICV normalisation methods. This package has been designed using an installation of and data generated from Freesurfer version 5.3 - it may function with older/newer versions but it is untested.

[1] https://cran.r-project.org/web/packages/rsurfer/

6.1 Related Work

This work was prompted by the previous development of a plug-in for the data mining software KNIME (Berthold et al., 2008), the plug-in was titled K-Surfer[2] (Sarica et al., 2014a) and it provided the functionality of manipulating Freesurfer generated data in KNIME. K-Surfer provides more broad functionality than rsurfer; but the functionality rsurfer provides is more in-depth. The package AnalyzeFMRI (Bordier et al., 2011) is a package built to handle MRI formats (discussed in Section 2.3.1): Analyze and Nifti. It provides functionality to read the headers of the files, convert between the two formats, visualise the data and analyse it too. While this is an MRI analysis program it is very different to rsurfer as this focuses more on the lower level handling of raw MRIs whereas rsurfer is aimed at operations on processed MRIs. fmri (Tabelow and Polzehl, 2010) is a package which provides similar functionality to AnalyzeFMRI, so it is aimed at an analysis of a single MRI. RNiftyReg (Clayden, 2011) is a package providing an interface between R and the tool NiftyReg[3] which is an open source software used for image registration of MRI data. The package mritc (Feng and Tierney, 2011) provides tools to classify brain tissue, again this is a low level step and this will be performed in the Freesurfer processing step.

The most similar R package to rsurfer is "freesurfer"[4], which was published on CRAN while rsurfer was being developed and provides similar functionality to read output from Freesurfer (to avoid confusion between the software suite Freesurfer and the R package "freesurfer", the latter will always be referred to in lowercase and wrapped in quotation marks). The difference between the two packages is that "freesurfer" provides functions in R to run Freesurfer processing on MRI files, with "freesurfer" you can control the individual steps of the Freesurfer pipeline; whereas rsurfer does not have this functionality and its aim is to manage already processed data and provide functionality to manipulate that data.

[2]https://sourceforge.net/projects/ksurfer/
[3]https://sourceforge.net/projects/niftyreg/
[4]https://cran.r-project.org/web/packages/freesurfer/index.html

6.2 Importing Data Generated by Freesurfer

This step will assume that the MRI data has already been downloaded and is ready to be processed with Freesurfer. rsurfer has been built using the following reconstruction command on the MRI data:

```
recon-all -i /input_data/MRI_IMAGE.nii.gz -sd /output_data/ -subjid SUBJECT_ID
    -all -hippo-subfields
```

Once this process has completed successfully then the MRI data will be ready to be imported with rsurfer. The following R code will install and load the rsurfer package:

```
install.packages("rsurfer")
library(rsurfer)
```

The data generated by Freesurfer can be imported with two commands, the first command will point rsurfer to the Freesurfer installation location as it calls some scripts which are included within a Freesurfer installation. The second command requires the directory of the subject data generated with Freesurfer and it will assume every directory inside belongs to a subject exported from Freesurfer. For example if the Freesurfer installation is located in "/Applications/freesurfer" then the function "setfshome" would be used with "/Applications/freesurfer" as its parameter. If the output data is located in "/output_data/" then the "fsimport" function would require "/output_data/" as its parameter.

```
setfshome("/Applications/freesurfer")
mri_data <- fsimport("/output_data/")
```

Note that the directories in the functions will likely have to be changed due to Freesurfer and the output subjects being stored in a different location. The second function returns a data frame with each row representing a single subject and the columns representing the measurements computed by Freesurfer. Importing a large number of processed subjects into R may cause the import function to take a few minutes to run, so rsurfer has an in-built function which will run the import function and then serialise the output to a file. When this function is run again, it will see the existence of the serialised file and just load that; and if

it fails to find the serialised file then it will import the data again and create the serialised file.

```
setfshome("/Applications/freesurfer")
mri_data <- fsimport.serialise("/output_data/", "serialised.rds")
```

The import and serialise function takes an additional argument, this argument determines the location where the serialised file will be saved.

When importing generated data, some of the rows and columns of the data table have missing values or columns where all values are zero. There is a function provided to clean up the imported data.

```
mri_data <- eliminateabnormalities(mri_data, verbose = T)
```

The verbose parameter is an optional argument, here it is set to true because it will warn the user of any columns or rows that are removed so they could manually check the Freesurfer output to try and determine what went wrong. In some cases, the Freesurfer data may fail to import then potentially something has gone wrong in the Freesurfer processing step. rsurfer provides a function which checks for any missing files which cause the import function to fail. The functions require the file path of the subjects' directory and will check every subfolder there thus it can perform error checking on every subject in one call (assuming every subfolder belongs to a subject).

```
fsdirectorycheck("/output_data/")
```

rsurfer also has the ability to generate random data if this is needed in order to test rsurfer and no MRI data is available to import. The random generation utilises a seed which can be changed using the in-built R function "set.seed". This function has an optional function argument to determine how many subjects to generate, the default is 40; the first line generates 40 subjects, and the second generates 100 subjects.

```
generaterandomsubjects()
generaterandomsubjects(100)
```

6.3 Manipulating Data Generated by Freesurfer

rsurfer provides a variety of functions to extract groups of measurements from the data. The first function extracts all MRI data from the data frame, this is useful when the data has been augmented with additional features such as gender and sex and only the raw MRI data is desired:

```
extract.brain.features(mri_data)
```

Extract all volume measurements:

```
extract.volumes(mri_data)
```

Extract all hippocampal volumes, these are the features that are generated with the -hippo-subfields flag with the Freesurfer cortical reconstruction process:

```
extract.hippocampalvolumes(mri_data)
```

Extract all the cortical measurements (includes areas, volumes, thicknesses and standard deviations of thicknesses) from the data, the different types of measurements can also be specified:

```
extract.cortical(mri_data)
extract.corticalvolumes(mri_data)
extract.corticalsurfaceareas(mri_data)
extract.corticalthicknesses(mri_data)
extract.corticalthicknessstddevs(mri_data)
```

The following function extracts all the subcortical volumes:

```
extract.subcorticalvolumes(mri_data)
```

We can use the following code to augment a random age feature between 50 and 80, and a random gender ("Male" or "Female") using an rsurfer function with the MRI data as the parameter:

```
mri_data$Age <- runif(nrow(mri_data), 50, 80)
mri_data <- addrandomgender(mri_data)
```

All of the extraction functions have an additional optional argument which enables any additional fields to be extracted, so we could extract the newly added age and gender as well as the cortical thickness using:

```
extract.corticalthicknesses(mri_data, c("Age", "Gender"))
```

Note that while the cortical thickness extraction is used, any other extraction function rsurfer has would work. Another function exists which can extract a set of features specified by a string thus writing a for loop which can iterate through the different sets of features to extract is simple:

```
fieldGroups <- c("corticalvolumes", "subcortical", "hippocampal", "
    corticalareas", "corticalthicknesses", "corticalthicknessstds")
for (fieldGroup in fieldGroups) {
    extract.byname(mri_data, fieldGroup)
}
```

rsurfer provides methods to discover information about a specific feature based on its name. For these examples the area of the banks of the superior temporal sulcus in the left hemisphere, with the feature name "lh.bankssts.area". To determine which hemisphere a feature belongs to (note that if it belongs to neither left or right then central is returned); the following would return "left":

```
feature <- "lh.bankssts.area"
get.hemisphere.side(feature)
```

To get information about what type of measurement the feature is (cortical volume, hippocampal volume, subcortical volume, area, thickness, or standard deviation of thickness) use the following function:

```
getfieldgroup(feature)
```

This function has an optional parameter, which if set determines how specific the returned information is. The default value of 1 is the most specific the returned information can be; the value of 2 is the least specific it can be (returns volume, area, thickness or thickness

standard deviation).

```
getfieldgroup(feature, 2)
```

There is an rsurfer function to return the opposite measurement (if it exists) so given a left hemispherical measurement, will return the corresponding right hemispherical measurement; and vice-versa. The following function returns "rh.bankssts.area".

```
get.opposite.hemisphere.measurement(feature)
```

6.4 Intracranial Volume Normalisation

rsurfer implements all the ICV normalisation methods discussed in Section 5.5. All ICV normalisation methods use the same function, they are passed a string to determine which method to apply to the subjects. The following code will apply the following ICV normalisation methods: proportional, residual, residual with a separate model for males and females, covariate and power proportional.

```
normalise(mri_data, "normalisation.proportional")
normalise(mri_data, "normalisation.residual")
normalise(mri_data, "normalisation.residualgender")
normalise(mri_data, "normalisation.covariate")
normalise(mri_data, "normalisation.powerproportion")
```

6.5 Merging Freesurfer Data with Subject Information

rsurfer provides a number of a functions to aid the importing of data from various data sources, it allows information about the subject such as age, diagnosis and gender to be automatically merged with the imported Freesurfer data.

6.5.1 Alzheimer's Disease Neuroimaging Initiative

The Alzheimer's Disease Neuroimaging Initiative provides the information about their subjects (age, gender and diagnosis) in two comma separated value files: ADNIMERGE.csv and DXSUM_PDXCONV_ADNIALL.csv. rsurfer automates this process for baseline scans:

```
adni.setfiles("DXSUM_PDXCONV_ADNIALL.csv", "ADNIMERGE.csv")
mri_data <- adni.mergewithfreesurferoutput(mri_data)
```

The first line points rsurfer to the locations of the files (these will likely have to be changed depending on where the files are stored), and the second line merges the data from the files assuming that the row names of the subjects are their ADNI subject IDs.

6.5.2 Information Extraction from Images

Similar to the ADNI functionality, rsurfer provides methods to merge the subject's age and gender with their structural MRI data. Since IXI contains only HC subjects, it does not contain the diagnosis of the subject as they are guaranteed to be healthy. The code to merge the data is:

```
ixi.setfile("IXI.csv")
mri_data <- ixi.mergewithfreesurferoutput(mri_data)
```

6.5.3 CAD Dementia

The CAD Dementia challenge provides a set of structural MRIs diagnosed as HC or AD; rsurfer provides functionality to merge subject information with the imported data for both the training and test data:

```
caddementia.setfiles("train.txt","test.txt")
mri_data <- caddementia.mergewithfreesurferoutput(mri_data)
```

6.6 Conclusion

rsurfer has provided a way such that code written for other parts of this research may be used by other researchers to help speed up their research as they will not have to write this functionality themselves. Future improvements to rsurfer would involve testing it on different versions of Freesurfer, such as the latest version, and if it fails, to implement a method where it can determine the version of Freesurfer which was used to process the data and then to import accordingly. Other improvements would be make the ICV normalisation functions more flexible - currently they assume the ICV feature is named "EstimatedTotalIntraCranialVol", and if it differs from this then the function will fail. Creation of an additional parameter to specify the name of the ICV field will enable it to be more flexible and the user would not have to spend time organising their data to work for rsurfer.

The correctness of the functions was determined manually, however, future work could involve developing unit tests for rsurfer to ensure any future changes do not break any existing functionality.

Chapter 7

Data Selection and Preliminary Analysis

This chapter will describe preliminary experiments performed to analyse the data and derive information about it. The data used for these experiments are from the ADNI dataset and the subject distribution of diagnosis, gender and age are shown in Table 7.1. Note that only a single cohort was used for the analysis in this chapter due to the potential adverse effects of combining multiple cohorts such as differences in protocols causing a different distribution of values for features.

7.1 Data Sources

This section will cover the locations where the data from which this work was sourced. The subjects from each cohort were filtered by the following criteria: the MRIs were the baseline scans for each subject - this is the initial scan taken and initial diagnosis given to the subject; slice thickness of the scan was 1.5mm; and the MRI was weighted in T1.

Alzheimer's Disease Neuroimaging Initiative

The Alzheimer's Disease Neuroimaging Initiative is an initiative to provide researchers with various subject data that they can use to aid their research into the progression of AD. Multiple different types of data are available, some examples of which are: MRIs, positron emission tomography (PET) images, gene samples and cognitive test results. The data on ADNI is sourced from participants in North America. The ADNI study has three phases: ADNI1, ADNI GO, and ADNI2. The different phases involve different subjects and different data is captured in the different phases. ADNI is not a population based study and thus the results obtained from ADNI data may fail to generalise to other populations.

The initial design for the ADNI1 phase was for it to be a five year partnership to test whether MRI, PET, neuropsychological tests and other biomarkers could be combined to be able to measure the progression of MCI and early AD. The requirements for it were that there are uniform standards with regards to data collection and that the repository needed to be accessible and contain the aforementioned biomarkers. This study included 200 AD positive subjects and 400 MCI subjects (with a high risk for AD) scanned every 6 months, 25% of subjects were scanned using a 3T scanner while the rest were scanned using a 1.5T scanner; the scanners were performed at a six-monthly interval until (and inclusive of) the 24th month.

For the ADNI GO phase, the goal was to see whether there are relationships between clinical, cognitive imaging, biochemical and genetic biomarkers of AD. It built upon the original ADNI1 phase. The subjects used in this phase were 200 early MCI subjects scanned at 3T every 3 months for a year; and a continuation of the monitoring of around 500 subjects from the initial ADNI phase (using the same weighted scanner as before, so if they were scanned with a 1.5T scanner in ADNI1, they would again be scanned by a 1.5T scanner).

Based on the results from the initial ADNI phase being positive, the ADNI2 phase was introduced collecting data from 550 subjects. The goal was the same as the goal for the ADNI GO phase and the subject data collection methods were the same: new subjects scanned at

Table 7.1: Distribution of subjects in the ADNI dataset

Diagnosis	**#Subjects**	**%Male**	**Age** ($\mu \pm \sigma$)
HC	145	45.5	73.9 ± 6.32
MCI	216	53.2	71.2 ± 7.31
AD	145	55.2	75.1 ± 7.75

3T; and existing subjects scanned at the same weight as they have been scanned already.

Information Extraction from Images

Information Extraction from Images was a three year project with the aim to collect healthy MRIs to be used for drug discovery, medical research and to improve decision making in healthcare. IXI collected nearly 600 MRIs of HC subjects collected from three different hospitals in London. Both 1.5T and 3T scanners were used depending on the hospital.

Computer-Aided Diagnosis of Dementia Challenge

The Computer-Aided Diagnosis of Dementia (CAD Dementia) challenge collected MRI data of HC, MCI and AD subjects from three different hospitals/medical centres. It provides 30 subjects as training data whose diagnosis is available. The rest of the test data does not have the diagnosis available so it is unusable except for a submitting a result to the CAD Dementia challenge.

7.2 The Relationship Between Volumes of Regions of Interest and the Intracranial Volume

This experiment investigates the proportionality between ICV and volumes of regions in the brain. Since the ICV is the total volume of the brain, based on common sense we would expect volumes inside the brain (such as the total volume of the cerebral cortex) to be proportional to the ICV. The volumes of various ROIs were compared with the total ICV of the brain, the Pearson correlation values of these comparisons can be found in Figure 7.2. The total

Table 7.2: Pearson values of the correlation between ICV and other volumes

Subject Group	Total Grey Matter Volume	Subcortical Grey Matter Volume	Supra Tentorial Volume	Cortical WM Volume
HC Male	0.719	0.630	0.903	0.743
AD Male	0.639	0.611	0.894	0.746
HC Female	0.478	0.366	0.633	0.405
AD Female	0.698	0.451	0.825	0.636

GM volume is plotted against the ICV and a graph is shown for males with HC or AD in Figure 7.1 and females with HC or AD in Figure 7.2. There is notably less correlation for HC females than the other groups. The same process was repeated with the subcortical GM volume, the graphs of which are in Figures 7.3 and 7.4. In this case the female volumes are relatively less correlated than the male volumes. For the supratentorial volume, the graphs are in Figures 7.5 and 7.6 where the correlation values are all high apart from HC females which is slightly lower.

The difference in correlation between males and females could be due to men having 6.5 times the amount of GM voxels than women, and women have roughly 9 times the amount of WM than men (Haier et al., 2005). This would mean that in men, the GM volume makes a larger proportion of the ICV, whereas for women, the WM volume makes a larger proportion of the ICV. To verify this, we must also look at how WM relates to the ICV, the correlation results of this are shown in Table 7.2, this demonstrates there is not much correlation in females between the WM of the cortex and ICV. This lack of correlation could potentially mean that ICV normalisation can only benefit measurements which correlate with it; as if a non-correlating value is normalised to ICV then errors could be induced. However, in this experiment only four features have been investigated, as due to the large set of features it is infeasible to manually analyse and discuss them individually.

