

Probabilistic prediction of Alzheimer's disease from multimodal image data with Gaussian processes

Jonathan Young

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
of the
University of London.

Department of Medical Physics and Biomedical Engineering
University College London

January 27, 2015

To Laura, who has been very patient.

Abstract

Alzheimer's disease, the most common form of dementia, is an extremely serious health problem, and one that will become even more so in the coming decades as the global population ages. This has led to a massive effort to develop both new treatments for the condition and new methods of diagnosis; in fact the two are intimately linked as future treatments will depend on earlier diagnosis, which in turn requires the development of biomarkers that can be used to identify and track the disease. This is made possible by studies such as the Alzheimer's disease neuroimaging initiative which provides previously unimaginable quantities of imaging and other data freely to researchers.

It is the task of early diagnosis that this thesis focuses on. We do so by borrowing modern machine learning techniques, and applying them to image data. In particular, we use Gaussian processes (GPs), a previously neglected tool, and show they can be used in place of the more widely used support vector machine (SVM). As combinations of complementary biomarkers have been shown to be more useful than the biomarkers are individually, we go on to show GPs can also be applied to integrate different types of image and non-image data, and thanks to their properties this improves results further than it does with SVMs.

In the final two chapters, we also look at different ways to formulate both the prediction of conversion to Alzheimer's disease as a machine learning problem and the way image data can be used to generate features for input as a machine learning algorithm. Both of these show how unconventional approaches may improve results.

The result is an advance in the state-of-the-art for a very clinically important problem, which may prove useful in practice and show a direction of future research to further increase the usefulness of such methods.

Acknowledgements

First of all I would like to thank my supervisors, Professors Sébastien Ourselin and John Ashburner, for their guidance and encouragement during the last few years.

I am also very grateful to Ged Ridgway for helping me get started early on and for statistical help, and to a number of people at the Dementia Research Centre, especially Kelvin Leung for helping me find my way through huge quantities of data from the ADNI study, and Dave Cash and Jonathan Schott for providing clinical perspective.

My colleagues at the Centre for Medical Image Computing have been indispensable. In particular, Marc Modat and Jorge Cardoso have not only provided some excellent software tools but have been extremely patient with my questions about how they should be used and have always been helpful in providing advice on my publications. Everyone at the UCL machine learning for medical imaging reading group has been a great source of ideas and suggestions.

Finally, I must express my deepest gratitude to my parents for supporting me throughout my education, and to my wife Laura for her companionship and patience in the last four years.

Contents

1	Introduction	15
1.1	Alzheimer's disease biology and biomarkers	15
1.2	Diagnosis of Alzheimer's disease	15
1.3	Neuropathology of Alzheimer's disease	16
1.4	Treatment of Alzheimer's disease	17
1.5	Alzheimer's Disease biomarkers	18
1.5.1	The Alzheimer's Disease Neuroimaging Initiative	20
1.5.2	Imaging biomarkers	21
1.5.3	CSF biomarkers	21
1.5.4	Other biomarkers	22
1.6	My contribution	22
1.6.1	Motivation	22
1.6.2	Outline of the thesis	23
2	Medical imaging and image processing	25
2.1	Introduction	25
2.2	Structural MRI	25
2.3	Positron emission tomography	27
2.4	Image registration	30
2.4.1	Transformation models	31
2.4.2	Interpolation methods	34
2.4.3	Objective functions	35
2.4.4	Optimisation	36
2.5	Anatomical segmentation	37
2.5.1	Manual segmentation	37
2.5.2	Automatic segmentation	38
2.6	Tissue segmentation	40

2.6.1	Expectation maximisation	40
2.6.2	Extensions to the expectation maximisation model	41
3	Machine learning	43
3.1	Introduction	43
3.2	Machine learning taxonomy	43
3.3	Preprocessing of data	45
3.3.1	Feature extraction	45
3.3.2	Feature selection	45
3.4	Performance measurement and validation in machine learning	46
3.4.1	Performance measures for classification	46
3.4.2	Performance measures for probabilistic classification	49
3.4.3	Performance measures for regression	49
3.4.4	Validation strategies	49
3.5	Pitfalls of machine learning experiments	51
3.5.1	Overfitting	51
3.5.2	Double dipping	51
3.6	Support vector machines	53
3.6.1	Linear SVMs	54
3.6.2	Soft-margin SVMs	55
3.6.3	Kernels and nonlinear SVMs	56
3.7	Gaussian processes	57
3.7.1	Gaussian process priors	58
3.7.2	Gaussian process regression - weight space view	59
3.7.3	Gaussian process regression - function space view	60
3.7.4	Gaussian process classification	61
3.7.5	Gaussian process regression and classification in practice	63
3.8	Precomputed kernels	65
4	Literature review	67
4.1	Introduction to the existing literature	67
4.2	Review of the existing literature	68
4.2.1	Voxel-based features	68
4.2.2	Region-based features	69
4.2.3	Cortical thickness features	70

4.2.4	Hippocampal features	72
4.2.5	Side-by-side assessment of features	74
4.2.6	Multi-MRI features	75
4.2.7	Multimodal classification	77
4.2.8	Other approaches	81
4.3	Summary	82
5	Classification of Alzheimer’s disease patients and controls with gaussian processes	84
5.1	Introduction	84
5.2	Materials and methods	85
5.2.1	Images	85
5.2.2	Image processing	85
5.2.3	Gaussian process regression and classification	86
5.2.4	SVM calculations	87
5.3	Results	87
5.4	Discussion	87
5.5	Conclusions	89
6	Multiple kernel learning for prediction of conversion to AD	90
6.1	Introduction	90
6.2	Materials and methods	93
6.2.1	MRI data	93
6.2.2	PET data	93
6.2.3	ApoE data	94
6.2.4	CSF data	94
6.2.5	Subjects	94
6.2.6	MRI image processing	96
6.2.7	PET image processing	97
6.2.8	Gaussian process classification	98
6.2.9	Gaussian process classification as a multimodal kernel method	98
6.2.10	SVM classification	100
6.2.11	Classification strategy	101
6.2.12	Validation	102
6.3	Results	103
6.3.1	Accuracy of binary classification	103

6.3.2	Accuracy of probabilistic classification	105
6.4	Discussion	106
6.5	Conclusion	111
7	Continuous proxies for AD diagnosis and prognosis	112
7.1	Introduction	112
7.2	Materials and methods	114
7.2.1	Subjects	114
7.2.2	MRI data	116
7.2.3	MRI image processing	116
7.2.4	PET image data	117
7.2.5	PET image processing	117
7.2.6	CSF data	118
7.2.7	Boundary shift integral	118
7.2.8	MMSE scores	118
7.2.9	Gaussian processes	119
7.2.10	Classification and validation in BSI experiment	119
7.2.11	Classification and validation in MMSE experiment	120
7.3	Results for BSI experiment	120
7.4	Results for MMSE experiment	122
7.4.1	Effect of MRI field strength on results	122
7.4.2	Accuracy of MMSE predictions	123
7.5	Discussion	123
7.6	Conclusion	127
8	Anatomical regional kernels	128
8.1	Introduction	128
8.2	Image and biomarker data	129
8.3	Image processing	130
8.3.1	Groupwise registration	130
8.3.2	Image segmentation	130
8.3.3	Image parcellation	130
8.3.4	Atlas construction	130
8.4	Gaussian process classification	131
8.4.1	Gaussian processes as multimodal kernel methods	132

8.5	Results	132
8.5.1	Binary accuracy	132
8.5.2	Information scoring	133
8.5.3	Individual predictions	134
8.5.4	Effects of scanner field strength	135
8.6	Discussion	135
8.6.1	Interpretation of hyperparameters	136
8.7	Conclusion	137
9	Conclusions	139
9.1	Overall conclusions	139
9.2	Future research	141
A	Running times and computational complexity	145
A.1	Experiment and results	145
A.2	Discussion	146
B	Lists of subjects in experiments	148
B.1	Subjects in experiment in chapter 5	148
B.2	Subjects in experiment in chapter 6	149
B.3	Subjects in experiments in chapter 7	151
B.4	Subjects in experiment in chapter 8	164

List of Figures

1.1	Formation of plaques from APP	17
1.2	Formation of neurofibrillary tangles and their interaction with neurons	18
1.3	Model of biomarker trajectories and ordering	20
2.1	Magnetisation in MRI.	26
2.2	Basic Fourier transform MR sequence	27
2.3	Molecular structure of glucose and fluorodeoxyglucose	28
2.4	PET pipeline	29
2.5	Outline of registration algorithms.	31
2.6	Rigid transformation for registration	32
2.7	Affine transformation for registration	33
2.8	Nonlinear transformation for registration	34
2.9	Manual segmentation of brain structures	37
2.10	Automatic anatomical segmentation by atlas propagation	38
2.11	Multiatlas segmentation	39
2.12	Segmentation of a structural MRI brain image into three tissue types	41
3.1	Clustering unlabelled data	44
3.2	Classification of labelled data	44
3.3	Receiver operating characteristics (ROC) curves	48
3.4	Leave-one-out cross validation (LOOCV) in a set of ten data points	50
3.5	Overfitting and underfitting of a function to a set of points	51
3.6	Divergence of training and testing error curves shows overfitting	52
3.7	Many hyperplanes can divide two linearly separable groups of points.	53
3.8	The SVM chooses the hyperplane that maximises the margin.	54
3.9	Use of a kernel to separate data that are not separable in the input space	56
3.10	Function space view of GP prior and posterior	60
3.11	Sigmoidal link functions	62

3.12	Gaussian process classification with two dimensional toy data	63
3.13	Maximum likelihood selects the model of the appropriate complexity	65
5.1	Pipeline to produce modulated GM images in a common space	86
5.2	AUC curves for classification of 40 AD and control subjects with GP and SVM.	88
6.1	Pipeline for multiple kernel learning with MRI, FDG-PET and ApoE data	100
6.2	Relationship between AD and MCI classification	101
6.3	Empirical risk vs. corrected predicted risk for the PET group.	107
6.4	Empirical risk vs. corrected predicted risk for the PET-CSF group.	107
7.1	Measured BAR across groups	121
7.2	Predicted BAR across groups	121
7.3	Predicted vs actual MMSE for FDG-PET data in PET. group	124
7.4	Predicted vs actual MMSE for MRI data in PET. group	124
7.5	Predicted vs actual MMSE for MRI data in MRI. group	124
7.6	Measured and predicted MMSE across clinical groups in the PET group	126
8.1	Pipeline for constructing atlas in groupwise space	131
8.2	Differences between individual predictions of AD versus control status by the ARK and voxel methods	134
8.3	Differences between individual predictions of AD versus control status by the ARK and regions methods	134
8.4	Differences between individual predictions of MCI conversion by the ARK and voxel methods	135
8.5	Differences between individual predictions of MCI conversion by the ARK and regions methods	135
8.6	Spectrum of regional weights in AD/HC classification	136
8.7	Maps of regions with more than 1% of total weight	137

List of Tables

1.1	Biomarkers for AD	19
3.1	Confusion matrix for classification	47
6.1	Demographics of PET group	95
6.2	Demographics of PET-CSF group	95
6.3	Times of conversion t , in months, for subjects in the PET group.	96
6.4	Regions included in GM segmentations	97
6.5	Accuracy of methods in the PET group with GP classification	103
6.6	Accuracy of methods on the PET group with SVM classification	104
6.7	Accuracy of methods on the PET-CSF group with GP classification	104
6.8	Accuracy of methods on the PET-CSF group with SVM classification	105
6.9	Side-by-side statistical comparison of GP and SVM classification for different groups and modalities.	106
6.10	Reported results from a variety of studies for predicting MCI conversion on ADNI data	110
7.1	Subject groups and demographics for the BSI experiment.	114
7.2	Subject groups and demographics for the PET group in the MMSE experiment.	115
7.3	Subject groups and demographics for the MRI group in the MMSE experiment.	115
7.4	Accuracy of discrimination between MCI-s and MCI-c with predicted brain atrophy rate	121
7.5	Accuracy of discrimination between MCI-s and MCI-c with training on binary diagnostic class labels	121
7.6	Accuracies for predicting conversion to AD in MCI subjects in the MMSE experiment	122
7.7	Accuracies for predicting conversion to AD in MCI subjects in the MMSE experiment using the entire MRI group.	122

7.8	Breakdown of accuracy of predicted MMSE in MCI conversion by MRI field strength	123
7.9	Accuracy of predicted MMSE compared to ground truth	123
8.1	Subject groups and demographics	130
8.2	Demographics of subjects for ARK experiments	130
8.3	Accuracy of classification between control and AD subjects with ARK, voxels and regions methods	133
8.4	Accuracy of classification between MCI-s and MCI-c subjects with ARK, voxels and regions methods	133
8.5	Results for ARK classification of MCI-s and MCI-c, broken down by MRI scan field strength	136
9.1	Best overall classification with MRI data only	139
9.2	Best overall classification with PET or PET and MRI data	141
A.1	Running times for training a model on 50 AD and 50 control subjects	145
B.1	List of subjects used in the control/AD classification experiment in chapter 5	148
B.2	List of subjects used in the PET group of the MKL experiment in chapter 6	149
B.3	List of subjects used in the BSI experiment in chapter 7	151
B.4	List of subjects used in the MRI group of the MMSE experiment in chapter 7	153
B.5	List of subjects used in the ARK experiments in chapter 8	165

Chapter 1

Introduction

1.1 Alzheimer's disease biology and biomarkers

Alzheimer's disease (AD) is a condition causing dementia, primarily in the elderly population. The condition is named for Alois Alzheimer, a German psychiatrist who was the first to identify and describe the condition, and link its symptoms and pathology in 1906 [Berchtold and Cotman, 1998]. While a number of other conditions are known to cause dementia, AD remains by far the most common, although it may often occur alongside other dementia-causing conditions such as vascular dementia, the second most common such disease [Zekry et al., 2002]. A small minority of AD cases are inherited familial AD, but the vast majority of cases occur sporadically and among these, age is by far the most important risk factor. The prevalence among people over 84 years old is estimated to be up to 42% [Hebert et al., 2003]. As a consequence of ageing populations worldwide, due to improved healthcare and living conditions, the number of people living with AD is expected to rise to a global total of more than 100 million in 2050 [Brookmeyer et al., 2007], which would represent a quadrupling since 2006. This will translate into a huge economic impact; as AD cannot be cured and gradually progresses, producing increasingly severe symptoms, it results in huge costs from patient care alongside lost productivity of patients and carers. The consequent costs were estimated at \$100 billion annually in the US in 1998 [Meek et al., 1998]. The early stages of AD are marked by short term memory loss, with the symptoms progressing to loss of longer term memories and other cognitive domains. AD ultimately leads to death, with no cure currently in existence.

1.2 Diagnosis of Alzheimer's disease

Typically, in the clinic a diagnosis of probable AD is made based on a set of consensus criteria which are regularly reviewed [McKhann et al., 2011]. Such a diagnosis may be based on examining the patient and their medical history, and interviewing them and those they are in regular contact with, as well as cognitive testing such as mini-mental state examination

(MMSE) [Folstein et al., 1975] or the Alzheimer's disease assessment scale cognitive subscale (ADAS-cog) [Rosen et al., 1984], and imaging. This is known as probable AD because the gold standard for AD diagnosis is based on histology and so can only be made at autopsy [Kist and Hastie, 1995]. The correspondence between a diagnosis in the clinic and subsequent confirmation by autopsy has been studied and found to be high [Ranginwala et al., 2008] although this may vary substantially between different AD centres [Beach et al., 2012]. In more recent years, the emphasis has shifted heavily to diagnosing the condition in its very early stages [Chong and Sahadevan, 2005], as the disease process is thought to begin long before symptoms become obvious, and future disease modifying treatments will be of most use to patients at this stage. This has led to the introduction of the concept of mild cognitive impairment (MCI) [Petersen et al., 1999], defined as a memory impairment greater than would be expected from normal ageing, but less than that of AD, and which does not affect a patient's ability to carry out routine tasks from day to day. Diagnostic criteria for MCI match this definition. MCI can be seen as a risk state for AD because the annual rate of conversion from MCI to AD is 10-15%, as opposed to only 1-2% for the general population. However MCI cannot be seen as equivalent to actual prodromal AD, as MCI is in fact quite heterogenous and can be the manifestation of a variety of different conditions [Dubois and Albert, 2004]. However a subtype of MCI is recognised as being the early stages of AD, known as MCI due to Alzheimer's disease [Albert et al., 2011].

1.3 Neuropathology of Alzheimer's disease

The effect of AD pathology on the gross scale is characterised by atrophy caused by loss of neurons, most marked in several structures in the brain's temporal lobes and in enlargement of the ventricles. However, such changes are also present in normal ageing [Raz et al., 2005] but proceed at a much slower pace. A more specific effect of AD is seen in amyloid plaques and neurofibrillary tangles.

Plaques are dense aggregates of insoluble protein that form around neurons. In AD, their main constituent is beta amyloid ($A\beta$); however plaques have also been observed in the brains of undemented elderly people and it is the specific distribution of plaques, not their mere presence, that is indicative of AD [Bouras et al., 1994]. $A\beta$ exists in two common forms, $A\beta_{40}$ and $A\beta_{42}$. Both are formed from sequential cleavage of the amyloid precursor protein by the enzymes β - and γ - secretase. The more common $A\beta_{40}$ form is soluble and is found in cerebrospinal fluid (CSF) [Ghisso and Frangione, 2002] whereas the insoluble $A\beta_{42}$ form is produced when cleavage by γ secretase occurs at the $A\beta_{42}$ rather than $A\beta_{40}$ residue [Selkoe, 2004]

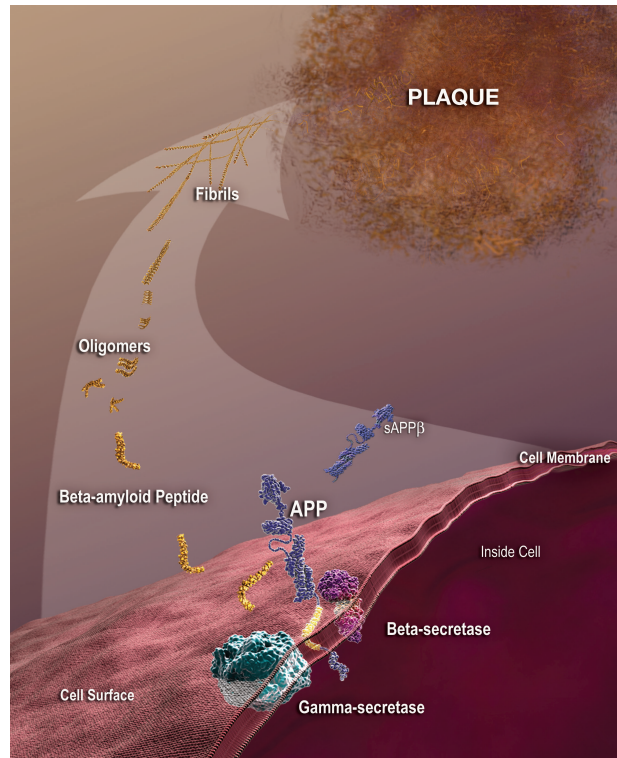


Figure 1.1: Formation of plaques from APP. Courtesy of National Institute on Aging/National Institutes of Health, <http://www.nia.nih.gov/alzheimers/scientific-images>.

and is strongly implicated in AD [Yin et al., 2007]. The resulting plaques are toxic to neurons [Yankner et al., 1990].

Neurofibrillary tangles (NFTs) are bundles of insoluble protein that accumulate inside neurons. Similarly to plaques, different types of NFT are associated with different conditions, and they are also sometimes found in otherwise healthy brains, so it is the pattern and type of NFT rather than their presence that is indicative of AD [Bouras et al., 1994]. The primary protein in NFTs is tau protein. AD disrupts the dephosphorylation of tau protein, leading to hyperphosphorylation. This means that the tau can no longer perform its role in assisting the stabilisation of microtubules within the neuron, causing them to begin to disintegrate. The unbound tau instead clumps together to form tangles [Lee et al., 2005].

1.4 Treatment of Alzheimer's disease

Currently available treatments for AD remain entirely targeted at treating the symptoms rather than interrupting the disease process. Of the five drugs, four are targeted at the reduction in activity of cholinergic neurons which marks AD [Geula and Mesulam, 1995]. They do this by acting as acetylcholinesterase inhibitors which slow down the rate at which acetylcholine is broken down, which partially compensates for the loss of cholinergic neurons [Stahl, 2000].

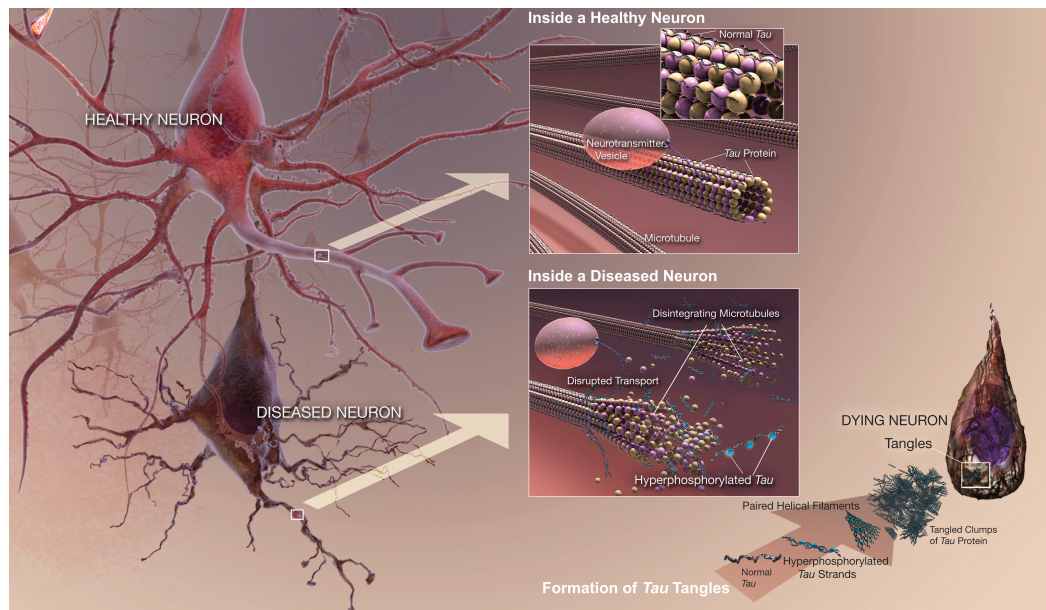


Figure 1.2: Formation of neurofibrillary tangles and their interaction with neurons. Courtesy of National Institute on Aging/National Institutes of Health, <http://www.nia.nih.gov/alzheimers/scientific-images>.

The fifth drug instead blocks overstimulation of NMDA receptors by glutamate, which can lead to cell death [Lipton, 2006]. However the effects of both types of treatment are modest [Birks and Harvey, 1996, Reisberg et al., 2003], offering small benefits in terms of cognitive function and daily living.

Future treatment will be aimed more at disrupting the underlying disease process, essentially preventing neurodegeneration rather than allowing AD patients to make better use of their remaining brain tissue. This can be done by targeting the formation of β amyloid plaques [Lashuel et al., 2002] or aggregation of tau protein [Wischik et al., 2008]. However to maximise the effectiveness of these approaches, treatment would have to begin earlier in the disease process, which is a major motivation for this work.

1.5 Alzheimer's Disease biomarkers

Biomarkers are measurable quantities that are indicative of the presence or progress of some underlying disease. A number of these are associated with AD and are summarised in table 1.1.

Based on longitudinal study of a large cohort of elderly subjects, it was hypothesised that the various biomarkers do not all begin to depart from normal levels simultaneously. Rather, the level of abnormality in each biomarker follows a sigmoidal trajectory, initially rising steeply before levelling off, with a distinct ordering [Jack et al., 2010]. An updated version of the model

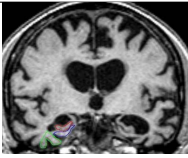
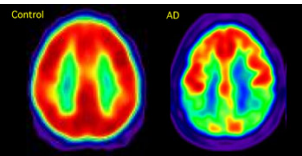

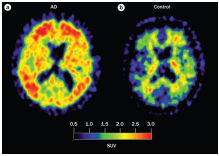
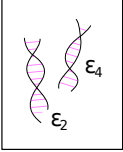
Biomarker	Measured with	
Brain atrophy/volume		Structural MRI
Brain metabolism		FDG-PET
A β and tau protein		CSF sampled with spinal tap
Amyloid plaques		Amyloid PET
ApoE genotype		Genetic testing

Table 1.1: Biomarkers for AD. Images from <http://www.esciencenews.com>, <http://www.alzforum.org/>, www.lymphomajournal.com/, and [Nordberg et al., 2010]

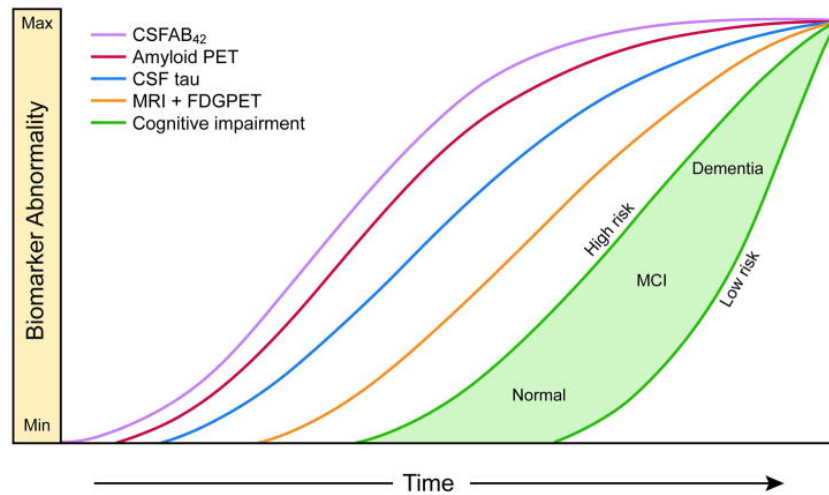


Figure 1.3: Model of biomarker trajectories and ordering. Courtesy of [Jack et al., 2013].

[Jack et al., 2013] is shown in figure 1.3. In accordance with the notion of amyloid levels being the cause of AD, levels of $A\beta$ become abnormal long before any other biomarker, possibly many years before symptoms become apparent. This is closely followed by abnormality in levels of tau protein. After this come the downstream effects of the AD process, in the form of tissue loss and reduced metabolism, as measured by structural MRI and FDG-PET respectively. Finally the effect of AD on cognition becomes apparent when the condition becomes symptomatic. This has obvious implications in the choice of data used to attempt early diagnosis. Clearly, symptoms or cognitive scores will be of relatively little help as they will be close to normal early in the disease process. $A\beta$, and tau, conversely, will already be abnormal well before any cognitive decline is noticed. This means that in combination they can distinguish AD subjects from controls as accurately as clinical examination [Sunderland T et al., 2003]; however they plateau early which means they may be less effective to track disease progression.

1.5.1 The Alzheimer's Disease Neuroimaging Initiative

A great deal of research into AD biomarkers during the previous decade has been based on data from the Alzheimer's disease neuroimaging initiative (ADNI). This was launched in 2003 by the National Institute on Ageing (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organisations, as a \$60 million, five-year public/private partnership. The primary goal of ADNI has been to test whether serial MRI, PET, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild MCI and early AD. Data was obtained on subjects from more than 50 centres in the USA and Canada. All subjects were given structural MRI scans, and partially overlapping subsets of

the subjects also had FDG-PET scans and samples of cerebrospinal fluid taken with a lumbar puncture for CSF biomarkers. Initial recruitment for ADNI was for a total of approximately 800 adults in the age range of 55 to 90 years. This comprised 200 cognitively normal older individuals to be followed for three years, 400 people with MCI to be followed for three years and 200 people with early AD to be followed for two years.

Due to the great success of ADNI, two extensions to the original initiative were begun more recently. ADNI-Grand Opportunity (ADNI-GO) enabled extended follow-up of about 500 MCI and healthy subjects from the original ADNI cohort, and enrolled another 200 new MCI subjects, adding amyloid PET imaging to the protocol for these new subjects. ADNI 2 aimed to recruit a further 550 subjects, with a similar proportion of healthy, MCI and AD subjects as the original ADNI. Advanced MRI modalities including diffusion tensor imaging (DTI), resting-state functional MRI (fMRI) and arterial spin labelling (ASL) were added to the protocol at centres where the scanner permitted, and amyloid PET scans were performed on all subjects. Imaging protocols for the original modalities of structural MRI and FDG-PET for both ADNI 2 and ADNI-GO were designed to ensure compatibility with the original ADNI data.

1.5.2 Imaging biomarkers

The primary imaging modalities used in this thesis are those available for subjects in the initial ADNI cohort: structural MRI and FDG-PET. Broadly speaking, structural MRI gives information about anatomy while PET yields functional information; FDG-PET is used to assess the uptake of glucose which is seen as an indicator of metabolic activity. Both are described in detail at the start of chapter 2.

1.5.3 CSF biomarkers

As previously stated, $A\beta$ and tau proteins are heavily implicated in the process of AD and so their concentration in the body is a promising biomarker for AD. Levels of both can be measured in blood or CSF. In ADNI, the biomarker levels are calculated from CSF samples. As these require a lumbar puncture to obtain, CSF testing is more invasive than that of blood samples. However, it was preferred to blood testing as the CSF is in direct contact with extracellular spaces in the brain and should most directly reflect the brain's biochemistry [Blennow, 2004] and biomarkers for CSF are better established than those for blood [Cedazo-Minguez and Winblad, 2010]. In ADNI, levels of $A\beta$, total tau and phosphorylated tau are measured. Levels of $A\beta$ fall in AD as it accumulates in the brain, whereas levels of tau in CSF rise with the onset of disease.

1.5.4 Other biomarkers

The rare familial form of AD is entirely genetic, but sporadic AD also has a genetic component. The apolipoprotein E (ApoE) gene has long been known to be involved in AD. It exists in three alleles, known as $\epsilon 2$, $\epsilon 3$ and $\epsilon 4$. The most common allele is $\epsilon 3$. However, the second most common allele, $\epsilon 4$ is associated with increased risk of developing AD [Corder et al., 1993] whereas the rarest allele, $\epsilon 2$, appears to be protective against AD [Corder et al., 1994]. ADNI tests ApoE genotype for almost all subjects.

A number of newer imaging modalities have also produced promising biomarkers for AD. As a complement to measuring levels of $A\beta$ in CSF, the radiotracer Pittsburgh compound B (PiB) was developed specifically to bind to amyloid plaques in the brain. PiB retention has been shown to be similar in AD and converting MCI patients, and greater in the converting than the stable MCI patients [Forsberg et al., 2008]. Alternative MR imaging modalities such as DTI which is sensitive to microstructural changes in white matter, and fMRI which can show which brain regions are active in specific cognitive tasks have also shown promise in yielding new AD biomarkers, either individually [Nir et al., 2013, Li et al., 2013] or jointly [Wee et al., 2012]. However most such studies have been small, and amyloid-PET, DTI and fMRI data are currently available for only a few subjects in the ADNI database.

1.6 My contribution

1.6.1 Motivation

In the previous decade, the AD biomarkers described previously have been combined with automated medical image processing (chapter 2), and machine learning (chapter 3) to automatically diagnose AD. Such methods can now attain an accuracy as good as experienced clinicians; however, exceeding this accuracy is difficult as the classifiers are ultimately reliant on ground truth diagnoses provided by clinicians for training data. While automated methods do not have a clear advantage over more traditional methods of diagnosis, there is a very clear application for them in predicting the onset of disease in subjects with less severe (or even no) symptoms. By definition, prognosis based on test scores will be very difficult as at this stage there will be little or no difference between subjects who will go on to develop dementia and those who will not. Moreover, distinguishing between these is a problem that has gained increasing levels of clinical relevance as the desired time of diagnosis is pushed earlier and earlier.

This is largely due to the development of new types of treatment for AD, which will require identification of AD while symptoms are still very mild in order to be effective. Early identification will also make recruitment for future clinical trials for these treatments much easier and

cheaper.

In this thesis, we add to the literature on applying machine learning methods to neuroimaging data to study and classify healthy, MCI and AD subjects. In particular, our focus is on prediction of AD, rather than simple diagnosis. This is primarily accomplished through classification of the MCI population into converting and stable subjects.

1.6.2 Outline of the thesis

We begin by giving technical background on the major computational methods of all original work in the chapters discussing image processing (chapter 2) and machine learning (chapter 3).

Next the literature review (chapter 4) puts the work presented here in the context of research performed by others working to perform early diagnosis of AD. As the first three chapters of this thesis imply, there are three major stages at which choices can be made in building such a classifier. They are the type(s) of image or other data used, the features extracted and/or selected from that data, and the choice of classification strategy and algorithm. The literature review discusses these stages and their interaction further before examining the state-of-the-art in depth.

The thesis contains some material relating to each of these three stages. However, the first experiment we present, in chapter 5 focuses on the last. In it, a comparison is made of GP and SVM classification of a small set of subjects from the ADNI study into AD patients and controls. Exactly the same features are used for both classifiers, as the aim of this study was purely a proof of concept to show that GPs could produce results with comparable accuracy to more widely used methods. This was the first application of GP classification to AD although it had previously been applied for classification of structural MRI data in Huntington's disease [Chu et al., 2010].

Chapter 6 is more broad-based. It is a study of the utility of multimodal data in prediction of conversion in MCI subjects, making use of the multiple-kernel learning (MKL) paradigm discussed in the literature review (section 4.2.7). The advantages of GPs over SVMs - chiefly, in this case, automatic parameter setting from training data alone - are more fully exploited. Again, this is the first use of multimodal GP classification to medical image data. The classification results obtained from using MRI, FDG-PET, CSF and ApoE data are compared to each other, and to a combination of all of them. It also compares GPs and SVMs. While there is little to choose between them on grounds of accuracy for monomodal data, for multimodal data the GP performs much better as it has a superior mechanism for optimally weighting the types of data.

The next two chapters of research, 7 and 8, attempt to improve results further by taking

a more unconventional approach to classification strategy and feature extraction. Firstly, the standard classification approach is compared to using regression to predict a continuous proxy that is known to relate to AD. This produces vastly superior results to classification when only a small amount of data is available, although the advantage is much smaller when more data is available. Secondly, the MKL framework for GPs is applied to combine data from different regions of the brain, rather than data from different image modalities. This combines the strengths of defining features at the voxel level and at the regional level, offering better classification than either alone.

Finally chapter 9 briefly summarises the conclusions that can be drawn from the previous four chapters and suggests some areas of future research.

Chapter 2

Medical imaging and image processing

2.1 Introduction

The processing of medical image data is an extremely broad topic, and I do not go in to great depth for every technique in this chapter. In particular, there are a wide variety of specialised algorithms that can be seen as feature extraction methods. These are described in the articles referenced in the literature review, and in chapters 5 to 8. However there are also lower level procedures that are fundamental in the analysis of MR and PET images and are used in virtually every study using image data to diagnose and predict AD. The first is image registration, which finds the transformation that optimally aligns one image with another. This is necessary to establish a correspondence between images so information can be transferred from one to the other. The second is anatomical parcellation, in which one of a set of labels indexing particular anatomical structures is assigned to every voxel in an image, dividing it up into regions. The third is tissue segmentation, where voxels are assigned a label or labels representing the type of tissue they contain. The first image process topic described is image registration, as both anatomical parcellation and tissue segmentation rely on it. However before that comes an introduction to structural MRI and PET, the two primary imaging modalities used in this thesis.

2.2 Structural MRI

MRI is an extremely flexible imaging modality that can be employed to study contrasts between a very wide variety of tissue types, or measure many physical properties of tissues. Its initial function, however, was in the study of anatomy and this is what is most widely used in the study of AD, in what is known as structural MRI.

MRI is dependent on the phenomenon of nuclear magnetic resonance (NMR). The physics of NMR can only be fully described by quantum mechanics, however for these purposes the property of nuclear spin can be seen as a physical rotation, with an orientation described by a vector. In an MRI machine, the nuclei of objects inside the scanner (such as hydrogen nuclei

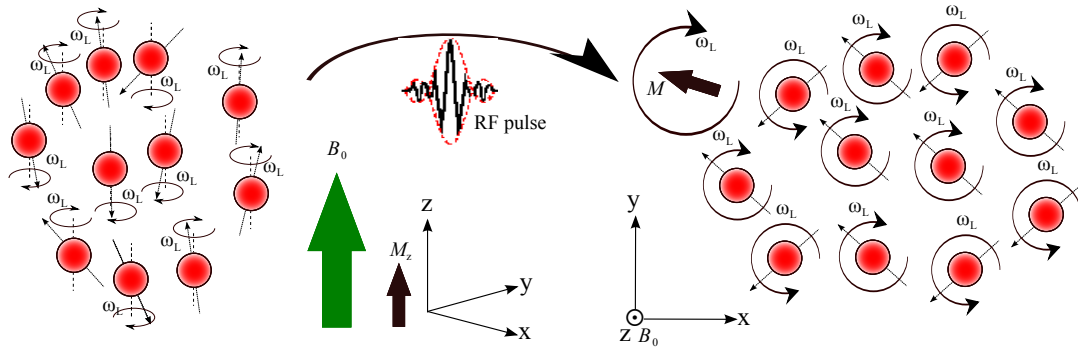


Figure 2.1: Magnetisation in MRI.

in the human body) are subjected to a strong, static magnetic field. A field strength (B_0) of 1.5 or 3T is typical in scanners available for clinical use; the higher the value of B_0 , the higher the image quality. The nuclear spins align parallel or anti-parallel to B_0 . A small majority of spins lie in the parallel configuration as it has slightly lower potential energy, resulting in a small net magnetisation M_z . The atoms also precess around B_0 with an angular frequency $\omega_L = \gamma B_0$, known as the Larmor frequency, where γ is a property of particular atomic nuclei. This is shown in figure 2.1.

The phase of the atoms' precession is random, so the net magnetisation in the xy plane transverse to B_0 is zero. However, according to quantum mechanics, the spins can absorb photons with an energy $E = h\nu$, where h is Planck's constant and ν is frequency corresponding to the the Larmor frequency, $\nu = \frac{\omega_L}{2\pi}$. For a typical B_0 this is in the radio frequency (RF) range, so the spins can be excited by an RF pulse. If this pulse is applied perpendicular to the z axis, the spins can be made coherent in phase and the net magnetisation is rotated through 90 degrees into the xy plane. After the pulse is stopped, the spins resume precession in the xy plane, but as they are in now in phase and the magnetisation is in the xy plane also, this produces a rotating net magnetisation. This induces a voltage in the scanner's receiver coils, which constitutes the MR signal. The time varying signal is converted into a frequency spectrum via Fourier transforms. The signal immediately begins to decay, which is a results of two processes: return of net magnetisation to the z direction, with time constant T1, and decoherence of phase in the xy plane, with time constant T2. Both time constants are dependent on tissue type, which gives MRI its excellent contrast between tissues.

A three dimensional image is built up by spatially localising the signal source, using gradients of static magnetic field B_0 . A slice along the z axis is chosen by applying a slice selection gradient G_s that varies the magnitude of B_0 along the z axis. This means that the Larmor frequency will vary along z , so only nuclei in one particular slice will be excited by an RF pulse.

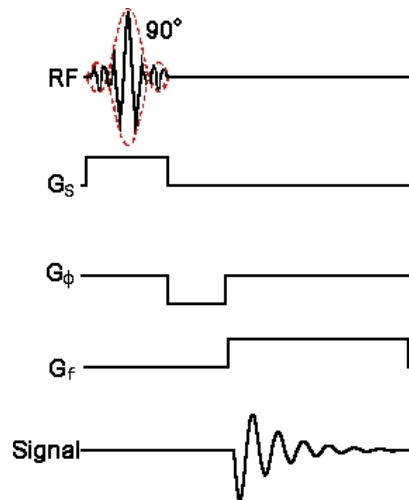


Figure 2.2: Basic Fourier transform MR sequence. From <http://www.cis.rit.edu/htbooks/mri>.

Two further gradients are applied to localise the signal within the slice in the x and y directions. A phase encoding gradient G_ϕ is applied in one of these directions. This means that the frequency of precession alters along the direction of the phase encoding gradient, meaning that spins lose phase coherence in the phase encoding direction. When the phase encoding direction is turned off, the spins will all precess at the same frequency once more, but with phase varying along the phase encoding direction. Finally, while the signal is being measured, a frequency encoding gradient G_s is applied in the remaining direction, causing the frequency of precession to vary in that direction. This is shown in figure 2.2

This means the each location in the xy plane within the selected slice has a unique combination of signal phase and frequency. As a result, the strength of signals from different locations in the slice can be identified from the Fourier transform of the signal.

It should be emphasised that this example is probably the simplest possible useful MR sequence. The versatility of MR imaging, in terms of its ability to show many types of different tissue contrast or even dynamic processes in the human body, stems from the enormous variety of RF pulses, gradients, and timings that can be applied.

2.3 Positron emission tomography

Positron emission tomography (PET) is a method of imaging a specific physiological process within the body. The patient is injected with a radioactive tracer compound, and then when the patient is inside the scanner, the concentration of the tracer can be mapped by detecting the effects of its radioactivity. The concentration can then be used to make inferences about processes such as rate of uptake or total amount of particular compounds.

The tracer molecule is designed to chemically mimic a compound that occurs naturally as

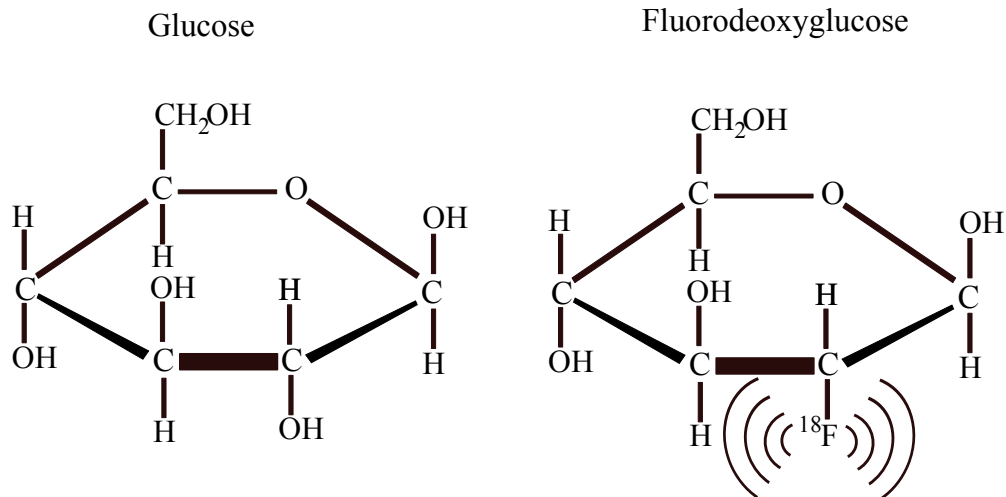


Figure 2.3: Molecular structure of glucose and fluorodeoxyglucose. A hydroxyl group on the glucose is replaced with a radioactive ^{18}F atom to make it a radiotracer analogue of the original molecule.

closely as possible, so that the behaviour of the tracer is as similar as possible to that of the natural molecule within the body. Different tracers are designed to mimic (be an analogue for) different compounds, and the one which is chosen depends on what physiological function the scan is meant to measure. In fluorodeoxyglucose-PET (FDG-PET), the tracer is an analogue for glucose, which is a primary source of energy for cellular respiration. FDG-PET can therefore be used to image the metabolism of cells with high glucose uptake, which includes those of the brain as well as cancer cells. This makes it a common choice in both neurological and oncological applications. However, the tracer must be radiolabelled by replacing part of the molecule with a positron emitting radionuclide. This may be done by changing the isotopes of an atom in the compound, such as substituting a positron emitting ^{11}C carbon atom for a ^{12}C stable one. This has the desirable property as the altered molecule is chemically identical with naturally occurring glucose. However the ^{11}C atom has a half life of only about 20 minutes, making tracers using it impractical in hospitals that do not have an on site cyclotron. Hence, it is more common to use a radiotracer where one of the hydroxyl groups normally present on a glucose molecule is replaced by an ^{18}F fluorine atom (figure 2.4).

As the ^{18}F has a half-life of 110 minutes, the resulting fluorodeoxy glucose (FDG) can be manufactured off site and shipped to where it is needed. This, combined with wide applicability, makes FDG-PET imaging by far the most common type of PET.

All radionuclides used in PET tracers decay by emitting a positron. The positron travels a short distance from where it was emitted, losing kinetic energy by interactions with the sur-

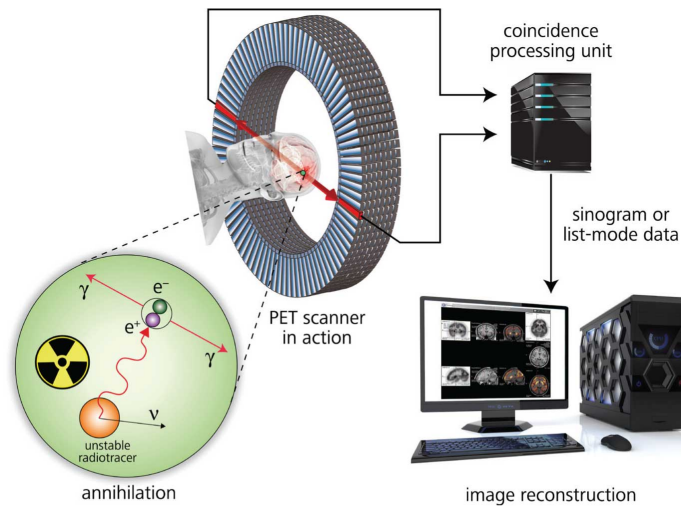


Figure 2.4: PET pipeline. Image courtesy of <http://www.sepscience.com>

rounding tissue. At this point, having typically travelled a few millimetres, it has a sufficiently small kinetic energy to interact with an outer shell electron of a nearby atom. The resulting annihilation reaction produces two 511keV gamma rays which travel in exactly opposite directions. If these are not absorbed in the body, they will activate a pair of detectors, arranged in a cylinder of concentric rings around the patient. If a pair of detectors in the same ring both detect a gamma ray within a short window of time of each other (of the order of a few nanoseconds) then the gamma rays are considered to come from a single annihilation (event), which is assumed to have occurred on the line linking that pair of detectors (line of response). In practice, neither of these assumptions is always correct, as gamma rays produced from distinct but almost simultaneous annihilations can be counted as an event, and gamma rays can be deflected by Compton scattering. Both of these are sources of noise in PET images.

Events on each line of response are counted, and then the counts for parallel lines of response grouped together. These can then be used to form a projection image called a sinogram, which includes information from all projections in a ring.

Image reconstruction techniques are then used to generate a map of the estimated concentration of the tracer in the body from the sinogram. Filtered back projection [Natterer, 1986] is a frequently used approach. The sinograms are used to create intensity profiles for each projection angle, which can then be back-projected to reconstruct the original image. More recently, iterative approaches based on expectation maximisation have emerged as an alternative. These try to find the distribution of tracer most likely to have produced the observed data. This can produce images with less severe noise and artefacts, but at much greater computational cost [Vardi et al., 1985].

The probability of a gamma ray interacting with an atom in the body before being detected depends on the distance it travels through tissue. Therefore structures deeper inside the body are assigned an artificially low activity in image reconstruction. This can be addressed with attenuation correction, based on a map of the attenuation coefficient in tissue across the body. The map can be obtained from a CT image if the PET image is from a PET/CT scanner. Alternatively a transmission image from a radioactive source may be used. While visual assessment of PET scans by a radiologist is still widespread, the scan may undergo further processing steps to produce more objective measures that allow fairer comparison between subjects. The standardised uptake value (SUV) [Zasadny and Wahl, 1993] normalises uptake by the subject's dose and body weight. For specific conditions, normalisation may also be done with reference to a region where uptake remains relatively unaffected by disease, such as the cerebellum, or to a customised reference cluster [Yakushev et al., 2009].

2.4 Image registration

Medical image registration is the process of aligning two medical images, so that a one to one mapping between the coordinate systems exists and shows correspondence between equivalent locations. This is generally done by defining one image as a target or reference image, which remains fixed, and the other as a floating image which is allowed to transform. Registration finds the transformation such that the the transformed floating image is optimally aligned with the target image. Algorithms for registration therefore consist of three components: a transformation model that defines in what ways the floating image is allowed to move, an objective function comprising some measure of the quality of the alignment between the images, with a regularisation term for complex transformations, and an optimiser that determines the best parameter values for the transformation. The floating image is deformed according to the transformation model, and the resulting image is compared to the target image with the optimiser iteratively updating the transformation. This is shown schematically in figure 2.5.

Image registration may be between different subjects to compare anatomy, or between scans taken at different timepoints for a single subject to track anatomical change, or between scans taken close in time for a single subject but with different imaging modalities to perform multimodal analysis or assist in further analysis of data such as PET images. These different applications generally require particular choices for the three components listed above to obtain the best results.

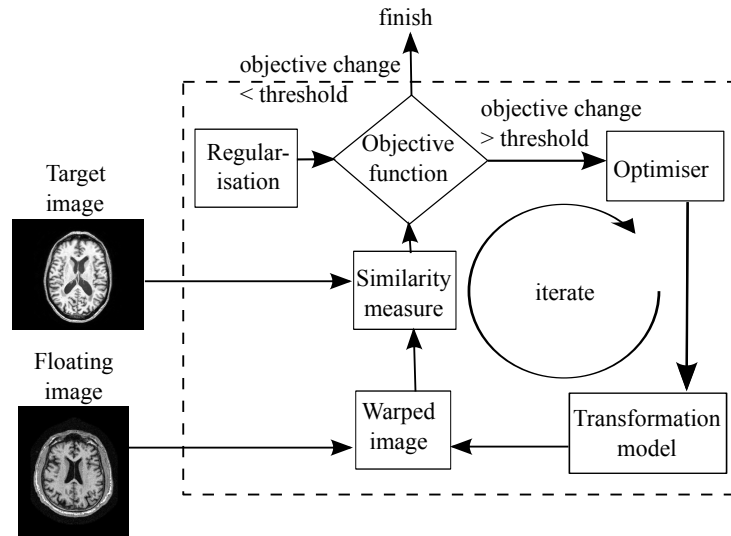


Figure 2.5: Outline of registration algorithms.

2.4.1 Transformation models

A transformation model \mathbf{T} defines a mapping for a voxel in the reference image to a coordinate in the floating image.

The simplest transformation model is that of rigid transformations, which allows translation and rotation of the floating images but no changes of shape. This might be used to provide a good initialisation for a more flexible registration, or when very little shape difference between images is expected.

In the two dimensional model shown in figure 2.6, a rotation about the z axis by an angle θ may be represented by a matrix $\mathbf{R}(\theta)$:

$$\mathbf{R}(\theta) = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \quad (2.1)$$

and a translation can be represented by a vector \mathbf{t} . The two may be simply combined: The effect of a rotation $\mathbf{R}(\theta)$ and a translation \mathbf{t} on a point $\mathbf{x} = (x, y)$ is given by $\mathbf{T}(\mathbf{x}) = \mathbf{R}\mathbf{x} + \mathbf{t}$. This can be expressed in homogeneous coordinates as a single matrix, which can be decomposed into separate rotation and translation matrices:

$$\mathbf{T}(\mathbf{x}) = \begin{bmatrix} \mathbf{R} & \mathbf{t} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ 1 \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{t} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & \mathbf{R} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ 1 \end{bmatrix} \quad (2.2)$$

In the three dimensional case, common in medical image registration, there are three rotation matrices: A rotation in the yz plane by θ_1 about the x axis, a rotation in the xz plane by θ_2 about the y axis, and a rotation in the xy plane by θ_3 about the z axis:

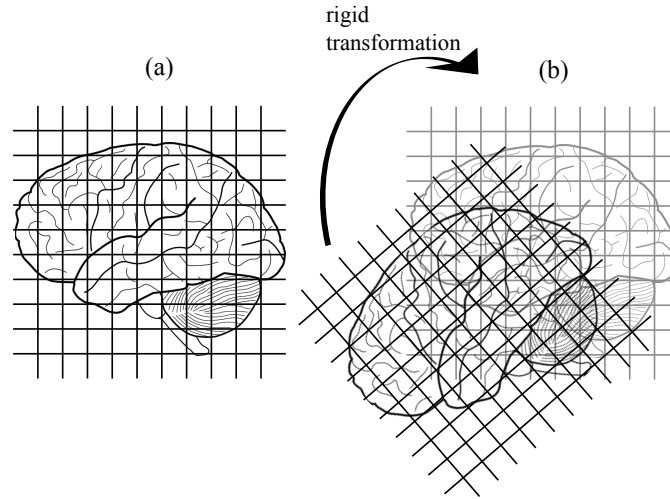


Figure 2.6: Rigid transformation for registration. (a) shows a target image; (b) shows a floating image brought into alignment with (a) by a rigid transformation consisting of a translation and a rotation.

$$\mathbf{R}_x(\theta_1) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos(\theta_1) & -\sin(\theta_1) & 0 \\ 0 & \sin(\theta_1) & \cos(\theta_1) & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad \mathbf{R}_y(\theta_2) = \begin{bmatrix} \cos(\theta_2) & 0 & \sin(\theta_2) & 0 \\ 0 & 1 & 0 & 0 \\ -\sin(\theta_2) & 0 & \cos(\theta_2) & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (2.3)$$

$$\mathbf{R}_z(\theta_3) = \begin{bmatrix} \cos(\theta_3) & -\sin(\theta_3) & 0 & 0 \\ \sin(\theta_3) & \cos(\theta_3) & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (2.4)$$

The three rotations can be combined by multiplying the three separate rotation matrices together so $\mathbf{R} = \mathbf{R}_x(\theta_1)\mathbf{R}_y(\theta_2)\mathbf{R}_z(\theta_3)$. This can then be combined with a three dimensional translation vector, as shown in the two dimensional case, to represent a general rigid transformation in a single matrix. This will consist of six parameters - rotation angles about and translations in the x , y and z directions.

The rigid transformation model may be generalised to affine transformations by the addition of scaling and shearing. This gives a set of transformations that allow changes to shapes and volumes but preserves collinearity and coplanarity of points.

Affine transformations may also be used to initialise a nonlinear registration, or when registering a PET image to an MRI image of the same subject, as this may involve some scaling. A scale factor can be applied independently to each of the x , y and z axes:

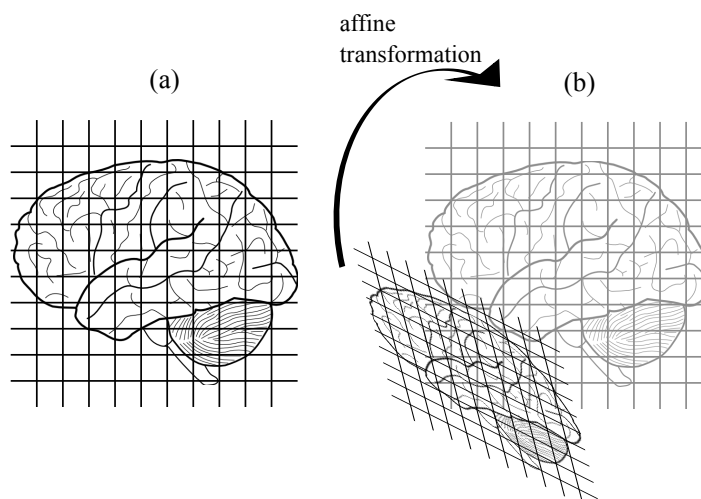


Figure 2.7: Affine transformation for registration. (a) shows a target image; (b) shows a floating image brought into alignment with (a) by an affine transformation consisting of a scaling and shear plus rigid transformation.

$$\mathbf{T}_{sc} = \begin{bmatrix} sc_x & 0 & 0 & 0 \\ 0 & sc_y & 0 & 0 \\ 0 & 0 & sc_z & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (2.5)$$

Shears are parameterised by off diagonal components of a similar matrix. These components represent a displacement along a particular axis by an amount proportional to the coordinate in another axis. In the following matrix, for example, the term sh_{xy} is the part of shear along the y axis dependent on the x coordinate and sh_{yz} is the displacement parallel to the z axis that depends on y .

$$\mathbf{T}_{sh} = \begin{bmatrix} 1 & sh_{xy} & sh_{xz} & 0 \\ sh_{yx} & 1 & sh_{yz} & 0 \\ sh_{zx} & sh_{zy} & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (2.6)$$

Rigid, scaling and shearing matrices may be combined to form a general affine transformation matrix by multiplication, so $\mathbf{T}_{aff} = \mathbf{T}_{sh} \mathbf{T}_{sc} \mathbf{T}_{rigid}$. The resulting matrix would at first glance appear to have fifteen components (three translation, three rotation, three scaling, six shear) but these are not in fact independent and general affine transformations in three dimensions are represented by twelve components. In practice, affine transforms are parameterised by these twelve components rather than separate rotation, scaling, et cetera transformations.