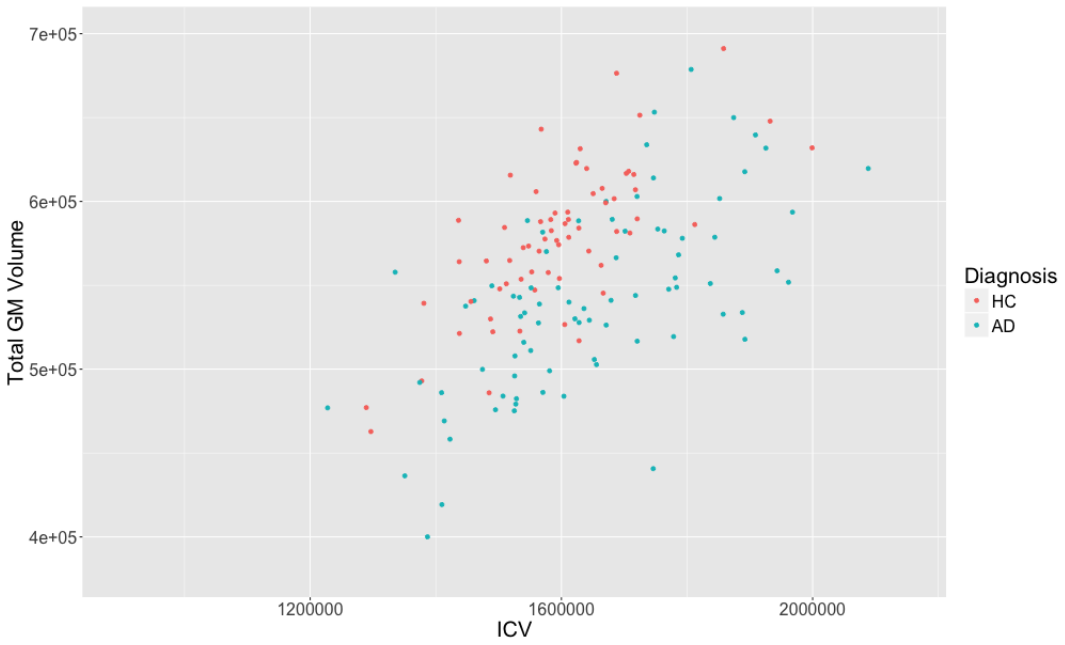

Figure 7.1: Total GM Volume (of males) versus their ICVs

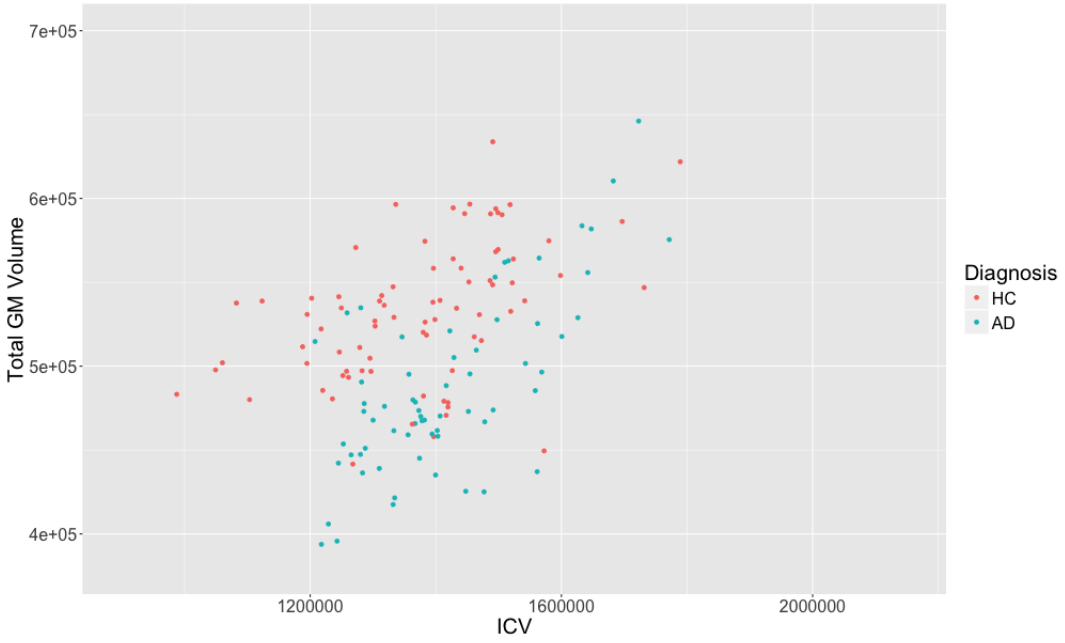

Figure 7.2: Total GM Volume (of females) versus their ICVs

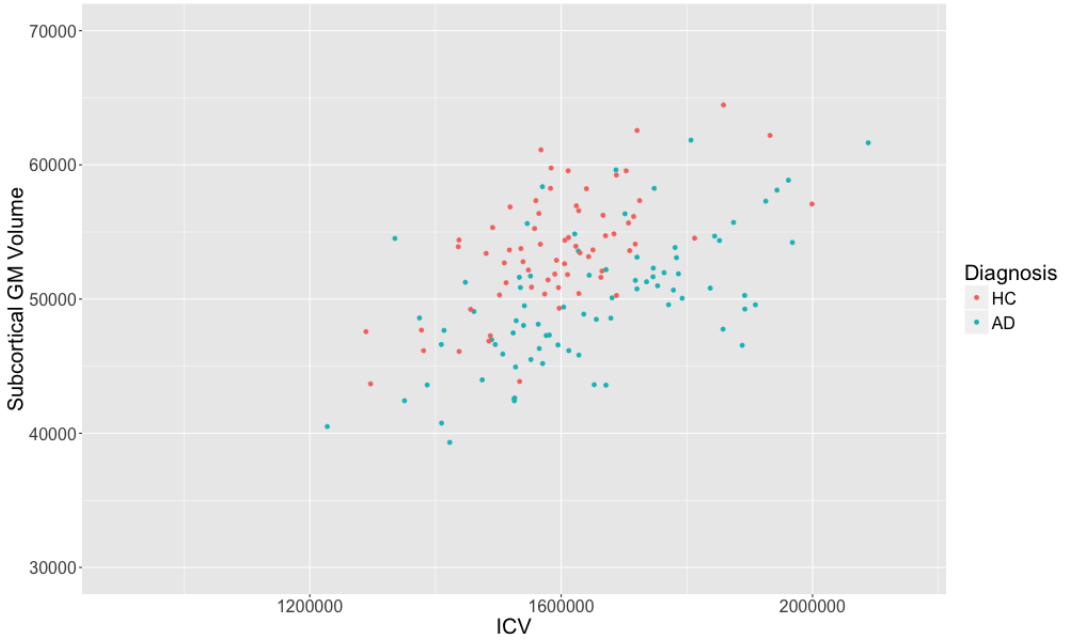

Figure 7.3: Subcortical GM Volume (of males) versus their ICVs

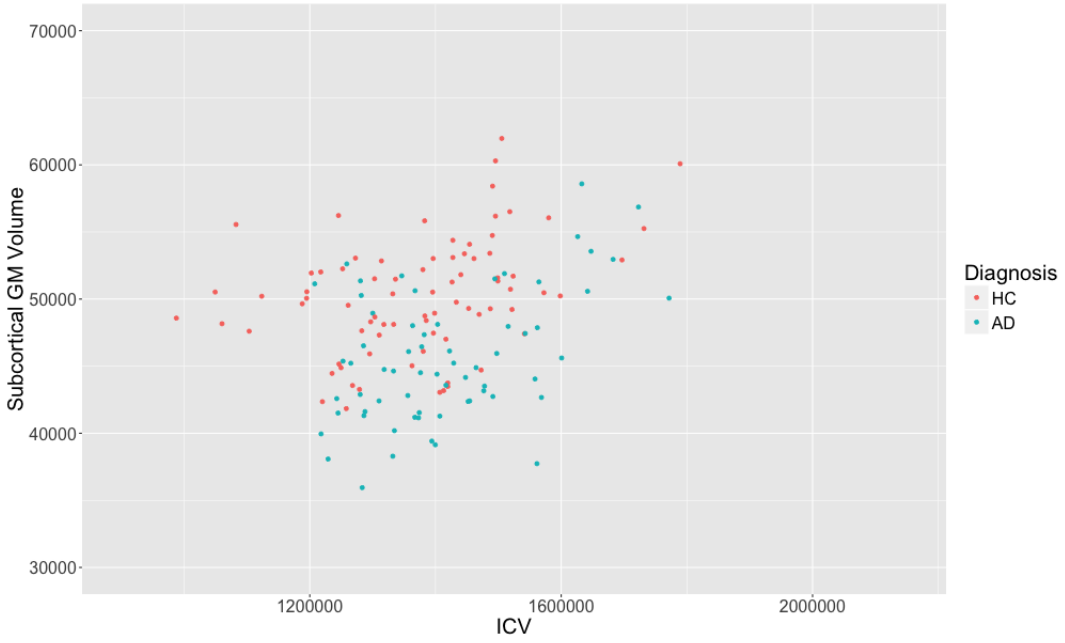

Figure 7.4: Subcortical GM Volume (of females) versus their ICVs

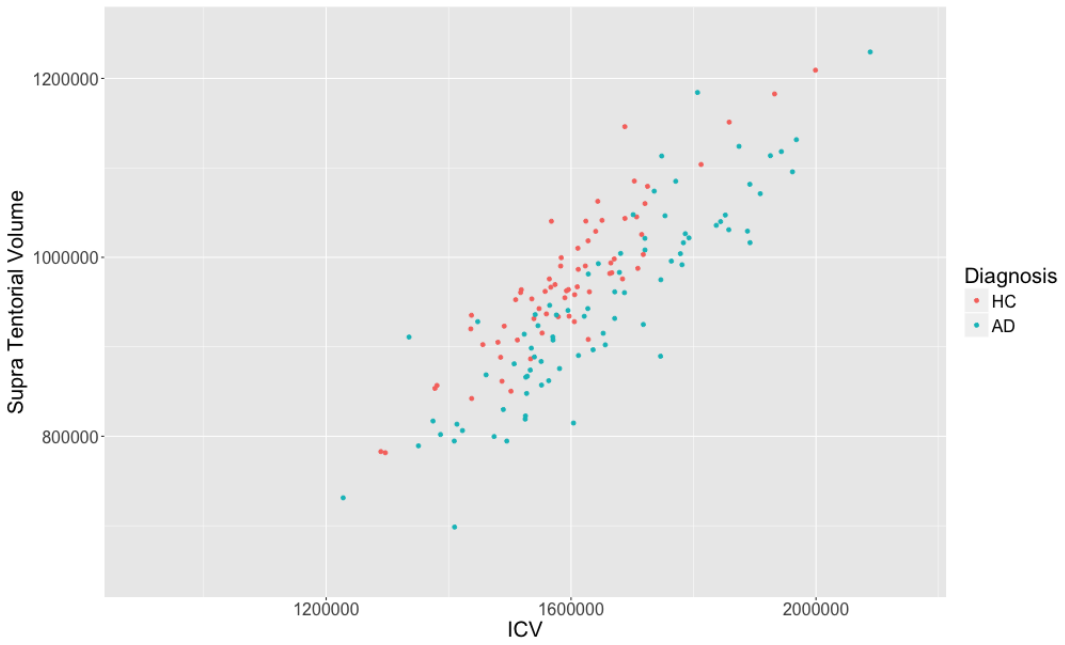

Figure 7.5: Supra Tentorial Volume (of males) versus their ICVs

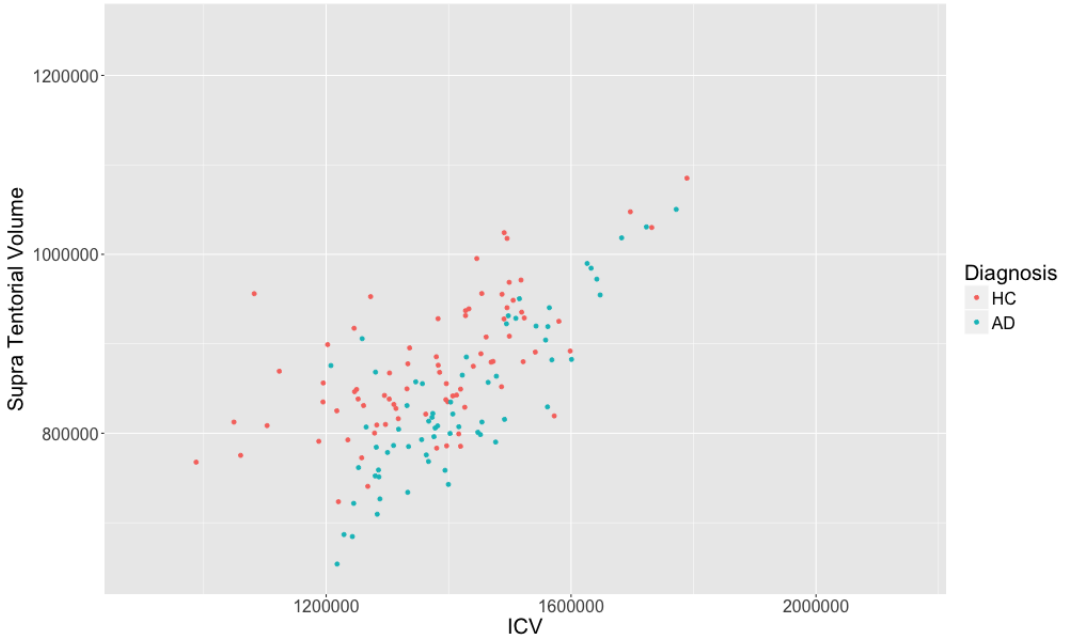

Figure 7.6: Supra Tentorial Volume (of females) versus their ICVs

Grey matter structures in the brain have been found not to scale proportionately with the ICV of the brain (Barnes et al., 2010), thus as the brain size increases, the grey matter structures will not increase at the same rate. The graph of the supratentorial volume against

ICV for males in Figure 7.5 shows a linear relationship between the two; however, in the graph of the subcortical GM volume against the ICV for females, see Figure 7.4, there is a large amount of variance and it appears as though there is almost no relationship between the two. One way to analyse this objectively, would be to build a linear regression model where the volumes of a given ROI are predicted by ICV, and measure how close the data is to the regression line. The metric which measures this is known as the coefficient of determination or the R^2 value of a regression model and it is computed as:

$$R^2 = \frac{\sum_i (\hat{y}_i - \overline{y})^2}{\sum_i (y_i - \overline{y})^2}, \tag{7.1}$$

where $\hat{y}_i$ is the predicted value of y_i. The value of R^2 is between 0 and 1 (inclusive of both), a higher R^2 is an indicator of goodness of fit of the observations. In this case, a better goodness of fit means that there is a higher proportion of variance in the dependant variable that is predictable from the independent variable. An R^2 of 0 means that the regression line represents none of the data (no information can be inferred about the data from the regression line), whereas an R^2 of 1 means that the regression line represents all of the data (the regression line explains the relationship between the two variables perfectly). The R^2 can be used as a measure of how linear the relationship between a volume and the ICV is, a low R^2 would mean that there is not a linear relationship between the two, and a high R^2 means there is a linear relationship between the two.

The R^2 was computed for between all linear regression models between each ROI volume and ICV for HC subjects (MCI and AD were excluded due to being abnormal states). The subjects were grouped by gender, thus males and females each had a separate regression model per ROI volume, and the density plots are shown in Figure 7.7. The density plots show that very few of the models have an R^2 greater than 0.5, the specific values are for females there is only one feature and for males it is 12 volumes (out of a total 145 volumes). These density plots show that very few brain ROI volumes are proportional to ICV; the proportional ICV normalisation assumes a linear relationship between the volume of the ROI and ICV, but the

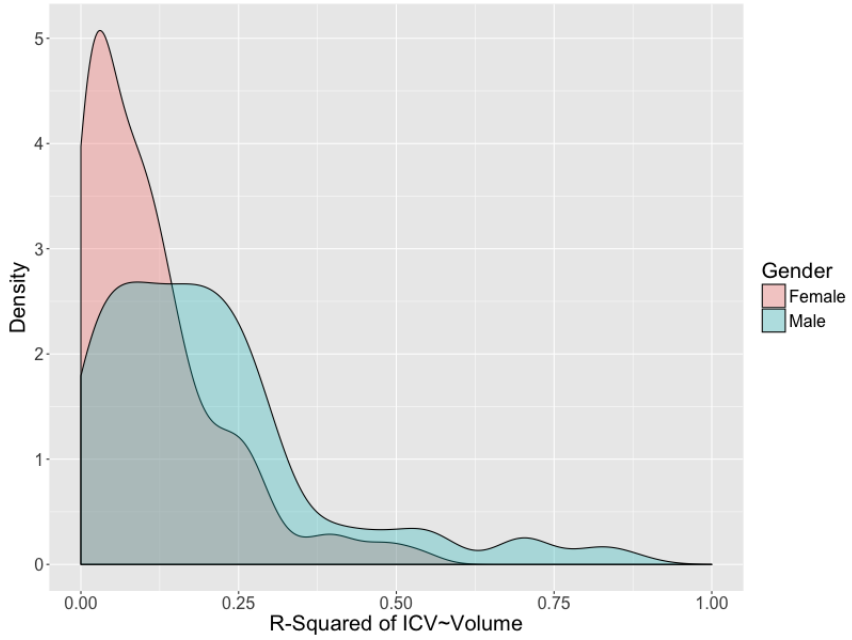

Figure 7.7: The density of R^2 values of the linear regression model between ICV and individual volumes

data demonstrates this relationship is non-existent for the majority of ROI volumes. This may mean the proportional ICV normalisation technique may be inducing error into the data.

7.3 The Effect of ICV Normalisation Methods on Classification Accuracy

Various ICV normalisation methods have been defined, however a single best method has not been discovered. One way to evaluate the best performing method would be to use each ICV normalisation method on the data, and then run a classifier to compute the accuracy. The ICV normalisation method which achieves the highest accuracy is likely to be the best method to use (at least in this context). This will be tested on a three-class problem: HC vs. MCI vs. AD. To compute the sensitivity and precision of a three-class problem, each class will have their own sensitivity and precision computed as follows:

Table 7.3: The effect of ICV normalisation methods on classification accuracy, where Sen_c and Pre_c refer to the sensitivity and precision of class c respectively

Normalisation	**Accuracy**	**Sen$_{\mathbf{HC}}$**	**Pre$_{\mathbf{HC}}$**	**Sen$_{\mathbf{MCI}}$**	**Pre$_{\mathbf{MCI}}$**	**Sen$_{AD}$**	**Pre$_{\mathbf{AD}}$**
None	54 ± 1.22	50.4 ± 1.22	49.2 ± 1.22	51.7 ± 1.22	49.9 ± 1.22	60.1 ± 1.22	64.8 ± 1.22
Proportional	54.9 ± 1.63	51.6 ± 1.63	51 ± 1.63	53 ± 1.63	50.9 ± 1.63	60.4 ± 1.63	64.8 ± 1.63
Residual	52.5 ± 1.46	49.5 ± 1.46	47.9 ± 1.46	50 ± 1.46	48.6 ± 1.46	58.7 ± 1.46	62.8 ± 1.46
Residual Gender	54 ± 1.07	50.7 ± 1.07	49.8 ± 1.07	52.1 ± 1.07	49.9 ± 1.07	59.4 ± 1.07	64.3 ± 1.07
Covariate	53.6 ± 1.77	50.3 ± 1.77	47.7 ± 1.77	51.7 ± 1.77	50.4 ± 1.77	59.1 ± 1.77	64.3 ± 1.77
Power Proportional	32 ± 2.77	23.4 ± 2.77	14.6 ± 2.77	41.5 ± 2.77	25.3 ± 2.77	28.5 ± 2.77	59.4 ± 2.77

$$Sensitivity_c = \frac{TP_c}{TP_c + FN_c}, \tag{7.2}$$

$$Precision_c = \frac{TP_c}{TP_c + FP_c}, \tag{7.3}$$

where c is the class the metric is being computed for, TP_c are the true positives of class c (i.e. the number of observations of class c predicted correctly), FP_c are the false positives of class c (i.e. the number of observations classified incorrectly as class c), FN_c are the false negatives of class c (i.e. the number of observations of class c predicted incorrectly).

The workflow for this is shown in Figure 7.8. The results are shown in Table 7.3, and these show that proportional ICV normalisation performs only slightly better than no ICV normalisation, and many of the ICV normalisations perform worse, especially power proportional ICV normalisation which produces a worse classifier model than random guessing. The conclusion that can be drawn from these results are that certain ICV normalisation methods perform slightly better than no ICV normalisation, however it is not a notable performance increase. From these results we cannot be certain whether or not ICV normalisation should definitely be used - potentially it may depend on the classifier or feature set used thus in the work both ICV normalisation being used, and not being used will be tested, to try and determine if there is a noticeable performance boost.

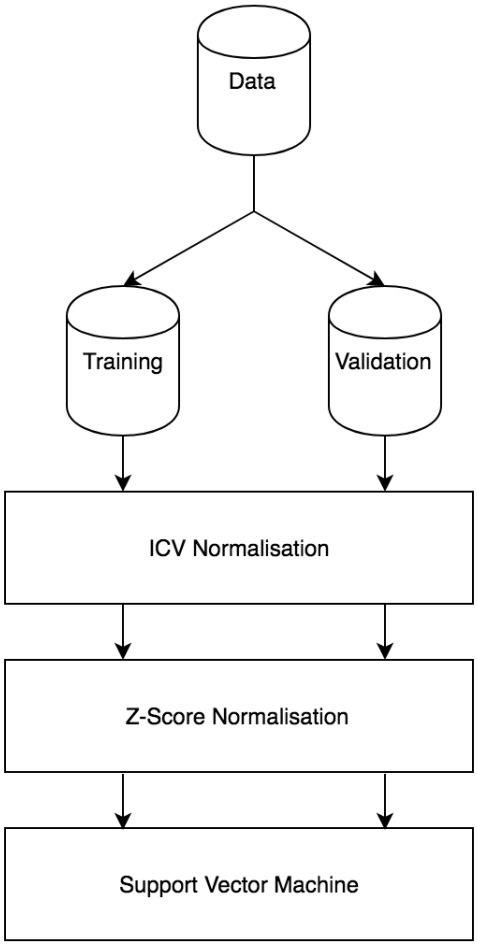

Figure 7.8: The workflow for classifying the data

7.4 A Genetic Algorithm for Feature Selection

The rest of this chapter uses a genetic algorithm (GA) as a feature selection method to identify key features for the classification of MCI and AD using an SVM. SVMs are a state-of-the-art technique to classify high dimensional data as they are generally one of the best classifiers at obtaining a high accuracy. Research has found (Nilsson et al., 2006) that feature selection does not improve the accuracy of the SVM due to its robustness at handling the features. While reducing the features of the input data may not increase the accuracy of the classifier the advantage it would have is producing a more intuitive model of the brain with AD as the features which are not affected by AD would be removed. A smaller feature set being used to distinguish HC, MCI and AD subjects would mean that there is a model of the brain with a smaller number of features showing the main areas affected by MCI or AD. The advantage of this model would be to aid a medical centre in the diagnosis of patients who potentially suffer from MCI or AD. An MRI scan of the patient's brain can be taken as input to this automated classifier, this would then produce a diagnosis of the subject with a given confidence of what they suffer from (or if they are healthy), this can then be used to refer them onto further specialist treatment.

As in the previous analysis, Freesurfer is used to extract measurements regarding different regions of interest (ROIs) throughout the brain and these measurements are then used to diagnose the subject as being healthy or suffering from MCI or AD. The curse of dimensionality is a term referring to a set of problems which occur when handling high-dimensional data: as the number of dimensions increase, the data becomes more sparse and therefore patterns are harder to find; it can also lead to overfitting of the training data making the classifier useless when applied to new data; SVMs tend to be more robust than other classifiers with their ability to handle large feature sets, thus the aim of using an SVM as the classifier will be to eliminate any irrelevant features. If a classifier was used which could not handle a large number of features then it would likely bias the GA to find small feature sets rather than just eliminate irrelevant ones. Another benefit of performing feature selection, even with an SVM

is that to achieve a model which can be easily interpreted by a human doctor the number of features must greatly be reduced otherwise there is too much information for a human to process such that the data can be displayed on a two-dimensional graph, or showing how certain areas of the brain are affected by AD as it would be difficult for a human to keep track of over three hundred variables.

By using a GA for feature selection the aim will be to develop an understanding of which features in the brain are the most useful for predicting MCI and AD without being biased by any previous knowledge of the functionality of the brain nor by any restrictions in other feature selection methods as the majority are based on finding a locally optimal solution which could potentially mean that other important but less obvious patterns are not discovered.

7.5 Method

7.5.1 Data Acquisition

A set of 435 MRI scans from ADNI were used with volumes from the subcortical, cortical and hippocampal areas. The hippocampal subfields were included since AD has been found to be prevalent in the hippocampal region. This meant that including the class label, there were a total of 358 features per subject. This method will also be tested with and without ICV normalisation to see what affect it has on the results as the previous chapter showed a negligible increase in accuracy when ICV normalisation was used, as well as previous literature which looked at feature selection for SVMs for MRIs having better results without ICV normalisation in some cases (Sarica et al., 2014b). Z-score normalisation (Equation 7.4 where μ is the mean of the data and σ is the standard deviation) is performed since SVM algorithms typically assume that the data is within a unified range; if the normalisation is not performed then the SVM can be adversely affected and misclassify the data.

$$z = \frac{x - \mu}{\sigma} \tag{7.4}$$

Table 7.4: Information about the subjects used

Dataset	Diagnosis	#Subjects	%Male	Age
Train	HC	117	49.1	74.4 ± 6.10
	MCI	117	50.0	73.9 ± 6.54
	AD	117	52.3	75.4 ± 7.93
Test	HC	28	50.0	72.6 ± 6.67
	MCI	28	46.7	73.7 ± 6.74
	AD	28	57.7	74.1 ± 7.53

The MRI data will be split into training data and test data - the classifier will be trained on the training data thus it will be able to learn to diagnose these subjects, then its prediction of the diagnosis of the subjects will be evaluated based on the test data. Stratified sampling will be used so that both sets have a similar proportion of the three classes. 351 subjects will be used as the training data and the remaining 84 will be used as test data, each data set is class balanced meaning that there are an equal number of subjects with one class as the other two classes, further information can be found in Table 7.4.

7.5.2 An Introduction to Genetic Algorithms

Genetic algorithms were pioneered by John Holland, he provided a framework (Holland, 1992) for viewing all adaptive systems and showing how the evolutionary process could be applied to them. Once a problem is defined in these genetic terms it can be solved by a GA. A GA simulates the evolutionary processes defined by Darwin (1859) where genetic operations are applied to chromosomes. In nature, a chromosome is a DNA molecule with genetic material of the organism to which it belongs and can be represented as a string where each letter can have one of four possibilities representing the nucleotide base: “A” for adenine, “C” for cytosine, “G” for guanine and “T” for thymine. The arrangement of the letters within the string determines the genome of the organism, for example “GATC” would be different to “CGAT”.

DNA molecules are able to self-replicate and they are able to translate the nucleotide bases into proteins and enzymes to control behaviour in biological cells. This means that the

structure of the DNA of an organism will affect the behaviour and its ability to perform tasks: the better at performing tasks an organism is, the more likely it is to survive and reproduce. The offspring produced by the reproduction will have chromosomes which are a combination of its mother and father. Thus an organism with a DNA structure which helps it survive and reproduce a lot is more likely to have this behaviour passed onto the next generation than an organism with a DNA structure that does not help it to survive long. This means that over time a stronger DNA structure will be passed to subsequent generations allowing them to survive for longer and produce more children. This can lead to an issue where a DNA structure may not be able to be improved on and spread through the population losing any diversity in DNA structures. However, there may be other existing structures which are even better performing than the structure which has been spread through the population. This more optimal DNA structure may never be found because the population is stuck in a "local maxima". Chromosomes have the ability to mutate, this is an unpredictable change which can affect the number of chromosomes in a cell or change the DNA structure. The latter enables populations to break free from the local maxima and potentially find the global maxima.

The implementation of a GA replicates this behaviour that is seen naturally. Once a problem is defined in genetic terms a solution can be represented by a chromosome. To explain this a very simple problem will be defined, find x such that $x^2 - 2x - 2 = 0$ where $0 <= x < 32$ and $x \in \mathbb{Z}$. Since the solutions are defined as $1 \pm \sqrt{3}$ and $x \in \mathbb{Z}$, the GA will find the value of x which finds the best result. A solution to the problem is simply the value of x which can be represented as a binary bit-string, two examples of this are:

13: 0 1 1 0 1

18: 1 0 0 1 0

Next a method is needed to evaluate the performance of the chromosomes, this is known as a fitness function. A fitness function is generally unique to each problem, and in this case we want to find x where $x^2 - 2x - 2$ is as close to zero as possible. Thus we can define the

fitness function as:

$$\text{fitness}(x) = \frac{1}{|x^2 - 2x - 2|} \tag{7.5}$$

Note that since $x \in \mathbb{Z}$, the problem of division by zero will not occur. fitness(x) will return a higher value the closer $x^2 - 2x - 2$ is to zero. Based on this we can evaluate the previous two example bit-strings, fitness(13) = 0.007 and fitness(18) = 0.003. Both of these values produce a very low fitness level so they will have less chance of reproducing for the next generation. Like the fitness function, reproduction and mutation depend on the solutions, if the solutions can be formatted as a bit-string, the reproduction and mutation are simple. Nonetheless there are multiple types of reproduction which can be implemented, for this example a single point crossover will be used. A random index is chosen between zero (inclusive) and the length of the bit-string (exclusive), and the two bit-strings are split at this index. Then the contents to the right of this split are swapped with the right-hand side of the other bit-string resulting in two children based on the structure of their parents. This is demonstrated here:

0	1	1	0	1	Crossing over to produce	0	1	0	1	0:	**10**
1	0	0	1	0		1	0	1	0	1:	**21**

Mutation of a chromosome in this case will involve choosing a random index and flipping the bit at that index:

9:	0	1	0	0	1	Mutating to produce	0	1	1	0	0:	**12**
			M		M							

A population size of a GA determines how many chromosomes are in a single generation, if the population size is too low then there are only a few possible crossover options and this will result in the GA potentially not being able to search the entire solution space (Pettit and Swigger, 1983). The higher the population size, the longer the GA will run (Grefenstette,

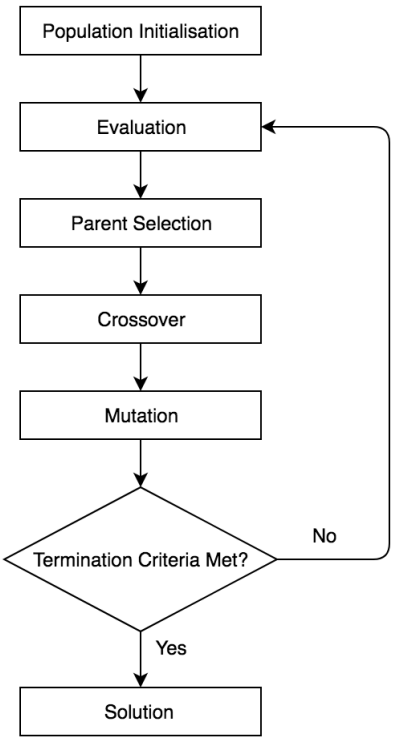

Figure 7.9: Workflow of a genetic algorithm

1986). A GA ends when a certain termination criterion is met, this can be defined by the user. Termination criteria usually relate to the amount of generations simulated, or the highest fitness value reached. A summary of the GA process is shown in Figure 7.9.

7.5.3 Feature Selection Algorithm

In this application of a GA to feature selection, each potential solution is represented by a bit string and length is equal to the number of features available. The value of the bit depends on whether the feature has been included in or excluded from this solution. For example, if the i^{th} bit is 1, then the i^{th} feature is passed to the classifier and it trains using that feature; and if the j^{th} bit is 0, then the j^{th} feature will be ignored by it. Initially the chromosomes will be initialised randomly: given a probability, p_I, if a random number between 0 (inclusive) and 1 (exclusive) is greater than p_I then that bit will be a 1. The crossover method used will be one-point crossover where two parents produce two children: a random index is chosen

in the bit strings and the data beyond that index is swapped between each string producing two children. Mutation will be implemented via bit flipping, each bit in each string will be flipped with a probability, p_M.

The fitness function (the function which is used to evaluate how well each chromosome performs at solving the problem) will be set to the classification accuracy achieved since the aim of the feature selection is to increase the successful classification rate of the problem; a second GA will be run with a modified fitness function whereby the chromosomes are penalised for having more than 20 features, for every feature over 20, the fitness is decreased by a value of 5 (this value of 5 is equivalent to a 5% drop in accuracy). This penalty is used to keep the number of features low as this is the aim of this research - to create a model with a small number of features. The value of 20 was chosen based on it being slightly higher than the count of the hippocampal features to determine whether the GA would select a feature set the size of the hippocampal feature set or if features from other parts of the brain contained additional information. Parent selection is the mechanism which chooses which parents breed together to produce the offspring for the next generation, using stochastic universal sampling which removes the bias fitness proportionate methods have towards only selecting solutions with the highest fitness (Baker, 1987).

For the development and testing of the GA described in this chapter, R was used along with external packages for the GA (Scrucca, 2013) and the SVM (Karatzoglou et al., 2004), as well as the rsurfer package.

7.5.4 Classifying the Data

The classification problem is a three-class problem, Galar et al. (2011) found that the classifiers perform better when multi-class problems were split into binary-class problems, the classification result for each binary-class problem is then combined using an aggregation method to obtain a classification for the multi-class problem. Following on from this research, the ternary-class problem will be split into three binary-class problems and a GA

will be trained for each of these binary-class problems. The Weighted Voting Strategy (WV) aggregation method (Galar et al., 2011) for the SVM will be tested to see how well it performs at combining the three two-class SVMs against the single three-class SVM, where the three-class SVM uses a One vs. Rest approach.

The fitness of each chromosome of the GA will be calculated from the accuracy of an SVM with an RBF kernel, $k(x, x') = \exp(-\sigma||x - x'||^2)$, using the given feature set, the SVM will be trained on the training data using 10-fold cross validation. Once the termination criteria is reached, the feature set which gave the highest accuracy will be trained on the entirety of the training data and then tested on the test data, and the accuracy of this will be a measurement of how well the final feature set chosen performs.

7.6 Results

A GA-based feature selection method was used in each of the 24 classification problems (e.g. one of these 24 problems is: HC vs. MCI with ICV normalisation using the cortical fields). The 24 cases include binary tests with and without ICV normalisation, multi-label classification with and without ICV normalisation, over three different initial sets of features. The All Fields subset contains both the cortical subfields and hippocampal subfields, gender and age of the subject are not included in this subset as the aim is to select the best features of the MRI data not MRI data augmented with other features. The Cortical Fields subset contains all the fields generated by the recon-all command of Freesurfer with the -all flag. The Hippocampal Subfields subset contains only the fields generated by the -hippo-subfields flag.

The parameters of the GA used are: a crossover rate of 0.6, a mutation rate of 0.02 (the probability that each bit of the bit string representation will flip), a population size of 50, single-point crossover and roulette wheel parent selection, these parameters were determined

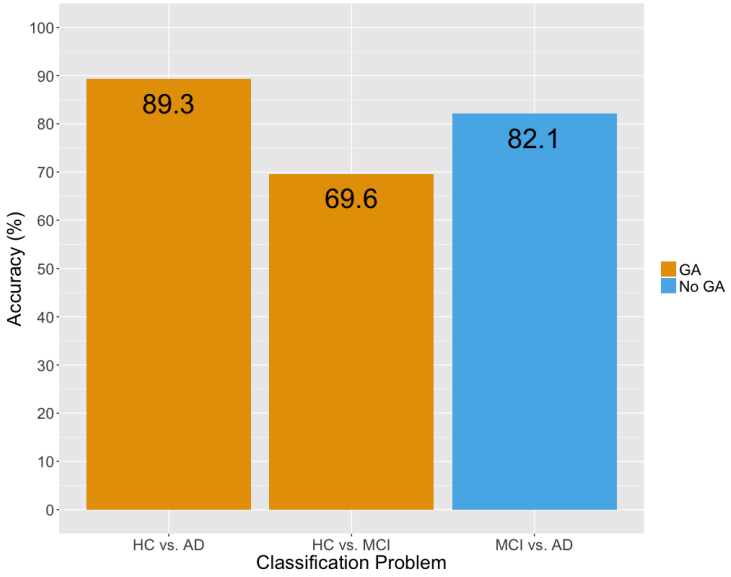

Figure 7.10: A summary of the two-class problem results. These bar charts show the best performing results on each two-class problem.

based on a preliminary parameter search. At the start of the GA, each chromosome is initialised with a probability of 0.5 that a bit is a 1 instead of 0. The results of the GA feature selection for the two-class problems are in Table 7.5 - the best results summarised in Figure 7.10; and the results of the GA feature selection for three-class problems and also the results of the two-class SVMs combined with WV are in Table 7.6 - the best results summarised in Figure 7.11. Within these tables, "Train" refers to the accuracy of the best performing genotype found, when calculating its accuracy using 10-fold cross validation on the training dataset (351 subjects); and "Test" is the accuracy found when the best features found from the GA are used to train an SVM on the entire training dataset and then used to evaluate the test dataset - a holdout method (a single split in the entire data set dividing it into two sets, one for training and one for testing). The second GA - using the fitness function with the penalty for over 20 features, behaves similarly to the former GA; other than the fitness function the only other difference is the initialisation of the chromosomes. Every chromosome has a random ten features selected, all of which are set to be included (every other feature is excluded).

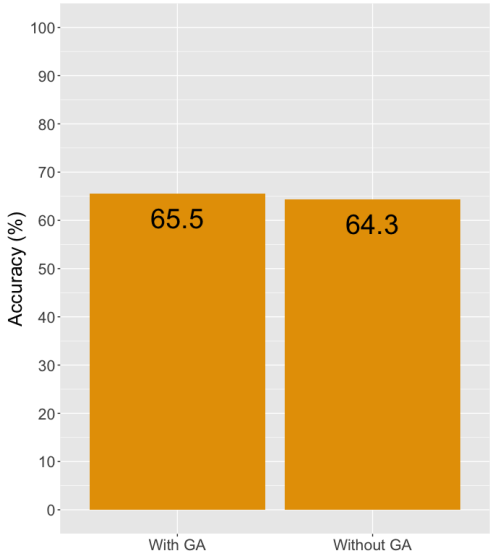

Figure 7.11: A summary of the three-class problem results. These bar charts show the best performing results with and without feature selection by a GA.

Table 7.5: Results of the two-class SVMs to solve the binary classification problems using a GA for feature selection. Classification problems marked with a (p) is when the GA is run with a penalty for a feature set with more than 20 features; problems marked with (NoFS) are the result of the classification using no feature selection. "#F" is the number of features in the most optimal run of the GA (the value excludes the age and sex which were included every time).

Classification Problem	**All Fields**						**Cortical Fields**						**Hippocampal Subfields**					
	ICV			*No ICV*			*ICV*			*No ICV*			*ICV*			*No ICV*		
	Train	*Test*	#F	*Train*	*Test*	#F	*Train*	*Test*	#F	*Train*	*Test*	#F	*Train*	*Test*	#F	*Train*	*Test*	#F
HC vs. MCI	65.7	62.5	186	64.6	67.9	184	64.7	58.9	182	66.6	67.9	169	64.5	71.4	7	67.3	48.2	11
HC vs. MCI (p)	67.6	67.9	16	69.7	66.1	20	67.6	64.3	10	73.9	55.4	11	65.9	57.1	7	68.0	37.5	7
HC vs. MCI (NoFS)	56.8	64.3	356	58.8	69.6	356	56.8	60.7	340	58.1	67.9	340	57.7	60.7	16	56.7	67.9	16
HC vs. AD	88.8	82.1	196	90.0	80.4	168	85.8	76.8	178	87.0	76.8	182	88.6	87.5	6	87.2	78.6	8
HC vs. AD (p)	89.3	78.6	15	88.8	75.0	20	88.3	80.4	13	83.8	73.2	7	88.5	89.3	6	86.2	76.8	8
HC vs. AD (NoFS)	85.7	82.1	356	86.1	80.4	356	82.7	82.1	340	84.0	76.8	340	84.7	87.5	16	83.0	75.0	16
MCI vs. AD	83.1	76.8	165	85.3	75.0	169	83.0	76.8	170	82.8	76.8	175	80.9	73.2	6	80.4	60.7	8
MCI vs. AD (p)	82.8	73.2	20	86.0	75.0	20	84.8	78.6	15	83.7	75.0	19	80.5	69.6	7	80.4	57.1	9
MCI vs. AD (NoFS)	80.6	78.6	356	81.0	75.0	356	79.2	82.1	340	79.7	73.2	340	75.3	71.4	16	76.2	62.5	16

Table 7.6: Performance of the three-class SVM and the two-class SVMs combined using WV. The best results are shown in bold. The number of features for the combined binary classifiers is the length of the union of the features from the individual classifiers. Classification problems marked with a (p) are results from the GA run with a penalty. Results marked with a * are when the SVM predicted all subjects to be of one class (for example, all subjects were predicted to be AD). "#F" is the number of features in the most optimal run of the GA (the value excludes the age and sex which were included every time).

Classification Problem	**All Fields**						**Cortical Fields**						**Hippocampal Subfields**					
	ICV			*No ICV*			*ICV*			*No ICV*			*ICV*			*No ICV*		
	Train	*Test*	#F	*Train*	*Test*	#F	*Train*	*Test*	#F	*Train*	*Test*	#F	*Train*	*Test*	#F	*Train*	*Test*	#F
HC v. MCI v. AD	62.1	59.5	170	62.8	58.3	187	60.9	65.5	171	65.2	33.3*	178	62.4	54.8	9	62.4	40.5	9
HC v. MCI v. AD (p)	61.5	58.3	19	65.3	52.4	9	61.0	63.1	20	61.6	33.3*	20	60.4	48.8	11	61.8	39.3	7
HC v. MCI v. AD (NoFS)	63.2	63.1	356	64.9	60.7	356	63.2	64.3	340	61.2	33.3*	340	61.5	50.0	16	62.1	39.3	16
Binary Combined	61.5	61.9	309	65.2	64.3	302	62.3	60.7	311	61.8	60.7	298	61.3	53.6	13	62.7	51.2	15
Binary Combined (p)	61.5	61.9	47	65.2	64.3	58	62.3	60.7	38	61.8	60.7	34	61.3	53.6	11	62.7	51.2	14
Binary Combined (NoFS)	61.5	61.9	356	65.2	64.3	356	62.3	60.7	340	61.8	60.7	340	61.3	53.6	16	61.5	39.3	16

7.7 Analysis and Reflection

7.7.1 Binary Classification Problems

Feature selection proved beneficial when the feature set was restricted to the hippocampal subfields; whereas for the restrictions of cortical fields and searching the entire set of exported features, the feature selection did not have as great of an effect. For the classification between HC and MCI, the highest accuracy on the test set was 71.4% where the GA without a penalty was used, restricted to the ICV normalised hippocampal subfields; this was an improvement of over 10% when the ICV normalised hippocampal subfields were used without feature selection - an accuracy of 60.7%. The worst result for this classification problem was the non-ICV normalised GA with a penalty restricted to the hippocampal subfields which performed worse than a random guess with an accuracy of 37.5%, in this case it had likely overfitted to the training data as the accuracy on that was 68.0%.

The best performance at distinguishing between HC and AD subjects was again reached by the GA without penalty applied to the ICV normalised hippocampal subfields, achieving an accuracy of 89.3%. There is a notable decrease in accuracy when the hippocampal subfields are not ICV normalised: the GA without penalty applied to the non-ICV normalised subfields achieves an accuracy of 76.8%; this is a greater than 10% decrease in accuracy caused by the volumes not being ICV normalised.

For the MCI and AD classification the GA did not yield any improvement over the control experiment where only domain knowledge feature selection of the cortical fields was applied, an accuracy of 82.1% was reached. In this case the volumes were ICV normalised, and it showed an increase of nearly 9% in accuracy over the same features when they were not ICV normalised - this accuracy was 73.2%. The reason the control experiment outperformed the GAs could be caused by a couple of reasons. The first is that it could have been a run of the GA which was stuck in a local maxima and the mutation rate was not high enough for the solutions to leave this maxima. The second is that the robustness of the SVM to handle

high dimensionality data in this classification problem was enough to be able to achieve the highest accuracy by giving precedence to the most important features.

In both the feature selection restricted to the cortical fields and feature selection without any initial field restriction, the accuracy was not improved by feature selection. This is evidenced by the investigation in Nilsson et al. (2006) where feature selection was tested with an SVM and very high dimensional data, they found that when feature selection was performed, a lower accuracy was achieved. The results of these two class classification problems show that ICV normalisation of the volumes yields a notable increase in accuracy.

7.7.2 Ternary Classification Problems

Of the entire 3-class classification problem, the best accuracy was achieved by a 3-class SVM with feature selection performed by a GA using the cortical fields with ICV normalisation, this achieved an accuracy of 65.5%. The 3-class SVM generally performed better than the combined 2-class SVMs, except in the cortical fields without ICV normalisation where the SVM always predicted one class for every subject. Feature selection again performed better for the hippocampal subfields with the best accuracy being achieved by the GA being used for feature selection on the ICV normalised data.

7.8 Conclusion

This chapter has investigated applying a GA-based feature selection in conjunction with an SVM to classify HC, MCI and AD patients from structural MRI brain data. The results have shown that a high accuracy can be achieved using just the hippocampal fields as a feature set. An SVM performs better with feature selection when feature selection is applied to cortical fields and also when it is applied to hippocampal fields; however, when applied to all of the fields, an SVM without feature selection performs better. Feature selection performed better than no feature selection for the hippocampal subfields and cortical fields; this could be due

to these feature sets containing features which do not help the SVM make a decision and the feature selection is able to remove these features allowing the SVM to perform better.

The hippocampal subfields are great predictors for distinguishing between HC, MCI and AD; only a small number of features are needed to achieve a fairly high accuracy, showing that the effects of AD and MCI are prevalent in the hippocampus (in agreement with other literature on this topic (Jin et al., 2004; Ball et al., 1985)). From this finding, there is potential for the hippocampus subfields to be used to create a simple model to provide an understanding of why a patient was classified as HC, MCI or AD.

The accuracy obtained by the combined two-class SVMs was slightly lower than the single three-class SVMs, this could be down to the randomness of the GA in which feature subsets it evaluates - as in previous literature (Galar et al., 2011) the combined two-class classifiers performed better; another reason is that for this problem, a three-class SVM performs better than combined two-class classifiers. In general, the usage of ICV normalisation results in a classifier that performs better as from the results regarding the hippocampal subfields without ICV in Table 7.5 show that the accuracy on "Test" is greatly lower than the accuracy on "Train" suggesting that an SVM which overfits the training data has been created. In summary, a predictive model to distinguish between the classes HC, MCI and AD, can be trained with a small number of features that provides near the same accuracy as the entire set of 356 features. This is useful for creating a report to show which areas of the patient's brain are most affected by MCI and AD.

Future work would involve using other metrics for the fitness function not just accuracy. Precision and sensitivity could be used, and then it could be investigated into how the GA performs when it is biased towards predicting AD cases correctly (and as a result, HC and MCI cases are misclassified). This leads to the question as to whether it would be better to classify an AD positive subject as being HC or classify an HC subject as having AD.

Chapter 8

A Probability-based Classifier for the Diagnosis of Alzheimer's Disease

8.1 Introduction

The work in Chapter 7 showed the hippocampal subfields as being a very small set of features that can distinguish between HC and AD patients with a high accuracy. This chapter shows the development of a simple yet efficient classification algorithm which achieves similar results to the two tested state-of-the-art classifiers: an SVM and the Naive Bayes classifier. The classifier takes advantage of the atrophy of the hippocampus due to AD (Henneman et al., 2009), where lower volumes throughout hippocampal regions of interest are likely to correspond to AD. While capable of a high accuracy, the classifier also provides an insight to the ROIs in the hippocampus affected by AD and can easily explain why a classification decision was made.

8.2 Data Pre-processing

The data used in this chapter were 287 MRIs from ADNI processed with Freesurfer. Just the hippocampal volumes were used in this chapter, so including the diagnosis and age there were

Table 8.1: Statistics of the data used in this chapter

Diagnosis	**Number of subjects**	**% Male**	**Age (mean ± std)**
HC	145	45.5	73.9 ± 6.3
AD	142	54.2	75.0 ± 7.8

a total of 18 features. The hippocampal subfields consist of 16 volumetric measurements: the entire left hippocampus, the entire right hippocampus, the left and right presubiculum, the left and right subiculum, the left and right CA 1, the left and right CA 2-3, the left and right dentate gyrus, the left and right fimbria, and the left and right hippocampal fissure. The criteria used to refine the search to find these 287 subjects was: that they were the baseline scan for each subject - this is the initial scan taken and initial diagnosis given to the subject; the slice thickness of scan was 1.5mm and it was weighted in T1; and that Freesurfer was able to analyse the MRI (in rare cases Freesurfer crashes during its execution if it is unable to process the MRI). A breakdown of the information of the subjects used in this study can be found in Figure 8.1, which shows the distribution of attributes of the subjects such as diagnosis, age and gender.

In the data used there is not a linear relation between the hippocampal volumes and the ICV of the patient, thus using ICV normalisation may cause loss of relationships between variables as two of the three methods previously described require a linear relationship between the volumes and ICV. Also ICV normalisation has been shown to lead to bias in volume measurements based on the age or gender of the subjects (Nordenskjöld et al., 2013). While there are multiple arguments against the usage of ICV normalisation on the data used, it will be verified as it may improve the accuracy of the method regardless of the potential downsides. The proportional method will be used, $v' = \frac{v}{ICV}$, to compute the ICV normalised volumes as was the best method evaluated in Chapter 7.

8.3 Probability-based Classifier

8.3.1 Linear Discriminant Analysis to Partition the Feature Space of a Single Region

Linear Discriminant Analysis is a method used to find the optimal linear hyperplane to separate two classes with the least margin of error. In this case it will be applied to a single region to determine an optimal threshold where the values that lie below the threshold belong to one class, and the values that lie above the threshold belong to another class. Figure 8.1 shows the probability density functions for each of the hippocampal volumes for HC and AD patients and also the threshold for each attribute generated by LDA. From this it can be seen that generally the volumes for HC subjects are greater than the volumes for AD patients except in the left and right hippocampal fissures; however, the fissures are measurements of a gap rather than the volume of a mass in the hippocampus. It also shows the thresholds created when LDA is used on the attributes to classify the diagnosis of the subject. As the hippocampal fissure measurements do not provide any predictive power for the diagnosis of a subject these measurements will be ignored and only the other 14 attributes will be used.

Using the hyperplanes found during the LDA process, each subject's hippocampal attributes are tested to verify whether the value lies in the HC-sized side of the hyperplane or the AD-sized side. A value belongs to the HC-sized side if it lies above the threshold, and the AD-sized side otherwise. If the subject lies in the HC-sized side of the hyperplane then a -1 is recorded for that attribute, and for the AD-sized side, a 1 is recorded; this generates a binary feature vector for each subject as the attributes have been transformed from having a continuous value to having one of two possible values. An example of this process on a subject is shown in Table 8.2. Intuitively if a higher count of 1 is found it should mean the subject is more likely to have AD as more of the hippocampal attributes indicate a likelihood that AD is present; and vice versa - if a higher count of -1 was found it means there is a likelihood of HC.

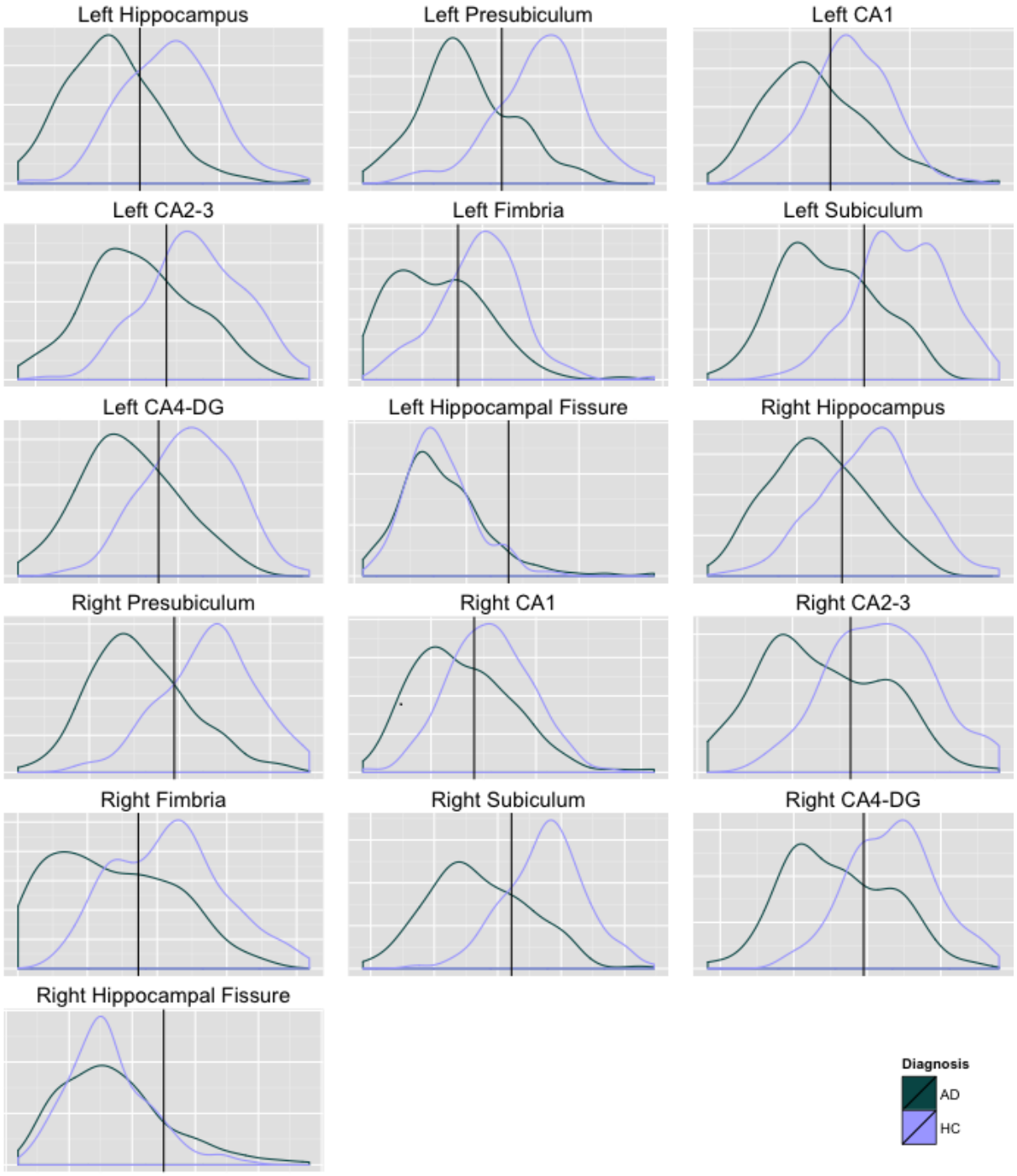

Figure 8.1: Density graphs of the hippocampal volumes and thresholds found using LDA (the axis labels have been removed for clarity)

This is shown in Figure 8.2, where the bar chart showing the distribution of subjects with the number of positive attributes and their actual diagnosis. This bar chart shows a correlation between the number of positive attributes and a diagnosis of AD; and also a

correlation between the number of negative attributes and HC subjects. However, there are subjects with a low number of positive attributes which are yet still classified with AD and vice versa for HC patients with a low number of negative attributes - in particular there are two HC subjects with all positive attributes; and two AD subjects with all negative attributes.