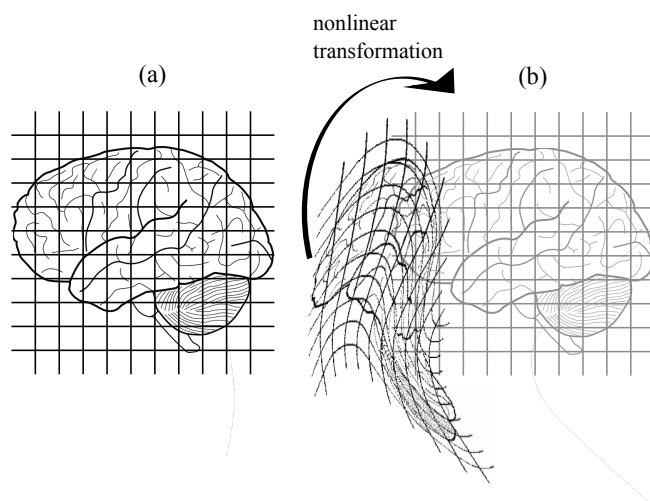


Figure 2.8: Nonlinear transformation for registration. (a) shows a target image; (b) shows a floating image brought into alignment with (a) by an affine, followed by a nonlinear registration.

Nonlinear transformation models are the most flexible used in registration, with far more parameters than affine transformations. Nonlinear registrations are typically used to accurately align images of different subjects, or scans of a single subject taken with a large time interval between so the anatomy has changed substantially.

A wide variety of nonlinear transformation models exist. The transformation model may be physically inspired, by representing the anatomy as able to deform like a viscous fluid [Bro-Nielsen and Gramkow, 1996] or an elastic material [Rohr, 2000]. This allows the transformation to be highly nonlinear while hopefully remaining anatomically plausible. However the approach used in most of our work is based on free-form deformation (FFD) [Rueckert et al., 1999]. This defines a large set of control points arranged on an axis aligned grid; the deformation is then a set of three dimensional vectors at each control point, so the number of parameters is three times the number of control points. To define the deformation field over the entire image, the control points are convolved with cubic B-splines, and the bending energy of these is included as a regularisation term to favour smooth solutions. The updated implementation we use [Modat et al., 2010] is accelerated using graphics processing units (GPUs) and yields a diffeomorphic transformation [Modat et al., 2012], meaning that the resulting mapping is one-to-one with no tissue disappearing via folds in the displacement field.

2.4.2 Interpolation methods

When the floating image is resampled to produce a version in the same space as the target image, in general the positions of voxels in the new image will not correspond exactly to voxels in the original floating image - instead they will coincide with points arbitrarily between voxel

centres. So to find the intensities for voxels in the new image, it is necessary to interpolate between voxels. There are a number of methods for doing so. Cubic spline interpolation offers the most flexibility; however, the resulting interpolated intensities may go outside the range of intensities in the original image. This may be undesirable, for example if the image being resampled is a probabilistic segmentation so intensities should be between zero and one. In this case trilinear interpolation can be used. If the image being resampled is an anatomical atlas or parcellation where structures are labelled with integer value intensities, nearest neighbour interpolation is generally used.

2.4.3 Objective functions

The role of the objective function is to quantify how good the alignment between the transformed floating image and the target image is. If the target image is designated \mathbf{I}_t , the floating image \mathbf{I}_f and the transformation \mathbf{T} , then the objective function is a function of \mathbf{I}_t and $\mathbf{I}_f(\mathbf{T})$. Each image can be regarded as a set of N corresponding voxels, $(\mathbf{t}_1 \cdots \mathbf{t}_N)$ for the target image and $(\mathbf{f}_1 \cdots \mathbf{f}_N)$ for the floating image. The choice of similarity measure is largely motivated by the modalities of the images being registered. Simpler measures are less computationally demanding, but assume a straightforward relationship between voxel intensities which does not hold for intermodal registration, and are less robust to noisy images or variations in scan parameters.

The simplest, sum of squared differences (SSD), directly measures the mean-squared difference between corresponding voxels.

$$SSD = \frac{1}{N} \sum_{n=1}^N (\mathbf{t}_n - \mathbf{f}_n)^2 \quad (2.7)$$

It is an appropriate metric only when the intensities of corresponding voxels in properly aligned images are the same. A rather more robust metric is normalised cross-correlation (NCC), which is able to register images with a linear relationship between voxel intensities. If the mean intensities of \mathbf{t} and \mathbf{f} are given by \bar{t} and \bar{f} respectively, NCC is defined as

$$NCC = \frac{\sum_{n=1}^N (t_i - \bar{t})(f_i - \bar{f})}{\sqrt{\sum_{n=1}^N (t_i - \bar{t})^2} \sqrt{\sum_{n=1}^N (f_i - \bar{f})^2}} \quad (2.8)$$

To define a similarity metric robust enough for intermodal registration, where the relationship between the intensities of corresponding voxels can be complex, ideas from information theory are used. This leads to the metric of mutual information (MI) [Wells et al., 1996]. MI can be seen as a measure of how much information is shared by two variables, or how much

having one variable reduces uncertainty in the other. If the variables were completely independent, MI would be zero, whereas if one is a function of the other MI would be maximised. More formally, it measures the difference between the joint entropy of two variables and their conditional entropies.

To compute MI from pairs of images, a joint histogram of the images is usually constructed. This is done by putting the intensities of the target image \mathbf{t} and the floating image \mathbf{f} into a set of bins, each containing a range of intensity values. The joint histogram \mathbf{H} consists of entries representing pairs of binned intensities (t, f) , and each entry counts the number of times a particular intensity pair (t, f) has co-occurred at voxels in the same location in the two images. Then the joint probability distribution for intensity pairs $p(t, f)$ may be estimated as $p(t, f) = \frac{\mathbf{H}(t, f)}{N}$ where N is the total number of entries in \mathbf{H} . We may then estimate the marginal probabilities of binned intensity t in \mathbf{t} and f in \mathbf{f} , $p(t)$ and $p(f)$, and use them to calculate the Shannon entropies H [Shannon, 1948] of the image intensities:

$$H(\mathbf{t}) = - \sum_t (p(t) \log(p(t))), \quad H(\mathbf{f}) = - \sum_f (p(f) \log(p(f))) \quad (2.9)$$

Similarly, we can define the image pair's joint entropy:

$$H(\mathbf{t}, \mathbf{f}) = - \sum_t \sum_f (p(t, f) \log(p(t, f))) \quad (2.10)$$

The MI can then be defined as

$$MI(\mathbf{t}, \mathbf{f}) = H(\mathbf{t}) + H(\mathbf{f}) - H(\mathbf{t}, \mathbf{f}) \quad (2.11)$$

which is often made more robust by using the normalised mutual information (NMI) [Studholme et al., 1999].

2.4.4 Optimisation

An optimisation algorithm is required to find the set of transformation parameters giving the best registration. Typically this is done by finding the maximum of an objective function, which is a combination of a similarity term derived from one of the metrics described in the previous section and, for nonlinear registration, a penalty term to stop the transformation from becoming overly complex. As stated before, the penalty term in FFD is based on the bending energy of the transformation, whereas in other nonlinear registration methods the transformation is constrained by the physical model it is based on. Most optimisation algorithms used in registration are gradient based, meaning that they require derivatives of the similarity metric and penalty

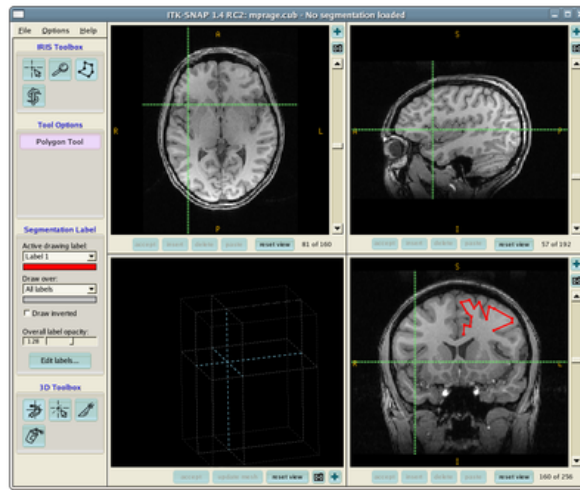


Figure 2.9: Manual segmentation of brain structures with ITK-snap software. From <http://www.itksnap.org/docs/viewtutorial.php?chapter=TutorialSectionIntroduction>

term, if any, with respect to the transformation parameters. The nonlinear registrations used in the work in this thesis use conjugate gradient ascent, which is faster than the simple steepest ascent approach [Modat et al., 2010].

2.5 Anatomical segmentation

To allow detailed analysis of medical images, it is often necessary to apply a label to all voxels in an image, parcellating it into anatomically defined structures. This can then be used to define regions used for normalising PET images, mask out regions that we do not wish to include in further analysis, or define features for classification or regression at a regional (as opposed to voxel or global) level. A number of methods can be used for anatomical segmentation, which we present here.

2.5.1 Manual segmentation

Manual segmentation of brain images by a trained human expert can produce high quality anatomical segmentations that are considered to be the gold standard for this process. Although experience plays a major role in the process, all raters must follow a carefully drafted protocol that exactly defines how each anatomical structure should be delineated, in order that intra- and inter- rater variability is kept as low as possible. The term manual segmentation may be considered something of a misnomer as software is often used to assist the process as shown in figure 2.9.

However even with the help of specialist segmentation tools, manual anatomical segmentation is a slow process. It is also subject to limited repeatability. To provide an objective,

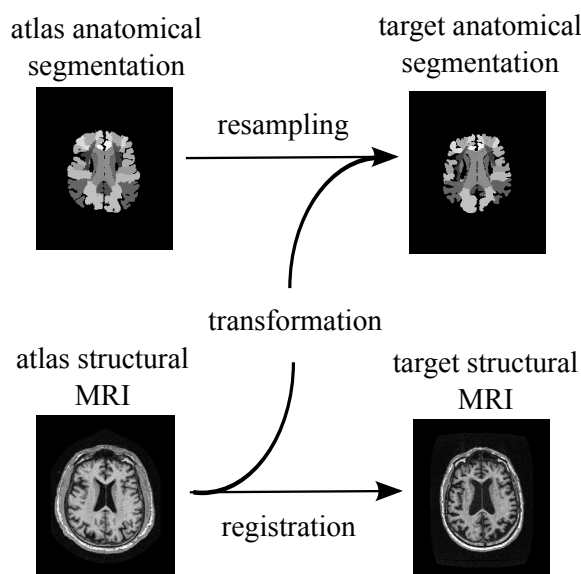


Figure 2.10: Segmentation propagation. An atlas structural image is registered to a target structural image, and the resulting transformation is used to warp the labels into the space of the target.

repeatable segmentation and to segment large numbers of images in a reasonable time, automated methods must be employed.

2.5.2 Automatic segmentation

Automatic anatomical segmentation methods are largely based on the concept of an atlas - a structural brain image, coupled with a set of anatomical labels for the image voxels in the same space. The labels in the atlas are generally provided by manual segmentation. Hence automatic segmentation still ultimately relies on the existence of *some* slow to produce manually labelled images. However, it provides a way to use a relatively small number of these to rapidly and accurately segment a much larger number of images.

An example of a widely used brain atlas was produced by the international consortium for brain mapping (ICBM) [Mazziotta et al., 2001]. To be more representative of the variation in human brain anatomy, this is based on a template produced by the Montreal Neurological Institute (MNI) by affinely aligning and averaging a set of 152 healthy brain MR images. The resulting template is known as MNI space and is also widely used as a standard space for brain image analysis.

The atlas can be used to accurately segment a new brain image automatically by segmentation propagation. First of all, the structural MR image associated with the atlas is accurately registered to the brain image to be segmented. This will typically be initialised with a rigid and then affine registration, followed by a nonlinear step to align structures locally. The resulting

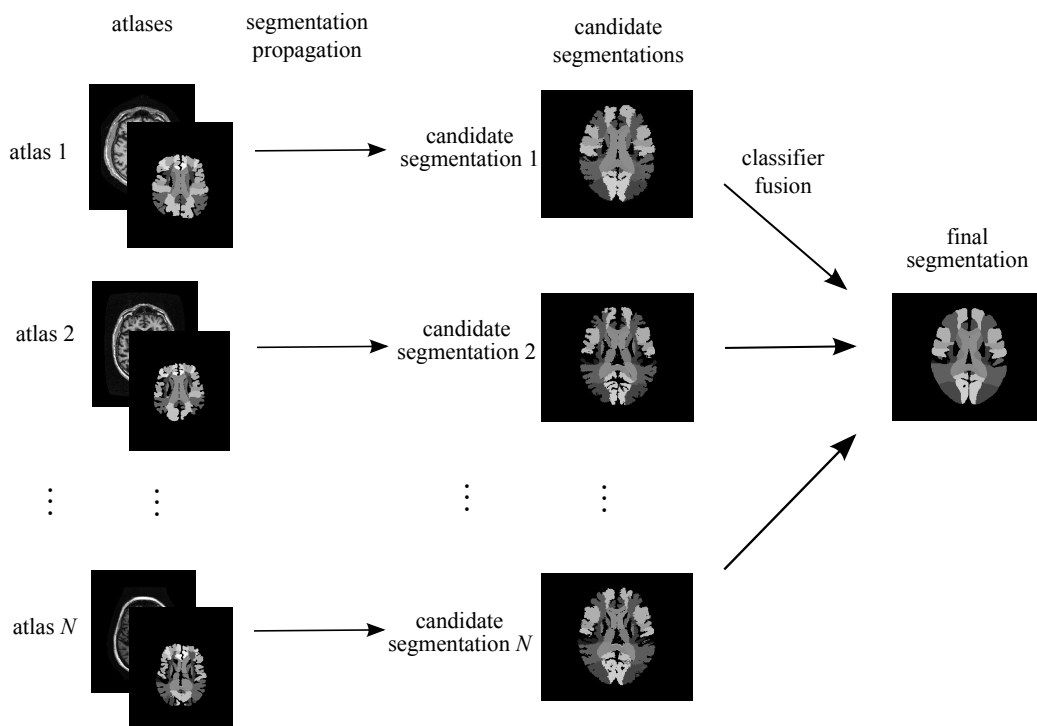


Figure 2.11: Multiatlas segmentation. Anatomical segmentations from a set of N atlases are propagated to the space of the target image. The resulting N candidate segmentations in the target image space are then fused to produce a final anatomical segmentation.

final transformation is then applied to the label image, resampling with nearest neighbour interpolation to maintain the correct integer labels. The result is a set of anatomical labels in the space of the target image. The segmentation has been propagated from the space of the atlas to the space of the target image. This process is shown in figure 2.10.

Results can be improved by registering a 'library' consisting of multiple atlases to the target image. Our work is based on a set of 20 atlases of healthy subjects. These were initially parcellated into 49 anatomically defined regions. The regions, and the protocols used to define them for their initial manual segmentation, are described in [Hammers et al., 2003]. A subset of the regions were further subdivided to create a set totalling 83 regions [Gousias et al., 2008].

This generates multiple anatomical segmentations in the space of the target image. Each of these can be seen as a classifier, assigning a class to each voxel of the target image based on the label of its corresponding voxel. To produce a single, final segmentation classifier fusion methods must be applied. The simplest method is majority voting, where the label of each voxel in the final segmentation is the most common label in the N corresponding voxels across the N segmentations being fused. This approach was shown in [Heckemann et al., 2006] to produce a final segmentation of a higher quality than that made from any single atlas in the library.

Both the segmentation propagation and label fusion steps may be improved. If many atlases are available, the final segmentation can be improved by using only a subset of atlases, whose anatomy is most similar to the target image by some metric [Aljabar et al., 2009]. The segmentations can also be propagated across a manifold representing the variability of brain anatomy, using intermediate anatomies between the atlas and target as stepping stones [Wolz et al., 2010], or the manifold can be used to improve selection of a subset of atlases [Hoang Duc et al., 2013].

For improved label fusion, the simulated truth and performance level estimation (STAPLE) algorithm [Warfield et al., 2004] has been used to form a probabilistic estimate of a correct underlying segmentation from a set of candidate segmentations of the same structure, using the expectation-maximisation algorithm to update the weight and estimated performance level of each candidate segmentation. However STAPLE is designed for single label (background or structure of interest) fusion, and so cannot be used to combine segmentations of multiple anatomical regions. Multi-label similarity and truth estimation for propagated segmentations [Cardoso et al., 2012] extends the STAPLE model to multiple labels and adds extra smoothness constraints and resistance to bias due to structure size, further improving label fusion.

2.6 Tissue segmentation

Whereas anatomical segmentation is intended to parcellate the brain into anatomically meaningful structures, tissue segmentation segments the brain into its major constituent tissue types. These are normally considered to be grey matter (GM), white matter (WM) and cerebrospinal fluid (CSF). For applications in AD, atrophy in the cortex means that alterations to the structure of GM have been the primary object of study. More recently, alterations to WM have become of increasing interest although these are more usually studied with diffusion imaging.

2.6.1 Expectation maximisation

Tissue segmentation is almost always done by modelling the three tissue types as Gaussian distributions of voxel intensities, and thus the entire image as a Gaussian mixture model. The model parameters are thus the mean and standard deviation for each class of tissue. The parameters are set using the expectation-maximisation (EM) algorithm [Dempster et al., 1977]. This is a probabilistic method that alternates between an expectation step, which calculates a function giving the expectation for the likelihood of the data based on the current estimated model parameters, and a maximisation step that updates the parameters by maximising the the expected likelihood found according to the previous expectation step. This continues until the change in log likelihood between steps falls below a predefined threshold. As the result of the

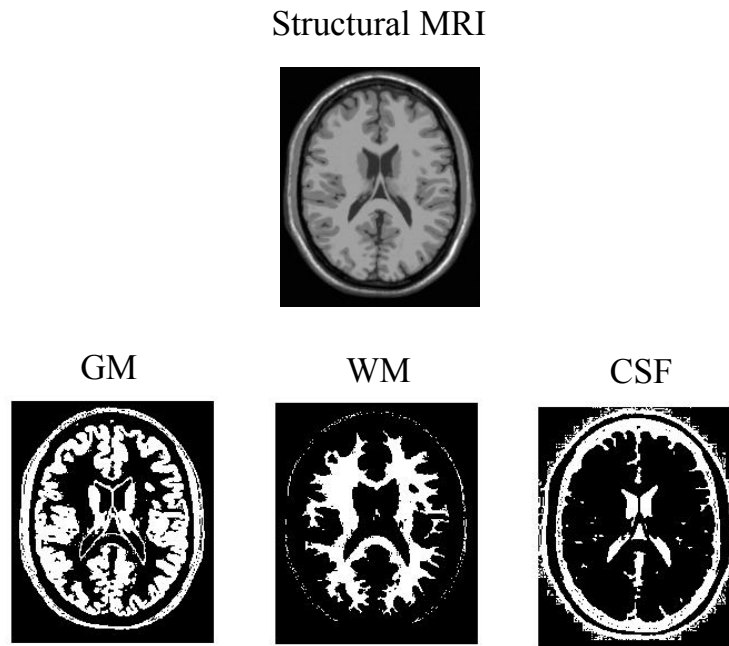


Figure 2.12: A structural MR brain image (top) is segmented into three tissue components shown on the bottom row: GM (left), WM (centre) and CSF (right). As the segmentations are probabilistic, their voxel intensities range from 0 to 1.

EM algorithm is a Gaussian distribution for the intensities of each tissue class, the resulting segmentation for a tissue class is often shown as a probability map, where the intensity in each voxel represents the probability of the voxel belonging to that class. Hence segmentation into GM, WM and CSF components gives three probability maps as shown in figure 2.12. These are sometimes combined into a single hard segmentation by assigning each voxel to the tissue type with the highest probability.

2.6.2 Extensions to the expectation maximisation model

As figure 2.12 shows, pure EM is not sufficient to produce high quality brain tissue segmentations. It is clear that a large amount of non brain material such as the skull, skin, dura and so forth has been included as brain tissue. This is because the intensity distributions of these strongly overlap with the tissue type we wish to include, so they cannot be distinguished with intensity information alone. Non brain material can be removed in a preprocessing step, or alternatively it can be excluded from the segmentations, and the overall quality of the segmentations improved, by introducing spatial information in the form of priors. These are derived from a brain atlas that contains probabilistic information on the spatial distribution of GM, WM, and CSF. The prior is applied by multiplying the estimated segmentations for each tissue class by

the corresponding prior in each iteration of the EM algorithm. This means, for example, that a voxel containing skull that would be classified as GM by intensity alone would instead be assigned a very low GM probability as it is in a location that is very unlikely to be GM according to the prior. To establish the spatial correspondence in order for this to work, the priors must be registered to the image being segmented before the algorithm is applied. In cases where the anatomy is considerably different from the priors, they can be relaxed to prevent them from having too strong an influence [Cardoso et al., 2011].

An early application of EM to brain image segmentation [Van Leemput et al., 1999b, Van Leemput et al., 1999a] included two additional enhancements to the EM algorithm alongside the use of priors. The first of these is the use of a Markov random field (MRF) to promote smoothness in the segmentations, by incorporating information from the estimated segmentation of each voxel's local neighbourhood into the EM model. This reduces the effects of noise in the image being segmented. The second is correction for a bias field due to EM signal inhomogeneities, which can result in voxels containing identical tissue having different intensities in a manner that varies smoothly across the image. Both of these are introduced fully into the EM model as an extra step interleaved with the expectation and maximisation steps rather than post hoc alteration of the segmentations. Additionally, the separate segmentation, registration of priors to the target image and bias field correction steps can be combined in a single generative model [Ashburner and Friston, 2005].

Chapter 3

Machine learning

3.1 Introduction

Machine learning is a branch of artificial intelligence (AI) that deals with algorithms that can learn from data. Whereas classical AI was concerned with developing general intelligence and emulating human thought [Turing, 1950], machine learning is largely stripped of any philosophical baggage and is instead focused on solving particular problems [Mitchell, 1997]. Such problems can be from a very wide variety of areas, such as optical character recognition, face recognition, and, as in the rest of this document, medical diagnosis. The usual approach is to have a large body of example items from which the algorithm can learn, referred to as the training data. Each example consists of a vector of d features, so each example can be seen as a point in a d dimensional space. If the features consist of the voxels of a large, high resolution image then d can potentially be in the millions; however in the following toy examples the dimensionality is two to allow the results of algorithms to be visualised.

3.2 Machine learning taxonomy

There are many types of machine learning algorithm; however, there is a fundamental distinction between unsupervised and supervised learning. In the former, each training example is presented to the algorithm without any extra information about it, and the task is to discover some structure in the training data from its distribution. A classic unsupervised learning task is clustering - uncovering natural groupings in the data. This is shown in figure 3.1.

In supervised learning, each point in the training data also has an attached value. The object of the algorithm is to use the training data to learn a function that, when presented with a hitherto unseen sample, can accurately predict its corresponding value. If the value to be predicted is continuous, then this is a regression problem, and the attached values are called targets. Simple least squares is an example of an algorithm that can solve some regression problems. If the value to be predicted indicates which of a set of discrete groups each example

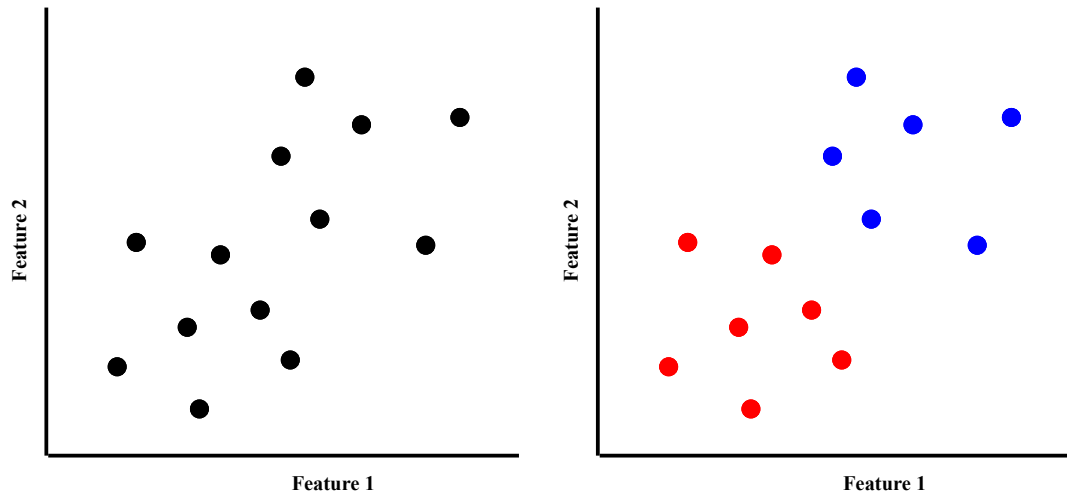


Figure 3.1: Clustering unlabelled data

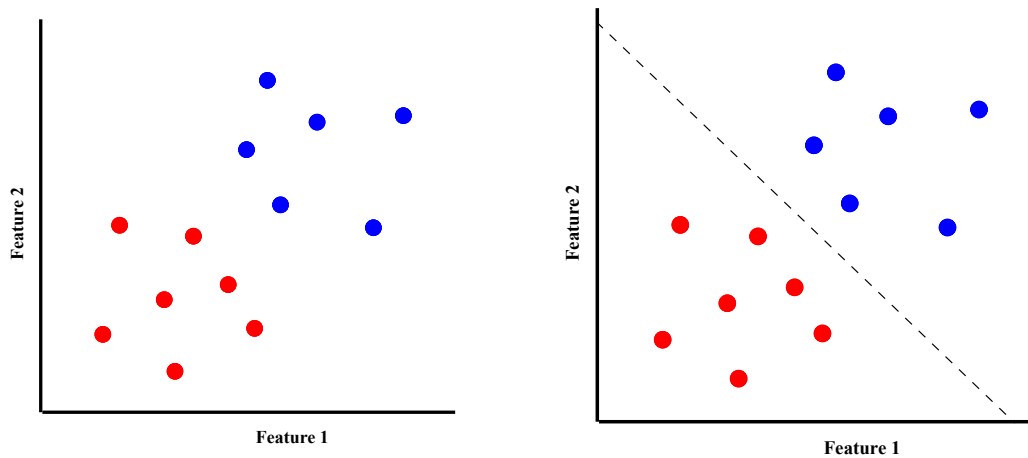


Figure 3.2: Classification of labelled data

belongs to, this is a classification problem. Most commonly, there are only two such groups, in which case it is a binary classification problem and the labels are generally $\{-1, 1\}$. However multiclass classification with three or more groups is also possible; although algorithms and performance measures are more complex in this case. Binary classification is illustrated in figure 3.2.

It is also possible to use a small quantity of labelled data mixed with a much larger quantity of unlabelled data to improve classification performance; this is known as semi-supervised learning and has also been applied to the problems addressed in this thesis.

Finally, a relatively rare approach to some machine learning problems is known as transfer learning. This is generally defined as applying the knowledge gained in one problem domain, and applying it to a different but related domain, with the aim of learning representations that generalise across the problems. This thesis uses the term to describe the application of a classifier trained on AD and control subjects to MCI subjects, in order to predict conversion from

MCI to AD.

3.3 Preprocessing of data

A variety of operations can be performed on data to improve classification or regression performance, before a machine learning algorithm is used. These generally fall into two categories: feature extraction, where some transformation is applied to the existing data to generate a new set of features from the original ones, and feature selection, where a subset of the original features are selected to be used in classification or regression. The two types can also be combined sequentially in either order.

3.3.1 Feature extraction

When the data being used are medical images, the line between image processing and feature extraction can be very blurry. For example, averaging or summing over voxel intensities within different anatomical regions could be seen as feature extraction, as could calculating cortical thickness from a structural MRI scan, although this would normally be seen as part of the image processing pipeline. More general feature extraction techniques involve changing the representation of the data. For example, it is common to use a z -transform to standardise all features to zero mean and a standard deviation of unity, to prevent features with a large range from dominating others which have a smaller range but may be more informative. Dimensionality reduction techniques are frequently used as well. Principal component analysis (PCA) [Jolliffe, 2005] is the most popular linear technique, seeking to represent the data in a lower dimensional linear subspace that retains most of the variance, while hopefully reducing the noise level. Recently nonlinear techniques such as manifold learning [L J P Van Der Maaten, 2007] have also become popular. These see the data as lying on a nonlinear manifold of low dimension, embedded in the higher dimensional original data space. Manifold learning algorithms (themselves a type of unsupervised learning) attempt to recover the structure of the manifold, and then data points are represented by a position in the manifold coordinate system.