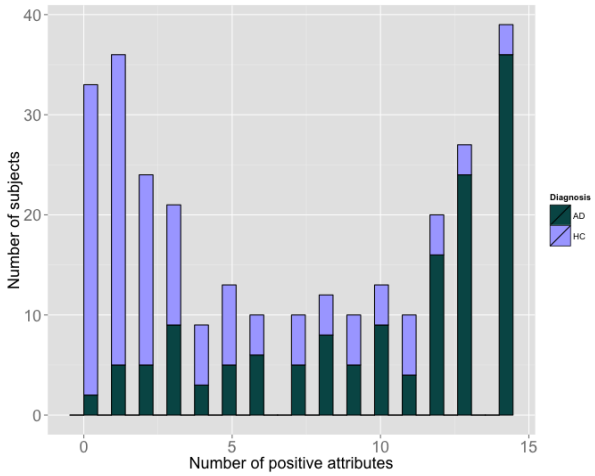

Figure 8.2: Bar chart showing the number of positive attributes in subjects

8.4 Probability Weight-based Classification Algorithm

Using the LDA thresholds, an attribute is defined as being negative if $x_n > t_n$, where t_n is the n^{th} threshold value, and positive otherwise. We will define a function $\phi(x_n)$ which will classify a single attribute as being negative or positive which will be given values of -1 and 1 respectively, the function is defined in Equation 8.1, where 1_A is the indicator function and returns 1 if the conditional statement input to the function evaluates as true, and it returns 0 if the statement evaluates as false. Based on the number of positive attributes for a subject, a classification can be made as to whether the subject is HC or AD, as for an AD patient, the number of positive regions found would be expected to be greater than the number of negative regions found, this algorithm will be referred to as the Binary Region Classification (BRC) algorithm. BRC works by summing the binary feature vector (which is comprised of

Table 8.2: LDA process for a subject with AD. A -1 in the "Binary Feature" column means that the region has been identified as having a negative attribute, and a 1 means a positive attribute.

Attribute Name	**Attribute Value**	**Binary Feature**
Left Hippocampus	2353.56	-1
Left Presubiculum	3979.98	1
Left CA1	2101.93	1
Left CA2-3	5403.29	1
Left Fimbria	207.403	1
Left Subiculum	3979.98	1
Left CA4-DG	3257.18	1
Right Hippocampus	2350.65	1
Right Presubiculum	2433.25	1
Right CA1	2877.41	-1
Right CA2-3	6835.59	1
Right Fimbria	127.686	1
Right Subiculum	4262.02	-1
Right CA4-DG	3956.99	-1

-1 and 1 rather than 0 and 1) and if the value is above 0 a classification of HC is made; if it is 0 or below then an AD classification is made. BRC is shown in Equation 8.2.

$$\phi(x_i) = 1_A(x_i \leq t_i) - 1_A(x_i > t_i) \tag{8.1}$$

$$BRC(\vec{x}) = \sum_{i=1}^{dim(\vec{x})} \phi(x_i) \tag{8.2}$$

One downside to the BRC is that each region is given the same weight as every other region (1 or -1) - it assumes all regions have the same predictive power. This is not the case, some attributes will be better at identifying AD in a subject given that the attribute is positive. This probability can be written as $P(AD \mid x_i \leq t_i)$. The probability of a patient being HC given the attribute is negative is given by $P(HC \mid x_n > t_i)$. The predictive powers for each volume measurement are shown for HC and AD in Table 8.3 and they can be used to weight the attributes: rather than each attribute having an integer value of -1 or 1. Table 8.3 also shows some additional interesting information: the top four predictors for both AD

and HC are the same (though in a slightly different order of ability for each diagnosis): Left Subiculum, Right Subiculum, Left Presubiculum and Right Presubiculum; this matches with the findings of Carlesimo et al. (2015) where it was found that atrophy of the subiculum and presubiculum were the best hippocampal markers for detection of AD.

Table 8.3: Ability of single regions to diagnose a subject. An attribute is HC-sized (HC_{sized}) if it is negative, and AD-sized (AD_{sized}) if it is positive.

Region Name	$P(HC \mid HC_{sized})$	**Region Name**	$P(AD \mid AD_{sized})$
Left Subiculum	0.81	Left Subiculum	0.82
Right Presubiculum	0.80	Right Subiculum	0.80
Left Presubiculum	0.79	Right Presubiculum	0.78
Right Subiculum	0.76	Left Presubiculum	0.77
Right CA4-DG	0.76	Left Hippocampus	0.75
Left Hippocampus	0.75	Right CA4-DG	0.75
Left CA2-3	0.74	Left CA2-3	0.74
Left CA4-DG	0.74	Left CA4-DG	0.74
Right CA2-3	0.73	Right CA2-3	0.72
Right Hippocampus	0.70	Left CA1	0.69
Left CA1	0.68	Right Hippocampus	0.69
Right Fimbria	0.67	Left Fimbria	0.68
Left Fimbria	0.66	Right Fimbria	0.66
Right CA1	0.60	Right CA1	0.60

Using this weighting system which will be referred to as Probability-Weight Classifcation (PWC), two variants of the algorithm are proposed and tested. $\text{PWC}_1(X)$ and $\text{PWC}_2(X)$ are the variants and they adopt two equations which use the summation of the probability weights, and the resultant value is used to classify the subject as HC if it is less than 0 and AD otherwise. The variants, $\text{PWC}_1(X)$ and $\text{PWC}_2(X)$, are shown in Equations 8.3 and 8.4 respectively.

Built-in feature selection can be implemented for these algorithms by having a threshold where if the predictive ability of a region falls below this threshold then it is discarded. The reasoning behind this is that regions which are worse at predicting HC or AD in a subject and thus the information they provide may lead to more misclassifications. Therefore if a threshold P_T is provided such that all regions with a predictive power lower than P_T in the

$$\mathrm{PWC}_1(\vec{x}) = \sum_{i=1}^{dim(\vec{x})} \left(\frac{\phi(x_i) - 1}{2} \Big(P(HC \mid x_n > t_i) - P(AD \mid x_n > t_i) \Big) + \frac{\phi(x_i) + 1}{2} \Big(P(HC \mid x_n \leq t_i) - P(AD \mid x_n \leq t_i) \Big) \right) \quad (8.3)$$

$$\mathrm{PWC}_2(\vec{x}) = \sum_{i=1}^{dim(\vec{x})} \Big(\frac{\phi(x_i) - 1}{2} P(HC \mid x_n > t_i) - \frac{\phi(x_i) + 1}{2} P(AD \mid x_i \leq t_i) \Big) \quad (8.4)$$

context of $P(HC \mid x_i > t_i)$ and $P(AD \mid x_i \leq t_i)$ are ignored (note that the minimum number of features that can remain are 1 no matter the value of P_T).

8.4.1 Intracranial Volume Normalisation

Two methods of ICV normalisation will be tested, the proportional method discussed earlier where $v' = \frac{1}{ICV} v$ and a custom method which involves only normalising the volumes which lie outside the mean of all the subjects' values of that volume plus or minus the standard deviation of the values for the volumes. This is shown in Equation 8.5, where $sd(ICV)$ is the standard deviation of the volume values and will be referred to as ICV normalisation SD (where SD stands for standard deviation). The ICV normalisation SD method was designed because in the initial testing of the algorithm, many of the false positives (HC misclassified as AD) and false negatives (AD misclassified as HC) were from patients who had a very high or very low ICV in comparison to the other patients (such that their ICV was outside the range of the mean ± the standard deviation of the ICVs of all subjects).

$$v' = \alpha v + (1 - \alpha) \frac{\overline{ICV}}{ICV} v$$

$$\alpha = 1_A \Big(\big(\overline{ICV} - sd(ICV) \big) > ICV < \big(\overline{ICV} + sd(ICV) \big) \Big) \quad (8.5)$$

8.4.2 Naive Bayes and Support Vector Machine Classification

This chapter describes a custom method being used to classify the data, to evaluate the performance of this method it will be compared to two state-of-the-art classification methods: a Naive Bayes classifier and a Support Vector Machine (SVM). The Naive Bayes classifier is a probabilistic classifier which applies Bayes' theorem to a set of independent features, in this case the independent features are the hippocampal volume measurements for each subject. An SVM is a non-linear classifier using a kernel to transform data into a higher dimensional feature space and then classify the data in the new feature space. The kernels which will be tested are the linear kernel $k(x, x') = \langle x, x' \rangle$ and the Gaussian RBF kernel $k(x, x') = \exp(-\sigma \|x - x'\|^2)$.

8.5 Experimental Setup

After the initial pre-processing of the data with Freesurfer, it was imported into R using the package, *rsurfer*. The following R packages were used to implemented the algorithms: *MASS* (Venables and Ripley, 2002) for the LDA; *kernlab* (Karatzoglou et al., 2004) for the SVM; *e1071* (Meyer et al., 2014) for the Naive Bayes classifier. The results were evaluated using 10-fold cross validation.

8.6 Results

Initial testing in Figures 8.3 and 8.4 show that the optimal probability thresholds are found between 0.7 and 0.8 (Figure 8.3), so those set the range for the probability thresholds which will be tested further. The results for the BRC algorithm are found in Table 8.4 where BRC is used on data where ICV normalisation has not been applied; Table 8.5 shows BRC applied to ICV normalised data and Table 8.6 is BRC applied to data which uses the ICV normalisation SD method. The results for the PWC algorithm are found in Table 8.7 where PWC_1 and

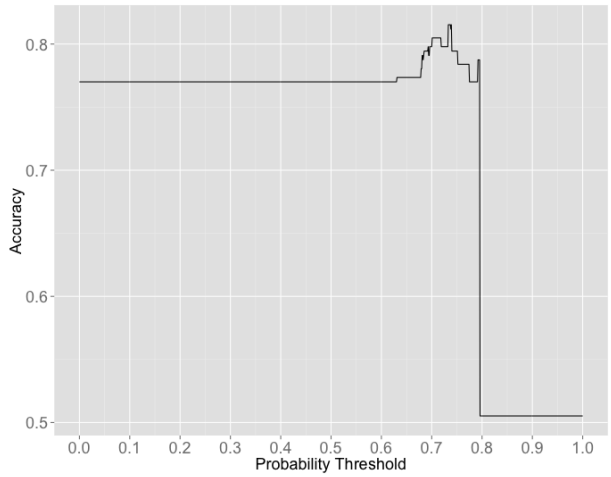

Figure 8.3: Accuracy achieved via PWC_1 on the non-ICV normalised entire dataset as the probability threshold is changed

PWC_2 are used on data which has not been ICV normalised; PWC_1 and PWC_2 are applied to data which have been ICV normalised and the results are in Table 8.8; and in Table 8.9, PWC_1 and PWC_2 have been applied to data with the ICV normalisation SD method used. A Naive Bayes classifer has been tested on the data in Table 8.11 as it, like PWC, also uses probability to classify data so they both share a similarity. SVMs were also applied to the data in Table 8.10 because out of various classifiers applied to the data, SVMs outperformed them all thus it would be useful to compare the classifiers developed in this chapter to what might be one of the highest accuracies achievable on the data with commonly used classifiers. For all the results 10-fold cross validation was repeated ten times using a different set of folds for each repetition. The 10-fold cross validation was repeated ten times for all classifiers in order to try and eliminate any bias towards the data set being tested. All of the classifiers tested used the same sets of folds so they were training and testing on the same data sets. The best results for each classifier are shown in Figure 8.5.

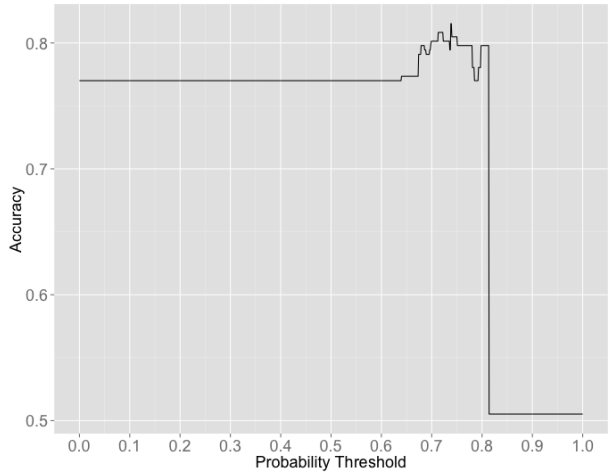

Figure 8.4: Accuracy achieved via PWC_1 on the ICV normalised entire dataset as the probability threshold is changed

Table 8.4: BRC without ICV normalisation and 10-fold cross validation repeated ten times

P_T	**Acc**	**Sen**	**Spe**
0.7	0.785	0.777	0.792
0.71	**0.795**	0.795	0.795
0.72	**0.795**	0.8	0.79
0.73	0.793	0.799	0.787
0.74	0.789	0.798	0.781
0.75	0.782	0.792	0.774
0.76	0.784	0.787	0.781
0.77	0.775	0.792	0.758
0.78	0.765	0.781	0.749
0.79	0.695	0.613	0.775
0.8	0.606	0.331	0.876

Table 8.5: BRC with ICV normalisation and 10-fold cross validation repeated ten times

P_T	**Acc**	**Sen**	**Spe**
0.7	0.835	0.805	0.863
0.71	0.836	0.804	0.867
0.72	0.838	0.803	0.872
0.73	0.84	0.801	0.879
0.74	0.838	0.799	0.876
0.75	0.832	0.792	0.872
0.76	0.838	0.8	0.874
0.77	0.849	0.813	0.883
0.78	**0.858**	0.832	0.883
0.79	0.852	0.837	0.868
0.8	0.832	0.825	0.839

Table 8.6: BRC with ICV normalisation SD with 10-fold cross validation repeated ten times

P_T	**Acc**	**Sen**	**Spe**
0.7	0.841	0.821	0.86
0.71	0.842	0.82	0.862
0.72	0.843	0.818	0.867
0.73	0.844	0.812	0.874
0.74	0.84	0.813	0.866
0.75	0.839	0.815	0.862
0.76	0.842	0.827	0.857
0.77	**0.846**	0.841	0.85
0.78	**0.846**	0.849	0.843
0.79	0.843	0.851	0.835
0.8	0.837	0.84	0.833

Table 8.7: PWC_1 and PWC_2 with no ICV normalisation with 10-fold cross validation repeated ten times

P_T	**No ICV Normalisation**					
	PWC_1			**PWC_2**		
	Acc	**Sen**	**Spe**	**Acc**	**Sen**	**Spe**
0.7	0.785	0.777	0.792	0.786	0.779	0.792
0.71	**0.795**	0.795	0.795	0.795	0.796	0.795
0.72	**0.795**	0.8	0.79	0.795	0.8	0.79
0.73	0.793	0.799	0.787	0.793	0.799	0.787
0.74	0.789	0.798	0.781	0.789	0.798	0.781
0.75	0.782	0.792	0.774	0.782	0.791	0.774
0.76	0.784	0.787	0.781	0.783	0.786	0.781
0.77	0.775	0.792	0.758	0.775	0.792	0.758
0.78	0.765	0.781	0.749	0.765	0.781	0.749
0.79	0.695	0.613	0.775	0.695	0.613	0.775
0.8	0.606	0.331	0.876	0.606	0.331	0.876

Table 8.8: PWC_1 and PWC_2 with ICV normalisation with 10-fold cross validation repeated ten times

P_T	**ICV Normalisation**					
	PWC_1			**PWC_2**		
	Acc	**Sen**	**Spe**	**Acc**	**Sen**	**Spe**
0.7	0.828	0.812	0.845	0.841	0.821	0.86
0.71	0.831	0.816	0.845	0.842	0.82	0.862
0.72	0.832	0.815	0.848	0.843	0.818	0.867
0.73	0.835	0.818	0.85	0.844	0.812	0.874
0.74	0.838	0.823	0.854	0.84	0.813	0.866
0.75	0.835	0.815	0.854	0.839	0.815	0.862
0.76	0.839	0.823	0.854	0.842	0.827	0.857
0.77	0.847	0.837	0.857	0.846	0.841	0.85
0.78	**0.856**	0.854	0.859	0.846	0.849	0.843
0.79	0.85	0.84	0.859	0.843	0.851	0.835
0.8	0.833	0.799	0.866	0.837	0.84	0.833

Table 8.9: PWC_1 and PWC_2 with ICV normalisation SD with 10-fold cross validation repeated ten times

P_T	**ICV Normalisation SD**					
	PWC_1			**PWC_2**		
	Acc	**Sen**	**Spe**	**Acc**	**Sen**	**Spe**
0.7	0.829	0.813	0.845	0.843	0.823	0.862
0.71	0.83	0.815	0.845	0.842	0.822	0.862
0.72	0.831	0.814	0.848	0.843	0.818	0.867
0.73	0.834	0.818	0.85	0.844	0.812	0.874
0.74	0.838	0.823	0.854	0.84	0.813	0.866
0.75	0.834	0.811	0.857	0.838	0.816	0.859
0.76	0.84	0.824	0.856	0.842	0.827	0.857
0.77	0.847	0.837	0.857	0.846	0.841	0.85
0.78	**0.856**	0.854	0.859	0.846	0.849	0.843
0.79	0.85	0.84	0.859	0.843	0.851	0.835
0.8	0.832	0.799	0.865	0.834	0.842	0.827

Table 8.10: SVM with 10-fold cross validation repeated ten times

	Kernel	**Acc**	**Sen**	**Spe**
SVM without ICV normalisation	Linear	0.806	0.815	0.797
	RBF	0.823	0.831	0.814
SVM with ICV normalisation	Linear	**0.849**	0.847	0.852
	RBF	0.835	0.85	0.819
SVM with ICV normalisation SD	Linear	0.835	0.837	0.833
	RBF	0.841	0.866	0.815

Table 8.11: Naive Bayes with 10-fold cross validation repeated ten times

Data Transform Method	Acc	Sen	Spe
NB without ICV normalisation	0.777	0.732	0.821
NB with ICV normalisation	0.833	0.81	0.855
NB with ICV normalisation SD	**0.847**	0.817	0.876

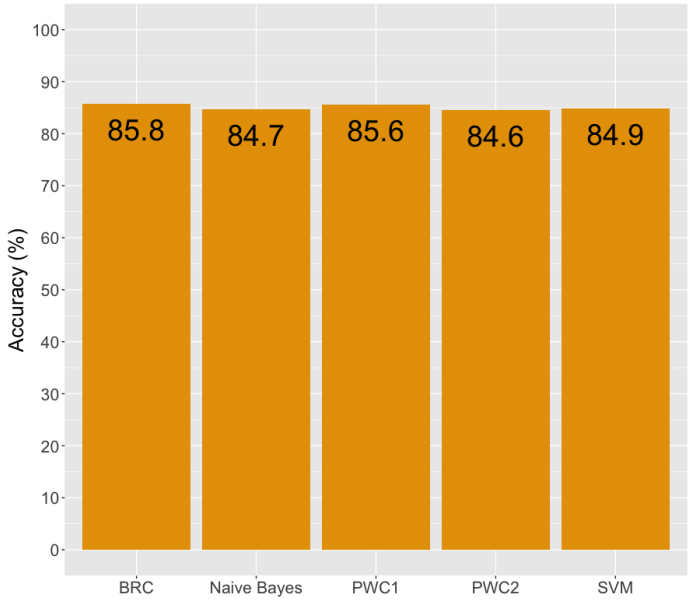

Figure 8.5: A summary of the classifiers showing the best accuracy that they were each able to achieve.

8.7 Discussion

The BRC achieved the best result with an accuracy of 85.8% using a probability threshold of 0.78 on ICV normalised data. This is a similar accuracy to the SVM which obtained an accuracy of 84.9% using a linear kernel on ICV normalised data and also outperformed the Naive Bayes classifier with a maximum accuracy of 84.7%. The BRC performed the same as the PWC, the latter achieving 85.6% showing that the additional complexity incurred in the implementation of the PWC is not worthwhile. PWC_1 always achieved a higher maximum accuracy than PWC_2, however, at lower probability thresholds (0.7 - 0.75), PWC_2 performs better than PWC_1 so it is better able to cope with less relevant features.

Intracranial volume normalised data far outperformed non-ICV normalised data; in BRC and PWC using ICV normalisation improved the maximum accuracy by over 5% in both cases, and all the accuracies at each probability threshold used were improved. With the Naive Bayes classifier and the SVM, ICV normalisation also improved the results, with the SVM being the classifier able to cope best with the non-ICV normalised data and achieved the highest accuracy of 82.3%. The ICV normalisation SD method described in this chapter in all cases bar one performs worse than standard ICV normalisation when accuracy is the evaluation metric, and it always outperforms ICV normalisation. For the Naive Bayes classifier however, ICV normalisation SD outperforms ICV normalisation by 1.5%, though this is only a marginal amount and since ICV normalisation SD was only better than standard ICV normalisation in one case, the SD method is not worth using.

The descriptive model produced by this algorithm is shown in Table 8.2, each region of a subject's hippocampus is given a binary feature of 1 or -1 depending on the value of the volume measurement for that region of the hippocampus. If a subject has more negative attributes than positive attributes then a diagnosis of HC is made; and if more positive attributes than negative attributes then an AD diagnosis is made. There is a case which could occur for certain subjects where the number of negative attributes is equal to the number of positive attributes; in this case, a classification of AD will made - the reasoning

behind this is that it would be better (regarding the patient's interest) for a healthy patient to receive treatment for AD rather than an AD patient to go ignored. An alternative would be to not make a decision and leave the diagnosis unknown which could potentially lead to doctor's monitoring the development of the patient's brain over the foreseeable future.

8.8 Conclusion

This work has created a classifier which performs similarly to an SVM and a Naive Bayes classifier when the accuracy of diagnosing a patient as HC or AD based on a baseline structural MRI scan is measured. As well as attaining a similar accuracy, it also produces an easily understood description of why a diagnosis was made, whereas an SVM does not have this advantage as an SVM works by transforming the data into a higher dimensional space where descriptive information about the data is lost. The descriptiveness of this model could aid a medical professional by determining the regions in the hippocampus that the doctor should look at - if we let all negative features be a warning sign of AD, then say the left presubiculum is negative, the classifier could advise the doctor to manually check the left presubiculum of the subject's MRI to see if there is anything about that region that could correlate with AD. This work has also shown that the best features in the hippocampus are the left presubiculum, right presubiculum, left subiculum and right subiculum matching with current literature on this topic (Carlesimo et al., 2015).

There is also potential for the algorithm to have a parallel execution in both the training and predicting parts of it. The training of the algorithm could be parallelised by: creating the thresholds for the decision boundary in parallel and evaluating different probability thresholds in parallel. Regarding the prediction part of the algorithm: the proposed methods in this chapter are based on the summation of weights and each of the weights are independent of the other weights, this means that the weights can be computed in parallel.

Chapter 9

Generation of a Descriptive Apparent Brain Age Feature

9.1 Introduction

This chapter presents a data workflow based on structural magnetic resonance imaging (MRI) to detect brain morphologic changes beyond normal ageing, which could be ascribed to AD and, ultimately, help to improve classification accuracy.

Healthy control subjects are used to define statistical properties of brain regions of interest and to guide feature selection to guarantee appropriate statistical robustness. A feature selection approach is adopted, which is based on a statistical test aimed at maximising the discriminative ability of the selected features with respect to the classification task. A LASSO regression is applied to the selected features to build a model for the subject's age. The predicted age is used as an apparent brain age and compared to the actual age of the subject. The predictive ability of the apparent brain age (in conjunction with the real age) is discovered to be great as used by itself it can achieve a high accuracy. It also improves the classification accuracy of state-of-the-art predictive models when it is combined with other MRI features.

The rest of the chapter is organised as follows: Section 9.2 presents the datasets adopted in this study and the pre-processing applied to them; Section 9.3 discusses the workflow designed to compute the apparent brain age and use it to classify subjects; Section 9.4 details the regression model and data used to compute the apparent brain age; Section 9.5 details the classification method to classify HC (healthy control) and AD subjects using the apparent brain age; Section 9.6 displays and discusses the results of this workflow; and Section 9.8 concludes the work.

The main contribution is the creation of a new descriptive feature, the age deviation score, which can be used in conjunction with the features from the MRI, and this new feature is a good summary of a larger number of features as it can improve the classification accuracy when it is being augmented with the hippocampal subfields. For this purpose, this work introduces the Differential Two-step Unequal Variances t-Test Feature Selection method, which is a novel feature selection technique that is used to merge MRI data collected from heterogeneous sources (different MRI scanners using different protocols). Moreover, the coefficients created by the LASSO regression model of the apparent age can be used to evaluate which features mostly affect the predicted age of a subject.