3.3.2 Feature selection

Feature selection methods come into three broad categories: filters, wrappers and embedded methods [Guyon and Elisseeff, 2003]. Filters apply a simple criterion to each individual feature in turn, and reject the features where the resulting test statistic is over a predetermined threshold. For example, a t -test might be used to check if a feature separates the two groups to be classified with $p < 0.05$. For regression, a feature might be selected if it correlates with the target variable sufficiently strongly. Wrappers make use of the learning algorithm: a set of features is used to learn a model, and the results are assessed and compared to other sets of

features. As the number of possible sets of features is given by 2^D , D being the dimensionality, an exhaustive search rapidly becomes intractable for more than a trivially tiny set of features. Instead, heuristics are used to progressively add features to an empty set or remove them from a full set, or a combination of the two. A more complex wrapper method is recursive feature elimination [Guyon et al., 2002] which uses the weights of a trained linear model to infer and rank the importance of corresponding features. Both filters and wrappers have the disadvantage of needing target or label information to select features, which means great care must be taken to avoid double dipping. This can be avoided by using other types of feature selection. Embedded methods are machine learning algorithms that are designed to drive most of the weights of a linear classifier to zero. Most frequently, this is done with a sparsity favouring regularisation term, such as the ℓ_1 norm of the weight vector w as in the LASSO [Tibshirani, 1994], or for Bayesian methods, a sparsity inducing prior can be applied to the weights. Such methods combine the feature selection and training steps into a single operation. Perhaps the broadest method, however, is to remove features based on prior knowledge of the problem domain. For example, for AD classification, we know which regions of the brain are affected by the disease processes, so features representing other regions may be eliminated. This is probably the most effective method of feature selection for this type of problem [Chu et al., 20].

3.4 Performance measurement and validation in machine learning

Clearly, when a classification or regression model has been built, it is important to estimate how well the model will perform in practice. To do this, the model must be applied to a set of data for which the ground truth is known so it is possible to assess the performance of the model. This is known as the test set. To minimise any bias, the test set must not contain any subjects used to train the classifier, as inclusion of these will obviously falsely inflate the performance of the model. Simply dividing a set of data into training and testing points is easy, and there are many ways to do this which are discussed in the section on validation strategies. However, there are subtleties in which this separation can unintentionally be violated in the experimental design, which we discuss in a later section.

3.4.1 Performance measures for classification

The simplest and most widely used performance measure for classifiers is the accuracy. This just expresses the fraction of data points in the test set which are correctly classified. In a typical classification experiment with a medical application, the test subjects will be divided into patients and controls. Based on this and the classification results, we can further split the test set into patients who are classified as patients (true positives or TP), patients who are

results

Ground truth	Prediction	
	patient	control
patient	TP	FN
control	FP	TN

Table 3.1: Confusion matrix for classification

incorrectly classified as controls (false negatives or FN), controls who are classified as controls (true negatives or TN), and controls who are wrongly identified as patients (false positives or FP). This is summarised in table 3.1.

So the accuracy can be expressed as

$$accuracy = \frac{TP + TN}{TP + FN + FP + TN} \quad (3.1)$$

This can be broken down into individual accuracies for the ground truth patients and the ground truth controls. In medical applications, the resulting quantities are known as the sensitivity (proportion of ground truth patients, or more generally proportion of ground truth positive class subjects, which are correctly classified) and specificity (proportion of ground truth controls, or more generally proportion of ground truth negative class subjects, which are correctly classified):

$$\begin{aligned} sensitivity &= \frac{TP}{TP + FN} \\ specificity &= \frac{TN}{TN + FP} \end{aligned} \quad (3.2)$$

If we know the sensitivity and specificity, and the proportion of ground truth positive and negative class subjects in the test set, we can calculate the accuracy:

$$accuracy = \left(sensitivity * \frac{P}{N + P} \right) + \left(specificity * \frac{N}{N + P} \right) \quad (3.3)$$

This shows the danger in reporting accuracy alone. If the test data is very unbalanced (having many more subjects from one class than the other) a very high accuracy may mask a very high sensitivity and very poor specificity, or vice versa. To take an extreme case, if a classifier does not classify at all but simply assigns all test data to the positive class, then a test data comprising 99 points from the positive class and only one from the negative class would produce an accuracy of 99%, despite having a specificity of zero. Best practice is to report sensitivity and specificity as well as accuracy. To give a single overall statistic, it is also

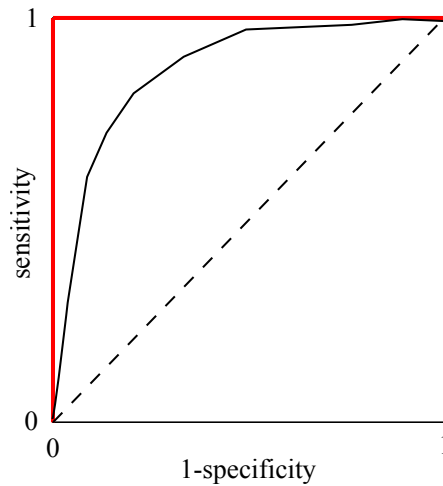


Figure 3.3: ROC curves. The red line shows the tradeoff between sensitivity and specificity for a hypothetical perfect binary classifier. A random classifier would be represented by the diagonal line, and most real binary classifiers would have an ROC curve resembling the intermediate solid black one.

possible to report the balanced accuracy, which is often defined as the mean of sensitivity and specificity.

Other alternatives are based on the idea of varying thresholds. All classifiers produce a decision value (DV). This is a scalar which is thresholded to determine to which class a test data point belongs. For example, support vector machines have a DV which is the signed distance to the separating hyperplane. Normally the class of a test subject is decided by the sign of this DV, i.e. it is thresholded at 0. However, we may trade off sensitivity against specificity by changing this threshold. We make extensive use of this, providing an alternative balanced accuracy as the accuracy obtained at the threshold value where the difference between sensitivity and specificity is smallest. The same ideas are used in receiver operating characteristic (ROC) curve. This summarises the tradeoff between sensitivity and specificity, by calculating them for all possible thresholds.

The ROC curve may be used where there are particular costs attached to false negatives and false positive, to find a threshold that minimises the total expected cost. The area under the ROC curve (AUC) is also widely used as a measure of classifier performance as it is aggregated over all thresholds. It can be interpreted as the probability that a randomly chosen test subject from the positive class will have a greater dv than a randomly chosen subject from the negative class [Fawcett, 2006]. In a perfect classifier, all subjects in the positive class will have a greater dv than subjects in the negative class so this is equal to one.

3.4.2 Performance measures for probabilistic classification

All the measures defined in the previous section can be applied to probabilistic classification. However the probabilistic predictions also enable some more options. In particular, we may want to know if the probabilities are well calibrated - in other words, are predictions made with a higher degree of confidence more likely to be correct? This can be measured with the Brier score [Brier, 1950]. This was originally proposed for assessing the quality of probabilistic weather forecasts and is given by

$$BS = \frac{1}{N} \sum_{i=1}^N (p_i - c_i)^2 \quad (3.4)$$

where N is the number of test subjects, and p_i is the predicted probability of the i th subject belonging to its true class c_i , $c \in \{0, 1\}$. Another method is to plot an error-reject curve. If the confidence threshold below which we reject a probabilistic prediction is increased, then the error rate among the retained predictions should smoothly decrease if the predictions are well calibrated. We also introduce some novel metrics for probabilistic predictions in later sections.

3.4.3 Performance measures for regression

There are two simple and widely applied performance metrics for regression. They are the Pearson correlation r between the actual target values \mathbf{y} and predicted target values $\hat{\mathbf{y}}$

$$r = \frac{\sum_{i=1}^N (y_i - \bar{y}_i)(\hat{y}_i - \hat{\bar{y}}_i)}{\sqrt{\sum_{i=1}^N (y_i - \bar{y}_i)^2} \sqrt{\sum_{i=1}^N (\hat{y}_i - \hat{\bar{y}}_i)^2}} \quad (3.5)$$

where \bar{y} and $\hat{\bar{y}}$ are the sample means of the actual and predicted targets values, and N is the number of data points, and the root mean squared error (RMSE)

$$RMSE = \frac{\sum_{i=1}^N (y_i - \hat{y}_i)}{N} \quad (3.6)$$

3.4.4 Validation strategies

As previously stated, we must measure the performance of our learning algorithms on a test data set that is separate from the training data. However, an independent source of test data may not be available, or the amount of data available may be so small that splitting it into training and testing sets would result in both being very small. The solution is to adopt cross validation. In this, the available data points are split into k roughly equally size folds, or groups. Then, each fold in turn is left out as a test set, and the remaining $k - 1$ folds together form the test set. After this procedure has been repeated k times, each data point has been used for testing once and for training $k - 1$ times. The results on the different test sets may then be averaged together to

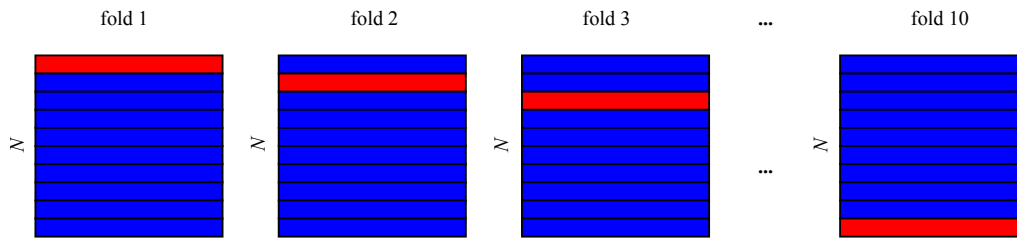


Figure 3.4: LOOCV in a set of ten data points. Each data point in turn is left out as testing data (red) while the other nine are used for training (blue). The accuracy of a classifier built using all ten points to train on new data is estimated by the accuracy across the ten left out data points.

form an overall estimate of accuracy. This is known as k -fold cross validation. If k is equal to the number of data points N this is known as leave-one-out cross validation (LOOCV). As the training set for each fold is only one smaller than the entire data set, this is a nearly unbiased estimator for the accuracy of a classifier constructed from the entire data set on unseen data [Cawley and Talbot, 2004]. However it is subject to greater variance than cross validation with fewer folds.

For classification, if the data set is unbalanced between classes, the folds may be stratified, so the proportion of the classes in each fold is roughly reflective of the entire data set, to improve accuracy estimation [Kohavi, 1995]. Cross validation can also be applied to estimate accuracy for the purpose of parameter tuning. In this case, tenfold cross validation is preferable due to its better lower variance than LOOCV. When doing this, however, care must be taken to avoid double dipping.

An alternative to k -fold cross validation is Monte Carlo cross validation [Xu et al., 2004]. This randomly partitions the data into training and test sets of a size determined by the user, and can be repeated as many times as desired. This has a lower variance than k -fold cross validation, but may be more biased as there is no guarantee each data point is used in training and testing.

Comparing, as opposed to estimating the accuracies of classifiers is still a difficult problem. McNemar's test [McNemar, 1947] may be used to calculate a p -value for the difference in accuracies for two classification methods applied to the same set of test data, by generating a χ^2 statistic from the number of datapoints switching from rightly to wrongly classified and vice versa between the two methods. Alternatively confidence intervals may be constructed around the estimated accuracy [Newcombe, 1998].

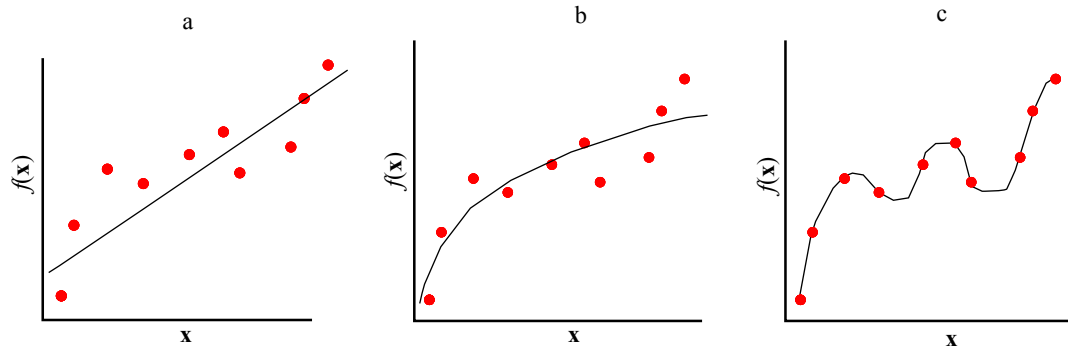


Figure 3.5: Levels of fitting of a function $f(x)$ to a sample of points. Plot (a) shows underfitting; the linear model is insufficiently flexible to capture the variation in the underlying function that generates the points. Plot (b) shows a correct level of fit. Plot (c) shows overfitting; error on the sampled points is zero but the model will likely generalise very poorly to new data.

3.5 Pitfalls of machine learning experiments

Two common errors in performing machine learning experiments are discussed here; the first concerns an error which leads to poor predictive performance and the second to positively biased predictions of performance. The two are, however, quite closely related. In the neuroimaging domain, which typically has relatively few data points of possibly very large dimensionality, the potential for both to cause problems is potentially grave.

3.5.1 Overfitting

Overfitting describes a situation in which an overly complex model no longer fits the (unknown) underlying function that describes the training data, and instead fits the noise characteristics of the training data. It is illustrated in figure 3.5.

Many successful machine learning techniques are designed around a method to combat overfitting, such as maximising the margin in support vector machines, or model averaging and maximum likelihood parameter setting in Gaussian processes. Determining when overfitting is happening is challenging as it is difficult to detect in training data alone. In fact the hallmark of overfitting is regime where increasing the model's flexibility continues to reduce training error, but error on an independent test set begins to rise.

3.5.2 Double dipping

As previously stated, good estimates of generalisation performance in supervised learning come from applying a model to a set of test data that were not used in training. While this statement is self evident and apparently simple, it is possible for it to be violated in surprisingly subtle ways, as some experimental designs accidentally introduce circularity into the analysis by allowing

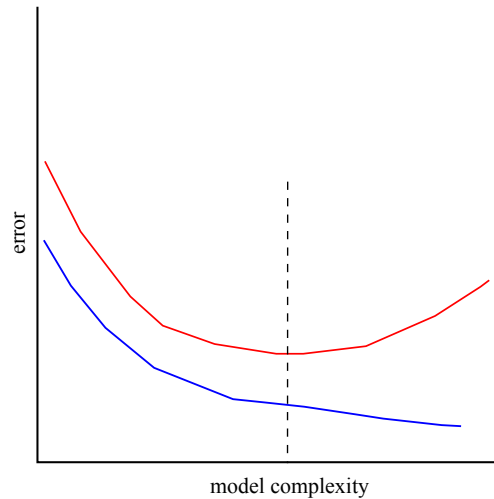


Figure 3.6: As the complexity of a model is increased, the training error (blue curve) may approach or reach zero. However after initially decreasing, the testing error (red curve) will begin to rise as an overly complex model begins to overfit the training data. Testing error is minimised at the appropriate level of complexity (dashed vertical line).

the labels or targets of test data to be used in generating the model, allowing information to leak from the test set to the training set. Most commonly, this occurs when a method requires a number of parameters to be set, or when supervised feature selection or extraction methods are applied. An experimenter may, for example, apply a feature selection method to choose features based on all data points and their labels. After this is done, the accuracy of the resulting classifier is assessed in 10-fold CV. This introduces circularity to the data, as the labels of data points used to evaluate the classifier are also used in constructing the classifier. The result is inflated accuracy estimates as it results in testing a hypothesis suggested by the data, which can lead to the detection of spurious effects [Kriegeskorte et al., 2009]. The same effect has been noticed in a number of papers directly relevant to this thesis, inflating the accuracy of AD classification [Eskildsen et al., 2013]. Double dipping may be understood by expanding the definition of training to also include parameter setting feature selection or extraction, or training data selection (which may be grouped under model selection), as well as model optimisation. Seen in this way, it clearly violates separation of training and testing data.

In practice, the problem can be avoided by modifying the experimental design. For example, an experimenter may use three groups of disjoint data, for training, tuning and testing an algorithm, rather than training and testing groups alone. If cross validation is being used, then the situation is a little more complex. A common approach is to perform a separate CV loop within the training set of *each* fold of a CV loop. The inner CV loop can be used for parame-

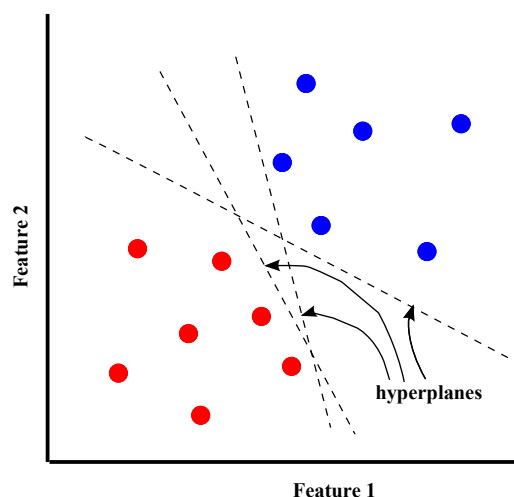


Figure 3.7: Many hyperplanes can divide two linearly separable groups of points.

ter tuning et cetera and the outer loop for accuracy estimation, a method known as nested CV loops [Varma and Simon, 2006]. Many different validation strategies may even be combined in a complex pipeline, as long as separation of model selection and training data from testing data is maintained.

3.6 Support vector machines

The support vector machine (SVM) is possibly the most widely used algorithm for classification. If the training subjects represent points in a d -dimensional space, the SVM constructs a $d - 1$ dimensional surface (a hyperplane) that separates the two classes of training points. In general, as shown in figure 3.7, there will be an infinite set of possible hyperplanes that do this.

SVMs select a separating hyperplane based on the principle of structural risk minimisation [Cortes and Vapnik, 1995]. This controls the tradeoff between fitting the training data well and also offering good generalisation behaviour, that is making high quality predictions on samples outside of the training set. In the framework of SVMs, this translates to a simple geometrical intuition: the best hyperplane is the one which maximises the distance to the closest points in each class. This distance is known as the margin; hence SVMs are maximum-margin classifiers. The hyperplane is a function only of the subset of training points which lie on the margin. These are known as the support vectors, from which the SVM takes its name. This is illustrated in figure 3.8.

More formally, a training data point may be seen as vector $\mathbf{x} = (x_1, x_2, \dots, x_d)$ in a d dimensional space. The aim of the SVM is to produce a function $y(\mathbf{x})$, so that the sign of $y(\mathbf{x})$ indicates on which side of the separating hyperplane \mathbf{x} lies, and hence its class.

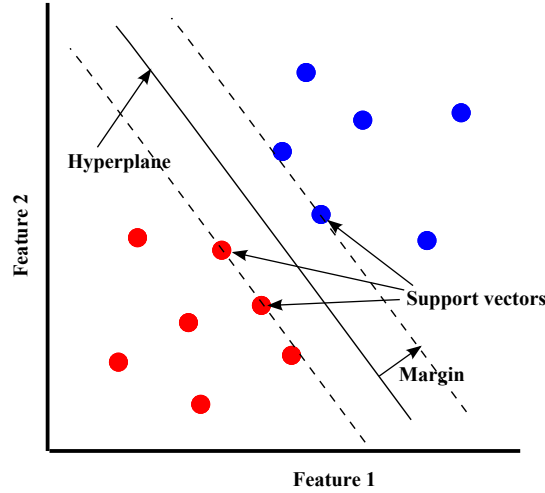


Figure 3.8: The SVM chooses the hyperplane that maximises the margin.

3.6.1 Linear SVMs

Support vector machines were originally conceived as a purely linear classifier for separable data [V. Vapnik, 1963]. The equation for the separating hyperplane is $y(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b = 0$, where w is a vector normal to the separating hyperplane and b a bias term. The margin is then defined by the hyperplanes $y(\mathbf{x}) = 1$ and $y(\mathbf{x}) = -1$. We can then see that the distance between these planes, which is the size of the margin, is given as $2/\|\mathbf{w}\|$. Hence we want to minimise $\|\mathbf{w}\|$ without allowing any points to fall inside the margin. If N training data points of dimensionality D $\mathbf{x}_i \in \mathbb{R}^D$ have corresponding labels $\mathbf{y} = (y_1, y_2, \dots, y_N)$, $y_i \in \{-1, 1\}_{i=1}^N$ this gives the following optimisation problem:

$$\begin{aligned} \min_{\mathbf{w}, b} \quad & \|\mathbf{w}\| \\ \text{subject to} \quad & y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 \end{aligned} \quad (3.7)$$

which is equivalent to

$$\begin{aligned} \min_{\mathbf{w}, b} \quad & \frac{1}{2} \|\mathbf{w}\|^2 \\ \text{subject to} \quad & y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 \end{aligned} \quad (3.8)$$

By introducing Lagrange multipliers α this can be expressed as

$$\begin{aligned} \min_{\mathbf{w}, b} \max_{\alpha} \quad & \left\{ \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{i=1}^N \alpha_i \left[y_i(\mathbf{w}^\top \mathbf{x}_i + b) - 1 \right] \right\} \\ \text{subject to} \quad & \alpha \geq 0 \end{aligned} \quad (3.9)$$

Which can be solved using standard quadratic programming methods, yielding a solution for \mathbf{w} as a linear combination of the training data points, $\mathbf{w} = \sum_{i=1}^N \alpha_i y_i \mathbf{x}_i$. This identifies

which training points are the support vectors, as they are the only ones whose corresponding α is nonzero. The bias b can be found from a single support vector, but a more reliable method is to average over all of them:

$$b = \frac{1}{N_{SV}} \sum_{i=1}^{N_{SV}} (\mathbf{w}^\top \mathbf{x}_i - y_i) \quad (3.10)$$

By substituting the expression for \mathbf{w} back in to equation 3.9 we can write the optimisation problem in its dual form:

$$\begin{aligned} \max_{\alpha} \tilde{L}(\alpha) = & \max_{\alpha} \left\{ \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) \right\} \\ \text{subject to} & \quad \alpha_i \geq 0 \text{ and } \sum_{i=1}^N \alpha_i y_i = 0 \end{aligned} \quad (3.11)$$

Here the kernel function k is just the dot product so $k(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^\top \mathbf{x}_j$.

3.6.2 Soft-margin SVMs

If the distributions of the two classes in training data overlap, then the classes are not linearly separable and the SVM algorithm will fail. A modified formulation, the soft margin SVM [Cortes and Vapnik, 1995], was introduced. This was done by adding slack variables ξ_i which measure the amount of error in classification of training data. Clearly allowing a greater error in the training set will allow a larger margin among the training points that are *not* misclassified, and this tradeoff between margin size and training error is controlled by a parameter C . The optimisation then becomes

$$\begin{aligned} \min_{\mathbf{w}, \xi, b} & \left\{ \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^N \xi_i \right\} \\ \text{subject to} & \quad y_i (\mathbf{w}^\top \mathbf{x}_i) - b \geq 1 - \xi_i \text{ and } \xi_i \geq 0 \end{aligned} \quad (3.12)$$

Applying Lagrange multipliers and converting to dual form as before, the problem is then expressed as

$$\begin{aligned} \max_{\alpha} \tilde{L}(\alpha) = & \max_{\alpha} \left\{ \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) \right\} \\ \text{subject to} & \quad 0 \leq \alpha_i \leq C \text{ and } \sum_{i=1}^N \alpha_i y_i = 0 \end{aligned} \quad (3.13)$$

Conveniently, the slack variables ξ_i disappear from the dual problem and the constant C appears only in the constraints. C is however a free parameter than can be difficult or costly

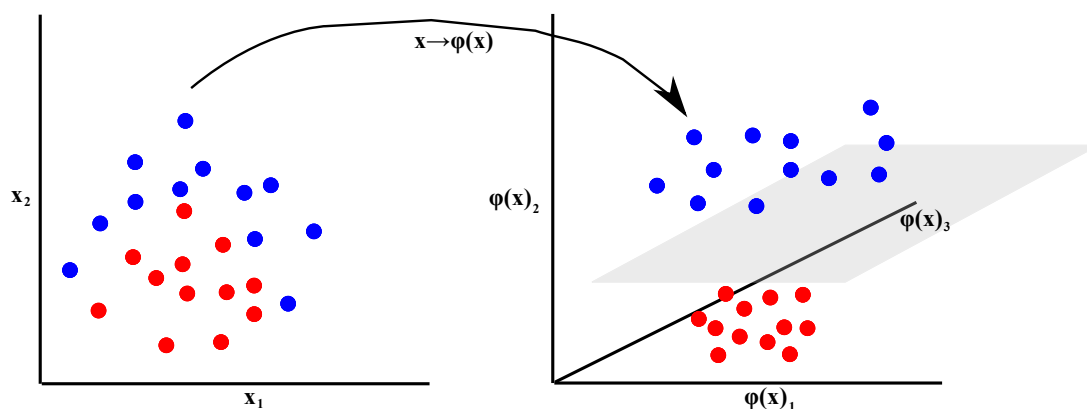


Figure 3.9: Data that are not linearly separable in the input space \mathbf{x} become linearly separable in the feature space $\phi(\mathbf{x})$. However we do not have to explicitly calculate $\phi(\mathbf{x})$ as only dot products between transformed feature vectors appear, forming the kernel matrix.

to optimise. Often it is set by a heuristic or a grid search for the value providing the best performance.

3.6.3 Kernels and nonlinear SVMs

Note that in the dual problems, the data points \mathbf{x} only appear in the form of a kernel function $k(\mathbf{x}_i, \mathbf{x}_j)$. This means that an SVM built from N training data points is based on an N by N kernel matrix of dot products. However the dot product can be replaced with a nonlinear kernel function $\phi(\mathbf{x})$ [Aizerman et al., 1964]. This means that the hyperplane is fit in some higher dimensional feature space, enabling data which are not linearly separable in their original form (the input space) space to be easily separated in the transformed space (the feature space) with a nonlinear SVM [Boser et al., 1992].

As the data only affects the optimisation problem via the kernel matrix, we do not need to explicitly calculate the feature space, but only the dot product of training samples in the feature space. Any function $k(\mathbf{x}_i, \mathbf{x}_j)$ that produces a valid (symmetric positive definite) kernel matrix can be used as a kernel function. This also means that any linear combination of valid kernels is itself a valid kernel, a fact which we make extensive use of. More generally, the kernel can be seen as a matrix of pairwise similarities between data points. The radial basis function (RBF) kernel for example, a widely used choice, is based on a function of the Euclidean distance between points:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right) \quad (3.14)$$

This implicitly makes use of an infinite dimensional feature space. Although it can be very

powerful, the flexibility of nonlinear classifiers can make them more prone to overfitting when there is little training data available. In the case of SVMs, the RBF kernel also adds another parameter to be tuned, the kernel width σ .

3.7 Gaussian processes

Gaussian Processes (GPs) provide an alternative method for machine learning that, like SVMs, can be used for both classification and regression. GPs have some features in common with SVMs - they provide similar predictive accuracy, and are also well suited to very high dimensional data. This is because, again like SVMs, they are based on a kernel, making them very efficient in the domain where the data dimensionality is much greater than the number of samples (as is very common in neuroimaging applications). SVMs also make use of the principle of structural risk minimisation [Cortes and Vapnik, 1995], where the criterion of finding the largest possible margin helps to prevent overfitting.

Furthermore, the interpretation of the kernel matrix is different; rather than summarising the similarity between points it represents the covariance matrix of a multivariate Gaussian prior over the classification or regression model parameters. This implies that GPs are a probabilistic method, which distinguishes them from SVMs. Prevention of overfitting in GPs is done by probabilistic model averaging, rather than optimisation of a single model as in SVMs.

Bishop [Bishop, 2007] describes probabilistic methods as having four advantages over non probabilistic ones:

- Easy risk minimisation when the costs of making a mistake (the loss) change;
- Allowing a reject option where a classifier only gives a decision for test data it is can categorise with confidence;
- Easy compensation for class priors, where one class is much more common than the other(s);
- Models can be combined in an existing framework - the laws of probability.

To this we can add automatic setting of parameters from training data only, which we make extensive use of.

What follows is a brief introduction to the application of GPs to machine learning problems. For a much more rigorous treatment, [Rasmussen and Williams, 2006] is recommended. We begin by showing how GPs are used in regression, as this is the simplest case, and then extend the model to binary classification. But first of all we must introduce Bayes' rule, which

describes conditional probability. Consider two random variables, A and B . From the laws of probability, we know that the joint distribution of A and B , $p(A, B)$, can be written as either $p(A, B) = p(A|B)p(B)$ or $p(A, B) = p(B|A)p(A)$. So we can write

$$p(A|B)p(B) = p(B|A)p(A) \quad (3.15)$$

which by dividing through gives us

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)} \quad (3.16)$$

Which relates the conditional probabilities $p(A|B)$ and $p(B|A)$. $p(A|B)$ is known as the posterior (as it describes our beliefs about A after having observed B), $p(A)$ the prior (as it describes our beliefs about A prior to observing B), $p(B|A)$ the likelihood and $p(B)$ the marginal likelihood.