9.2 Data Selection and Pre-processing

9.2.1 Data Acquisition

The initial data set used in this work consists of 813 subjects of either HC or AD though this is refined to a smaller amount based on some criteria which will be defined later. The 813 subjects were obtained from the collections of two previous studies namely: ADNI, which includes HC and AD subjects; and the IXI study, which included only HC subjects. There were a total of 290 subjects from ADNI and the other 523 subjects were from IXI, and a distribution table of the subjects is shown in Table 9.1.

Table 9.1: The distribution of subjects from the two cohorts used

Cohort	Diagnosis	Number of subjects	Proportion of subjects who are male	Age			
				Mean	SD	Min	Max
ADNI	HC	145	0.455	73.9	6.32	56.2	88.6
ADNI	AD	145	0.552	75.1	7.75	55.9	90.9
IXI	HC	523	0.445	48.6	16.5	20.0	86.3

Both ADNI and IXI datasets are used as it is a necessary to have a large number of healthy subjects to build the regression model. Without the additional HC subjects from IXI then the majority of the HC subjects would be used for the regression model leading to not many being left for training and testing the classification model; there would be a large class imbalance as the regression model would not have used any of the AD subjects. As with the other chapters, Freesurfer is used to segment the MRIs, with rsurfer allowing for easier manipulation of the preprocessed data.

9.3 The Data Workflow

A diagram of the workflow used in this chapter is given in Figure 9.1, which shows the pipeline of the methods used to classify the subjects based on their apparent brain age. The first step is the selection of the data where the set of subjects is refined based on criteria such as their age, gender and whether they are an outlier compared to other subjects. The next step is a simple normalisation step, where z-normalisation is applied to the generated features: $z = \frac{\overline{x}-\mu}{\sigma}$, where μ and σ are the mean and standard deviation of the variable being scaled respectively. The feature selection step consists of a novel feature selection algorithm specifically designed to leverage on the heterogeneity of the data sources. Each step of the pipeline is discussed in detail in the following subsections.

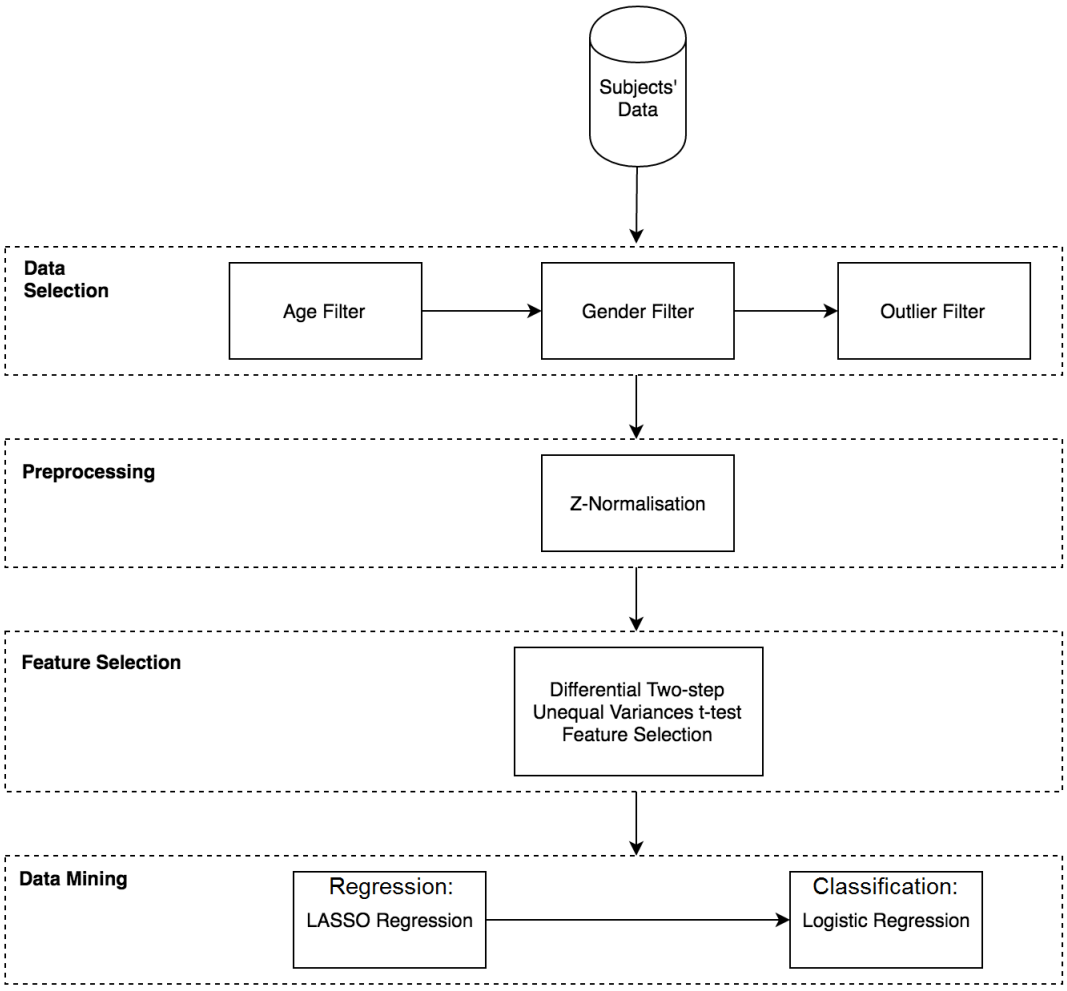

Figure 9.1: The workflow to estimate age and use it for classification

9.3.1 Data Selection

This part of the workflow involves selecting patients to be included in the generation of the classification model. It removes subjects who are likely to hinder the accuracy of the model.

Age Filter

Grouping the data by their cohort (IXI or ADNI), Figure 9.2 compares the distributions of the age of the subjects between these cohorts. As shown, the IXI cohort contains many younger subjects than the ADNI cohort: the ADNI cohort - at least the ADNI subject data

selected for this workflow - only contains subjects of 55 years or older and the IXI cohort contains subjects of 20 years and above. For the solution proposed in this chapter it was decided to investigate how inclusion of patients above a certain age would affect the results.

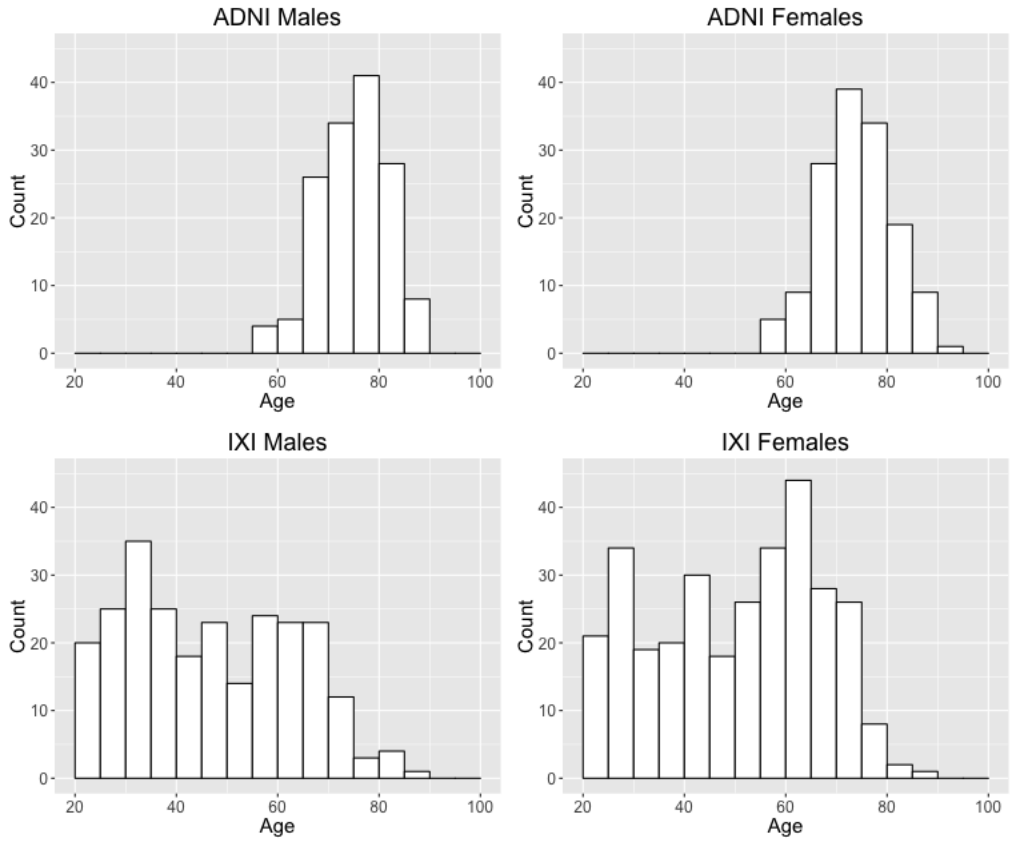

Figure 9.2: Histograms showing the distribution of the subjects ages across the datasets

Gender Filter

A separate regression model and classification model will be built for each gender, thus there will be two accuracy results: one for males and one for females. The reasoning behind this is due to the differences in the morphometry of male and female brains and this may have an adverse effect on the regression model being created to predict a subject's apparent brain age.

Outlier Filter

In this step outliers are filtered as their inclusion will result in a worse performing classification model. In a statistical population an outlier is defined as an observation that differs significantly enough from the other members in the population such that there can be doubt that it has been generated in the same way as the rest of the population (Hawkins, 1980). While outliers could have been created from an error in the recording of the value; it could also be the case that it is just a deviation from the normal variability in the population (Grubbs, 1969).

Due to the high dimensionality of the data generated by Freesurfer, outlier detection is difficult; using Principal Component Analysis (PCA) the data will be transformed into a lower dimensional space retaining the majority of the information contained in the original data. PCA is a technique to reduce the number of variables in the data and detect structure in the relationships between the variables. PCA computes the eigenvectors and eigenvalues of the covariance matrix of the mean adjusted data. Each eigenvector represents a principal component and the associated eigenvalue represents how much information of the original data is represented by the eigenvector with larger eigenvalues meaning that the eigenvector contains more information. When PCA is applied to the data we must define the number of principal components that are going to be kept, two methods were investigated in an attempt to find out the optimal number to keep: the Kaiser criterion and the scree test.

The Kaiser criterion (Kaiser, 1958) selects those principal components which have eigenvalues greater than 1; principal components whose eigenvalues are less than 1 are discarded. The scree test is a graphical method proposed by Cattell (1966) where the eigenvalues of the principal components are sorted from highest to lowest and then plotted on a graph with the x-axis being the index of the eigenvalue and the y-axis being the eigenvalue itself. Cattell suggests locating the part of the graph where the decrease of the eigenvalues begins to plateau. Both of these methods will be analysed to see how they differ in the number of principal components selected.

Data observations can be classed as high leverage points if their independent variable is significantly different from the independent variables of other observations. These high leverage points adversely affect the regression model especially when non-robust methods are used such as least squares.

The Mahalanobis distance is a common method used to identify outliers in multivariate data. Given a p-dimensional multivariate observation $\overline{x_i}$, and the multivariate means of the population $\overline{\mu}$, the Mahalanobis distance can be computed as:

$$MD_i = \sqrt{(\overline{x_i} - \overline{\mu})^T \Sigma^{-1} (\overline{x_i} - \overline{\mu})}, \tag{9.1}$$

where Σ is the covariance matrix of the data. Once the data is transformed using PCA into a lower number of dimensions (how this number is determined is explained in Section 9.6.2), then the MD is computed for each data observation resulting in the Mahalanobis distances which are denoted as $\overline{d}$. The set of outliers O is defined as:

$$O = \{x_i | \overline{d} > \mu + 3\sigma\}, \tag{9.2}$$

where μ and σ are the mean and standard deviation of $\overline{d}$ respectively. This is based on the three-sigma edit rule for outlier detection (Maronna et al., 2006). Every observation that is included in O is marked as an outlier and removed from the data set.

9.3.2 Differential Two-step Unequal Variances t-test Feature Selection

A custom feature selection method is introduced with the aim of identifying statistically different features between two cohorts of subjects. The method is applied twice (two steps). First, it is used to eliminate the features that are statistically different (specifically distributed features) between the two cohorts of HC subjects. Differences may be induced because of unknown, non controllable factors, e.g., the differences in the protocols used in data

acquisition for IXI and ADNI datasets, and a bias induced by the limited size of the cohorts. Second, the method is then used in the opposite way to identify and retain the specifically distributed features between AD and HC subjects in the ADNI cohort, emphasising the discriminative ability of the feature set. In particular, the Welch's unequal variances t-test (Welch, 1947) has been adopted in both steps. Welch's t-test is a variant of the Student's t-test that is more reliable when the two samples have unequal variances and unequal sample sizes.

Removing dissimilar features between HC subjects of different cohorts

There is a concern that since ADNI subjects and IXI subjects have their data generated using different protocols, then some features may have artificially different distributions for the two cohorts. To find and remove these features, a t-test is used and the resultant p-value is used to determine whether the attributes of the subjects between the two cohorts are equivalent or not, if the p-value is less than 0.05 then the attributes are considered to have different distributions. The version of the t-test used is the Welch unequal variance t-test, as shown in Equation 9.3, where μ_i is the mean of the region of interest of subjects in cohort i, σ_i^2 is the variance of region of interest of subjects in cohort i, and n_i is the number of subjects with the region of interest in cohort i which in this case is the number of subjects in each cohort.

$$t = (\mu_1 - \mu_2)\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)^{-\frac{1}{2}} \tag{9.3}$$

$$p = P(t \geq 2.311) \tag{9.4}$$

To eliminate bias from the t-test, 10 samples are used resulting in ten t-tests being performed for each feature. The number (out of these ten) of t-tests which compute the samples as being different will be denoted as the function $DIFF(f_i)$, where f_i is the feature being tested. A noise threshold T_N is used: if $DIFF(f_i) > T_N$ then the feature is regarded

as a noise and discarded. This process is run independently for male and female subjects, where t-tests are performed for each gender, and the final feature set output is the input feature set with the removal of the features classed as noise. Table 9.2 demonstrates how the number of output features changes with respect to T_N.

Table 9.2: How the noise threshold affects the number of features

T_N	Number of features for males	Number of features for females
−1	0	0
0	166	130
1	203	160
2	232	181
3	242	192
4	255	203
5	260	216
6	269	233
7	283	243
8	295	258
9	308	269
10	354	354

Retaining dissimilar features between classes

The feature set is further refined by selecting only the features which are useful in differentiating between HC and AD subjects. A similar threshold approach is used to retain dissimilar features between the classes; the new threshold used is the retain threshold T_R, it determines which features are kept. This method also uses ten t-tests with random samples to determine the set of features to keep. The output set of features are the ones determined to be dissimilar dependent on the class of the subjects. Table 9.3 demonstrates how the number of output features changes with respect to T_R.

Table 9.3: How the retain threshold affects the number of features

T_R	Number of features for males	Number of features for females
−1	354	354
0	182	172
1	158	155
2	138	145
3	131	135
4	126	130
5	113	123
6	99	116
7	94	104
8	89	93
9	75	84
10	0	0

9.4 Apparent Brain Age Model

9.4.1 LASSO Regression

The regression method used is LASSO regression due to the ability to handle data with a large number of features. It is described in detail in Section 4.5.2.

9.4.2 Choice of LASSO Regression Regularisation Parameter

To choose the regularisation parameter λ the following steps have been followed: first a list of candidate λ values, 100 logarithmically spaced values from 10^{-2} to 10^{10} (i.e. such that the exponents are linearly spaced between -2 and 10), has been created. The list can be defined formally as $\{10^n | n \in \{x_1, x_2 ... x_{100} | x_1 = -2, x_{100} = 10, x_i - x_{i-1} = \frac{12}{100}\}\}$. Next, a LASSO model is built for each of the values in the list, based on predicting the subject's age. Then each of the models are evaluated using cross validation on the dataset and the cross validation errors are calculated for each of them, and the value of λ which yields the

minimum cross validation error is chosen as value of the regularisation parameter.

9.4.3 AD Effect on Brain Age

LASSO regression can be applied to the features of the data set to predict the age of an HC subject. The hypothesis is that due to the degradation of volume in the brain from AD, when an AD subject's age is predicted by this model the apparent brain age will be older than the real age, whereas an HC subject's apparent age should be similar to the real age. This difference between the actual age and predicted age will be referred to as the age deviation score, where:

$$\text{Age Deviation Score} = \text{Apparent Brain Age} - \text{Real Age} \tag{9.5}$$

If the hypothesis is true then higher age deviation scores will correlate with AD subjects, and near zero age deviation scores will correlate with HC subjects.

9.5 Classification

9.5.1 Logistic Regression

Logistic regression is a classification model which can be used to find a decision boundary between two classes. The logistic function is defined as:

$$\sigma(t) = \frac{e^t}{e^t + 1} = \frac{1}{1 + e^{-t}} \tag{9.6}$$

The logistic function is used in logistic regression to estimate the probability of a binary class occurring based on one or more predictors. This probability can then be rounded to 0 or 1 to convert the probability estimate into a predicted class response. t can be expressed as a linear combination of the input features, thus the logistic function can be written as:

$$F(x) = \frac{\exp(\beta^T x)}{\exp(\beta^T x) + 1} = \frac{1}{1 + \exp(-\beta^T x)} \tag{9.7}$$

The above logistic function gives the probability of the success case, in this chapter, the success would be defined as a subject having AD. Thus based on a feature set x and a coefficient set β the probabilities of a subject being AD and HC can be defined as below:

$$P(Diagnosis = AD|X = x) = \frac{\exp(\beta^T x)}{1 + \exp(\beta^T x)} \tag{9.8}$$

$$P(Diagnosis = HC|X = x) = \frac{1}{1 + \exp(\beta^T x)} \tag{9.9}$$

From these equations we can then derive a decision boundary to classify the data:

$$\hat{D}(x) = \begin{cases} 0 & \text{if } -\beta x \geq 0 \\ 1 & \text{if } -\beta x < 0 \end{cases} \tag{9.10}$$

In this work we will apply logistic regression to two inputs (actual age and age deviation), thus: $\beta = \beta_0, \beta_1, \beta_2$ and $x = x_1, x_2$, the decision boundary can be represented by:

$$-\beta x = 0 \text{ therefore } -\beta_0 - \beta_1 x_1 - \beta_2 x_2 = 0 \tag{9.11}$$

The equation for the line of the decision boundary is given by:

$$x_2 = -\frac{\beta_1}{\beta_2} x_1 - \frac{\beta_0}{\beta_2} \tag{9.12}$$

After using the logistic regression each subject (provided it is not filtered as an outlier) will be classified as either HC or AD. From that, the number of correct predictions can be computed and we can find the accuracy of the model.

Examples of the data created by the logistic regression are shown in the following four graphs: the first set of two graphs shows males and females respectively, the subjects' ages

and their apparent brain ages, see Figures 9.3 and 9.5; the second set of two graphs shows males and females respectively, the subjects' ages and their age deviation scores, see Figures 9.4 and 9.6.

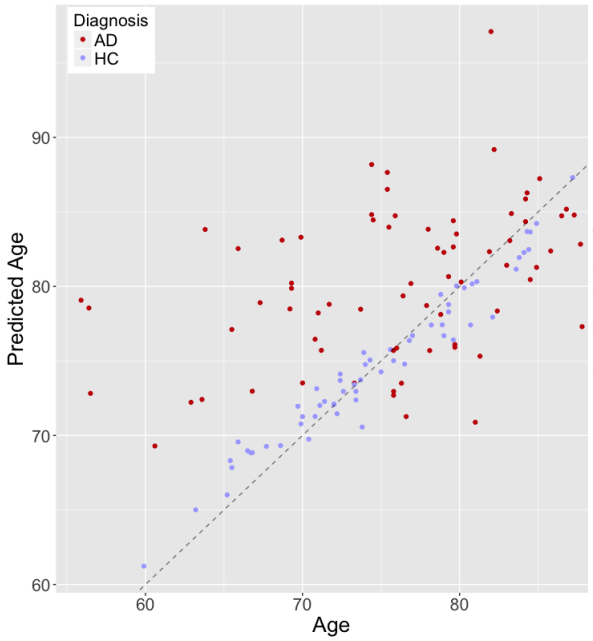

Figure 9.3: An example of Age vs. Predicted Age for male subjects. The dotted line represents $y = x$.

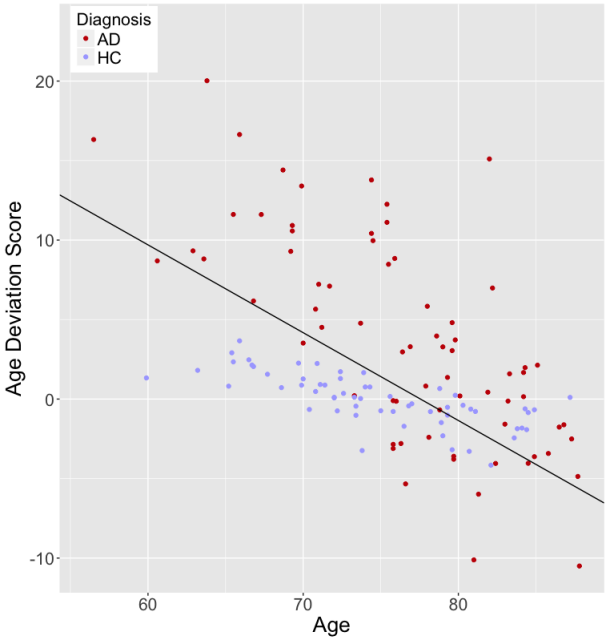

Figure 9.4: An example of Age vs. Age Deviation Score for male subjects. The line is the logistic regression decision boundary.

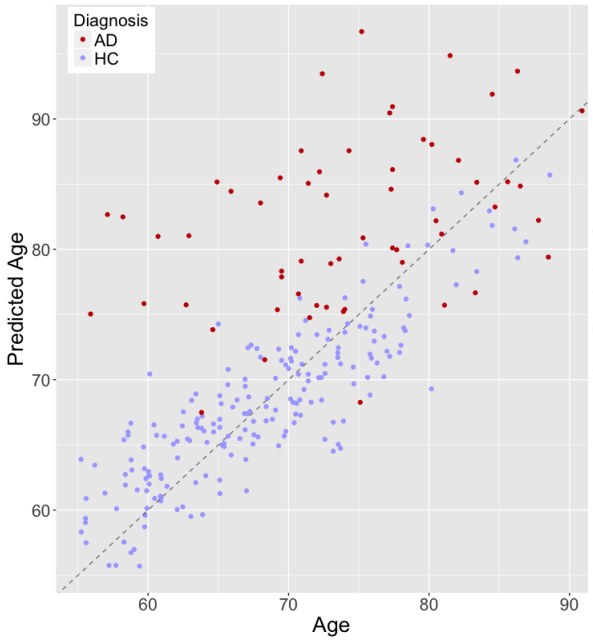

Figure 9.5: An example of Age vs. Predicted Age for female subjects. The dotted line represents $y = x$.

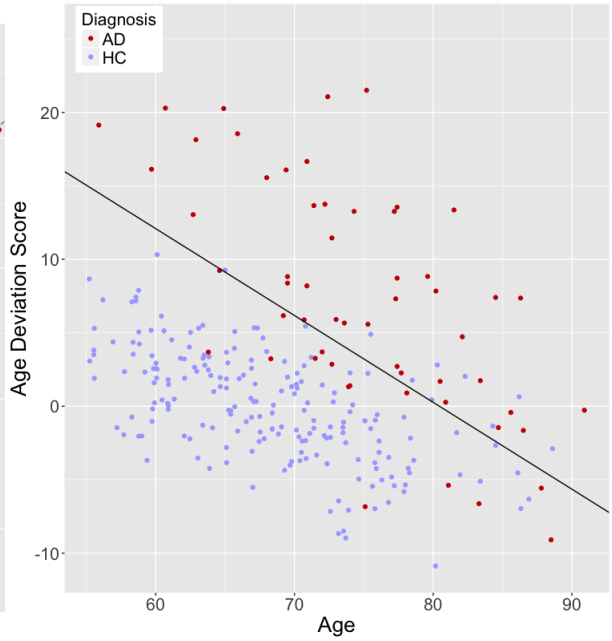

Figure 9.6: An example of Age vs. Age Deviation Score for female subjects. The line is the logistic regression decision boundary.