3.7.1 Gaussian process priors

Formally speaking, a GP is a generalisation of ordinary multivariate Gaussian distributions to the case of an infinite number of variables, with the condition that any finite subset of the variables form a multivariate Gaussian. It is this latter property - that any marginal distribution of the GP is also Gaussian - which is the key to GPs' applicability to machine learning problems, given the somewhat abstract definition. Again generalising from familiar multivariate Gaussians, which are parameterised by a mean vector and a covariance matrix, a GP can be described only by its mean function $m(\mathbf{x})$ and covariance function $k(\mathbf{x}, \mathbf{x}')$. So we can write

$$GP \sim \mathcal{N}(\mathbf{m}, \mathbf{K}) \quad (3.17)$$

where

$$\mathbf{m} = \begin{bmatrix} m(\mathbf{x}) \\ m(\mathbf{x}') \end{bmatrix}, \mathbf{K} = \begin{bmatrix} k(\mathbf{x}, \mathbf{x}') & k(\mathbf{x}, \mathbf{x}') \\ k(\mathbf{x}', \mathbf{x}) & k(\mathbf{x}', \mathbf{x}) \end{bmatrix} \quad (3.18)$$

The matrix \mathbf{K} is a kernel matrix identical in form to those used in SVMs, and the allowed kernel functions are the same as can be used with an SVM. A linear (dot product) covariance function can be used, but we can also apply the kernel trick to use kernel functions such as radial basis functions (sometimes referred to in GP literature as squared exponentials) to perform nonlinear regression or classification, again like an SVM. However, the interpretation of the kernel function is different in SVMs and GPs. In an SVM, the kernel function measures the similarity between vectors in an inner product space, whereas in a GP it measures a component

of the covariance of a prior over functions. Hence the terms kernel function and covariance function are equivalent in the GP literature.

By itself, the GP cannot be used for machine learning. However, it can be applied as a prior to models of regression and classification. There are two ways of seeing how this can be done, which are different conceptually but can be shown to be equivalent mathematically [Rasmussen and Williams, 2006]. These are introduced in sections 3.7.2 and 3.7.3.

3.7.2 Gaussian process regression - weight space view

One way of understanding GP regression is known as the weight space view. This is introduced first as it involves using the GP as a prior on the weights \mathbf{w} of a familiar linear regression model.

In ordinary linear regression, the target vector \mathbf{y} is modelled as a linear function¹ of data \mathbf{X} with zero mean Gaussian noise of standard deviation σ : $\mathbf{y} = \mathbf{w}^\top \mathbf{X} + \epsilon$, with $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Now consider placing a GP prior (with, for simplicity, zero mean) over the weights \mathbf{w} . By substituting this GP for the prior, and Gaussian noise as the likelihood into Bayes' rule (equation 3.16) we obtain the following equation for the posterior for the weights:

$$p(\mathbf{w}|\mathbf{X}, \mathbf{y}, \boldsymbol{\theta}) = \frac{p(\mathbf{w}|\boldsymbol{\theta})p(\mathbf{y}|\mathbf{X}, \mathbf{w})}{p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})} \quad (3.19)$$

where the (GP) prior is $p(\mathbf{w}|\boldsymbol{\theta})$, the (Gaussian) likelihood is $p(\mathbf{y}|\mathbf{X}, \mathbf{w})$ and $\boldsymbol{\theta}$ is a set of hyperparameters that determine the exact form of the prior. The denominator $p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})$ represents a marginal likelihood equal to $\int p(\mathbf{w}|\boldsymbol{\theta})p(\mathbf{y}|\mathbf{X}, \mathbf{w})$. The Gaussian form of the prior provides a form of regularisation, favouring weights that are not too large. To make predictions, we do not usually select a single set of weights but instead integrate over the posterior to average over all possible values for \mathbf{w} . If we are presented with a new data point \mathbf{x}^* , we make a prediction for the corresponding y^* by integrating over 3.19, effectively averaging across all possible values for \mathbf{w} , weighted by their posterior probability.

$$p(y^*|\mathbf{X}, \mathbf{y}, \mathbf{x}^*, \boldsymbol{\theta}) = \int p(y^*|\mathbf{w}, \mathbf{x}^*)p(\mathbf{w}|\mathbf{X}, \mathbf{y}, \boldsymbol{\theta})d\mathbf{w} \quad (3.20)$$

As the prior and likelihood are both Gaussian, the marginal likelihood and posterior are also Gaussian. This means a closed form solution for equation 3.20 can be found [Rasmussen and Williams, 2006]:

$$p(y^*|\mathbf{X}, \mathbf{y}, \mathbf{x}^*, \boldsymbol{\theta}) \sim \mathcal{N}\left(\frac{1}{\sigma^2}\mathbf{x}^{*\top}\mathbf{A}^{-1}\mathbf{X}^\top\mathbf{y}, \mathbf{x}^{*\top}\mathbf{A}^{-1}\mathbf{x}^*\right) \quad (3.21)$$

¹Note that for simplicity, the bias term, b , has been subsumed into the weights, \mathbf{w} .

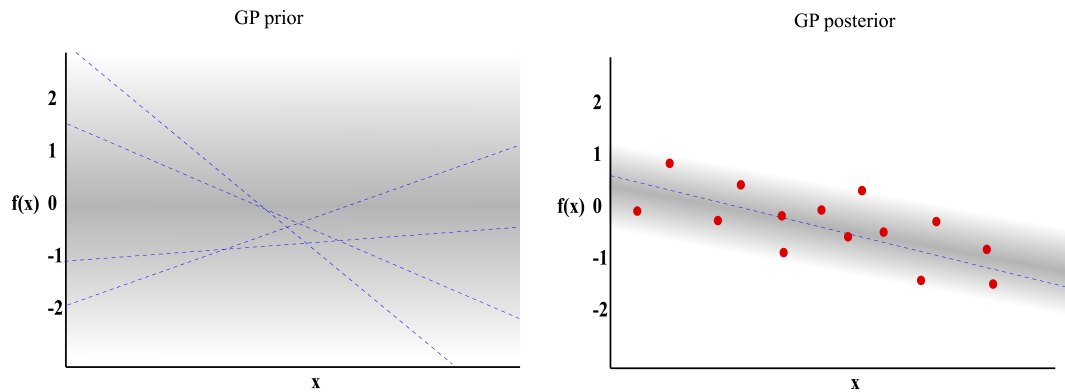


Figure 3.10: The GP prior (left) is vague and permits a wide range of linear functions admitted by the prior (blue) and high uncertainty (grey). After introducing data points (red), the posterior allows a much narrower range of linear functions and reduced uncertainty.

Where $\mathbf{A} = \frac{1}{\sigma^2} \mathbf{X}^\top \mathbf{X} + \mathbf{K}^{-1}$ and \mathbf{K} is the kernel matrix representing the covariance of the GP prior.

3.7.3 Gaussian process regression - function space view

The function space view of GP regression initially seems more unfamiliar as it dispenses with the entirely use of weights, apparently breaking the link with linear regression. However it provides a more natural and concise way to describe GP regression, by instead applying the GP as a prior directly over *functions* $f(\mathbf{x})$. This means that the mean vector \mathbf{m} has infinite length and the kernel matrix \mathbf{K} has infinite size, as it forms a prior over all possible linear functions. However by introducing a set of training data vectors \mathbf{X} and a vector of corresponding target values \mathbf{y} we can calculate a posterior distribution for $f(\mathbf{x})$. This is illustrated for the simple case of one dimensional, linear regression in figure 3.10. In the left hand diagram, the GP prior is across the space of all possible linear functions, shown in blue, and is correspondingly broad. In the right hand figure, some training (one dimensional) data vectors \mathbf{X} and target values \mathbf{y} have been introduced, shown as the red dots. Pairs of data vectors and their corresponding target values (\mathbf{x}, y) can be seen as samples from the distribution of $f(\mathbf{x})$. By the properties of GPs, this finite sample is a (finite sized) multivariate Gaussian. It can therefore be used with Bayes rule to calculate a posterior for $f(\mathbf{x})$, giving a much tighter distribution which can be used to make high quality predictions.

For the regression case, as in the weight space view we model the targets \mathbf{y} as being a linear function of the data vectors \mathbf{x} plus zero mean Gaussian noise. The difference between the weight space and function space view is that we apply the prior directly to the value of $f(\mathbf{x})$ rather than the weight. Again as in the weight space view, for simplicity we assume the GP is

zero mean. We apply the GP as a prior and Gaussian noise as the likelihood, so the model is given by

$$\begin{aligned} y_i &= f(\mathbf{x}_i) + \epsilon \\ f &\sim \mathcal{N}(\boldsymbol{\mu} = 0, \mathbf{K}) \\ \epsilon &\sim \mathcal{N}(0, \sigma^2) \end{aligned} \quad (3.22)$$

As before, we can substitute these terms into equation 3.16 and simplify. The resulting posterior gives the core predictive equations for GPs. The posterior is, again, also a multivariate Gaussian. For training data \mathbf{X} and training targets \mathbf{y} , and an unseen test data vector \mathbf{x}^* the predictive distribution for the value of $f(\mathbf{x}^*)$ is given by

$$\begin{aligned} p(f(\mathbf{x}^*)|\mathbf{X}, \mathbf{y}, \mathbf{x}^*, \boldsymbol{\theta}) &\sim \mathcal{N}(\mu^*, \sigma^{2*}) \\ \mu^* &= \mathbf{k}^{*\top} \mathbf{C}^{-1} \mathbf{y} \\ \sigma^{2*} &= k(\mathbf{x}^*, \mathbf{x}^*) = \mathbf{k}^{*\top} \mathbf{C}^{-1} \mathbf{k}^* \end{aligned} \quad (3.23)$$

Where $\mathbf{C} = \mathbf{K} + \sigma^2 \mathbf{I}$. \mathbf{K} is the covariance matrix derived from the covariance function k , training data \mathbf{X} and covariance hyperparameters $\boldsymbol{\theta}$, so $\mathbf{K}_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j, \boldsymbol{\theta})$, and \mathbf{k}^* is a vector of covariances between the test data point \mathbf{x}^* and all the training data points.

As previously stated, the weight space and function space views of GP regression can be shown to be equivalent. However, a subtle difference is that the function space formulation makes use of the kernel trick. Compare the predictive equations for the weight space and function space views, equations 3.21 and 3.23 respectively. In the former, predictions are made by inverting the matrix \mathbf{A} which is size $d \times d$ where d is the data dimensionality. In the latter, making predictions involves inverting \mathbf{C} , which is size $N \times N$, where N is the number of subjects. For neuroimaging applications, where very frequently d is much larger than N , this can be a great advantage, especially if precomputed kernel matrices are used.

3.7.4 Gaussian process classification

In order to perform classification, rather than regression, we must modify the likelihood function of the GP. If we are given a set of data points \mathbf{x} and corresponding vector of labels y : $\{(\mathbf{x}_i, y_i)_{i=1}^N\}$ with binary class labels $y_i \in \{-1, +1\}$ then we can use a sigmoidal function σ such as the logistic or probit function as the likelihood, so $p(y_i = 1|\mathbf{x}_i) = \sigma(f(\mathbf{x}_i))$. This then maps $f(\mathbf{x}_i)$ to the $[0, 1]$ interval, representing the probability that the label for \mathbf{x}_i , y_i , is equal to 1.

Gaussian process classification could therefore be seen as similar to logistic or probit regression with a GP prior. In binary classification, the class probabilities must sum to unity so

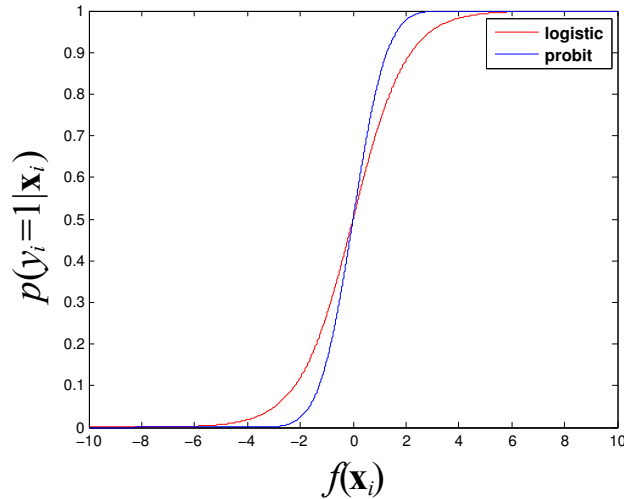


Figure 3.11: Sigmoidal functions link the value of the latent function $f(\mathbf{x})$ to class membership probabilities.

$p(y = 1|\mathbf{x}_i) = 1 - p(y = -1|\mathbf{x}_i)$, and the sigmoid functions σ are symmetric so $\sigma(x) = 1 - \sigma(-x)$. This means the likelihood terms can be rewritten as $p(y_i|\mathbf{x}_i) = \sigma(y_i f(\mathbf{x}_i))$. Plugging this into equation 3.16 we can rewrite Bayes' rule as

$$p(\mathbf{f}|\mathbf{X}, \mathbf{y}, \boldsymbol{\theta}) = \frac{p(\mathbf{f}|\boldsymbol{\theta})}{p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})} \prod_{i=1}^N \sigma(y_i f(\mathbf{x}_i)) \quad (3.24)$$

whereas for regression $p(\mathbf{f}|\boldsymbol{\theta})$ is the prior over the latent function f and $p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})$ is the marginal likelihood. The likelihood over training samples is factorised as they can safely be treated as independent.

Unfortunately, the non-Gaussian likelihood function means that the marginal likelihood and posterior are also non-Gaussian in the classification case. An exact evaluation of the integrals of these distributions cannot be obtained analytically. However Monte-Carlo Markov Chain (MCMC) methods may be used to give an extremely accurate evaluation of them, which in the limit of an infinite number of samples would be exact. Alternatively, although the resulting posterior distributions are non-Gaussian, they can be shown to be unimodal and can be reasonably approximated by a Gaussian. A number of methods can be used to fit a Gaussian; these include the Laplace approximation [Williams and Barber, 1998], variational Bayes [Gibbs and MacKay, 2000] and expectation propagation (EP) [Minka, 2001]. All three of these were compared against an MCMC 'gold standard' in [Nickisch and Rasmussen, 2008], and EP was found to be almost as accurate as MCMC while being much faster to run in practice. We therefore use EP as the approximation method in all GP classification experiments. Briefly, it calculates the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ of the approximating multivariate Gaussian by approxi-

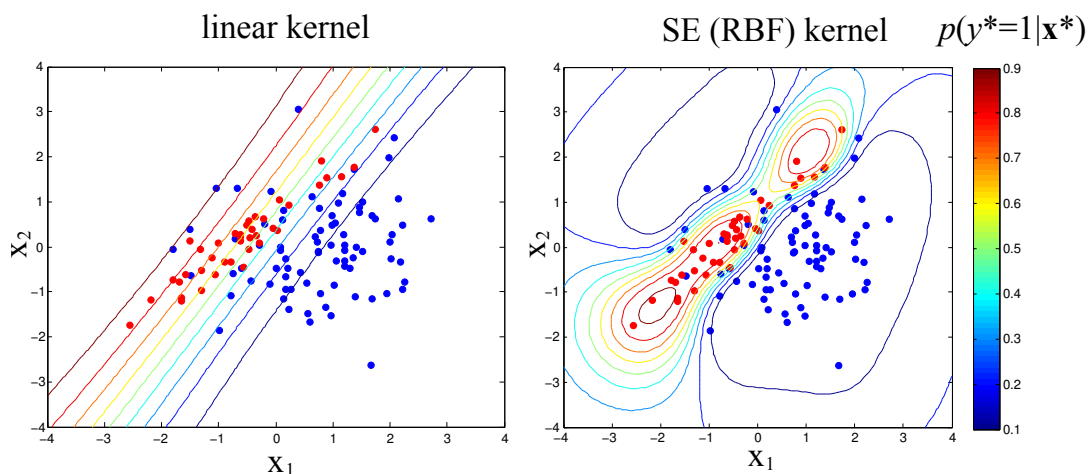


Figure 3.12: Gaussian process classification with two dimensional toy data, with linear (left) and squared exponential (right) kernels. Red dots indicates the $+1$ labelled training data and blue dots the -1 labelled ones. Contours show the locus of equal $p(y^* = 1|\mathbf{x})$ as indicated by the colour bar.

minating the likelihood hood terms with unnormalised Gaussian site functions, and then cycling between the site function, updating each in turn with moment matching, until a convergence criteria is met. No formal proof of convergence has been found for EP but it usually terminates successfully in practice [Nickisch and Rasmussen, 2008].

All methods are explained in more detail in the references above. A very detailed mathematical explanation of their application to GPs is given in [Rasmussen and Williams, 2006]. Once the posterior has been found, predictions can be made in a similar manner to GP regression. GP classification generally gives us similar accuracies to SVMs. However it has a number of advantages. As well as being easily extended to multiclass classification and offering automatic hyperparameter tuning (described in the next section) it can quantify the predictive uncertainty by providing a probability of class membership of test subjects. GP classification does not explicitly calculate a separating boundary between the two classes in the input space. However as figure 3.12 shows, for a linear kernel, the $p(y^*|\mathbf{x}^*) = 0.5$ contour very much resembles one. However unlike an SVM the class membership predictions will always be probabilistic. For a nonlinear kernel, an SVM would nonlinearly divide the input space into two regions, but GP classification with a nonlinear kernel produces a posterior distribution that tends to the class prevalence prior far away from training data.

3.7.5 Gaussian process regression and classification in practice

As described above, GPs for regression could be implemented only a few lines of code. However, we may want to use approximations for classification, use complex covariance and mean

functions, select different likelihood functions, and choose the best value for hyperparameters. Because of this, all of our experiments are done using the GPML toolbox ² for Matlab, which provides functionality for all of these. While a very simple GP classification might have no hyperparameters at all, almost all real problems will include some describing the covariance function (such as kernel width for an SE kernel), the posterior mean, and the noise variance in regression problems. A fully Bayesian way to deal with these is to define prior distributions on their values (hyperpriors) and integrate over the hyperpriors in a hierarchical model. However, the resulting posterior distributions can only be computed using MCMC methods, which can be very slow.

Fortunately there is a practical and only slightly less effective alternative. We can instead maximise the marginal log likelihood of the training data and labels with respect to the set of hyperparameters. The marginal log likelihood is given by

$$\ln(p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})) = \ln \int p(\mathbf{y}|\mathbf{X}, \mathbf{f})p(\mathbf{f}|\boldsymbol{\theta})d\mathbf{f} \quad (3.25)$$

where \mathbf{f} is a vector of predictive function values for all the training points. The above expression can be shown to be equal to $-\frac{1}{2}\mathbf{y}^\top \mathbf{C}^{-1}\mathbf{y} - \frac{1}{2} \ln |\mathbf{C}| + \text{const}$, where \mathbf{C} is again $\mathbf{K} + \sigma_n^2\mathbf{I}$. This expression can then be differentiated. The resulting derivative can be used with a standard gradient based optimisation to maximise the log likelihood with respect to the noise variance σ^2 (directly) or covariance and mean hyperparameters (via application of the chain rule, to the derivatives of the covariance matrix or mean vector elements with respect to the hyperparameters). The motivation for setting hyperparameters with this method is best understood by viewing it as a model selection problem: which model (set of hyperparameter values) best explains the observations (data)? The model that maximises the likelihood is the one with just enough complexity to describe the data, but no more. This is shown in figure 3.13. Three models are shown - a simple one in green, a complex one in blue, and one of intermediate complexity in red. The simple model can only describe a narrow range of possible data, so assigns the data we have (the vertical line) a very low likelihood. The most complex one gives a reasonably high likelihood to the data, but also to many other sets of data. The intermediate model gives the data a higher likelihood than either other model. Hence maximising the marginal likelihood gives the most parsimonious model that explains the data, obeying Occam's razor ³ and helping to avoid overfitting. This can also be seen in equation 3.25 - the first term measures the fit to the data, and the second penalises model complexity.

²<http://www.gaussianprocess.org/gpml/code/matlab/doc/>

³'Plurality must never be posited without necessity'

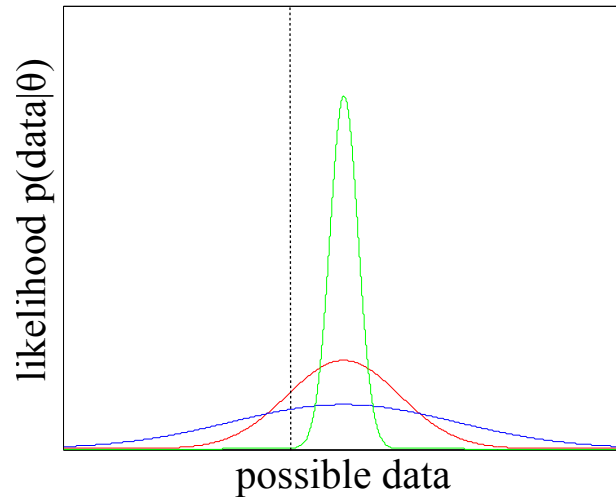


Figure 3.13: The most simple model in green is not flexible enough to describe the data and assigns it a low likelihood. The most flexible model (blue) can explain many possible data and so assigns each a low likelihood. The model with the correct degree of complexity (red) assigns the data a higher likelihood than either of the others.

The likelihood as a function of the hyperparameters is not convex, so there is a danger this method will find only a local maximum. However, in practice good results are usually obtained.

3.8 Precomputed kernels

As is noted above, both GP classification and regression and SVM classification are kernel based techniques. In neuroimaging problems, we typically have images for only a few dozen to at most about a thousand or so subjects. Meanwhile each image may consist over a million features, if we are using voxels as features, so $d \gg N$. The N by N kernel matrix \mathbf{K} is therefore much easier to work with than the N by d training data matrix \mathbf{X} . In an SVM, the data only appears in the optimisation problem as the kernel matrix. We can therefore greatly reduce memory requirements and speed up calculation by calculating \mathbf{K} once and then throwing \mathbf{X} away. If all the images we want to use cannot fit into memory at once, we can also calculate \mathbf{K} element by element, storing only a few images at a time. The SVM library we use, LIBSVM [Chang and Lin, 2011], allows us to do this as an option. For GPs, things are a little more difficult. Strictly speaking, the final kernel matrix is likely to be a function of covariance hyperparameters that must be set as well as of the data. However, in many cases the hyperparameters can be applied in the kernel space. For example, when the final kernel is defined as a linear combination of linear subkernels, the covariance hyperparameters represent the weight of each subkernel in the sum, and so we can precompute each subkernel from data. Then during hyperparameter optimisation, the subkernels are multiplied directly by the weights, rather

than multiplying the data for each subkernel by the weights and then recalculating the subkernel. GPML does not include a facility to do this so it was modified to add one. The resulting representation is very flexible. For example, cross validation or training and test sets can be performed by selecting the relevant rows and columns of master kernel matrices, rather than having to reload images and form new kernel matrices from scratch.

Chapter 4

Literature review

4.1 Introduction to the existing literature

Early diagnosis is often studied by applying classification methods to MCI subjects, trying to predict which subjects will convert to AD (designated MCI converters, or MCI-c), and which will remain stable (MCI-s) over some fixed period of time, usually 18 to 36 months. Comparing results between experiments predicting conversion in MCI subjects is particularly difficult as this follow up period varies and both a diagnosis of MCI and conversion status may be hard to verify.

Methods used in automated diagnosis and prediction of AD vary quite widely in the types of data and learning algorithms used; nevertheless the vast majority can be summarised as following the same basic pipeline. The starting point is a number of training subjects, usually of AD patients and healthy controls, together with a set of labels that indicate to which group each subject belongs. Next comes a feature extraction step, in which the images are processed so that each subject is represented by a vector of features, which (hopefully) are relevant to the classification problem. Feature extraction procedures vary greatly, and the features can be very similar to the original image (where the features are in fact voxel intensities) or quite abstract and far removed from the images they are derived from if complex transformation and dimensionality reduction procedures are applied. The training feature vectors and labels are then fed into a learning algorithm, where again a great deal of variation is possible in the choice of algorithm, as well as the methods to set parameters. Finally, to assess generalisation ability the learned model is applied to previously unseen testing subjects, which have been put through the same feature extraction procedure as the training subjects. The results are then evaluated, and again a choice of statistics can be used to summarise the performance of the pipeline.

The distinction between features and learning algorithm in the context of brain image classification for a variety of neurological conditions is the focus of a recent study

[Sabuncu and Konukoglu, 2014]. The authors divide the factors on classification into three. The first is what the authors call biological footprint, which is the effect size of the difference between classes. The second and third are the choice of measurements (features) and classifier algorithm. The authors find that the biological footprint and choice of feature have a much greater influence on accuracy than the choice of algorithm. A similar point, made in reference to machine learning problems in general, is made in [Domingos, 2012], which stresses the importance of feature engineering. This is perhaps not surprising in the context of the no free lunch theorem [Wolpert, 1996].

As this literature review looks only at diagnosis and prediction of AD, the biological footprint is fixed. Hence the existing studies discussed in this section are broadly grouped by the type of data and features used in classification, as this the stage of the pipeline having both the widest choice of possible options and the greatest effect on results.

4.2 Review of the existing literature

4.2.1 Voxel-based features

Methods using voxel-based features for classification are among the simplest in terms of feature extraction as the feature vectors can be viewed as a type of image. Such approaches are exemplified by [Klöppel et al., 2008]. This study was on three different cohorts, each of which was scanned at a single centre. Groups I and II each consisted of equal numbers of age and gender matched AD patients and healthy controls. Significantly, the AD patients in these two groups had neuropathological confirmation of their disease status. Group III was both larger overall and had a greater number of controls than AD patients, however the patients were diagnosed clinically leading them to be described as having probable mild AD. The image data used consisted of T1 weighted structural MRI scans of all subjects. These images were segmented using SPM5 into grey matter (GM), white matter (WM) and cerebrospinal fluid (CSF) components. The GM segments were used as features in classification. To enable this, it was necessary to establish correspondence among voxels across subjects, so that those in the same position represent similar anatomical areas. To do this, a custom template was constructed from the images of all subjects using the DARTEL [Ashburner, 2007] nonlinear registration algorithm. A Jacobian-scaling (“modulation”) step was incorporated to ensure the total amount of each tissue type remained the same after registration [Ashburner and Friston, 2000]. The resulting three dimensional grey matter maps were then treated as vectors and classification was performed with an SVM, using a linear kernel with default parameter settings. Generalisation accuracy was assessed with leave-one-out cross validation (LOOCV) for each experiment. Accuracy on groups

I and II was similar at 95% and 93% respectively; when the two groups were pooled accuracy increased to 95.6%, showing that having a larger overall training set overcame differences between the groups due to different centres and scanners. However accuracy in group III was only 85.6%, probably as a result of misdiagnosis due to lack of post-mortem confirmation.

A much more recent and very different voxel-based approach was introduced in [Coupé et al., 2012]. This study used all 834 baseline scans from the ADNI dataset. The images were pre-processed, which involved registration to the ICBM152 template and cross normalisation of MRI intensity between subjects. The hippocampi and entorhinal cortices of ten randomly selected AD and ten control subjects were then manually segmented by an expert. These segmentations were then propagated to the remaining AD and control subjects to generate a training set. Classification in test subjects was then performed by a grading process, which simultaneously segmented the structures of interest. For each test subject, the ten closest subjects in each disease group were selected by a sum of squared differences (SSD) measure. A nonlocal means filter was then used to assign a weight comparing each voxel in the test subject to each voxel in each training subject. If the voxels in the training subjects are labelled with 1 for structure and 0 for background, the weights can be used to create a weighted sum to label each test subject voxel, segmenting it. Following this, a very similar procedure was used to propagate a new label indicating the disease state of each training subject to each voxel in the newly segmented structure in the test subject. This label was averaged across the structure to create an overall score. Using this approach it was possible to distinguish between AD patients and controls with 91% accuracy, and between MCI-c and MCI-s with 74% accuracy.

4.2.2 Region-based features

As an alternative to using individual voxels as features, they can instead be grouped together into regions with some anatomical meaning. These regions can be defined adaptively or on a manually labelled atlas which is propagated onto individual MRI images. Features are defined as the average value of voxel level quantities such as intensities within each region. For the former approach of adaptive feature extraction, the most commonly used procedure is COMPARE (Classification Of Morphological Patterns using Adaptive Regional Elements) [Fan et al., 2007]. This is a complex process consisting of multiple steps. Briefly, the MRI images are first aligned non-rigidly to a template using HAMMER [Shen and Davatzikos, 2002]. For each voxel, the robust correlation with class label across tissue types is calculated along with a spatial consistency measure. These are combined to produce a map of the separability scores across voxels for each tissue type. These are then smoothed and regions are generated with a watershed algorithm. Within each region, features are generated using a region growing

method, starting at the most discriminative voxel and adding neighbours until no more can be added. Even after this, the number of features is still high, so a selection is made using a ranking based on correlation and a method based on SVM-RFE (Recursive Feature Elimination) [Guyon et al., 2002]. The selected regional features can then be used in SVM classification. The initial application was to schizophrenia patients, but the method has also been applied to a number of problems related to AD, including predicting conversion of MCI patients within 15 months [Misra et al., 2009] with about 80% accuracy. For ADNI data with a three year follow-up period for conversion, it provides a sensitivity of 94.7% but only 37.8% specificity [Davatzikos et al., 2008].