9.5.2 Age Deviation as an Additional Feature

To evaluate the predictive power of the apparent brain age: the age deviation score will be appended to the original data set and the accuracy evaluated. This approach was chosen in preference to a decision fusion as the performance of the brain age in summarising the data set will be evaluated, but by augmenting it to a feature set it can be compared to the performance of the entire feature set. As well as applying the domain knowledge feature selection which has shown positive results in previous research (Spedding et al., 2015). The process to accomplish this is to compute the predicted age and age deviation score as before, then append the age deviation score to the initial set of features (or a subset of the initial set of features). Then a linear SVM is built on training data to classify the diagnosis of the subjects given the initial features as well as the new age features; and then finally compute the accuracy of the linear SVM on the test data. This will use 10-fold cross validation repeated

ten times so it is comparable to previous results.

The new age features will be used to augment the entire set of features exported from Freesurfer to see if the classification accuracy is improved; then it will also be tested with just the 16 volumes of the hippocampus that Freesurfer exports, these 16 volumetric features are: the entire left hippocampus, the entire right hippocampus, the left and right presubiculum, the left and right subiculum, the left and right cornus ammonis (CA) 1, the left and right CA 2-3, the left and right dentate gyrus, the left and right fimbria, and the left and right hippocampal fissure.

9.6 Workflow Setup

This section discusses how some initial selection of the data is performed, the effect of the outlier removal and parameter selection for the feature selection and regression algorithms.

9.6.1 Experimental Setup

The programming language R was used along with the following packages to implement the algorithms: *glmnet* was used to implement the LASSO regression; *e1071* was used to implement the SVM; *glmnet* was also used to implement the Logistic regression. Freesurfer and rsurfer were used to preprocess and manipulate the data respectively.

9.6.2 Outlier Removal

There were a total of 6 outliers found in the data with 1 male outlier and 5 female outliers. These were all deemed to be non-classifiable for each classification method tested. Preliminary testing showed that removal of outliers was necessary for the apparent age estimation to work well. To show the decrease in performance the outliers have on the data, a linear SVM was applied to the data both with outliers included and with outliers excluded, the results of this are shown in Table 9.4, from these results outliers in females had a negligible effect

whether they were removed or not; and outliers in males once removed showing a very small increase the accuracy - though that is unexpected considering there are more outliers in the female data. These results were found using 10-fold cross validation repeated ten times. A scree plot is shown in Figure 9.7 where the highest 50 eigenvalues of the principal components are set out.

Table 9.4: Evaluating the effect of outliers on the linear SVM

Gender	**Outliers Removed?**	**Accuracy**
Male	F	88.5 ± 1.08
Female	F	89.0 ± 1.39
Male	T	90.4 ± 0.99
Female	T	88.6 ± 1.59

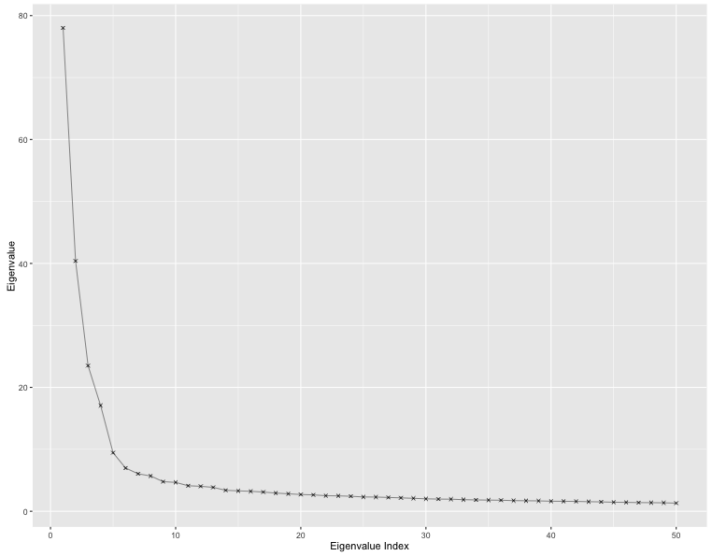

Figure 9.7: Scree plot of principal components on a sample of male subjects.

9.6.3 Age-based Subject Selection

A range of ages for an age filter were used. The results are in Table 9.5 for males and Table 9.6 for females. From both sets of results when the entirety of the subjects are used for the regression model the sensitivity is very low meaning that the classifier is worse at predicting HC subjects correctly. This may be due to the younger subjects from IXI all being HC leading

to bias making it unable to distinguish the older HC subjects from the younger HC subjects. From the various ages chosen to filter by, for females it appears to be 55 and males 50; for both genders accuracy remains fairly consistent, however the specificity is clearly affected by the chosen age to filter by. When this age is too low or too high then the specificity drops greatly. In both cases, precision and sensitivity only fall as more subjects are filtered showing that having younger subjects in the regression model does not affect these noticeably.

Table 9.5: The effect varying the age filter has on the results for males

Age Filter	**Gender**	**Accuracy**	**Sensitivity**	**Specificity**	**Precision**
>0	Male	89.4 ± 1.09	95.1 ± 1.00	65.3 ± 3.38	92.1 ± 0.69
>50	Male	90.9 ± 1.06	93.8 ± 1.15	84.5 ± 1.97	93.3 ± 0.80
>55	Male	89.7 ± 1.21	91.8 ± 1.38	85.3 ± 1.62	93.1 ± 0.72
>60	Male	88.0 ± 1.32	90.5 ± 1.95	83.4 ± 1.86	91.2 ± 1.04
>65	Male	86.5 ± 1.35	88.4 ± 2.33	83.4 ± 2.33	90.0 ± 1.38
>70	Male	81.2 ± 2.64	84.9 ± 3.76	76.9 ± 3.67	83.0 ± 2.90

Table 9.6: The effect varying the age filter has on the results for females

Age Filter	**Gender**	**Accuracy**	**Sensitivity**	**Specificity**	**Precision**
>0	Female	90.3 ± 0.92	95.9 ± 0.71	56.7 ± 3.62	93.1 ± 0.56
>50	Female	91.6 ± 1.09	95.5 ± 0.80	76.7 ± 4.56	94.2 ± 1.09
>55	Female	90.7 ± 0.74	95.3 ± 0.30	75.3 ± 2.79	93.1 ± 0.80
>60	Female	88.0 ± 1.15	93.3 ± 1.26	72.2 ± 2.61	91.4 ± 0.80
>65	Female	85.0 ± 1.71	92.7 ± 1.92	65.7 ± 4.42	87.6 ± 1.27
>70	Female	79.7 ± 2.05	87.7 ± 3.07	65.1 ± 4.17	83.1 ± 1.79

9.6.4 Feature Selection

To determine the optimal T_R and T_N for each gender, a grid search was performed with $-1 \leq T_R \leq 9$ and $0 \leq T_N \leq 10$ and for each pair of thresholds, the logistic regression was performed and with the training data resubstituted in as the testing data and the pair of thresholds which achieves the highest accuracy are determined to be the optimal thresholds and will be the thresholds used to select the features that the test data is classified with.

Another method was also tested working similarly to a holdout method: the t-test feature selection is performed on the entire training data, next the training data is partitioned into two smaller datasets with 70% of data in one of these datasets, and 30% of data into the other dataset. Then the LASSO and logistic regression models are built on the larger partitioned data set, and the accuracy is computed from the 30% of data which the models were not trained on. The pair of thresholds which attain the best accuracy are selected to be the optimal folds. This method was implemented in an attempt to eliminate bias from the first resubstitution method which may cause the optimal thresholds to be selected to overfit the data when the final model is applied to the test data. This method differs from a standard holdout method as here the t-test selection is performed on the entire data rather than each sample of 70% of the data; standard holdout was not implemented due to the length of time it would take to compute. Given no time constraints whatsoever, 10-fold cross validation on the training data would be used to select the optimal thresholds. Both of these methods will be tested to see which (if any) produces better results.

9.6.5 Age Regression Model

To show the descriptive ability of our model we can look at the coefficients generated by the LASSO model. In this case the thresholds are set such that they cause the LASSO model to produce the lowest MSE when it is applied to a dataset. For the males, a list of the coefficients is shown in Table 9.7; and for women they are in Table 9.8. The higher the absolute value of the coefficient, the more that feature affects the output of the model i.e. a change in the value of a feature with an absolute coefficient value of 10 would affect the predicted age more than a feature with an absolute coefficient value of 5. The sign of a coefficient relates to how that feature affects the predicted age; a feature with a positive coefficient means that the feature value is proportional to the subject's age: i.e. if the feature increases then the output predicted age is likely to be higher. The opposite is true for a negative coefficient, where if that feature increases then the predicted age will be lower. This means that the

coefficients can help to show why a decision was made and which features in a subject's brain are affected by the disease (or not in the case of an HC classification).

Table 9.7: Coefficients of the LASSO model generated for male patients which have an absolute value of greater than 0.5

Feature	**Coefficient**
(Intercept)	130.89
BrainSegVol to ICV Ratio‡	-76.72
Right Caudal Anterior Cingulate Thickness SD	9.59
Right Superior Frontal Thickness	-8.85
Right Superior Temporal Thickness	-7.22
Left Caudal Anterior Cingulate Thickness SD	6.79
Left Superior Frontal Thickness	-6.70
Left Caudal Middle Frontal Thickness	6.51
Left Fusiform Thickness	5.39
Left Insula Thickness SD	5.23
Left Rostral Middle Frontal Thickness	4.36
Right Superior Parietal Thickness	4.13
Left Inferior Temporal Thickness	-3.73
Left BankSSTS Thickness	2.40
Left Lateral Orbito Frontal Thickness	-2.23
Right Para Hippocampal Thickness	2.13
Right Temporal Pole Thickness	2.01
Left Middle Temporal Thickness	-1.30
Right Medial Orbito Frontal Thickness	-1.21
Left Temporal Pole Thickness	1.06
Right Middle Temporal Thickness	-0.762

Table 9.8: Coefficients of the LASSO model generated for female patients.

Feature	**Coefficient**
(Intercept)	119.28
Left Insula Thickness SD	13.51
Right Lingual Thickness SD	12.20
Right Paracentral Thickness SD	9.77
Left Superior Frontal Thickness	-9.45
Right Lateral Occipital Thickness	-3.98
Left Lateral Orbito Frontal Thickness	-3.92
Right BankSSTS Thickness	-3.83
Right Superior Frontal Thickness	-2.51
Right Middle Temporal Thickness	-2.36
Right Isthmus Cingulate Thickness	1.79
Right Temporal Pole Thickness SD	1.72
Left Parahippocampal Thickness	0.527

9.7 Classification Results

9.7.1 Evaluation Metrics

The five metrics used are: accuracy, sensitivity, specificity, precision and the NPV.

9.7.2 Results

Table 9.9 shows the progression of how the workflow was refined to produce a more optimal apparent age feature. The first results column, method ID 1, shows the absolute baseline result - an SVM applied on the ADNI data with no preprocessing. The aim of this was to create an apparent brain age feature which can achieve this accuracy or greater. The second results column, method ID 14, shows the baseline performance of the apparent brain age, this is the entire feature of the ADNI data being input into the LASSO regression, and the apparent brain age feature being used in conjunction with the actual age as input into a logistic regression classifier to predict the subjects' diagnosis. The result of this was that for all cases tried the accuracy was far lower than the baseline accuracy.

Table 9.9: The results demonstrating how features and data were added to the workflow to improve the accuracy of the apparent brain age feature.

Method ID			1	2	3	4	5
Apparent Age Used In Classificaton?			No	Yes	Yes	Yes	Yes
Data Used			ADNI	ADNI	ADNI	ADNI, IXI	ADNI, IXI
Feature Selection for Age Regression Model			-	-	Class-based	-	Class-based Cohort-based
Feature Selection for Classification			All Features	Real Age	Real Age	Real Age	Real Age
Classifier			SVM	Logistic Regression	Logistic Regression	Logistic Regression	Logistic Regression
No Class Balance	Male	Accuracy	86.5 ± 0.96	76.5 ± 1.86	84.0 ± 1.78	88.7 ± 0.63	90.9 ± 1.06
		Specificity	86.9 ± 1.69	72.7 ± 2.89	82.7 ± 2.75	91.9 ± 1.00	93.8 ± 1.15
		Sensitivity	86.2 ± 2.38	79.6 ± 3.18	85.3 ± 1.02	81.3 ± 2.54	84.5 ± 1.97
		NPV	87.6 ± 1.86	76.8 ± 2.82	84.2 ± 1.71	92.1 ± 0.99	93.3 ± 0.80
		Precision	87.3 ± 1.90	78.7 ± 2.09	86.3 ± 2.00	84.0 ± 1.28	86.9 ± 2.20
	Female	Accuracy	88.9 ± 2.28	73.4 ± 2.25	74.5 ± 1.69	89.3 ± 1.05	90.7 ± 0.74
		Specificity	89.5 ± 3.81	81.3 ± 2.07	80.1 ± 1.87	94.2 ± 1.07	95.3 ± 0.30
		Sensitivity	87.8 ± 2.36	63.9 ± 3.54	67.4 ± 2.34	72.7 ± 2.71	75.3 ± 2.79
		NPV	90.0 ± 2.63	74.3 ± 2.19	76.6 ± 1.91	92.4 ± 0.72	93.1 ± 0.80
		Precision	89.7 ± 3.11	75.7 ± 3.42	75.1 ± 2.79	80.8 ± 3.35	84.0 ± 1.68
Class Balance	Male	Accuracy	87.5 ± 1.37	77.6 ± 3.64	81.6 ± 2.16	83.3 ± 1.60	88.2 ± 1.25
		Specificity	89.8 ± 1.77	76.9 ± 5.92	81.9 ± 3.76	82.9 ± 2.30	90.1 ± 1.59
		Sensitivity	85.2 ± 1.83	78.2 ± 3.98	81.5 ± 2.09	83.6 ± 1.66	86.2 ± 2.01
		NPV	87.1 ± 1.38	78.6 ± 3.96	82.9 ± 2.45	84.8 ± 1.53	88.2 ± 2.10
		Precision	90.4 ± 1.43	78.6 ± 4.41	83.2 ± 3.37	84.2 ± 2.40	90.7 ± 1.65
	Female	Accuracy	88.3 ± 1.47	66.5 ± 2.45	72.1 ± 2.31	84.4 ± 2.88	88.9 ± 1.43
		Specificity	88.6 ± 2.14	67.9 ± 4.19	72.8 ± 3.07	87.3 ± 5.53	89.7 ± 1.19
		Sensitivity	87.9 ± 1.56	64.9 ± 2.82	71.5 ± 2.82	81.4 ± 3.58	87.8 ± 2.16
		NPV	88.9 ± 1.45	67.2 ± 2.24	73.5 ± 2.83	84.1 ± 2.29	89.4 ± 1.69
		Precision	89.6 ± 1.92	68.2 ± 3.66	73.9 ± 2.86	88.1 ± 5.25	90.6 ± 1.51

The next step to try and improve the predictive power of the apparent brain age feature was to use feature selection on the input features to the regression model in case the shrinkage methods of the model were not able to handle the large feature set. Using the t-test feature selection on the input of the regression model, the accuracy of the feature was increased but it was still lower than that of the baseline accuracy.

To further improve on the accuracy, more data was added, this time from the IXI cohort (which only contains HC subjects). Using this data in conjunction with the ADNI data increased the accuracy of the apparent age above the baseline. However, since the IXI cohort was shown to have certain features with different distributions of values compared to ADNI, there could have been the possibility this was affecting the result negatively and holding it back from its full potential. Thus another step was introduced into the t-test feature selection to handle this. The result of when this feature selection was applied further increases the accuracy of the apparent brain age. This shows that the age deviation score by itself is a powerful indicator of AD.

Other tests showing that the apparent brain age when augmenting the base feature set shows no improvement as expected as it is a summary of the features thus adds no predictive power.

Table 9.10 shows the performance of the hippocampal subfields selected as features, augmented with the apparent age feature. The most promising aspect of this work is that when the age score filter is used to augment the hippocampal subfields (or even the age score and age on their own as predictors with logistic regression) the accuracy this can achieve, is very close (or higher in some cases) to what the entire feature set can achieve. For males, the hippocampal subfields with the apparent brain age feature achieves 90.3 ± 1.15, whereas the entire feature set (with the apparent brain age feature) only reaches 86.5 ± 0.96. This shows that the difference between a subject's actual age and their predicted age can contain enough information to nearly have the same predictive power as over 350 features being used.

Table 9.10: The results demonstrating the predictive power of the apparent brain age when it is augmented with the hippocampal subfields

Method ID			6	7	8	9	10	11
Apparent Age Used In Classificaton?			No	Yes	Yes	No	Yes	Yes
Data Used			ADNI	ADNI	ADNI	ADNI, IXI	ADNI, IXI	ADNI, IXI
Feature Selection for Age Regression Model			N/A	None	Class-based	N/A	None	Class-based Cohort-based
Feature Selection for Classification			Hippocampus Subfields	Hippocampus Subfields	Hippocampus Subfields	Hippocampus Subfields	Hippocampus Subfields	Hippocampus Subfields
Classifier			SVM	SVM	SVM	SVM	SVM	SVM
No Class Balance	Male	Accuracy	81.7 ± 1.60	86.2 ± 1.80	87.3 ± 1.44	87.3 ± 0.52	90.3 ± 0.88	90.3 ± 1.15
		Specificity	83.1 ± 3.67	87.2 ± 3.80	89.1 ± 2.50	92.4 ± 0.59	92.8 ± 0.98	93.0 ± 0.83
		Sensitivity	80.1 ± 5.28	85.2 ± 4.54	85.8 ± 3.73	76.0 ± 1.55	84.7 ± 0.99	84.2 ± 2.44
		NPV	82.0 ± 3.96	87.4 ± 3.10	87.4 ± 2.72	89.9 ± 0.58	93.3 ± 0.43	93.2 ± 0.99
		Precision	83.9 ± 2.15	88.0 ± 2.76	89.7 ± 1.76	83.1 ± 1.56	85.4 ± 2.06	85.8 ± 1.53
	Female	Accuracy	79.1 ± 1.26	80.3 ± 1.28	82.6 ± 1.50	90.3 ± 0.58	91.4 ± 1.09	91.4 ± 0.79
		Specificity	79.5 ± 3.78	81.2 ± 2.21	82.1 ± 4.70	94.8 ± 0.58	95.7 ± 1.19	95.4 ± 0.62
		Sensitivity	78.0 ± 3.28	78.4 ± 2.89	82.1 ± 4.87	74.9 ± 1.70	77.0 ± 2.80	77.9 ± 2.02
		NPV	81.9 ± 2.88	81.7 ± 2.66	85.2 ± 3.30	93.0 ± 0.46	93.6 ± 0.80	93.8 ± 0.57
		Precision	79.0 ± 3.18	81.4 ± 1.57	82.8 ± 3.08	83.2 ± 1.76	85.2 ± 3.30	84.6 ± 2.42
Class Balance	Male	Accuracy	79.2 ± 1.71	88.4 ± 2.49	88.7 ± 1.99	83.8 ± 0.66	85.1 ± 2.10	88.7 ± 1.33
		Specificity	80.0 ± 2.04	88.8 ± 2.83	90.4 ± 2.76	84.7 ± 1.18	85.0 ± 2.86	90.6 ± 1.49
		Sensitivity	78.5 ± 2.81	88.1 ± 2.74	86.9 ± 2.43	82.9 ± 1.23	85.0 ± 2.01	86.9 ± 2.28
		NPV	80.5 ± 2.53	89.5 ± 2.48	89.0 ± 1.81	84.5 ± 1.00	86.4 ± 1.76	88.5 ± 1.94
		Precision	80.8 ± 1.91	90.0 ± 2.72	91.3 ± 2.43	85.8 ± 1.47	86.4 ± 2.25	90.9 ± 1.25
	Female	Accuracy	78.3 ± 1.89	78.4 ± 1.23	80.7 ± 1.92	82.1 ± 1.04	86.8 ± 2.20	87.6 ± 1.51
		Specificity	79.3 ± 3.23	79.8 ± 2.86	82.2 ± 2.81	86.4 ± 1.54	89.6 ± 4.01	90.6 ± 2.94
		Sensitivity	77.3 ± 1.52	77.1 ± 1.76	79.0 ± 2.22	77.6 ± 1.61	83.8 ± 2.68	84.3 ± 2.10
		NPV	79.7 ± 1.80	79.6 ± 1.38	81.9 ± 1.44	81.0 ± 1.39	86.3 ± 1.97	86.8 ± 1.43
		Precision	80.2 ± 3.34	80.2 ± 2.19	83.1 ± 2.80	86.7 ± 1.52	90.4 ± 3.33	91.0 ± 2.89

Overall Results Analysis

The results show that the addition of the IXI data to augment the ADNI data is always beneficial especially when the age feature is used which is likely to be due to the IXI data providing additional healthy subjects which the regression model can use to build the model of HC subjects. Even when the age feature is not used it still improves the classification

accuracy. The only downside of using the IXI data as well is that in certain cases - such as the SVM on male ADNI subjects - when the IXI data is added the specificity will drop, in the aforementioned case it drops from 86.2% to 84.9%. However, there are two points to note about this drop in specificity, the first is that it is significantly within the bounds of the standard deviation; the second is that when the data is class balanced then the sensitivity increases greatly. Therefore the drop in sensitivity could be caused by the large number of IXI HC subjects causing a bias to the classifier.

The t-test feature selection applied to the data is beneficial, especially when the age is used as the only predictor with the logistic regression, Figures 9.8 and 9.9 show the improvement the feature selection and addition of IXI can provide. When age is the only feature, the feature selection increases the accuracy by nearly 10% for males, but there is a negligible increase in the accuracy for females; with class balanced data there is an increase of 4% for the accuracy for males and nearly 6% for females. Similar accuracy increases occur when the t-test feature selection is used for the workflow applied to both ADNI and IXI data. When an SVM is used the feature selection also proves to be beneficial but not to the same extent, this may be due to the additional features in the data set meaning the classifier relies less on the age score feature.

The best results were all achieved by the SVM on both the ADNI and IXI data with the new age score feature: for non-class balanced males the highest was an accuracy of 92.4 ± 1.10 with the t-test feature selection used; for non-class balanced females it was an accuracy of 91.6 ± 1.08 without the t-test selection used (though this was only a 0.1% increase over when t-test selection was used so the effect is negligible here); for class balanced males an accuracy of 92.0 ± 0.90 is achieved by the SVM when both the ADNI and IXI data without the age feature or the feature selection; and for class balanced females the highest accuracy is 92.6 ± 1.49 with the SVM using ADNI and IXI data and the age feature - however there is fairly negligible difference in accuracy between this and whether no age feature and no feature selection is used.

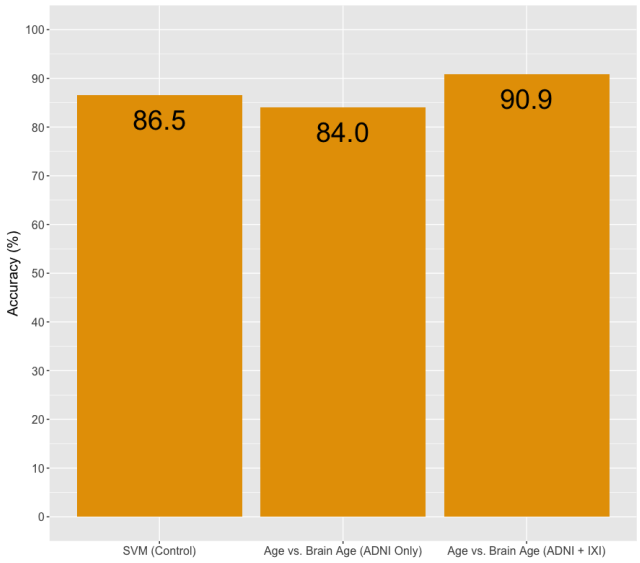

Figure 9.8: The classification results for non-class balanced males. "SVM (Control)" is the baseline result, the SVM applied to ADNI data. The "Age vs. Brain Age" is when logistic regression is applied between the real age and the brain age.

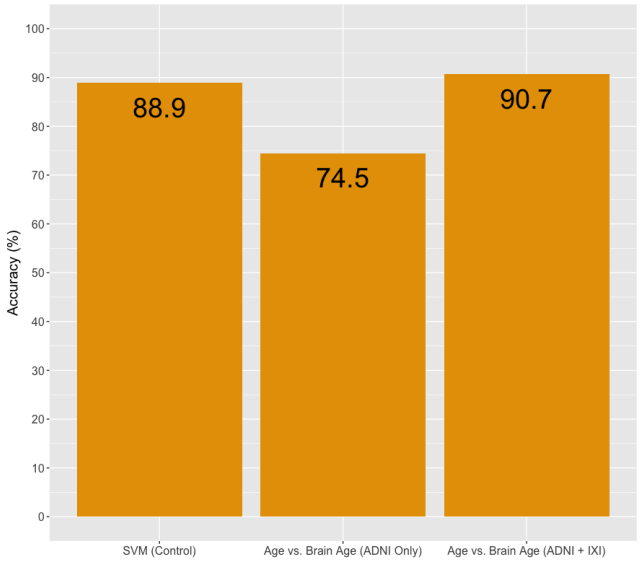

Figure 9.9: The classification results for non-class balanced females. "SVM (Control)" is the baseline result, the SVM applied to ADNI data. The "Age vs. Brain Age" is when logistic regression is applied between the real age and the brain age.

9.8 Conclusion

The workflow described in this chapter has demonstrated the ability of a custom feature selection method to handle feature selection when subjects are from multiple sources. It has also demonstrated that LASSO regression can be used on data generated by Freesurfer to be able to predict an apparent brain age for subjects and use this age to both classify a subject as either HC or AD and also provide descriptive information about a subject. The new apparent brain age feature generated has shown it is capable of reducing the number of features necessary for classification by summarising the information of many other features within it.

Chapter 10

Conclusion

The work presented in this thesis has demonstrated promising results in meeting the objectives of the thesis. Each objective defined in Section 1.3 will be analysed to measure how well the thesis met the objective.

"To identify and use some state-of-the-art and black-box approaches as baseline performance for the accuracy, such that newly designed predictive models have a baseline to be evaluated against."

Each body of work uses at least one state-of-the-art method to compare against. Chapter 7 uses an SVM for its baseline result. As this work focuses on feature selection rather than creation of a model, the SVM is analysed without any feature selection as a control to provide the baseline accuracy to evaluate against. Chapter 8 compares the novel predictive model against two state-of-the-art classifiers: an SVM and a Naive Bayes classifier. Chapter 9 has many methods of evaluation all of which are compared to an SVM as the baseline result, it is also analysed using an SVM with the generated apparent brain age feature.

The difficulty with meeting this objective stems from it being difficult to objectively compare results as each body of work in the literature uses a unique data set and also there is no

Table 10.1: Summary of the highest accuracies achieved by the work in this thesis

Method Name	**Genetic Algorithm + SVM**	**Probability-based Classifier**	**Apparent Brain Age**
HC vs. AD	89.3%	85.8%	91.4%
HC vs. MCI	71.4%	-	-
MCI vs. AD	82.1%	-	-
HC vs. MCI vs. AD	65.5%	-	-
Chapter	7	8	9

standard evaluation method - some papers use holdout methods, others use cross-validation. The accuracies achieved for the methods in this thesis are shown in Table 10.1 which are comparable with the methods reviewed in Chapter 5, some methods in the review achieve a higher accuracy, but they are operating on far smaller data sets. Regardless of difficulty of comparisons to the literature, the objective is fully met by various bodies of work as control experiments are performed within this thesis.

"To design a descriptive novel predictive model achieving a state-of-the-art accuracy."

Chapter 7 demonstrated a genetic algorithm being used for feature selection for an SVM. While the feature selection helped to narrow down the most important features, the SVM is still a black-box approach and does not help to show why a decision was made meaning that this chapter did not meet the descriptive aim. The next body of work in Chapter 8 was a probability-based classifier operating on the volumes of the hippocampus, this method classified individual volumes of the hippocampus as being HC-sized or AD-sized and then counted the occurrences of each type of region. This classifier can easily show why a decision was made and thus it meets the descriptive aim. In Chapter 9 an apparent brain age was computed for each subject giving an indication of the actual age of their brain, this is a descriptive indicator as you can determine that a classification of AD was made because the age of their brain appears to be a number of years older than their actual age. It is untested but this method could easily be expanded to measure the progression or severity of AD. Two out of the three of the bodies of work build descriptive classifiers meeting the aim to produce

classifiers with a descriptive ability.

"To investigate different techniques for feature selection to avoid the curse of dimensionality and to allow the predictive models to have a greater descriptive expressiveness."

This objective is met by various methods in the thesis using both domain knowledge for feature selection as well as data-driven feature selection. Both domain-based and data-driven feature selection are used in Chapter 7 where the data-driven feature selection is the usage of the GA, the domain-based feature selection is the division of the initial feature set into subsets based around the location of the brain the features belong to. Chapter 8 uses domain-based feature selection as the predictive model is solely based around the hippocampal volumes of an MRI. It also uses data-driven feature selection where the hippocampal fissures are excluded as they provide no information, as well as investigating the exclusion of features if they provide too little descriptive ability. Chapter 9 uses domain-based feature selection by generating the apparent brain age from a subset of features based on their location within the brain. Therefore this objective is met multiple times in methods in this work.

"To develop a software library to support and speed up the data manipulation tasks which are a fundamental requirement of every predictive model in this work."

Chapter 6 meets this objective by developing an open-source, documented R package aimed at simplifying and speeding up the manipulation of Freesurfer data within R. The functionality in this package could be replicated in other programming languages (such as Python) given time.

"To define the apparent brain age as the difference between the expected age of a brain assuming the brain is healthy and the actual age of a brain, and then use this apparent brain

age to provide a predictive model with and excellent descriptive information."

This objective is met by Chapter 9 which provides a descriptive method computing the apparent age of a brain and achieves a state-of-the-art accuracy. The method utilises the atrophy caused to a brain by AD to infer that this brain has aged the subject more than it would be if they did not suffer from AD. The way by which the method produces its diagnosis can easily be understood by a medical professional and this can be used to gain their trust in the method.

All objectives that were defined at the start of the thesis have been fully met, thus all the hypotheses have been proven and the research questions answered. Nonetheless the thesis still had one major difficulty which was sourcing enough MRI data. Given the lack of open data for AD MRIs combined with the large amount of time Freesurfer takes to process an MRI it becomes quite a task just to collect enough data. The National Health Service in the United Kingdom is a publicly funded national healthcare service, and while they are beginning to provide various data to be accessed, AD positive MRIs are not one of them. The author hopes that within the next couple of decades all the data the NHS collects will be anonymised and then made freely available to researchers with the aim to help cure or prevent the disease as this will be of great benefit to the research community.

10.1 Future Work

10.1.1 Improving the Probability-based Classifier

In Chapter 8 a probability-based classifier was developed with the application to distinguish between HC and AD subjects. It could be further developed so as to be able to apply to a three-class problem (HC vs. MCI vs. AD). One way of implementing this would be an additional threshold which could be used to indicate an MCI-sized region; or with the single threshold approach, the distance to the threshold could be computed and the sum of these

distances for each feature determines which of the three classes the subject belongs to. In the current implementation after the thresholds are created for each feature, regions are given the diagnosis of negative or positive which converts a continuous variable into a binary variable; thus on these binary variables, rule mining to generate rules to further the understanding of AD in the hippocampus could be performed. In Chapter 8 only the hippocampal subfields are used, the performance of the classifier when applied to the cortical and subcortical fields of the brain could improve so it is worth seeing how well it works applied to the additional features. Another idea is to use subspace clustering to find a subspace where the subjects can be grouped into clusters and see if rules can be deduced from the clusters that can be merged with the rules found by this classifier to create a better classifier.

10.1.2 Improving the Apparent Brain Age

It would be beneficial if the apparent brain age method was adapted to be able to classify MCI patients as well, logistic regression and the linear SVM could be used in a combined one versus all approaches to achieve this, but it is also likely that the feature selection method will need to be adapted to enable the selection of distinguishing features for MCI patients as well. It may also be worth furthering the descriptive ability of the coefficients, perhaps by integrating them with the segmented MRI by highlighting the regions which are causing the subject to have a greater predicted age than their actual age, this would mean a medical expert could see quickly which area of their patient's brain is being affected by the disease. The role of visualising the information like this will be very useful in comparative diagnosis based on the feature space of the MRI. In this workflow, the age filter excludes patients under the age of 65, it would be worth investigating methods to allow this workflow to predict the age for younger patients so we can further apply this to early onset of AD, this method will probably require more data specifically younger patients suffering with AD.

10.1.3 Biomarkers to Determine the Progression of Alzheimer's Disease

This work has focused on finding biomarkers based on their ability to predict AD, however work in Wan et al. (2012) found their biomarkers were also capable of determining the progression of AD. The apparent brain age in Chapter 9 could easily be extended to see whether it is able to detect the progression of AD. Additional data sources would also be required, ADNI provides multiple scans of subjects over time so this could be used to see if it could further progress AD results in an even higher apparent brain age than the baseline scan.

10.1.4 Validating Model Performance on Different Populations

Based on the problems of data-driven diagnosis in Obermeyer and Emanuel (2016) state that to overcome the problem of overfitting is to validate the models on different populations. In databases such as ADNI and IXI, which were used for the work in the thesis, subjects' MRIs are scanned from various hospitals using different MRI scanners. Using this we could class all subjects processed with a certain MRI scanner as a population, and validating models where it is trained on a number of populations and applied to other populations to test the performance. An examination of the state of the art showed that no research takes this into account, but it is concluded that it is vital as in the real world.

10.1.5 Investigating the Applicability of the Designed Methods to Other Neurodegenerative Diseases

Currently all the work performed in this thesis has been specifically looking at MCI and AD. Many other neurodegenerative diseases exist (e.g. Parkinson's, Huntington's and Motor Neurone Disease), thus the methods developed in this thesis could also be applied to these other diseases. One idea would be to use the apparent brain age concept for other diseases:

the difference between a healthy brain and a brain with another disease may end up being determined via the same regression model used in Chapter 9. Parkinson's disease is caused by the loss of nerve cells in the substantia nigra and basal ganglia, however, extracting these structures from T1 MRIs is challenging. Messina et al. (2011) show that the automated segmentation by Freesurfer shows no difference between the segmented results of HCs and those suffering with Parkinson's disease. Johnson et al. (2015) use Freesurfer to segment MRIs of patients with Huntington's disease and determines that the thickness of the occipital cortex in Huntington's disease subjects was smaller than HCs. This shows that Freesurfer is capable of detecting regions in the brain affected by Huntington's disease making it an ideal disease to apply the apparent brain age method to.

10.1.6 Quality Assurance of Freesurfer Generated Data

Due to the Freesurfer processing being automated, it can introduce errors by processing the data incorrectly, for instance it could be segmented incorrectly. Freesurfer encourage you to manually inspect the results of each reconstruction. The simplest method of error checking would be to look at the signal-to-noise ratio for the WM, this is an integer value with a higher value meaning that WM voxels will have a higher intensity relative to the background noise and thus appear brighter. Thus a low signal-to-noise ratio will mean it is more difficult to distinguish actual WM in the brain from the noise. To calculate the signal-to-noise ratio, Freesurfer provides a quality assurance (QA) tool script which will compute the signal-to-noise ratio as well as other checks such as the checking the segmentation for outliers and checking that the output order of files was correct.

These QA values could be used to filter out badly processed subjects, potentially classify them as AD due to the deformity of the brain (Wachinger et al., 2014). As well as being used to filter, they could also be used as features to aid the prediction, as potentially the atrophy of AD may deform the brain meaning the QA values will show the MRI was processed less well than normal; thus bad QA may correlate with AD.

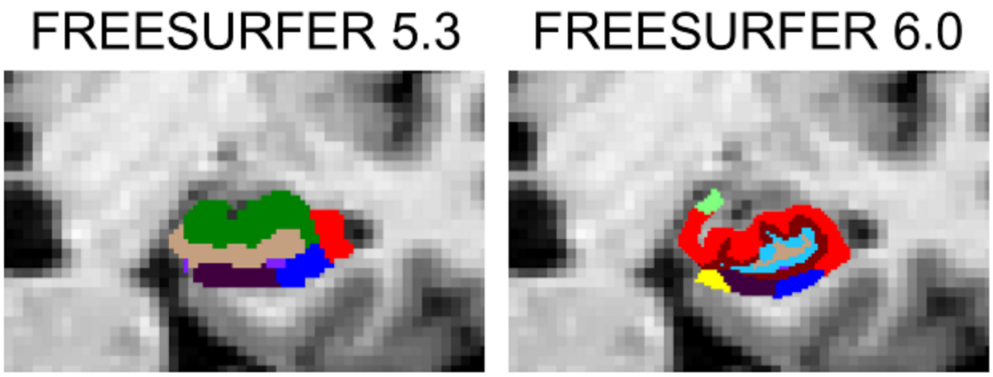

Figure 10.1: A comparison of the hippocampal segmentation of Freesurfer 5.3 and 6.0.
Source: https://surfer.nmr.mgh.harvard.edu/fswiki/HippocampalSubfields

10.1.7 Additional Functionality in Newer Freesurfer Versions

The MRIs used in this thesis were all processed with Freesurfer version 5.3, Freesurfer version 6 has been released and contains new functionality which would be relevant to the work in this thesis. The most relevant of the changes are a new hippocampal subfields generation method: the new version of Freesurfer uses a higher resolution statistical atlas to segment the hippocampal subfields which solve a number of problems with the previous atlas. These problems were: the image resolution of the MRIs used to build the atlas were at too low a resolution for humans to distinguish the subfields meaning that the segmentation operation relied on geometric data to find boundaries between the ROIs affecting the accuracy of the segmentations; the atlas lacked a molecular layer compromising the ability to segment high resolution data as this molecular layer is a key feature to describe the internal structure of the hippocampus; the previous atlas was lacking in its ability to segment the head and tail of the hippocampus. Figure 10.1 shows the difference between the segmentations of the hippocampus in the versions, and the new method to segment the hippocampus can be found in the work of Iglesias et al. (2015).

Appendix A

List of Publications

1. **A Genetic Algorithm for the selection of structural MRI features for classification of Mild Cognitive Impairment and Alzheimer's Disease** - Alexander Luke Spedding, Giuseppe di Fatta, Mario Cannataro, and the Alzheimer's Disease Neuroimaging Initiative. Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Workshop on Machine Learning in Decision Making for Biomedical Applications, Nov. 9-12, 2015, Washington D.C.

2. **An LDA and probability-based classifier for the diagnosis of Alzheimer's Disease from structural MRI** - Alexander Luke Spedding, Giuseppe di Fatta, James Douglas Saddy, and the Alzheimer's Disease Neuroimaging Initiative. Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Workshop on High Performance Bioinformatics and Biomedicine (HiBB), Nov. 9-12, 2015, Washington D.C.

Appendix B

Magnetic Resonance Image Weighting

The initial direction of the magnetic field of a proton is random until the magnetic field of the scanner is applied. Scans are created when the protons in the brain are brought into transverse coherence with each other by applying a short duration RF pulse along the x-axis. This RF causes the protons to resonant between alignment to the magnetic field and alignment perpendicular to the magnetic field. T1 is computed based on the time it takes for the magnetisation in the y-axis to decay to zero (or for the magnetisation on the z-axis M_Z to return to its original amount $M_{0,Z}$).

While the protons are resonating, they will exchange their energy with each other (known as the spin-spin interaction), these spins spread out across the xy-plane, and T2 is computed based on the time taken for the magnetisation on the xy-plane (M_{XY}) to return to its original value $M_{0,XY}$.

T1 and T2 can be mathematically defined using Bloch equations (Bloch, 1946), where TR is the length of time between each RF pulse and TE is the echo time which is the time from the centre of the RF pulse signal to the centre of the echo. For T1, TR must be short and TE must be as small as possible; whereas for T2, TR and TE must be longer. The magnetisation of a brain tissue is computed by:

$$\overline{M} = \overline{M_0}\left(1 - \exp\left(\frac{-TR}{T1}\right)\right)\exp\left(\frac{-TE}{T2}\right) \tag{B.1}$$

Where M is the net magnetisation of the proton (a three-dimensional vector), and M_0 is the initial magnetisation of the proton. T1 is computed by looking at the z component of the magnetisation to return to its initial state, and since $TE \approx 0$ then it can be defined using:

$$M_Z = M_{0,Z}\left(1 - \exp\left(\frac{-TR}{T1}\right)\right) \tag{B.2}$$

If we set TR to $T1$ then we can simplify the above to:

$$M_Z = \left(1 - \frac{1}{e}\right)M_{0,Z} \tag{B.3}$$

Showing that T1 measures the time taken for the z component of $\overline{M}$ to reach around 63% of its initial value. T2 can be similarity computed, since $TR \approx 0$, and T2 measures the xy-component of $\overline{M}$:

$$M_{XY} = M_{0,XY}\exp\left(\frac{-TE}{T2}\right), \tag{B.4}$$

setting the TE to $T2$ we get:

$$M_{XY} = \frac{1}{e}M_{0,XY}, \tag{B.5}$$

showing that T2 measures the time taken for the xy-component of $\overline{M}$ to reach roughly 37% of its initial value.

Bibliography

G. I. Allen, N. Amoroso, C. Anghel, V. Balagurusamy, C. J. Bare, D. Beaton, R. Bellotti, D. A. Bennett, K. L. Boehme, P. C. Boutros, and et al. Crowdsourced estimation of cognitive decline and resilience in alzheimer's disease. *Alzheimer's & Dementia*, 12(6): 645–653, 2016.

A. Alzheimer. Uber eine eigenartige erkrankung der hirnrinde. *Allgemeine Zeitschrife Psychiatrie*, 64:146–148, 1907.

American Psychiatric Association. Dsm-iv-tr: Diagnostic and statistical manual of mental disorders, text revision. *Washington, DC: American Psychiatric Association*, 75:78–85, 2000.

J. E. Baker. Reducing bias and inefficiency in the selection algorithm. In *Proceedings of the Second International Conference on Genetic Algorithms on Genetic Algorithms and Their Application*, pages 14–21, Hillsdale, NJ, USA, 1987. L. Erlbaum Associates Inc. ISBN 0-8058-0158-8. URL `http://dl.acm.org/citation.cfm?id=42512.42515`.

M. J. Ball, V. Hachinski, A. Fox, A. J. Kirshen, M. Fisman, W. Blume, V. A. Kral, H. Fox, and H. Merskey. A New Definition of Alzheimer's Disease: A Hippocampal Dementia. *The Lancet*, 325(8419):14 – 16, 1985. ISSN 0140-6736. doi: http://dx.doi.org/10.1016/S0140-6736(85)90965-1. URL `http://www.sciencedirect.com/science/article/pii/S0140673685909651`. Originally published as Volume 1, Issue 8419.

J. Barnes, G. R. Ridgway, J. Bartlett, S. M. D. Henley, M. Lehmann, N. Hobbs, M. J.

Clarkson, D. G. MacManus, S. Ourselin, and N. C. Fox. Head size, age and gender adjustment in MRI studies: a necessary nuisance? *NeuroImage*, 53(4):1244 – 1255, 2010. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2010.06.025. URL `http://www.sciencedirect.com/science/article/pii/S1053811910008694`.

C. F. Beckmann. Modelling with independent components. *Neuroimage*, 62(2):891–901, 2012.

A. L. M. Bergem, K. Engedal, and E. Kringlen. The role of heredity in late-onset alzheimer disease and vascular dementia: a twin study. *Archives of General Psychiatry*, 54(3):264–270, 1997.

M. R. Berthold, N. Cebron, F. Dill, T. R. Gabriel, T. Kötter, T. Meinl, P. Ohl, C. Sieb, K. Thiel, and B. Wiswedel. KNIME: The Konstanz Information Miner. In Christine Preisach, Hans Burkhardt, Lars Schmidt-Thieme, and Reinhold Decker, editors, *Data Analysis, Machine Learning and Applications*, Studies in Classification, Data Analysis, and Knowledge Organization, pages 319–326. Springer Berlin Heidelberg, 2008. ISBN 978-3-540-78239-1. doi: 10.1007/978-3-540-78246-9_38. URL `http://dx.doi.org/10.1007/978-3-540-78246-9_38`.

S. Beucher and F. Meyer. The morphological approach to segmentation: the watershed transformation. *OPTICAL ENGINEERING-NEW YORK-MARCEL DEKKER INCORPORATED-*, 34:433–433, 1992.

D. Le Bihan. Looking into the functional architecture of the brain with diffusion MRI. *Nature Reviews Neuroscience*, 4(6):469–480, 2003.

F. Bloch. Nuclear induction. *Physical review*, 70(7-8):460, 1946.

M. W. Bondi, A. U. Monsch, N. Butters, D. P. Salmon, and J. S. Paulsen. Utility of a modified version of the wisconsin card sorting test in the detection of dementia of the alzheimer type. *The Clinical Neuropsychologist*, 7(2):161–170, 1993.

C. Bordier, M. Dojat, and P. L. de Micheaux. Temporal and spatial independent component analysis for fMRI data sets embedded in the AnalyzeFMRI R package. *Journal of Statistical Software*, 44(9):1–24, 2011.

R. G. Boyes, J. L. Gunter, C. Frost, A. L. Janke, T. Yeatman, D. L. G. Hill, M. A. Bernstein, P. M. Thompson, M. W. Weiner, N. Schuff, G. E. Alexander, R. J. Killiany, C. DeCarli, C. R. Jack, and N. C. Fox. Intensity non-uniformity correction using n3 on 3-t scanners with multichannel phased array coils. *Neuroimage*, 39(4):1752–1762, 2008.

S. P. Brooks and B. J. T. Morgan. Optimization using simulated annealing. *The Statistician*, pages 241–257, 1995.

R. L. Buckner, D. Head, J. Parker, A. F. Fotenos, D. Marcus, J. C. Morris, and A. Z. Snyder. A unified approach for morphometric and functional data analysis in young, old, and demented adults using automated atlas-based head size normalization: reliability and validation against manual measurement of total intracranial volume. *Neuroimage*, 23(2): 724–738, 2004.

A. Burns and S. Iliffe. Alzheimer's disease. *BMJ*, 338, 2009. ISSN 0959-8138. doi: 10.1136/bmj.b158.

G. A. Carlesimo, F. Piras, M. D. Orfei, M. Iorio, C. Caltagirone, and G. Spalletta. Atrophy of presubiculum and subiculum is the earliest hippocampal anatomical marker of alzheimer's disease. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 1(1):24 – 32, 2015. ISSN 2352-8729. doi: http://dx.doi.org/10.1016/j.dadm.2014.12.001. URL `http://www.sciencedirect.com/science/article/pii/S2352872915000044`.

R. B. Cattell. The scree test for the number of factors. *Multivariate behavioral research*, 1 (2):245–276, 1966.

L. L. Chao, S. T. Buckley, J. Kornak, N. Schuff, C. Madison, K. Yaffe, B. L. Miller, J. H.

Kramer, and M. W. Weiner. ASL perfusion MRI predicts cognitive decline and conversion from MCI to dementia. *Alzheimer Dis Assoc Disord*, 24(1):19–27, 2010.

D. Charles, M. Furukawa, and M. Hufstader. Electronic health record systems and intent to attest to meaningful use among non-federal acute care hospitals in the united states: 2008–2011. *ONC Data Brief*, 1:1–7, 2012.

J. Chepkoech, K. B. Walhovd, H. Grydeland, and A. M. Fjell. Effects of change in freesurfer version on classification accuracy of patients with alzheimer's disease and mild cognitive impairment. *Human Brain Mapping*, 2016.

L. P. Clarke, R. P. Velthuizen, M. A. Camacho, J. J. Heine, M. Vaidyanathan, L. O. Hall, R. W. Thatcher, and M. L. Silbiger. MRI segmentation: methods and applications. *Magnetic resonance imaging*, 13(3):343–368, 1995.