The atlas based feature extraction procedures are much simpler. The method of [Magnin et al., 2009] propagates labels obtained from [Tzourio-Mazoyer et al., 2002] to their subjects. This was done by registering their subjects and the atlas to standard MNI space, and then inverting the transformations to create a labelled parcellation of 116 anatomical regions in the native space of each subject. In each region for each subject, the image histogram was separated into GM, WM and CSF components with the expectation maximisation (EM) algorithm and the relative weight of GM compared to the other tissue types was chosen as a feature to summarise each ROI. Classification was then performed using an SVM and bootstrap to assess generalisation accuracy; the result was an average accuracy of 94.5% for classifying AD subjects and healthy controls. However, this figure is most likely optimistically biased as testing data were used to select optimal SVM parameters within the bootstrap.

4.2.3 Cortical thickness features

Cortical thickness is a direct measure of the atrophy that is caused by AD [Dickerson et al., 2008] and is thus a powerful feature for disease classification that has been used in many studies. As in the previous section, features can represent the average thickness over anatomical regions or represent individual vertices of a cortical segmentation. Cortical thickness measurements are frequently derived from the FreeSurfer toolkit [Dale et al., 1999, Fischl et al., 1999a, Fischl et al., 1999b, Fischl and Dale, 2000]. [Querbes et al., 2009] however used a different approach, segmenting brain images into GM, WM and CSF. Within the cortical GM ribbon, the method based on Laplace's equation [Jones et al., 2000] was used to calculate cortical thickness in the native space of each subject's brain. These were then rigidly registered to MNI space. The cortical thickness maps were then parcellated into 96 areas using a Brodman area 3D map, which were then grouped into 22 zones. The final features were the mean cortical thickness in each zone. A normalised thickness index (NTI) was calculated from an optimally discriminative subset of these features as a linear discriminant. The reported AUC for discriminating

MCI-c and MCI-s subjects from the ADNI study was 0.76, with follow-up of 24 months.

An innovative method of using cortical thickness was employed by [Cho et al., 2012]. This study used a total of 491 MRI scans from the ADNI database. FreeSurfer was used to extract meshes, representing the inner and outer boundaries of the cortex, for each subject. To establish correspondence among subjects, all cortical surfaces were registered to FreeSurfer's cortical atlas. The meshes were then made isomorphic to each other, and thickness maps were derived from distances between corresponding inner and outer points. The resulting cortical maps were then denoised by removing high frequency components. Mapping onto a frequency domain was done using the manifold harmonic transform, which represents a signal by a linear combination of eigenfunctions of the Laplace-Beltrami operator. After removing high frequency components, 2400 frequency components were left for each brain hemisphere. Principal component analysis (PCA) was used to further reduce the dimensionality of the data, with components representing 70% of the variance retained. Finally, Fisher's linear discriminant analysis [Fisher, 1936] was used for classification. This method obtained a sensitivity of 63% and specificity of 76% for predicting MCI conversion within a follow-up period of 18 months.

[Eskildsen et al., 2013] also made use of cortical thickness features to predict conversion. Scans of converting and nonconverting MCI subjects from the ADNI database were used, including scans from all follow-up timepoints rather than baseline only. The MCI-c subjects were stratified by the time after baseline scan when they converted (up to 36 months). All images were corrected for noise and bias field, and then registered to MNI space and skull stripped. Cortical thickness was calculated with FACE (Fast Accurate Cortex Extraction) [Eskildsen and Ostergaard, 2006] and mapped to the cortical surface of a custom template. 51 scans were then removed for quality control. As each subject was represented by the cortical thickness at over 100,000 vertices, feature selection was then performed. This was done by computing t-tests at each vertex for the groups in question (eg MCI-s versus MCI-c at 12 months), finding local maxima in the resulting t-maps and then using these as seeds for a region growing procedure. Features were the mean thickness in each resulting region. The ten best features as chosen by maximum relevance, minimum redundancy (MRMR) [Peng et al., 2005] were retained for classification with LDA. Care was taken to perform the entire feature extraction and selection procedure independently in each iteration of an LOOCV loop, to avoid optimistic bias in the resulting estimated generalisation accuracy [Kriegeskorte et al., 2009, Varma and Simon, 2006]. The overall accuracy for predicting conversion was 67.3% with well balanced sensitivity and specificity when putting all converting subjects into a single group regardless of conversion timepoint. However, by constructing sep-

arate stratified classifiers for MCI-s subjects vs MCI-c for each conversion timepoint, and then classifying new subjects with the combined maximum posterior probability from all four stratified classifiers, this was increased to 73.5%. This improvement, however, did result in a less well balanced classifier.

4.2.4 Hippocampal features

As the hippocampus is one of the first brain structures to be affected by atrophy in Alzheimer's disease [Braak and Braak, 1995], many studies have focused on features derived from the hippocampus only to do early diagnosis of AD or to predict conversion to AD in MCI subjects. These studies all begin by segmenting the hippocampi in all subjects. Features as simple as the volume of the hippocampus, normalised by intracranial volume (ICV) can then be used. As hippocampal volume results in a feature dimensionality of only one, or two if both left and right hippocampi are used, advanced classifiers such as the SVM are not necessary. Hence hippocampal volumetry has been used for a relatively long period in early diagnosis of AD, with manual segmentation of the hippocampi [Jack et al., 1999]. More recently, automated hippocampal segmentation methods have become more widespread. The SACHA (Segmentation Automatique Competitive de l'Hippocampe et de l'Amygdale) method [Chupin et al., 2007] is based on competitive region growing to simultaneously segment the hippocampus and amygdala, constrained by anatomical and probabilistic priors. It was applied to a set of 210 MCI subjects from the ADNI database, 76 of whom converted to AD within the chosen 18 month follow-up period, in [Chupin et al., 2009]. Classification between MCI-s and MCI-c based on hippocampal volume in these subjects had an accuracy of 64%, which is somewhat disappointing - especially considering the short follow-up period. The authors suggest this may be improved by including shape analysis information rather than only volume. The same idea is suggested in [Csernansky et al., 2000], noting that in early AD the CA1 subfield of the hippocampus is particularly affected. Differential atrophy rates across the hippocampus mean it will change in shape as well as gross volume, allowing the authors to find areas of significant difference between AD patients and controls within the hippocampus; however, they did not test the diagnostic ability of shape for individual subjects. Parameterisations of hippocampal shape can, however be used as features for classification. For example, in [Gerardin et al., 2009], the SPHARM (SPherical HARMonics) toolkit [Gerig et al., 2001] was used to generate such a parameterisation, representing shapes as a sum of spherical harmonic basis functions in what can be considered a three dimensional analogue of Fourier analysis. The coefficients of the resulting series can be used as features themselves, or used to establish correspondence between sets of points representing the hippocampi, which are then used as features. The study used the coef-

ficients as the author claimed they offered better discrimination. Univariate t-tests and bagging were used to rank the coefficients as features, with a variable number of highest ranking features retained. Feature rankings and an optimal number of features and SVM parameters were determined by LOOCV on a cohort of locally scanned subjects, then the resulting optimised classifier was applied to a group of subjects from the ADNI database. Accuracy was 88% for classifying AD patients and controls, and 80% for classifying MCI patients and controls, but their method was not applied to predict conversion in MCIs. Other studies have applied shape features to predicting MCI conversion.

For example, [Costafreda et al., 2011] trained a classifier using AD and control subjects from the AddNeuroMed study [Lovestone et al., 2009]. Hippocampi were segmented using the method described in [Morra et al., 2008], which itself treats segmentation as a classification problem on a voxel by voxel basis. The segmentations were converted to 3D meshes, then a common triangulation with correspondence between points was obtained using direct hippocampal mapping (DHM) [Shi et al., 2007]. Radial distances from the hippocampal medial core to vertices, normalised by the cube root of ICV, were taken as features for training a SVM using a RBF kernel. The model was then tested on the same features derived from MCI subjects, also from the AddNeuroMed database. The assumption behind this approach was that MCI-c subjects will appear more AD-like and MCI-s ones more control-like, so a classifier trained on AD patients and controls could be applied to predict conversion in MCI subjects. The resulting accuracy was 80%, but the follow-up period for defining conversion was only one year.

The methods used in [Ferrarini et al., 2009] took a similar approach. The authors used a set of 50 locally scanned AD subjects and 50 controls, and a set of 15 MCI-c and 50 MCI-s subjects with a mean follow-up period of 33 months. All subjects' brain MR images were rigidly registered to a standard template, and both hippocampi were manually segmented. The most representative subject from the AD and control subjects was found and designated as a standard space; all other subjects were then rigidly registered to this and the transformations applied to the hippocampal surface points. The GAME (Growing and Adaptive MESHes) method [Ferrarini et al., 2007] was used to model the hippocampal shapes, making use of self-organising maps [Kohonen, 1990] to adapt the meshes, moving nodes and edges to increase the similarity to other subjects. Feature selection was done by permutation testing to assess the significance and consistency of each node in the final set, and then thresholded by p-value. This step was done using AD and control subjects only. To classify MCI-c and MCI-s, an LOOCV loop was used, with an RBF SVM. The SVM parameters were tuned with a grid search and a nested LOOCV loop within each fold. The resulting accuracy was similar to

[Costafreda et al., 2011] at 80%, although the result here is more impressive given the longer follow-up time.

4.2.5 Side-by-side assessment of features

Direct comparison between results obtained by the methods described up to this point is difficult, as they all involve different subjects, from both multicentre studies such as ADNI and local scanners, different criteria to define MCI converting and stable subjects, and different statistics used to assess predictive accuracy. To address this problem, Cuingnet et al [Cuingnet et al., 2010] conducted a large study assessing a number of methods alongside each other, using the same set of ADNI subjects and the same method of assessing the results. Furthermore, as the emphasis was on finding the best features, rather than classification methods and a linear SVM was used in all experiments except those using a single measure of hippocampal volume. The work compared the effects of varying preprocessing steps, such as registration algorithm, and whole brain versus volume of interest analysis. The methods assessed included variants of ones already discussed, including [Klöppel et al., 2008, Fan et al., 2007, Magnin et al., 2009, Chupin et al., 2007, Gerardin et al., 2009]. Also used were the STAND (SStructural Abnormality iNDex) score method [Vemuri et al., 2008], which is similar to [Klöppel et al., 2008] but with an extra image downsampling and features selection step, and the cortical thickness method described in [Desikan et al., 2009], which is similar to [Querbes et al., 2009]. Each method was used for three clinically relevant classification problems: AD versus control, MCI-c versus control, and MCI-converter versus MCI-s, with an 18 month follow-up period defining the conversion. For the first task, all methods performed statistically significantly better than chance and for the second task all but two did. By contrast, none at all significantly better than chance for predicting conversion in MCI subjects, although some did achieve over 50% in both specificity and sensitivity so this was probably a consequence of the small number of test subjects. Many methods offered 0% sensitivity and 100% specificity, or were almost as unbalanced. This was probably caused by the combination of the numerically unbalanced training set (39 MCI-c, 67 MCI-s) and highly overlapping feature distributions of the two classes forcing the classifiers to assign all training subjects to a single class as this offered the best overall accuracy. Overall, performance of methods in this comparative study was generally much lower than in the original papers where they were introduced. This may be caused by [Cuingnet et al., 2010] deliberately putting less effort into parameter tuning, feature selection, et cetera in order to focus on comparing feature extraction methods. It also may be because the authors were more rigorous in avoiding the type of optimistic bias described in [Kriegeskorte et al., 2009] than the originators of the methods were. Most likely it

is a combination of the two, as it is parameter tuning and feature selection steps that offer the opportunity to introduce this bias.

Because of these complicating factors, it is necessary to be cautious when deriving general conclusions from the results. However, based on the results from the AD versus control and MCI-c versus control, it appears that there is some benefit to using a more modern, accurate registration method. It also seems that whole brain methods are more effective for separating patients from controls, but for tasks involving MCI subjects, methods using the hippocampus only remain competitive. This is not unexpected as the hippocampus is one of the structures affected earliest in the disease process, but when AD is at a more advanced stage, atrophy is more widespread in the brain [Braak and Braak, 1995]. Methods incorporating data driven feature selection do not appear to bring any significant advantage and take much longer to run given the number of extra tunable parameters that they introduce. The authors suggest that making use of prior knowledge of the disease (such as focusing on a predetermined volume of interest) is a more robust way to reduce the dimensionality of the features. They also note that classifiers using combinations with other markers seem to be necessary to detect prodromal AD with high accuracy.

4.2.6 Multi-MRI features

This is the approach taken in [Westman et al., 2012], using a large set of control, AD and MCI subjects from the ADNI database. In the MCI subjects, conversion was defined with an 18 month follow-up period. The images were processed using FreeSurfer, rather than registering them to a template, which parcellated all the images into 68 cortical and 46 subcortical regions. A number of subcortical regions were excluded from the analysis, and the volumes of the remaining ones were averaged between hemispheres to leave a total of 21 subcortical volume features. For the cortical regions, FreeSurfer generated a large basket of features: thickness and volume, and also surface area, mean curvature, Gaussian curvature, folding index and curvature index. All these were averaged between left and right hemisphere parcellations to leave a total of seven measures for 34 cortical regions, giving 238 cortical features, and 259 features of eight different types in total. Classification was done by orthogonal partial least squares, also sometimes known as orthogonal projection onto latent structures (OPLS) [Trygg and Wold, 2002, Bylesjö et al., 2006]. This created a model with one predictive component and one or more orthogonal components representing variation in the training data that is not related to class differences. Before building the models, the data were preprocessed by centring and then scaling to unit variance. This is a common step, especially in models combining different types of data, as the different types may very well have quite different ranges.

Without the scaling, data with a large range could dominate even if it is not the most discriminative. Initially eight models were created, using each type of feature alone. Three types of feature (cortical thickness, cortical volume and subcortical volume) were shown to perform well for classification. Mean curvature and surface did a little less well, while Gaussian curvature, folding index and curvature index did not perform significantly better than chance and were dropped from further analysis. The next step was to assess the effectiveness of different types of feature in combination, which was done by creating hierarchical models from each possible pair of remaining types of features (except for the combination of the less well performing pair, mean curvature and surface area). Finally, a hierarchical model was created using all three of the best performing feature types (cortical thickness, cortical volume and subcortical volume). This entire exercise was done twice, once using raw features and the second time using all features normalised by the subjects' ICV. In both cases, all the models using two types of features outperformed any single feature type, and the model using three feature types outperformed all the ones using two. AD subjects and healthy controls could be separated with 90.5% accuracy in this way. This was further increased to 91.5% in a 'mixed model' when normalised thickness was used with raw volume features. This model could also predict conversion, when applied to the MCI subjects, with 68.5% accuracy.

A similar approach was used in [Wolz et al., 2011b]. The authors used all ADNI subjects that were available at the time their study was conducted, giving a very large group of 231 controls, 238 stable and 167 converting MCI patients (conversion being defined simply up to the point when the images were obtained in July 2011), and 198 AD patients. Four types of features were used for each subject. The first feature used was univariate hippocampal volume, as calculated from segmentations generated by label propagation from a set of atlases selected from a larger pool and merged [Lötjönen et al., 2010]. The second was cortical thickness measured by the CLASP (Constrained Laplacian-based Automated Segmentation with Proximities) algorithm [Kim et al., 2005], with the vertex-wise thickness measures smoothed with a Gaussian filter. The third feature was tensor-based morphometry (TBM) maps. A set of 30 randomly selected images (10 control, 10 AD, and 10 MCI) were chosen as templates and nonlinearly registered to all study images. The Jacobian determinants of the resulting deformation fields, representing local expansion or contraction in each voxel were calculated. To combine multiple results, all template images were registered to their own anatomical mean, and the resulting deformations applied to the appropriate Jacobian maps. The maps for each subject were then averaged to leave one overall Jacobian map per subject. The fourth feature was the coordinates of each subject image in a (relatively) low dimensional space [Wolz et al., 2011a], making use

of the Laplacian eigenmaps manifold learning algorithm [Belkin and Niyogi, 2003]. To reduce the dimensionality of the thickness and TBM data, both were aggregated into ROIs based on groupwise statistical tests in each vertex or voxel. This was done separately for each experiment, that is different regions were chosen for AD versus controls classification than MCI-s versus MCI-c. Three classification experiments were done: control versus AD, control versus MCI-c, and MCI-s versus MCI-c. For all of these, a combination of all features was more accurate than any single feature type when using LDA as a classifier, although this was not always the case when using an RBF SVM. LDA also tended to produce a better balance of sensitivity and specificity. The authors did not mention any feature normalisation step, so presumably the feature types were combined by simply concatenating the non-normalised feature vectors. For converting versus stable MCI subjects, sensitivity was 67% and specificity 69%.

4.2.7 Multimodal classification

An obvious extension to the idea of using multiple measurements derived from MRI is to combine MRI features with others from different imaging or non-imaging data. The most common other imaging modality is positron emission tomography (PET). Tracer radionuclides can be attached to molecules with a specific function in brain chemistry to image different aspects of brain function. For example, fluoro-deoxyglucose PET (FDG-PET) gives information on brain metabolism. As the degeneration caused by AD reduces brain metabolism, this can be a useful biomarker, and can also be used by itself in classification studies very much analogously to the ones already discussed, such as in [Herholz et al., 2002]. A variety of non-imaging data can also be used. The results of psychological tests that we have already introduced can be used. Additionally, there are genetic risk factors for the sporadic form of AD. In particular, it has been demonstrated that the variants of the apolipoprotein E (APOE) gene affect the chance of developing AD, with the $\epsilon 2$ allele conferring some degree of protection [Corder et al., 1994] whereas the $\epsilon 4$ allele increases risk [Corder et al., 1993]. Finally, levels of proteins can be measured in the cerebrospinal fluid (CSF), a liquid surrounding the brain and spinal cord, from which a sample can be drawn with a lumbar puncture. In particular, CSF levels of total tau protein (t-tau) and phosphorylated tau (p-tau) proteins, known to be implicated in the formation of neurofibrillary tangles that cause atrophy in AD, are elevated in AD patients, while levels of the amyloid- $\beta 42$ ($a\beta 42$) peptide in CSF fall [Fjell et al., 2010a, Holtzman, 2011].

The simplest method of combining features extracted from different types of data is to just concatenate the feature vectors for each subject into one long vector, possibly after rescaling all features to zero mean and standard deviation of one. This is the approach used in [Vemuri et al., 2009]. They compared the utility of the previously developed STAND score,

based on structural MRI [Vemuri et al., 2008], with that of CSF biomarker levels for predicting time to conversion of MCI subjects via Cox proportional hazard models. Their conclusion was that the CSF biomarkers and STAND scores are complementary. The two imaging types of structural MRI and FDG-PET were combined with CSF biomarkers in [Walhovd et al., 2010]. A large number of subjects from the ADNI database was used. FreeSurfer was used to parcellate the MRI images and the hippocampal volume, and the mean thickness in a number of cortical regions selected for having shown sensitivity to AD in other studies were retained as features. For the PET features, the same regions of interest were used; PET activity within each region was averaged, and then normalised by the activity within each subject's pons. For CSF features, $a\beta42$, t-tau and p-tau were included in the analysis alongside the ratios p-tau/ $a\beta42$ and t-tau/ $a\beta42$. Stepwise logistic regression was used to perform classification of AD versus controls, in three separate experiments using only features of each type. The selected MRI, PET and CSF features were then used in a multimodal stepwise regression to again classify AD versus controls. The final multimodal model showed a small improvement in accuracy and AUC when compared against the unimodal classifiers, although the final classifier did not select any PET features. The feature concatenation approach is also used in [Nho et al., 2010] to combine regional volumes and mean cortical thicknesses, mean regional grey matter densities (both derived from structural MRI with FreeSurfer and SPM5 respectively) and APOE genotype, performing classification with an RBF SVM. Results here were quite good, with an accuracy in predicting MCI conversion within three years of 72.3%. This was using a classifier trained on AD and control subjects, using a pool of all FreeSurfer, SPM and APOE features to which a feature selection procedure was applied. The optimal set of features contained both APOE $\epsilon2$ and $\epsilon4$ status and some regions from both SPM and FreeSurfer. Concatenation was also used in [Cui et al., 2011] to predict MCI conversion within a 24 month follow-up period. FreeSurfer was again used to generate features, with the volume of subcortical regions and the volume, mean and standard deviation of thickness, volume and surface area of cortical regions used. CSF features were the same protein levels and ratios as [Nho et al., 2010], and a variety of neuropsychological test scores were also used as features. A separate feature selection step was done for the CSF and for the MRI features using MRMR [Peng et al., 2005], and for the test scores using both MRMR and AUC in discriminating between AD and control subjects. Classification was performed using an RBF SVM, with parameters optimised on a training set of AD and control subjects. The optimised classifier was then applied to MCI subjects. The resulting accuracy was a 67.1%, but this masked quite poor balance as sensitivity was over 96% but specificity only 48%.

Structural MRI biomarkers were also combined with CSF in [Davatzikos et al., 2008], with a slightly different method. The COMPARE framework [Fan et al., 2007] was used to generate a set of regions sensitive to AD, and the result of classifying subjects using these regions was a score representing the degree to which they had AD-like patterns, called SPARE-AD (Spatial Pattern of Abnormalities for Recognition of Early AD) by the authors. These SPARE-AD scores were then combined with CSF biomarkers for the subset of ADNI subjects for which they were available. This results in a two or three dimensional feature set (SPARE-AD score and/or one or two CSF biomarkers) to which a second SVM classifier is applied. The results from the combination of SPARE-AD score and t-tau provide the best accuracy, but this is only 61.7% and again is a very unbalanced result, as sensitivity is much higher than specificity.

Better results have been obtained with a multi-kernel learning (MKL) framework. Many types of classifiers, including SVMs, make use of a kernel, which is a matrix of pairwise similarities between subjects. In a linear SVM, the kernel is made up of the inner products of pairs of feature vectors. As the sum of valid kernels is itself a valid kernel, this can be used to integrate multiple types of feature. Separate kernels are formed using the features from each data modality, and then an optimally weighted sum is used to produce a combined kernel. This combined kernel is then used to train a classifier. There are several methods to find the optimal weighting of the kernels, one of which is to find the weights that maximise the margin alongside other parameters, in an optimisation that is a modification of a standard SVM. This specific algorithm [Bach and Lanckriet, 2004] is also sometimes referred to as MKL. This is the method used in [Hinrichs et al., 2011]. Their study used subjects from the ADNI database and made use of a wide variety of features. For structural MRI images, SPM was used to segment all subjects' baseline and 24 month follow-up scans. A custom template was then created from all subjects' baseline scans, and the GM and WM segmentations warped into the template space and smoothed. Additionally, all subjects' baseline and 24 month scans were nonlinearly registered, and the Jacobian maps of the resulting deformations were also warped into the template space. FDG-PET images were registered to their corresponding structural MRI images, normalised by the mean activity within the pons, and then warped to the custom template spaces. CSF biomarker levels, APOE genotype, and cognitive scores were also used in the analysis. Three kernels were used: One using imaging data, one using biological measures (APOE and CSF), and one using cognitive scores. A classifier was trained on AD and control subjects and then applied to MCI subjects to predict conversion. The best result was an AUC of 0.791, which translates to an accuracy of 72% based on a leave-one-out loop of classifier scores for the MCI subjects. Interestingly, this was for longitudinal imaging information only.

The classifier using only baseline imaging performed less well; what is more surprising is that a classifier based on all modalities also failed to perform as well as longitudinal image data alone. Furthermore, although the follow-up period for MCI conversion here was three years, the requirement for 24 month follow-up imaging means that in effect prediction is only up to a year in advance. This also applies to other studies offering superficially excellent results such as [Vounou et al., 2012].

A much simpler variant of MKL was used in [Zhang et al., 2011]. Their method also made use of multiple kernel SVMs, but rather than simultaneously optimising the kernel weights and other SVM parameters, a (linear) kernel was explicitly generated for each modality. A grid search was then performed over kernel weights. At each point in the grid a combined kernel was generated, and used with a conventional SVM to assess accuracy in a cross validation loop. Three kernels were used, representing MRI, FDG-PET and CSF data, and as the kernel weights were constrained to be positive and sum to one the grid search was two dimensional only. MRI features consisted of volumes of GM tissue in 93 ROIs. These were generated by aligning all images rigidly, skull stripping, and registering with [Shen and Davatzikos, 2002] to a labelled atlas. The labels were then propagated to each subject's GM segmentation to calculate the features. For FDG-PET data, the images were rigidly aligned to the corresponding MRI for each subject. The FDG-PET images were then parcellated into the same 93 regions from the atlas, and the mean intensity within each region was taken as a feature. For the CSF data, $\alpha\beta 42$, t-tau and p-tau levels were used as features. All features were then normalised to zero mean and unit standard deviation. In addition to the MKL procedure just described, the authors attempted two other multimodal classification schemes: simple feature concatenating, and an ensemble approach where separate classifiers were constructed for MRI, FDG-PET and CSF data and the results were combined by majority voting. Validation was by tenfold cross validation, with a second tenfold cross validation in each iteration to determine the kernel weights. For classifying AD and control subjects, the MKL method performed substantially better than any single modality, and slightly better than feature concatenation or ensemble multimodal methods. When classifying MCI subjects versus controls, the same pattern was found, albeit with smaller margins. The authors went on to apply their method to predicting conversion in the MCI subjects, with 18 months of follow-up. Sensitivity was 91.5% and specificity 73.4%. However the authors do not state whether this was done in a cross validation using MCI subjects to train, or training a single classifier on AD and controls subjects and applying it to the MCI ones. They also do not say how many MCI subjects were converters so we cannot calculate the overall accuracy. However the MKL approach was successful enough to have been reused by

the authors in some of their subsequent work, such as integration of structural and functional connectivity data in [Wee et al., 2012].

4.2.8 Other approaches

A very common result in all the studies is that separating MCI subjects from AD patients or controls is much more difficult than separating AD patients from controls, and separating converting and stable MCI subjects is more difficult still. This is unsurprising given that we would expect smaller differences between the groups in disease severity to be reflected in smaller differences in image features. However the MCI subjects, and in particular the converters, form a very heterogeneous group. Many of them will in fact have conditions other than AD, but which ones these are cannot be ascertained given the general lack of postmortem confirmation of diagnosis. Meanwhile, the identification of stable MCI subjects is hampered by a limited follow-up period. Many subjects designated as stable would doubtless convert to AD within a slightly longer period. The authors of [Aksu et al., 2011] go so far as to say there are no definitively labelled examples of MCI converters. They proposed to circumvent this by constructing their own ground truth for MCI subjects. This was done by first training an AD versus controls classifier. This was then applied to subjects who were labelled MCI at baseline, for their MRI scans at every follow-up timepoint. This enabled a trajectory to be established for each baseline MCI subject, based on whether their follow-up images were consistently classified as normal or AD. The trajectories were used to label MCI subjects as MCI-c or MCI-s, which were then used to build a second classifier for the MCI subjects. Unfortunately the results were validated with respect to the 'by trajectory' definition of conversion or nonconversion, introducing a form of circular logic: MCI subjects were classified as converters because they look more like converters by trajectory, but converters by trajectory were labelled as such because they resemble each other. Introducing *validation* labels based on an actual clinical criterion would have made this much more valuable, and would still allow *training* labels to be based on something different such as conversion by trajectory.

The notion of using different criteria to train and test is taken to a greater extreme in [Gaser et al., 2013]. This abandons the idea of what the authors call a disease-specific pattern entirely. Instead of using discrete labels representing different disease states in a classification problem, they made use of their BrainAGE score [Franke et al., 2010] in a regression framework. This used a sparse Bayesian, kernel based method, the relevance vector machine (RVM) [Tipping, 2001]. The model was trained on 320 healthy subjects aged 50 or over, taken

from the IXI ¹ and OASIS ² databases. Test subjects comprised 195 MCI subjects from the ADNI database. The regression targets were the subjects' ages. All images were segmented into GM, WM and CSF and spatially normalised with affine registration. The GM segmentations were smoothed and retained for training and testing. The trained regression model was then applied to the test subjects' GM segmentations, to produce an estimated age for them based on the distribution of GM in their brain. This estimated age minus their chronological age is the subject's BrainAGE, indicating the degree to which they are aging abnormally. This has already been shown to correlate with disease severity and poorer cognitive function [Franke and Gaser, 2012]. By varying the threshold of BrainAGE scores for the MCI subjects, a classification accuracy of 75% was obtained for predicting conversion with a three year follow-up period. This was significantly more accurate than CSF, cognitive scores or hippocampal volumes for the same subjects. This may however include some optimistic bias as it appears that the threshold setting was not done inside an LOOCV loop.

4.3 Summary

As is made clear from the great variety of approaches discussed in the previous sections, the problem of predicting conversion of MCI patients to AD is one that has attracted wide interest due to both its challenging nature and clinical relevance. A major side-by-side comparison of some of these methods concluded that none could predict conversion with an accuracy significantly greater than chance [Cuingnet et al., 2010]. Nevertheless, a number of other publications have reported statistically significant accuracies, and while comparison of results is difficult, it appears that there has been an upward trend in accuracy. Excluding results from methodologically dubious procedures such as double dipping, it appears that for predicting conversion within three years from MRI data, the maximum accuracy is about 70-75% [Eskildsen and Ostergaard, 2006, Coupé et al., 2012, Ye et al., 2012]. It is notable that all three of these are somewhat unconventional in their approach, respectively using multiple classifiers stratified by conversion time, hippocampal grading, and sparse stability selection of features, but the classification algorithm used by all is standard or even very simple. This is in accordance with the conclusion in [Sabuncu and Konukoglu, 2014] that the type of classifier itself is relatively unimportant. Meanwhile, a number of recent publications have introduced much more sophisticated learning methods such as deep learning [Suk and Shen, 2013] and autoencoders [Suk et al., 2013] without noticeably increasing the resulting accuracy. Because of this, my thesis begins by introducing an application of GP classification to a simple problem as a

¹<http://www.brain-development.org>

²<http://www.oasis-brains.org>

proof-of-concept. However rather than applying ever more elaborate variants of the classifier, we show how we can use the same algorithm - or in the case of GP regression, a slightly simpler one - to achieve better results by framing the problem of predicting MCI conversion in novel ways. The following chapters show how GPs can be applied to multimodal classification, learning a continuous proxy for disease state rather than binary labels, or automatically weighting the importance of anatomical brain regions in classification. These are shown to advance the state-of-the art accuracy for predicting MCI conversion.

Chapter 5

Classification of Alzheimer's disease patients and controls with gaussian processes

5.1 Introduction

This chapter presents a comparison of GP and SVM methods for the classification of AD and control subjects. A variety of different imaging modalities have been used in previous attempts to perform this, including PET [Gray et al., 2012], and more recently both diffusion weighted and functional MRI [Wee et al., 2012], but the majority have used structural MRI [Barnes et al., 2004, Lerch et al., 2008, Klöppel et al., 2008]. This is because it is known that the early stages of AD are characterised by a pattern of atrophy in grey matter that is readily assessed and quantified on such images. For example, the hippocampus is known to be one of the structures most affected by the disease process so the volume of the segmented hippocampus, normalised by intracranial volume, can distinguish between AD patients and controls with high accuracy [Barnes et al., 2004]. Measurements of cortical thickness across the entire brain offer similar accuracy [Lerch et al., 2008]. This study follows [Klöppel et al., 2008] in using maps of grey matter density across the entire brain, as this approach includes all known areas of atrophy associated with AD.

Use of such images implies very high dimensionality in the data, as the dimensionality is equal to the number of voxels in the images. Various methods have been used to cope with this, such as using complex feature extraction procedures [Davatzikos et al., 2008] or well known methods such as principal components analysis to reduce the dimensionality of the images while attempting to preserve discriminative information, or selecting a subset consisting of the most discriminative features according to some statistical criterion. The most widely used classifier in these types of studies, however, is the support vector machine (SVM), which treats all the training images as points in a (potentially very) high dimensional space and attempts to find a hyperplane separating two labelled groups in the training data. It selects the hyperplane such

that the distance to the closest training data on either side is maximised. Such classifiers can deal directly with images of the entire brain without having to reduce dimensionality.

However, as discriminative classifiers, SVMs produce a simple decision value. Probabilistic predictions have a number of advantages. Firstly, they enable some clinically useful options such as a reject option for uncertain cases and use of decision theory to optimise classification rules. Furthermore, the paradigm of evidence based medicine can be viewed as an example of Bayesian reasoning [Ashby and Smith, 2000]. As a more practical consideration, in a probabilistic setting, classification parameters can be tuned via type-II maximum likelihood rather than computationally expensive cross-validation. This chapter describes the Gaussian process (GP) regression and classification method, which is fully Bayesian, and applies it to classification of AD from structural MRI images.

5.2 Materials and methods

5.2.1 Images

The data used in this study consisted of 60 T1 weighted structural MRI images of healthy controls, and 60 T1 weighted scans of subjects diagnosed clinically with probable AD. All images were obtained from the ADNI database. The two groups were matched for age and sex. This was done by first randomly selecting 60 subjects from the clinical group with fewer subjects overall (AD patients). Then for each selected AD patient, a matching control was selected as a subject from the pool of controls of the same sex and age. If there were multiple subjects with both sex and age matching, one of these was chosen at random, and if there were no subjects matching on both sex and age then the nearest age match was selected. Then the selected 60 subjects in each disease category was further split into 40 subjects for training the classifiers and 20 images to test them. The age and sex matching was preserved during this split by maintaining the pairing between AD and control subjects.

5.2.2 Image processing

To enable a classifier to be constructed using the training images, they were first transformed into the same space. The images were masked to remove non-brain material, the masks generated from brain MAPS [Leung et al., 2011] and then used to perform groupwise registration. All images were repeatedly registered to a target image in an iterative procedure. At the end of each iteration, all registered images were averaged together to create an updated target image, with a randomly chosen image serving as the target in the first iteration. Initially, all images were rigidly registered to avoid bias to the target image. This was followed by a round of affine registrations, and then by 10 rounds of nonrigid registrations. All registrations were performed

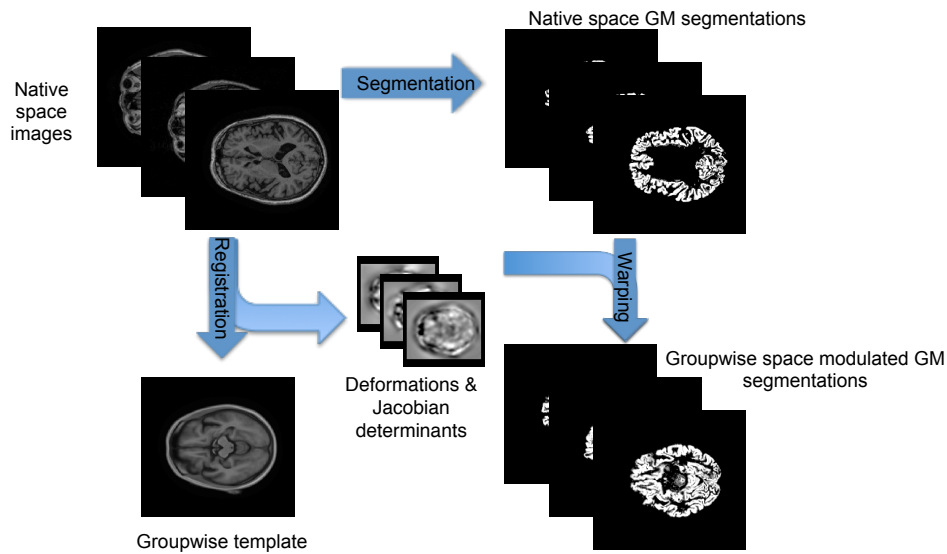


Figure 5.1: Pipeline to produce modulated GM images in a common space. Native T1 images are warped into a common space via a groupwise registration procedure. Native GM density maps are also produced from the native T1 images, and are then also warped in to the common space to generate images whose voxels serve as features in classification.

by NiftyReg [Modat et al., 2010], a registration toolkit that performs fast diffeomorphic non-rigid registrations. The native space images were probabilistically segmented using the NiftySeg [Cardoso et al., 2011] tool into five tissue types: white matter, cortical grey matter, external cerebrospinal fluid, deep grey matter and internal cerebrospinal fluid. The transformations from each image’s native space to the space of the final groupwise template were then applied to the segmented native space images to warp the segmentations to the template space. Finally, the segmentations were modulated by multiplying each voxel by the Jacobian determinant of the deformation field transforming it from its native space to the template space. This step ensures the total volume of tissue remains constant. Spatial smoothing of the image was not performed as part of this study.

The pipeline is summarised in figure 5.1.

5.2.3 Gaussian process regression and classification

The primary purpose of this work is to demonstrate that GP classification can provide equivalent results to an SVM. We make use of the GPML implementation of GP classification (<http://www.gaussianprocess.org/gpml/code/matlab/doc/>). A detailed explanation of GP regression and classification is given in an earlier section (3.7), and in the documentation accompanying the GPML software [Rasmussen and Williams, 2006]. GP classification was applied in two stages; firstly hyperparameters were learned from the training data, and then the

log probabilities of the test data having labels equal to 1 (that is, having AD) were calculated using GPs. These log probabilities were then exponentiated and thresholded at 0.5 to provide a hard classification.

5.2.4 SVM calculations

To provide a baseline classification accuracy for these data, classification was also performed with the widely used LIBSVM library (<http://www.csie.ntu.edu.tw/~cjlin/libsvm>) implementation of the support vector machine. These are also explained in more detail in section 3.6. All analysis was conducted in MATLAB (The MathWorks Inc., Natick, MA, 2011).

5.3 Results

Two sets of results are presented, those obtained from GP classification and those from SVM classification. Both sets of results were generated by training a classifier on the 80 designated training subjects, and then evaluating the accuracy of the resulting model in classifying the 40 subjects set aside for testing. The SVM correctly classified 31 out of the 40 test subjects, giving an accuracy of 77.5%. The GP model correctly classified 33 out of the 40 test subjects, equal to an accuracy of 82.5%. However, the difference in absolute classification accuracy is not statistically significant. Both classifiers appeared to be well balanced, with the SVM having a sensitivity (assuming AD = positive) of 75% and specificity of 80%, and the GP having a sensitivity of 80% and specificity of 85%. The results using SVM classification are also in line with those given in [Cuingnet et al., 2010], a large study comparing various methods of classifying Alzheimer's disease that included a procedure very similar to this one. ROC curves and the associated areas under the curve are given in figure 5.2.

As the figure shows, the ROC curves for the two classification methods are virtually identical. The areas under the two ROC curves are also very similar, at 0.890 for the SVM and 0.888 for the GP.

5.4 Discussion

This chapter presented the first application of Gaussian process classifiers to distinguish between healthy elderly controls and subjects with AD. While many previous studies have achieved similar accuracy on this type of data [Klöppel et al., 2008, Cuingnet et al., 2010], they have all used classification algorithms that simply produce a binary decision between two classes, mostly using support vector machines. While the output of SVMs can be converted to probabilities, the methods for doing so are somewhat ad-hoc due to the SVM's non proba-

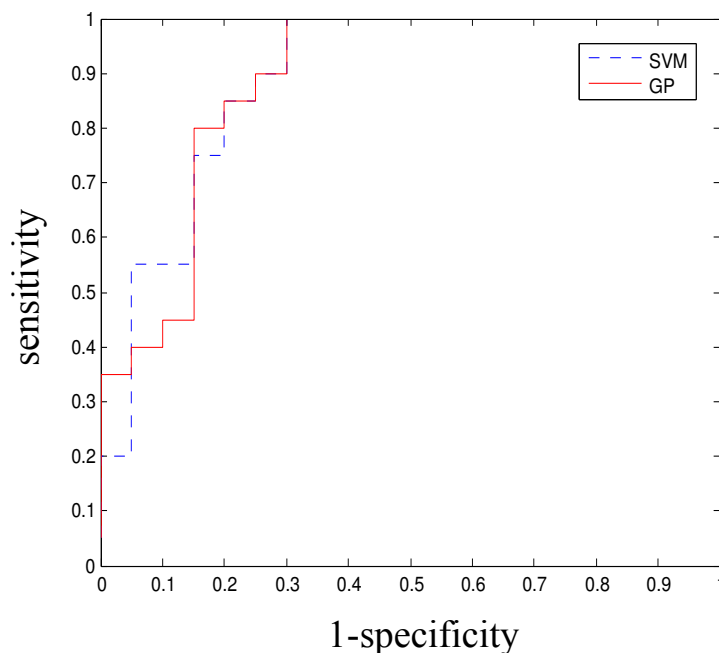


Figure 5.2: AUC curves for classification of 40 AD and control subjects with GP and SVM.

bilistic objective function [Platt, 2000]. The probabilistic predictions made by the GP classifier were thresholded at $p(y^* = 1) = 0.5$ to produce a binary decision that can be directly compared to results from an SVM. The results from the two methods are extremely similar, showing that it is possible to switch to the probabilistic framework without a significant loss of accuracy. Moreover, all the subjects wrongly classified by the GP were also wrongly classified by the SVM, suggesting that the decision boundaries applied by the two classifiers are similar. However, it would be a waste to limit ourselves to the use of probabilistic predictions in this way. This formulation allows automatic variable selection via maximum likelihood that avoids the effects of overfitting. The equivalent feature selection must be done via computationally expensive cross validation or inaccurate filtering for other methods. While it was not possible to fully apply such techniques in this method, using a labelled atlas to aggregate the grey matter within anatomical regions as in [Chu et al., 2010] would allow ARD to be used. Alternatively, the feature dimensionality could be reduced by focusing on a small region of interest such as the area around the hippocampus as was also done in [Barnes et al., 2004], or by aggregating voxels into anatomical regions [Chu et al., 2010]. In the context of multikernel learning, a probabilistic formulation also offers the automatic tuning of the kernel mixing weights. Probabilistic predictions also open up possibilities that cannot be easily done otherwise, such as a reject option in which more uncertain classifications are passed to a different classifier or human expert, and tuning of the probability threshold to maximise positive predictive value of the test in a clinical

context [Ashby and Smith, 2000].

5.5 Conclusions

This experiment achieved its aim of successfully performing GP classification of AD patients and controls. As the data used for classification were GM density maps across the whole brain, they were extremely high dimensional. While SVMs have traditionally been used in the resulting regime of high dimensionality and low sample size, this study has demonstrated that GP classification is equally able to cope with this problem and yields a binary accuracy statistically indistinguishable from an SVM. The following chapter goes on to show how we can further utilise the advantages of GPs over SVMs using more sophisticated techniques.

Chapter 6

Multiple kernel learning for prediction of conversion to AD

6.1 Introduction

This chapter presents a study of early diagnosis of AD with multiple kernel learning. As in many previous studies, it focuses on patients with mild cognitive impairment (MCI) [Petersen et al., 1999]. MCI is typically defined as a state where patients have isolated memory deficits that are not severe enough to affect normal living. Studies have shown that MCI patients convert to AD at an annual rate of 10-15% per year [Braak and Braak, 1995]. MCI patients who do not convert to AD either develop other forms of dementia, remain stable, or in a small minority, revert to a nondemented state. Therefore predicting which MCI patients will develop AD in the short term (i.e. within a few years) and which will remain stable is extremely relevant to future treatments. Although a definitive diagnosis of AD can be made only at autopsy, in practice expert clinicians diagnose AD based on clinical history and batteries of cognitive tests. However these standard clinical tests are not able to identify the more subtle patterns of the disease process at this early stage, so more advanced methods are required.

The automated methods used to discriminate between stable (MCI-s) and converter (MCI-c) patients are similar to those used for diagnosis of AD. These automated tests use imaging and other biomarker data, and can now diagnose AD with an accuracy of about 90%, as well as expert clinicians can using more traditional methods [Beach et al., 2012]. While a number of different imaging modalities have been proposed for this application, the majority have used structural MRI, as atrophy in specific brain regions is one of the most established hallmarks of AD. The features used in classification derived from structural MRI can take a number of forms, including voxel level maps of grey matter density [Nho et al., 2010, Klöppel et al., 2008, Fan et al., 2007], volume or shape [Barnes et al., 2004, Gerardin et al., 2009, Zhang et al., 2011], or cortical thickness measure-

ments [Desikan et al., 2009, Eskildsen et al., 2013, Lerch et al., 2008, Querbes et al., 2009]. These features can be calculated over the whole brain or specific structures known to be affected by AD, such as the hippocampus. A comprehensive review and comparison of these methods, focused mainly on the type of MRI-derived features used rather than which machine learning algorithm was implemented, is given in [Cuingnet et al., 2010].

Looking beyond structural MRI, FDG-PET is capable of measuring the level of glucose metabolism in the brain. Studies have shown that glucose metabolism is reduced in some regions in patients before they develop AD [Drzezga et al., 2003, Mosconi et al., 2010] and this may be used to classify AD patients from controls or predict conversion from MCI to AD [Gray et al., 2012]. Biomarkers extracted from cerebrospinal fluid (CSF) have shown utility in the diagnosis of AD or MCI. In particular, CSF levels of total tau protein (t-tau) and phosphorylated tau (p-tau) proteins, known to be implicated in the formation of neurofibrillary tangles that cause atrophy in AD, are elevated in AD patients, while levels of the amyloid-42 (a42) peptide in CSF fall [Fjell et al., 2010a, Holtzman, 2011]. Measurements of amyloid load in the brain using amyloid PET have shown similar results [Rowe et al., 2010]. Also, variants of the apolipoprotein E (ApoE) gene affect the risk of developing AD [Corder et al., 1993, Corder et al., 1994].

These different types of biomarker data have been shown to be complementary, meaning that they provide superior classification when used in combination than when either is used individually, even if they are correlated [Fjell et al., 2010b, Landau et al., 2010]. Thus a number of studies have sought to make use of multiple biomarker types in classification. Structural MRI is used in combination with genetic data in [Vemuri et al., 2008] and with CSF biomarkers in [Vemuri et al., 2009] and [Davatzikos et al., 2008]. Structural MRI data, FDG-PET and CSF data are used in [Hinrichs et al., 2011, Walhovd et al., 2010, Zhang et al., 2011]. A noteworthy disadvantage of multimodal methods is that the problem of missing data is multiplied, as a subject must be discarded entirely or the missing data must be synthesised if it is not present in any one of the modalities used. An approach to tackle this issue is presented in [Yuan et al., 2012].

The most popular classification method is the support vector machine (SVM), due to its accuracy and ability to cope with very high dimensional data. Another advantage of the SVM is its ability to use the kernel, a matrix of size N by N that summarises the distances or covariances among N training subjects. This can be applied to learn from multimodal data. Rather than simply concatenating the features from different modalities into a single vector, an individual kernel can be formed from each modality and then a combined kernel generated as a weighted sum of the individual ones. Both [Zhang et al., 2011] and [Hinrichs et al., 2011] use this approach, but find the individual kernel weights in a different fashion. The former chooses them by a grid

search for the weights giving the best accuracy in a nested cross validation loop. This method is reused in a number of subsequent publications by the same authors, generally being applied after a more sophisticated feature selection process. For example, in [Liu et al., 2014], the feature selection step is used to jointly select a sparse set of features from MRI and FDG-PET data using a multi-task objective function designed to preserve the inter-modality relationship. The resulting features are then used to generate separate MRI and FDG-PET kernels, which are optimally combined using the existing grid search method. By contrast, in [Hinrichs et al., 2011] the subkernel weights are set by optimising them alongside the standard SVM parameters and with the standard SVM objective function. This specific algorithm is sometimes referred to as multiple kernel learning [Bach and Lanckriet, 2004], but the term is often used more broadly to refer to all methods that combine multiple subkernels to produce a final kernel.

The study presented in this chapter is a different method using a combination of structural MRI, FDG-PET, CSF and ApoE data to classify MCI-s and MCI-c patients. Primarily, this uses Gaussian process (GP) classification, which is a probabilistic classification algorithm. Bishop [Bishop, 2007] lists four general advantages of a probabilistic framework, however, for this particular study we would add two more which we feel to be particularly relevant: firstly, the option to tune free parameters automatically from the training data, avoiding the need for computationally expensive cross-validation loops, and secondly, that the probabilistic decisions produced by GP classification allow a great deal of flexibility in their interpretation. Despite the fact that for convenience, disease is frequently characterised as a binary distinction, such as healthy or AD patient, each subject in fact occupies a point on a continuous spectrum of disease severity, as is reflected by the concept of MCI. Probabilistic classification allows us to identify the position of subjects on this spectrum, enabling disease staging, stratification, or event based modelling [Fontejn et al., 2012]. Probabilistic decisions can also be made into a binary classification simply by thresholding, and our previous work shows that this method offers accuracy as good as an SVM on voxel level data for the diagnosis of AD [Young et al., 2012]; hence no diagnostic information is lost by choosing a probabilistic classification algorithm. While an SVM's output can be interpreted probabilistically by transforming the decision value with a sigmoid function, this method is a rather ad hoc modification to a binary classifier, and does not offer the principled formulation and automatic parameter tuning of GP classification.

This previous work is, to my knowledge, the only previous application of GP classification to AD. GPs have been used previously in a regression context with fMRI data in [Marquand et al., 2010], and for classification of structural MRI data in Huntington's disease by [Chu et al., 2010]. They have not been previously applied for multimodal medical image

classification. Here we use four types of data are used, as well as a comparison of two methods of setting the kernel weight, one very similar to that given by [Zhang et al., 2011] and the other a probabilistic method that is more natural within the GP paradigm. Finally the results are compared to those obtained by an SVM on the same data, again using the method of [Zhang et al., 2011] for setting kernel weights in the multikernel paradigm.

The training population comprises healthy controls and AD patients, allowing us to interpret the results in the MCI population as a risk score for conversion to AD. We introduce a new method for the validation of probabilistic predictions, which show that the predicted probability of conversion is a good estimate of the actual chances of conversion. As well as interpreting the results probabilistically, we also obtained a binary classification into MCI-s and MCI-c by adaptively thresholding the probabilities, resulting in a highly accurate prediction of conversion.

6.2 Materials and methods

All data were obtained from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database. This is described in detail in section 1.5.1. For up-to-date information, see <http://www.adni-info.org>.

6.2.1 MRI data

Images were all T1 weighted structural MRI scans from 1.5T scanners acquired using a 3D MPRAGE sequence, taken at the baseline time point for each subject. Back-to-back scans were taken for each subject, and the best scan of the pair for each subject determined by visual inspection. The images were then post-processed to correct for gradient warping, B1 non-uniformity and intensity non-uniformity and underwent phantom based scaling correction. Postprocessed images were downloaded as DICOM files, and were then converted to NIfTI format for further processing.

6.2.2 PET data

Images were again all taken from the baseline scan for each included subject. Images were acquired by scanning 30-60 minutes post injection using scanner-specific protocols. Six five minute frames were acquired for each subject, and then co-registered and averaged. The average images were then rigidly registered to a standard space, and the individual native space frames registered to the standard space average and averaged and intensity normalised in the standard space. Finally, the average images in the standard space were smoothed with a scanner-specific kernel [Joshi et al., 2009] to a uniform isotropic resolution of 8mm FWHM, which is approximately the resolution of the lowest resolution scanners used in ADNI. The postprocessed scans were downloaded as DICOM images.

6.2.3 ApoE data

Variants of the Apolipoprotein E (ApoE) genotype are known to affect the risk of developing sporadic AD in their carriers. Each individual has two copies of this gene, one inherited from each parent. The most common allele is ApoE ϵ 3, but carriers of the ApoE ϵ 4 variant are at heightened risk of AD, whereas the ApoE ϵ 2 variant confers some protection on carriers [Corder et al., 1993, Corder et al., 1994]. The ApoE genotype of each subject was recorded as a pair of numbers indicating which two alleles were present. ApoE genotype is determined from a 10 ml blood sample taken at screening time, and sent overnight to the University of Pennsylvania AD Biomarker Fluid Bank Laboratory for analysis. ApoE genotype was available for all subjects for which we had imaging data.

6.2.4 CSF data

CSF samples of 20ml volume were obtained from subjects by a lumbar puncture with a 24 or 25 gauge atraumatic needle around the time of their baseline scan. All samples were sent on dry ice on the same day as they were drawn to the University of Pennsylvania AD Biomarker Fluid Bank Laboratory, where levels of the proteins (a42, total tau, and phosphorylated tau) were measured and recorded. By design, only a subset of ADNI subjects had measurement of CSF levels. All three measured protein levels (t-tau, p-tau, and a42) were used in constructing a CSF kernel.

6.2.5 Subjects

All ADNI subjects were between 55 and 90 years old, spoke English or Spanish, and had a study partner able to provide an independent assessment of functioning. All subjects were willing to undergo neuroimaging and agreed to longitudinal follow up, and a subset was willing to undergo lumbar punctures. Subjects with specific psychoactive medication were excluded. Inclusion criteria for healthy controls (HC) are MMSE scores between 24 and 30, a CDR of 0, non-depressed and non-demented. Ages of the HC subjects were roughly matched to those of the AD and MCI subjects. For MCI subjects, the criteria are an MMSE score between 24 and 30, a memory complaint, objective memory loss measured by education adjusted scores on the Wechsler Memory Scale Logical Memory II, a CDR of 0.5, absence of significant levels of impairment in other cognitive domains, essentially preserved activities of daily living, and an absence of dementia.

For AD subjects, the criteria are an MMSE score between 20 and 26, CDR of 0.5 or 1.0, and meeting NINCDS/ADRDA criteria for probable AD. Subjects are designated as HC, AD or MCI at the time of the baseline scan, and for the purposes of this study MCI conversion status

Disease status	<i>n</i> (<i>n</i> female)	mean age (sd)	mean MMSE (sd)
NC	73 (27)	75.9 (4.6)	28.9 (1.2)
MCI-s	96 (34)	75.6 (7.0)	27.2 (1.7)
MCI-c	47 (17)	74.5 (7.4)	26.9 (1.8)
AD	63 (24)	75.2 (6.6)	23.6 (2.0)

Table 6.1: Demographics of PET group. NC = normal control, MCI-s = stable MCI, MCI-c = converting MCI, *n* = number of subjects, sd = standard deviation

Disease status	<i>n</i> (<i>n</i> female)	mean age (sd)	mean MMSE (sd)
NC	36 (12)	74.2 (4.2)	28.8 (1.3)
MCI-s	42 (16)	75.4 (7.0)	27.3 (1.6)
MCI-c	30 (11)	75.5 (7.6)	26.5 (1.8)
AD	35 (12)	75.2 (6.7)	23.9 (2.0)

Table 6.2: Demographics of PET-CSF group. NC = normal control, MCI-s = stable MCI, MCI-c = converting MCI, *n* = number of subjects, sd = standard deviation

is decided by whether subjects who were MCI at baseline were subsequently diagnosed as AD at any stage during the subsequent 36 month follow-up period.

A total of 682 subjects with baseline 1.5T MRI scans were available. Of these, the image parcellation procedure was run on 679, the manually generated brain masks required for the parcellation being unavailable for three. Of these 679 subjects, FDG-PET scans were also available for 286. Seven of these were diagnosed as MCI at baseline but as healthy at follow-up time points and were excluded as reverts, leaving a total of 279 subjects available for the study. The demographics of this group (referred to as the PET group) are given in table 6.1.

This experiment also examined the effect of using CSF in the multimodal classification. As there was relatively little overlap between the groups of patients given CSF biomarker testing as well as FDG-PET scans, the subset of the PET group for which full CSF data was also available (referred to as the PET-CSF group) was much smaller at a total of 143 subjects. The demographics of the PET-CSF group are given in table 6.2

In the PET group, 47 out of 143 (33%) of MCI subjects are converters. As conversion is defined over a three year follow-up period, this is equivalent to an annualised conversion rate of 12.5% per year, in line with other studies. Subjects diagnosed as MCI at baseline in ADNI are reassessed after approximately 6, 12, 18, 24 and 36 months, which allows us to roughly find the time after which they converted. The conversion times for the 47 MCI-c subjects in the PET

t	n
$t < 6$	5
$6 < t < 12$	15
$12 < t < 18$	9
$18 < t < 24$	14
$24 < t < 36$	4

Table 6.3: Times of conversion t , in months, for subjects in the PET group.

group are listed in 6.3

6.2.6 MRI image processing

To produce GM probability maps in a common space for classification, roughly the same procedure as [Klöppel et al., 2008] was used. However the processing was done using different image processing software, and with an additional step of masking the images to include only regions known to be affected by AD.

Initially the native space, preprocessed scans were probabilistically segmented using the open source NiftySeg tool [Cardoso et al., 2011]. Based on the expectation maximisation algorithm, this method produces probabilistic maps for five tissue types: white matter, cortical GM, external cerebrospinal fluid, deep GM and internal cerebrospinal fluid.

The native space, preprocessed scans were also anatomically parcellated into 83 regions. This was with a multi atlas segmentation propagation algorithm [Cardoso et al., 2012]. A library of 30 atlases manually labelled with 83 anatomical regions [Gousias et al., 2008] was used as a basis for the segmentations. In order to segment a new image, all the atlases and respective manual labels were first nonrigidly registered to this image. After registration, the manual labels of the locally most similar atlases were fused using a label fusion strategy based on an extension of the STAPLE algorithm [Warfield et al., 2004] to produce a final parcellation. The regions used in the classification process were chosen according to Braak and Braak [Braak and Braak, 1995] and are listed in 6.4. These regions were then intersected with the GM tissue segmentations obtained above.