J. Clayden. Rniftyreg: Medical image registration using the niftyreg library. *R package version 0.3*, 1, 2011.

E. H. Corder, A. M. Saunders, W. J. Strittmatter, D. E. Schmechel, P. C. Gaskell, G. Small, A. D. Roses, J. L. Haines, and M. A. Pericak-Vance. Gene dose of apolipoprotein e type 4 allele and the risk of alzheimers disease in late onset families. *Science*, 261(5123):921–923, 1993.

M. Cottle, W. Hoover, S. Kanwal, M. Kohn, T. Strome, and N. Treister. Transforming health care through big data strategies for leveraging big data in the health care industry. *Institute for Health Technology Transformation, http://ihealthtran. com/big-data-in-healthcare*, 2013.

R. Cuingnet, E. Gerardin, J. Tessieras, G. Auzias, S. Lehéricy, M. Habert, M. Chupin, H. Benali, and O. Colliot. Automatic classification of patients with Alzheimer's disease from structural MRI: A comparison of ten methods using the ADNI database. *NeuroImage*, 56(2):766 – 781, 2011. ISSN 1053-8119. doi: http://dx.doi.org/10.1016/j.

neuroimage.2010.06.013. URL http://www.sciencedirect.com/science/article/pii/S1053811910008578. Multivariate Decoding and Brain Reading.

A. M. Dale, B. Fischl, and M. I. Sereno. Cortical surface-based analysis. I. Segmentation and surface reconstruction. *Neuroimage*, 9(2):179–194, Feb 1999.

C. Darwin. On the origin of species by means of natural selection: or, the preservation of favored races in the struggle for life. 1859.

C. V. Dolph, M. D. Samad, and K. M. Iftekharuddin. Classification of alzheimer's disease using structural MRI. *Proc MICCAI Workshop Challenge on Computer-Aided Diagnosis of Dementia Based on Structural MRI Data*, pages 31–37, 2014.

S. Duchesne, A. Caroli, C. Geroldi, C. Barillot, G. B. Frisoni, and D. L. Collins. MRI-based automated computer classification of probable AD versus normal controls. *IEEE Trans Med Imaging*, 27(4):509–520, Apr 2008.

B. Englund, A. Brun, L. Gustafson, U. Passant, D. Mann, D. Neary, and J. S. Snowden. Clinical and neuropathological criteria for frontotemporal dementia. *J Neurol Neurosurg Psychiatry*, 57(4):416–8, 1994.

A. C. Evans, D. L. Collins, S. R. Mills, E. D. Brown, R. L. Kelly, and T. M. Peters. 3d statistical neuroanatomical models from 305 MRI volumes. In *Nuclear Science Symposium and Medical Imaging Conference, 1993., 1993 IEEE Conference Record.*, pages 1813–1817. IEEE, 1993.

A. C. Evans, M. Kamber, D. L. Collins, and D. MacDonald. An MRI-based probabilistic atlas of neuroanatomy. In *Magnetic resonance scanning and epilepsy*, pages 263–274. Springer, 1994.

Y. Fan, S. M. Resnick, X. Wu, and C. Davatzikos. Structural and functional biomarkers of prodromal Alzheimer's disease: a high-dimensional pattern classification study. *Neuroimage*, 41(2):277–285, 2008.

U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37, 1996.

H. H. Feldman. *Atlas of Alzheimer's Disease*. Informa Healthcare, 2007. ISBN 0415390451.

D. Feng and L. Tierney. mritc: A package for mri tissue classification. *Journal of Statistical Software*, 44(1):1–20, 2011.

B. Fischl, M. I. Sereno, and A. M. Dale. Cortical surface-based analysis: II: inflation, flattening, and a surface-based coordinate system. *Neuroimage*, 9(2):195–207, 1999.

B. Fischl, D. H. Salat, E. Busa, M. Albert, M. Dieterich, C. Haselgrove, A. van der Kouwe, R. Killiany, D. Kennedy, S. Klaveness, A. Montillo, N. Makris, B. Rosen, and A. M. Dale. Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain. *Neuron*, 33(3):341 – 355, 2002a. ISSN 0896-6273. doi: http://dx.doi.org/10.1016/S0896-6273(02)00569-X. URL `http://www.sciencedirect.com/science/article/pii/S089662730200569X`.

B. Fischl, D. H. Salat, E. Busa, M. Albert, M. Dieterich, C. Haselgrove, A. van der Kouwe, R. Killiany, D. Kennedy, S. Klaveness, A. Montillo, N. Makris, B. Rosen, and A. M. Dale. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron*, 33(3):341–355, 2002b.

B. Fischl, A. van der Kouwe, C. Destrieux, E. Halgren, F. Ségonne, D. H. Salat, E. Busa, L. J. Seidman, J. Goldstein, D. Kennedy, V. Caviness, N. Makris, B. Rosen, and A. M. Dale. Automatically parcellating the human cerebral cortex. *Cerebral cortex*, 14(1):11–22, 2004.

R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188, 1936.

M. F. Folstein, S. E. Folstein, and P. R. McHugh. mini-mental state: a practical method for grading the cognitive state of patients for the clinician. *Journal of psychiatric research*, 12 (3):189–198, 1975.

M. Galar, A. Fernández, E. Barrenechea, H. Bustince, and F. Herrera. An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes. *Pattern Recognition*, 44(8):1761 – 1776, 2011. doi: http://doi:10.1016/j.patcog.2011.01.017. URL `http://www.sciencedirect.com/science/article/pii/S0031320311000458`.

M. Goedert and M. G. Spillantini. A century of alzheimer's disease. *science*, 314(5800): 777–781, 2006.

J. J. Grefenstette. Optimization of control parameters for genetic algorithms. *IEEE Transactions on systems, man, and cybernetics*, 16(1):122–128, 1986.

E. H. B. M. Gronenschild, P. Habets, H. I. L. Jacobs, R. Mengelers, N. Rozendaal, J. van Os, and M. Marcelis. The Effects of FreeSurfer Version, Workstation Type, and Macintosh Operating System Version on Anatomical Volume and Cortical Thickness Measurements. *PLoS ONE*, 7(6):e38234, 06 2012. doi: 10.1371/journal.pone.0038234. URL `http://dx.doi.org/10.1371%2Fjournal.pone.0038234`.

F. E. Grubbs. Procedures for detecting outlying observations in samples. *Technometrics*, 11 (1):1–21, 1969.

M. Grundman, R. C. Petersen, S. H. Ferris, R. G. Thomas, P. S. Aisen, D. A. Bennett, N. L. Foster, C. R. Jack Jr, D. R. Galasko, R. Doody, J. Kaye, M. Sano, R. Mohs, S. Gauthier, H. T. Kim, S. Jin, A. N. Schultz, K. Schafer, R. Mulnard, C. H. van Dyck, J. Mintzer, E. Y. Zamrini, D. Cahn-Weiner, L. J. Thal, and the Alzheimer's Disease Cooperative Study. Mild cognitive impairment can be distinguished from alzheimer disease and normal aging

for clinical trials. *Archives of Neurology*, 61(1):59–66, 2004. doi: 10.1001/archneur.61.1.59. URL http://dx.doi.org/10.1001/archneur.61.1.59.

R. J. Haier, R. E. Jung, R. A. Yeo, K. Head, and M. T. Alkire. The neuroanatomy of general intelligence: sex matters. *NeuroImage*, 25(1):320–327, 2005.

T. J. Haight, W. J. Jagust, and the Alzheimers Disease Neuroimaging Initiative. Relative contributions of biomarkers in alzheimers disease. *Annals of epidemiology*, 22(12):868–875, 2012.

H. Hampel, A. Mitchell, K. Blennow, R. A. Frank, S. Brettschneider, L. Weller, and H. J. Möller. Core biological marker candidates of alzheimers disease–perspectives for diagnosis, prediction of outcome and reflection of biological activity. *Journal of neural transmission*, 111(3):247–272, 2004.

J. C. Hanson and C. F. Lippa. Lewy body dementia. *International review of neurobiology*, 84:215–228, 2009.

D. M. Hawkins. *Identification of outliers*, volume 11. Springer, 1980.

W. J. Henneman, J. D. Sluimer, J. Barnes, W. M. van der Flier, I. C. Sluimer, N. C. Fox, P. Scheltens, H. Vrenken, and F. Barkhof. Hippocampal atrophy rates in Alzheimer disease: added value over whole brain volume measures. *Neurology*, 72(11):999–1007, Mar 2009.

R. R. Hocking. A biometrics invited paper. the analysis and selection of variables in linear regression. *Biometrics*, pages 1–49, 1976.

J. H. Holland. *Adaptation in Natural and Artificial Systems*. MIT Press, Cambridge, MA, USA, 1992. ISBN 0-262-58111-6.

W. Hsu, M. Lee, and J. Zhang. Image Mining: Trends and Developments. *Journal of Intelligent Information Systems*, 19(1):7–23, 2002. ISSN 0925-9902. doi: 10.1023/A:1015508302797. URL http://dx.doi.org/10.1023/A\%3A1015508302797.

C. A. Hunter, N. Y. Kirson, U. Desai, A. K. G. Cummings, D. E. Faries, and H. G. Birnbaum. Medical costs of Alzheimer's disease misdiagnosis among US Medicare beneficiaries. *Alzheimer's & dementia: the journal of the Alzheimer's Association*, 11(8):887–895, 2015.

M. A. Iftikhar and A. Idris. An ensemble classification approach for automated diagnosis of alzheimer's disease and mild cognitive impairment. In *Open Source Systems & Technologies (ICOSST), 2016 International Conference on*, pages 78–83. IEEE, 2016.

J. E. Iglesias, J. C. Augustinack, K. Nguyen, C. M. Player, A. Player, M. Wright, N. Roy, M. P. Frosch, A. C. McKee, L. L. Wald, et al. A computational atlas of the hippocampal formation using ex vivo, ultra-high resolution MRI: application to adaptive segmentation of in vivo MRI. *NeuroImage*, 115:117–137, 2015.

K. Jin, A. L. Peel, X. O. Mao, L. Xie, B. A. Cottrell, D. C. Henshall, and D. A. Greenberg. Increased hippocampal neurogenesis in Alzheimer's disease. *Proceedings of the National Academy of Sciences*, 101(1):343–347, 2004. doi: 10.1073/pnas.2634794100. URL `http://www.pnas.org/content/101/1/343.abstract`.

E. B. Johnson, E. M. Rees, I. Labuschagne, A. Durr, B. R. Leavitt, R. A. C. Roos, R. Reilmann, H. Johnson, N. Z. Hobbs, D. R. Langbehn, et al. The impact of occipital lobe cortical thickness on cognitive task performance: An investigation in Huntington's Disease. *Neuropsychologia*, 79:138–146, 2015.

P. Johnson, L. Vandewater, W. Wilson, P. Maruff, G. Savage, P. Graham, L. S. Macaulay, K. A. Ellis, C. Szoeke, R. N. Martins, C. C. Rowe, C. L. Masters, D. Ames, and P. Zhang. Genetic algorithm with logistic regression for prediction of progression to Alzheimer's disease. *BMC Bioinformatics*, 15(Suppl 16):S11(6):e38234, 12 2014. doi: 10.1371/journal.pone.0038234. URL `http://www.biomedcentral.com/1471-2105/15/S16/S11`.

H. F. Kaiser. The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23(3):187–200, 1958.

A. Karatzoglou, A. Smola, K. Hornik, and A. Zeileis. kernlab – an S4 package for kernel methods in R. *Journal of Statistical Software*, 11(9):1–20, 2004. URL `http://www.jstatsoft.org/v11/i09/`.

H. Klafki, M. Staufenbiel, J. Kornhuber, and J. Wiltfang. Therapeutic approaches to Alzheimer's disease. *Brain*, 129(11):2840–2855, 2006. ISSN 0006-8950. doi: 10.1093/brain/awl280.

S. Klöppel, C. M. Stonnington, C. Chu, B. Draganski, R. I. Scahill, J. D. Rohrer, N. C. Fox, C. R. Jack, J. Ashburner, and R. S. J. Frackowiak. Automatic classification of MR scans in Alzheimer's disease. *Brain*, 131(3):681–689, 2008. ISSN 0006-8950. doi: 10.1093/brain/awm319.

M. Kuo, T. Sahama, A. W. Kushniruk, E. M. Borycki, and D. K. Grunwell. Health big data analytics: current perspectives, challenges and potential solutions. *International Journal of Big Data Intelligence*, 1(1-2):114–126, 2014.

M. Larobina and L. Murino. Medical image file formats. *Journal of digital imaging*, 27(2): 200–206, 2014.

I. Lawrence and K. Lin. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, pages 255–268, 1989.

J. Lee, S. Yang, Y. Wai, J. Wang, W. Hsu, and J. Chien. Probability-based prediction model using multivariate and lvq-pnn for diagnosing dementia. *Neuropsychiatry*, 6(6), 2016.

R. Li, W. Zhang, H. Suk, L. Wang, J. Li, D. Shen, and S. Ji. Deep learning based imaging data completion for improved brain disease diagnosis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 305–312. Springer, 2014.

D. Liu, H. J. Johnson, J. D. Long, V. A. Magnotta, and J. S. Paulsen. The power-proportion method for intracranial volume correction in volumetric imaging analysis. *Frontiers in neuroscience*, 8, 2014.

N. K. Logothetis. What we can do and what we cannot do with fMRI. *Nature*, 453(7197): 869–878, 2008.

S. S. Long and L. B. Holder. Graph based mri brain scan classification and correlation discovery. In *Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2012 IEEE Symposium on*, pages 335–342, May 2012. doi: 10.1109/CIBCB.2012.6217249.

R. A. Maronna, R. D. Martin, and V. J. Yohai. Robust Statistics: Theory and Methods. *J. Wiley*, 2006.

G. McKhann, D. Drachman, M. Folstein, R. Katzman, D. Price, and E. M. Stadlan. Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. *Neurology*, 34(7):939–944, Jul 1984.

D. Messina, A. Cerasa, F. Condino, G. Arabia, F. Novellino, G. Nicoletti, M. Salsone, M. Morelli, P. L. Lanza, and A. Quattrone. Patterns of brain atrophy in Parkinsons disease, progressive supranuclear palsy and multiple system atrophy. *Parkinsonism & related disorders*, 17(3):172–176, 2011.

D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, and F. Leisch. *e1071: Misc Functions of the Department of Statistics (e1071) and TU Wien*, 2014. URL `http://CRAN.R-project.org/package=e1071`. R package version 1.6-4.

T. B. Murdoch and A. S. Detsky. The inevitable application of big data to health care. *Jama*, 309(13):1351–1352, 2013.

R. Nilsson, J. M. Peña, J. Björkegren, and J. Tegnér. Evaluating feature selection for SVMs in high dimensions. In Johannes Fürnkranz, Tobias Scheffer, and Myra Spiliopoulou, editors, *Machine Learning: ECML 2006*, volume 4212 of *Lecture Notes in Computer Science*, pages 719–726. Springer Berlin Heidelberg, 2006. ISBN 978-3-540-45375-8. doi: 10.1007/11871842_72. URL http://dx.doi.org/10.1007/11871842_72.

R. Nordenskjöld, F. Malmberg, E. Larsson, A. Simmons, S. J. Brooks, L. Lind, H. Ahlström, L. Johansson, and J. Kullberg. Intracranial volume estimated with commonly used methods could introduce bias in studies including brain volume measurements. *NeuroImage*, 83:355 – 360, 2013. ISSN 1053-8119. doi: http://dx.doi.org/10.1016/j.neuroimage.2013.06.068. URL http://www.sciencedirect.com/science/article/pii/S1053811913007064.

Z. Obermeyer and E. J. Emanuel. Predicting the future - big data, machine learning, and clinical medicine. *The New England journal of medicine*, 375(13):1216, 2016.

A. Ott, M. M. B. Breteler, F. van Harskamp, J. J. Claus, T. J. M. van der Cammen, D. E. Grobbee, and A. Hofman. Prevalence of alzheimer's disease and vascular dementia: association with education. the rotterdam study. *Bmj*, 310(6985):970–973, 1995.

H. Peng, F. Long, and C. Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(8):1226–1238, 2005.

R. Perneczky, S. Wagenpfeil, K. Komossa, T. Grimmer, J. Diehl, and A. Kurz. Mapping scores onto stages: mini-mental state examination and clinical dementia rating. *The American Journal of Geriatric Psychiatry*, 14(2):139–144, 2006.

R. C. Petersen, G. E. Smith, S. C. Waring, R. J. Ivnik, E. G. Tangalos, and E. Kokmen. Mild cognitive impairment: clinical characterization and outcome. *Arch. Neurol.*, 56(3): 303–308, Mar 1999.

E. Pettit and K. M. Swigger. An analysis of genetic-based pattern tracking and cognitive-based component tracking models of adaptation. *Structure*, 10:010110, 1983.

C. Plant, S. J. Teipel, A. Oswald, C. Böhm, T. Meindl, J. Mourao-Miranda, A. W. Bokde, H. Hampel, and M. Ewers. Automated detection of brain atrophy patterns based on MRI for the prediction of alzheimer's disease. *NeuroImage*, 50(1):162 – 174, 2010. ISSN 1053-8119. doi: http://dx.doi.org/10.1016/j.neuroimage.2009.11.046. URL `http://www.sciencedirect.com/science/article/pii/S1053811909012312`.

R. A. Poldrack, J. A. Mumford, and T. E. Nichols. *Handbook of functional MRI data analysis*. Cambridge University Press, 2011.

M. J. Prince. Predicting the onset of alzheimer's disease using bayes' theorem. *American Journal of Epidemiology*, 143(3):301–308, 1996. URL `http://aje.oxfordjournals.org/content/143/3/301.abstract`.

A. Puente. Wisconsin card sorting test. *Test critiques*, 4:677–682, 1985.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. URL `http://www.R-project.org/`. ISBN 3-900051-07-0.

P. Refaeilzadeh, L. Tang, and H. Liu. Cross-validation. In LING LIU and M.TAMER ZSU, editors, *Encyclopedia of Database Systems*, pages 532–538. Springer US, 2009. ISBN 978-0-387-35544-3. doi: 10.1007/978-0-387-39940-9_565. URL `http://dx.doi.org/10.1007/978-0-387-39940-9_565`.

The Ronald, Nancy Reagan Research Institute of the Alzheimer's Association, and National Institute on Aging Working Group. Consensus report of the working group on:molecular and biochemical markers of alzheimers disease. *Neurobiology of Aging*, 19 (2):109–116, 1998.

R. Rutenbar. Simulated annealing algorithms: an overview. *Circuits and Devices Magazine, IEEE*, 5(1):19–26, 1989.

J. J. Ryan and A. M. Paolo. Frequency of occurrence of a wais dementia pattern in schizophrenia and bipolar affective disorder. *The Clinical Neuropsychologist*, 3(1):45–48, 1989.

A. Sarica, G. di Fatta, and M. Cannataro. K-Surfer: A KNIME-based tool for the management and analysis of human brain MRI Freesurfer/FSL Data. *Frontiers in Neuroinformatics*, (3), 2014a. ISSN 1662-5196. doi: 10.3389/conf.fninf.2014.18.00003. URL `http://www.frontiersin.org/neuroinformatics/10.3389/conf.fninf.2014.18.00003/full`.

A. Sarica, G. di Fatta, G. Smith, M. Cannataro, and J. D. Saddy. Advanced feature selection in multinominal dementia classification from structural MRI data. *Proc MICCAI Workshop Challenge on Computer-Aided Diagnosis of Dementia Based on Structural MRI Data*, pages 82–91, 2014b.

T. M. Schouten, M. Koini, F. de Vos, S. Seiler, M. de Rooij, A. Lechner, R. Schmidt, M. van den Heuvel, J. van der Grond, and S. A. R. B. Rombouts. Individual classification of alzheimer's disease with diffusion magnetic resonance imaging. *NeuroImage*, 152:476–481, 2017.

L. Scrucca. GA: A package for genetic algorithms in R. *Journal of Statistical Software*, 53 (4):1–37, 4 2013. ISSN 1548-7660. URL `http://www.jstatsoft.org/v53/i04`.

D. J. Selkoe. Normal and abnormal biology of the beta-amyloid precursor protein. *Annual review of neuroscience*, 17(1):489–517, 1994.

S. M. Smith, M. Jenkinson, H. Johansen-Berg, D. Rueckert, T. E. Nichols, C. E. Mackay, K. E. Watkins, O. Ciccarelli, M. Z. Cader, P. M. Matthews, and T. E. Behrens. Tract-based spatial statistics: voxelwise analysis of multi-subject diffusion data. *Neuroimage*, 31 (4):1487–1505, 2006.

L. Sørensen, M. Nielsen, P. Lo, H. Ashraf, J. H. Pedersen, and M. de Bruijne. Texture-based analysis of copd: a data-driven approach. *IEEE transactions on medical imaging*, 31(1): 70–78, 2012.

L. Sørensen, A. Pai, C. Anker, I. Balas, M. Lillholm, C. Igel, and M. Nielsen. Dementia Diagnosis using MRI Cortical Thickness, Shape, Texture, and Volumetry. *Proc MICCAI Workshop Challenge on Computer-Aided Diagnosis of Dementia Based on Structural MRI Data*, pages 111–118, 2014.

A. L. Spedding, G. di Fatta, and M. Cannataro. A genetic algorithm for the selection of structural MRI features for classification of mild cognitive impairment and alzheimer's disease. In *Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference on*, pages 1566–1571. IEEE, 2015.

R. L. Spitzer, K. Kroenke, and J. B. W. Williams. Diagnostic and statistical manual of mental disorders. 1980.

K. Tabelow and J. Polzehl. *Statistical parametric maps for functional MRI experiments in R: The package fmri*. WIAS, 2010.

J. Talairach. Atlas d'anatomie stéréotaxique du télencéphale: etudes anatomo-radiologiques; atlas of stereotaxic anatomy of the telencephalon, 1967.

J. Talairach and P. Tournoux. Co-planar stereotaxic atlas of the human brain. 3-dimensional proportional system: an approach to cerebral imaging. 1988.

O. Tange. GNU Parallel - The Command-Line Power Tool. *;login: The USENIX Magazine*, 36(1):42–47, Feb 2011. URL `http://www.gnu.org/s/parallel`.

T. N. Tombaugh and N. J. McIntyre. The mini-mental state examination: a comprehensive review. *Journal of the American Geriatrics Society*, 40(9):922–935, 1992.

V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag New York, Inc., New York, NY, USA, 1995. ISBN 0-387-94559-8.

W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. URL `http://www.stats.ox.ac.uk/pub/MASS4`. ISBN 0-387-95457-0.

C. Wachinger, K. Batmanghelich, P. Golland, and M. Reuter. BrainPrint in the Computer-Aided Diagnosis of Alzheimer's disease. *Proc MICCAI Workshop Challenge on Computer-Aided Diagnosis of Dementia Based on Structural MRI Data*, pages 129–138, 2014.

J. Wan, Z. Zhang, J. Yan, T. Li, B. D. Rao, S. Fang, S. Kim, S. L. Risacher, A. J. Saykin, and L. Shen. Sparse bayesian multi-task learning for predicting cognitive outcomes from neuroimaging measures in alzheimer's disease. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 940–947, June 2012. doi: 10.1109/CVPR.2012.6247769.

D. Wechsler. Manual for the wechsler adult intelligence scale. 1955.

D. Wechsler. Wechsler adult intelligence scale–fourth edition (wais–iv), 2014.

B. L. Welch. The generalization ofstudent's' problem when several different population variances are involved. *Biometrika*, 34(1/2):28–35, 1947.

J. M. Whittaker. *Interpolatory function theory*, volume 33. The University Press, 1935.

J. L. Whitwell, W. R. Crum, H. C. Watt, and N. C. Fox. Normalization of cerebral volumes by use of intracranial volume: implications for longitudinal quantitative MR imaging. *AJNR Am J Neuroradiol*, 22(8):1483–1489, Sep 2001.

L. Yu and H. Liu. Feature selection for high-dimensional data: A fast correlation-based filter solution. In *ICML*, volume 3, pages 856–863, 2003.

Z. Zhao and H. Liu. Spectral feature selection for supervised and unsupervised learning. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, pages 1151–1157, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-793-3. doi: 10.1145/1273496.1273641. URL http://doi.acm.org/10.1145/1273496.1273641.