All images were transformed into the same anatomical space in order to provide consistent anatomy at each voxel for the classifier. The images were masked to remove non brain material, and then used to perform groupwise registration. All images were repeatedly registered to the same target image in an iterative procedure. At the end of each iteration, all registered images were averaged together to create an updated target image, with a randomly chosen image serving as the target in the first iteration. Initially, all images were rigidly registered to avoid

Label numbers	Regions
1, 2	Hippocampus (R and L)
3, 4	Amygdala (R and L)
5, 6	Anterior temporal lobe, medial part (R and L)
7, 8	Anterior temporal lobe, lateral part (R and L)
9, 10	Parahippocampal and ambient gyri (R and L)
11, 12	Superior temporal gyrus, posterior part (R and L)
13, 14	Middle and inferior temporal gyrus (R and L)
15, 16	Fusiform gyrus (R and L)
24, 25	Cingulate gyrus, anterior part (R and L)
26, 27	Cingulate gyrus, posterior part (R and L)

Table 6.4: Regions included in GM segmentations. Label numbers are taken from the atlas used to perform the parcellation [Gousias et al., 2008]. R and L designated the hemisphere (Right and Left).

bias towards the chosen target. This was followed by a single round of affine registration, and then by 10 rounds of nonrigid registrations. All registrations were performed using NiftyReg [Modat et al., 2010], a registration toolkit that performs fast diffeomorphic nonrigid registrations. When the registrations had all been completed, the resulting deformations from each image’s native space to the final average image were applied to the anatomically masked native space segmentations to bring them into the groupwise space. The registered, anatomically masked segmentations were modulated by the Jacobian determinants of this final deformation. This ensures the total volume of tissue remains constant [Ashburner and Friston, 2000]. As a final step, the registered, anatomically masked and Jacobian modulated cortical GM and deep GM segmentations were summed to produce an overall GM density map for the AD relevant regions in a common space for all subjects.

6.2.7 PET image processing

The PET images had already been through substantial postprocessing, as discussed above. After downloading, they were registered to the native space MRI image of the same subject, again using the NiftyReg software. Then masks generated from the structural MRI parcellations were overlaid on each subject to calculate the total activity within each region from the PET image. The mean activity within each region was then used as a feature for classification.

6.2.8 Gaussian process classification

The resulting high dimensional image and biomarker data were then used to construct a GP classifier based on HC and AD subjects. For a full explanation of GP classification, we refer the reader to section 3.7 and [Rasmussen and Williams, 2006]. Here, we give a brief recapitulation of GP classification and give further details on the aspects that pertain to multimodal classification. All learning of hyperparameters and GP calculations were done using the GPML toolbox for MATLAB which was also used to analyse results.

Gaussian process classification can be seen as kernelised Bayesian extension of logistic regression. A Gaussian process, essentially a multivariate Gaussian, forms the prior on the value of a latent function, which is then mapped to the (0,1) interval through a sigmoid, which represents the probability of a subject belonging to a particular class. The exact prior is a function of the training data and labels, and a set of hyperparameters that control the shape of the prior. During the training phase, the hyperparameters are learnt from the training data and labels by type-II maximum likelihood. The likelihood of the training data and labels with respect to the hyperparameters is maximised using a conjugate gradient descent optimisation method. Once the hyperparameters have been set, predictions on unseen data are made by integrating across this prior. In the regression case, this is analytically tractable, but for classification it is not, due to the sigmoidal response function, so an approximation must be made instead. A number of different approximation schemes can be used, but all our experiments use the expectation propagation algorithm [Minka, 2001]. This has been shown to offer a good compromise of accuracy and computation time for GP classification [Nickisch and Rasmussen, 2008].

6.2.9 Gaussian process classification as a multimodal kernel method

Note that the GP classifier is based on a kernel matrix, \mathbf{K} , representing the covariance among training subjects. This is a symmetric positive definite matrix where entry (i, j) is a covariance or some function of distance between training subjects i and j . As such, this means that GP classification belongs to the family of kernel methods, as do SVMs, and all the rules for constructing valid kernels apply: in particular, a positive sum of valid kernels is a valid kernel, and a valid kernel multiplied by a positive scalar is also a valid kernel. The covariance between the i th and j th subject, \mathbf{K}_{ij} , is a kernel function of the feature vectors for the i th and j th subject, \mathbf{x}_i and \mathbf{x}_j and a hyperparameter or set of hyperparameters θ , which are learnt from the training data by type-II maximum likelihood. For multimodal classification, the rules for producing new kernels mean that we can define our kernel function as the weighted sum of a number of subkernels, each of which has been calculated from the feature vectors representing a particular type of data or modality for each subject. Each subkernel is constructed from a linear kernel

function, which is the scalar product of \mathbf{x}_i and \mathbf{x}_j : $k_{ij} = \mathbf{x}_i \cdot \mathbf{x}_j$. Each subkernel has a scaling hyperparameter representing the modality's weight in the overall kernel, and there is also a single bias term. So in the case of multimodal classification using each subject's MRI, PET and ApoE data the overall kernel is

$$\mathbf{K}_{ij} = \alpha^{MR} \mathbf{x}_i^{MR} \cdot \mathbf{x}_j^{MR} + \alpha^{PET} \mathbf{x}_i^{PET} \cdot \mathbf{x}_j^{PET} + \alpha^{ApoE} \mathbf{x}_i^{ApoE} \cdot \mathbf{x}_j^{ApoE} + \beta \quad (6.1)$$

where α are hyperparameters representing the weight given to each modality subkernel, and β is a hyperparameter representing the bias in the combined kernel. Thus θ is now a set of four hyperparameters which are learnt from the training data by maximum likelihood. In this way we can automatically set the kernel weights without needing to resort to a grid search with cross validation. This is possible as the GPML software allows complex covariance functions to be specified. It allows us to apply masks to include only certain columns of the training data to be used in a covariance function, so we can learn separate covariance kernels for the MRI, PET and ApoE data. The ApoE kernel is based on representing each subject as a vector of length two, encoding each allele as an element of the vector, so for example a subject with one copy of the $\epsilon 3$ allele and one of the $\epsilon 4$ would be encoded as (3, 4). More sophisticated kernels have been developed for genetic data and these may improve results further.

For the PET group, we also do a grid search for the kernel weights to compare the results of this method of setting the kernel weights to the maximum likelihood method and to [Zhang et al., 2011]. Each MCI test subject in turn is left out, and a GP classifier is trained on all AD and control training subjects for each legitimate combination of α . The best values of α are chosen empirically as the ones offering the most accurate classification on the $n - 1$ remaining MCI test subjects. As accuracy is a coarse measure, any ties are broken with the information theory based metric of classification quality suggested in [Rasmussen and Williams, 2006]. Finally the classifier offering the best accuracy was tested on the left out MCI subject, and the process repeated until all MCI test subjects had been classified. Due to the leave-one-out loop and the need to do one tuning classification for every combination of parameters within each iteration of the loop, this method is very time consuming if more than a handful of parameters have to be tuned. Hence to make the whole classification procedure tractable, values of α are constrained to be positive and sum to one, with no bias term, as in [Zhang et al., 2011]. The resulting two-dimensional parameter space is searched with increments of 0.1 for both parameters. Figure 6.1 represents the multikernel approach.

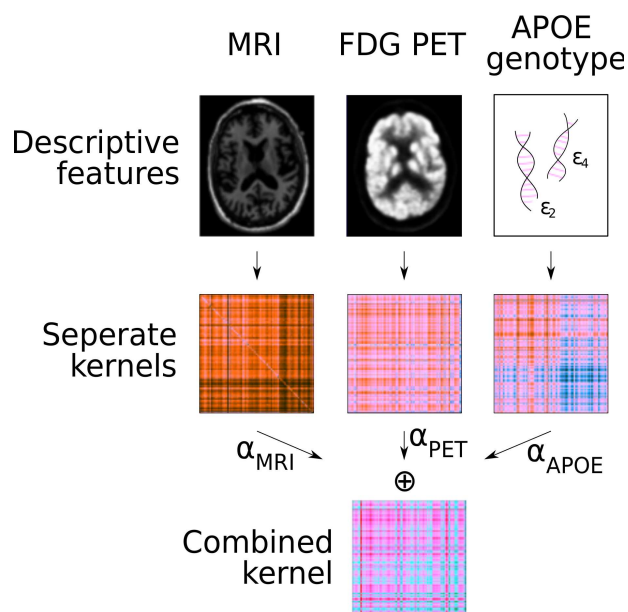


Figure 6.1: Pipeline by which kernels are constructed from features extracted from each type of data, before being summed to produce a combined kernel.

6.2.10 SVM classification

To put the results obtained by GP classification in context and compare them to a more widely used method, SVM classification on the same datasets was also performed. This was done with use of the open source libsvm library, with the C parameter left at its default setting and linear kernels, but used precomputed kernels both for the sake of speed and to facilitate multikernel classification. Training and testing kernels were constructed for all three modalities in the PET group (MRI, PET and ApoE) and all four in the PET-CSF group (MRI, PET, ApoE and CSF). Kernel weights are again set using the method of [Zhang et al., 2011] as described in section 6.2.9. The weight setting is done within a leave one out scheme, where the testing (MCI-s and MCI-c) subjects are repeatedly split into one subject used for testing and the remaining ones used for tuning the kernel weights until each MCI subject has been left out; in this way it is possible to tune on the training population and thus avoid introducing optimistic bias. We also tried to set the kernel weights using the training (NC and AD) subjects for tuning, by performing a leave-one-out cross validation on the training subjects at each legitimate combination of kernel weights. To break ties between parameter settings showing equal accuracy, we use the mean distance from the margin of correctly classified test subjects minus the mean distance from the margin of incorrectly classified test subjects as a metric of SVM quality. We also experimented with normalising training and testing data using a z-score to help combine different modalities on the same scale.

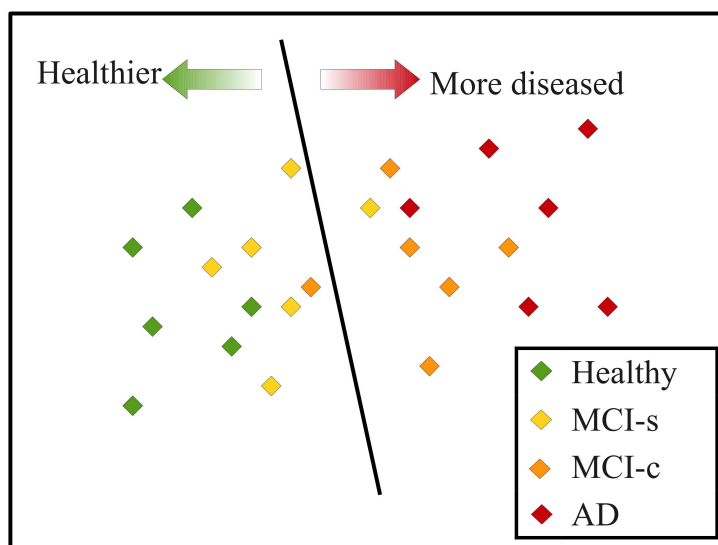


Figure 6.2: Relationship between AD and MCI classification. As AD, control and MCI subjects exist on a spectrum of disease severity, MCI-s subjects can be seen as control-like and MCI-c ones more AD-like. Hence a classifier trained to separate AD and control subjects may also be applied to separate MCI-c and MCI-s subjects.

6.2.11 Classification strategy

Rather than both training and testing the classifier on MCI-s and MCI-c subjects in a cross-validation loop, training data consists of AD and healthy subjects, and then results are obtained by applying the resulting classifier to the MCI population. This approach to classification of MCI subjects is widely used and was adopted here as it obtained substantially better results than those obtained by the training on MCI regime in all our preliminary work. The hypothesis justifying this is that the subpopulation of MCI subjects that are stable are more healthy-like (although some will go on to convert beyond the follow up period used for defining conversion, which is probably a factor in the limited accuracy of predictions of MCI conversion), while those who go on to develop dementia are more AD-like, as is consistent with our contention that discrete disease states are an approximation to a continuous disease spectrum.

This means a classifier that successfully separates AD and control subjects will also be able to distinguish between MCI-c and MCI-s to some degree. This notion, illustrated in figure 6.2 has been used with some success for this problem previously [Ferrarini et al., 2009, Singh et al., 2012]. As previously mentioned, however, when using a combined kernel with grid search we can use the MCI subjects not being classified to tune the kernel mixing parameters.

6.2.12 Validation

The results of GP classification are numbers between 0 and 1 representing the estimated probability that a test subject belongs to a particular class, in our case the class of MCI-c. A simple way to binarise these probabilities is to threshold them at 0.5. We do this, and report the resulting accuracy, sensitivity and specificity. However this approach has two disadvantages. Firstly, as the model is trained on one population (AD and control) and tested on another (MCI-s and MCI-c), this would be the correct threshold value if the test population were in some sense exactly half way between the two classes of the training population, but there is no reason to believe this is necessarily the case. Secondly, setting the cut point at 0.5 leads to varying balances between sensitivity and specificity among the different methods, making them hard to compare. Because of this, the test probabilities are used to determine the cut point that results in the closest possible value of sensitivity and specificity. Then the overall correct classification rate at this cut point is found and reported, as by definition it will be very close to both the sensitivity and specificity. This is done in a leave-one-out framework to avoid optimistic bias in the balanced accuracy. Finally, the probabilities are used to calculate the AUC, for easy comparison with results from other studies. For both PET and PET-CSF groups we also report the balanced accuracy for classification using each modality alone, except for the ApoE. This is left out because ApoE data consists of pairs of alleles labelled 2, 3 or 4. As order does not matter this means each subject can be at one of only six points in two-dimensional ApoE data space (in practice five points as one combination does not occur in our data), so an ApoE only classifier would produce probabilities that could only be one of five discrete values, making further analysis meaningless. The significance of the difference in balanced accuracy between multimodal classification and unimodal classification is assessed for both the PET group and PET-CSF group with McNemar's test [McNemar, 1947] if there appears to be a substantial difference. The balanced accuracies are derived from the probabilities before they are corrected for bias with the procedure described in section 6.3.2. We found that balanced binary accuracies derived from the corrected probabilities tended to be slightly lower.

However, to only do this would be to neglect the probabilistic information contained in the output of the GP. The probabilities can also be treated as risk scores for conversion to AD, and then used to determine how well they function as estimates of the actual chances of conversion. As each subject either does or does not convert to AD, this cannot be assessed at the individual level. Instead all MCI subjects are binned into eight equal intervals covering the range (0,1) by their risk score. For each of the eight intervals, the centre value of the interval is labelled the predicted risk. Then the empirical risk is calculated for each interval as the proportion of

Modality	acc (%)	sens (%)	spec (%)	bal acc (%)	AUC	p(M)	p(P)
MRI	64.30	53.20	69.8	61.50	0.643	-	-
PET	65.00	66.00	64.60	65.70	0.767	-	-
All (ML)	69.90	78.70	65.60	74.10	0.795	0.0162	0.0247
All (GS)	67.10	76.60	62.50	70.60	0.751	0.0865	0.2301

Table 6.5: Accuracy of methods in the PET group with GP classification. 'All' modalities indicates MKL with MRI, PET and ApoE kernels. ML and GS are, respectively, the maximum likelihood and grid search methods of setting the subkernel weights α . p values are of difference in classification vs. indicated single modality, M for MRI and P for PET.

patients in the interval that do in fact convert. Finally, the root mean square error between predicted and empirical risk is calculated as a measure of how well the risk scores from GP classification predict the actual risk of conversion. The number of intervals was chosen to provide the best balance between the demands for good statistics both within and between the bins.

The decision values obtained from SVM classification represent a signed distance from the optimal hyperplane determined from the training data, the sign indicating on which side of the hyperplane a test subject falls and thus to which class it is predicted to belong. We report the accuracy, sensitivity and specificity from the sign of the decision values (equivalent to thresholding the decision values at 0). We also perform a procedure to find the threshold producing the accuracy that best balances sensitivity and specificity in the same manner as we did for GP posterior probabilities, and finally calculate an AUC from the decision values.

6.3 Results

6.3.1 Accuracy of binary classification

The balanced accuracy, AUC, and p-value for comparison of multimodal methods with unimodal ones for the PET group are shown in table 6.5 for the GP results, and in table 6.6 for the SVM results.

The result in the last row of table 6.6 was obtained by using the MCI subjects as a tuning set, and normalising training data with a z-score, and then normalising testing data using the mean and standard deviations from the un-normalised training data. All other combinations of choices of tuning set and normalisation produced inferior results.

The same accuracy measures for the PET-CSF group are shown in table 6.7 for GP classification and table 6.8 for SVM. For the GP, we do not perform the grid search method due to

Modality	acc (%)	sens (%)	spec (%)	bal acc (%)	AUC
MRI	58.70	53.20	61.50	58.70	0.629
PET	69.90	55.30	77.10	67.10	0.762
All (GS)	65.70	68.10	64.60	67.80	0.731

Table 6.6: Accuracy of methods on the PET group with SVM classification. 'All' modalities indicates MKL with MRI, PET and ApoE kernels. Only the grid search (GS) method of setting the subkernel weights α can be used with SVM.

Modality	acc (%)	sens (%)	spec (%)	bal acc (%)	AUC	p(M)	p(P)	p(C)
MRI	63.9	76.7	54.8	61.1	0.682	-	-	-
PET	66.7	80.0	57.1	69.4	0.789	-	-	-
CSF	55.6	73.3	42.9	56.9	0.575	-	-	-
Im, ApoE	68.1	83.3	57.1	72.2	0.823	0.186	0.773	0.072
All	68.1	90.0	52.4	72.2	0.763	0.201	0.823	0.015

Table 6.7: Accuracy of methods on the PET-CSF group with GP classification. All modalities indicates MKL with MRI, PET, CSF and ApoE kernels, Im + ApoE is the image data (MRI and PET) and ApoE data without CSF. p values are of difference in classification vs. indicated single modality, M for MRI, P for PET and C for CSF.

the increased computational demands of having to do a three dimensional grid search for four modalities, rather than a two dimensional grid search for three modalities as in the previous experiment. However, the results for multimodal classification both with and without the CSF data are reported so it is possible to see its effect on classification with a consistent set of test subjects.

Again, the last two rows of table 6.8 present results obtained using MCI subjects for tuning the kernel weights, and with the data normalised with a z-score as these provided the best accuracy.

The results show a clear advantage in accuracy for multimodal imaging. In the larger PET group, both multimodal algorithms are better than any single modality alone. This advantage is statistically significant at the 5% level for the type-II maximum likelihood method with GP classification, which outperforms the grid search method and outperforms the best single modality by over 8%. The AUC measure of accuracy shows how results must be interpreted with caution, as the multimodal grid search method has a higher balanced accuracy than using PET alone, but offers a slightly lower AUC. In the smaller group for which both PET and

Modality	acc(%)	sens(%)	spec(%)	bal acc(%)	AUC
MRI	65.3	76.7	57.1	63.9	0.685
PET	69.4	63.3	73.8	65.3	0.782
CSF	56.9	73.3	45.2	55.6	0.575
Im + ApoE (GS)	68.1	76.7	61.9	68.1	0.745
All (GS)	66.7	76.7	59.5	69.4	0.727

Table 6.8: Accuracy of methods on the PET-CSF group with SVM classification. All modalities indicates MKL with MRI, PET, CSF and ApoE kernels, Im + ApoE is the image data (MRI and PET) and ApoE data without CSF.

CSF data were available in all subjects, the same pattern applied in that multimodal methods outperformed all single modality methods.

To enable a side-by-side comparison, table 6.9 shows the balanced accuracy for GP and SVM classification together with a p-value for the difference in accuracy. The p-value is generated by classifying all test subjects with the leave-one-out procedure used to generate the balanced accuracy figures, and comparing the resulting classifications, again using McNemar’s test.

6.3.2 Accuracy of probabilistic classification

The predicted risk figures produced in the manner described in section 6.2.12 exhibit some bias, in that the classifiers tend to overestimate the chances of conversion in general. This appears to be because of the transfer learning approach we use, where the classifier is trained on the AD and healthy population, and then applied to the MCI subjects. As the MCI subjects, in terms of the biomarker data we use, are not midway between the AD and control population but slightly closer to the AD subjects, this results in the classifier being somewhat biased in favour of predicting conversion. In order to remove this, we perform a correction procedure on the GP probabilities similar in approach to the one used to produce a balanced accuracy. We perform a logistic regression, using a leave-one-out approach again to avoid unduly optimistic results, on the GP probabilities and the labels indicating converter or stable status for the MCI subjects, with the label 0 indicating stable and 1 indicating converter. In this way we can learn the relationship between GP predicted risk and actual risk for the MCI subjects to correct for the bias. The resulting plots of empirical risk versus adjusted predicted risk for the PET and PET-CSF groups are shown in figures 6.3 and 6.4. Plotted points are labelled with the number of subjects in the corresponding bin. As not all the bins contain subjects, some empty bins are

Group	Modality	bal acc (%) (GP)	bal acc (%) (SVM)	p-value
PET	MRI	61.5	58.7	0.387
PET	PET	65.7	67.1	0.789
PET	MRI,PET, ApoE	74.1	67.8	0.151
PET-CSF	MRI	61.1	63.9	0.683
PET-CSF	PET	69.4	65.3	0.450
PET-CSF	CSF	56.9	55.6	1
PET-CSF	MRI, PET, ApoE	72.2	68.1	0.450
PET-CSF	MRI, PET, ApoE, CSF	72.2	69.4	0.803

Table 6.9: Statistical comparison of GP and SVM classification results for different subjects groups and combinations of modalities. MKL weights α are set by maximum likelihood for GP and grid search for PET. p-values are for significance of difference in accuracy between SVM and GP for a particular set of subjects and modalities used.

not plotted.

In these plots, a classifier producing accurate probabilities should have points plotted close to the diagonal. By inspection, the multimodal methods appear to perform well by this measure, and it is important to note that most points lying far away from the diagonal represent bins containing few subjects, making the empirical risk calculated for them less reliable. More broadly, the probabilities produced by the GP classification procedure appear to be accurate in the sense that increased predicted risk of conversion does generally imply an increased chance of conversion actually taking place. The adjustment appears to be effective, with little bias exhibited in the predicted risks. Note the only plotted points very far from the diagonal, and thus showing a large difference between empirical and predicted risk, are of risk bins containing only one or two subjects and are simply the results of outliers.

6.4 Discussion

As previously stated, a clear advantage can be seen both for multimodal classification, and for the use of GP classification over the more widely used SVM. This applied to results for both the PET and PET-CSF groups. Moreover, there appears to be quite a strong interaction between the utility of multimodal classification and the type of classifier used. Looking at the balanced accuracy of classification on single modalities of data, there is little to choose between GP and SVM classification, with differences of one or two per cent in accuracy in either direction. Thus, it seems reasonable to conclude by this measure that there is little difference in

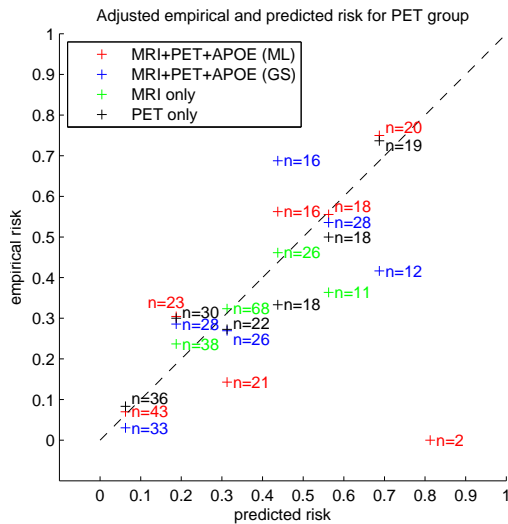


Figure 6.3: Empirical risk vs. corrected predicted risk for the PET group.

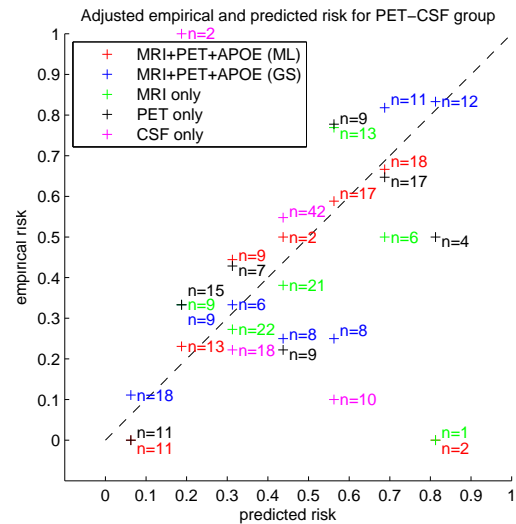


Figure 6.4: Empirical risk vs. corrected predicted risk for the PET-CSF group.

discriminative ability on identical sets of data. However, the GP framework appears to be able to take much greater advantage of the availability of multimodal data. GPs offer much larger gains for multimodal versus unimodal classification, with gains of eight per cent in the PET group against the best single data (PET) as against only a 0.7 per cent gain for the SVM approach. Similarly, the head-to-head comparisons between the GP and SVM methods using the same subjects and modalities, in table 9, show the greatest differences in classification accuracy and greatest statistical significance are for the multimodal methods. While the difference is not quite significant at the 0.05 level, due to the relatively small number of subjects in the study, the advantage for GP against SVM classification is clear and consistent across all three multimodal classification experiments and we plan to verify it with a larger dataset.

The improvement is most likely because the GP framework is better at finding a set of kernel weights for optimum classification. With an SVM we are restricted to finding these through a grid search, which has an inherently limited range and resolution if it is to be tractable, and is dependent on rather crude measures of accuracy to select an optimal parameter set. GPs offer tuning via the likelihood function, which seems to be more robust and also allows a wider search space - however this is not available for SVM classification, highlighting one of the advantages of a probabilistic framework mentioned in the introduction.

Adding CSF to multimodal classification did not increase the accuracy by any significant amount and in fact decreased the AUC, which is not surprising as CSF is the poorest single modality, offering accuracy little better than chance. The poor performance of CSF biomarkers

alone, and their failure to add diagnostic value when used alongside other biomarkers, is perhaps explained by the fact that about a third of controls have a high amyloid load, suggesting they may be in fact at a presymptomatic stage of AD. In this case CSF is still a potentially valuable biomarker, but our choice of defining AD and control subjects purely by current symptoms and cognitive test results limits its applicability. This again suggests the need to treat AD as a spectrum rather than a set of discrete states, or at least to very carefully define such states.

Comparing the results presented here to other attempts to predict conversion in MCI patients is difficult. This is because, while the problem has been addressed in a large number of studies, these vary widely in how MCI groups are defined, and the metric by which classification accuracy is assessed. However the method presented here certainly offers a high level of classification accuracy, especially considering studies that use ADNI data and offer higher accuracy make predictions over a time span of less than three years or make use of longitudinal data, which our algorithm does not need.

For MRI data, the most comparable methods are in [Cuingnet et al., 2010]. This study included a wide variety of types of feature, but those which used voxel level GM maps are quite similar to our work. Even within this definition, a wide range of options in image processing and feature extraction were used but the closest in methodology to ours is what they label as the Voxel-Direct-D-GM method. When applied to predicting MCI conversion this was found to have a specificity of 100% and sensitivity of 0%, i.e. The classifier simply assigned all subjects to the majority MCI-s class, possibly as a function of having trained on MCI-s and MCI-c rather than control and AD subjects. This paper did also find that the voxel based method in [Fan et al., 2007] achieved a sensitivity of 62% and specificity of 67%, although this was found not to be significantly greater than chance. Our method achieves much greater accuracy than any in [Cuingnet et al., 2010] for predicting MCI conversion, and moreover our accuracy is statistically significantly better than chance, which none of the methods assessed in that study managed to achieve.

Other studies, however, have had much greater success in predicting conversion. For example, [Coupé et al., 2012] and [Eskildsen et al., 2013] have presented methods capable of predicting conversion with accuracies similar to ours. The former uses a novel hippocampal grading biomarker. Using their most rigorous validation method, accuracy was slightly lower at 71% but their method needs no FDG-PET data and less computationally intensive image processing than the one presented here. The latter also achieves 74% accuracy by stratifying MCI-c subjects by conversion time and then combining the results of classifying each MCI-c subgroup against the MCI-s subjects. The classifier is rather unbalanced, with substantially higher speci-

ficity than sensitivity, a common problem with MCI classification, but again only structural image data is needed. Reported AUC values in [Ye et al., 2012] are up to 0.85 using MRI data, ApoE genotypes, and a variety of cognitive measures with a sparse logistic regression procedure but the authors do not list classification accuracy. In [Wee et al., 2012] features based on correlations between mean thicknesses of cortical regions of interest are used with SVM classification, and obtain 75% accuracy and an AUC of 0.8426. Among multimodal methods, [Zhang et al., 2011] reports a specificity of 91.5% and specificity of 73.4% for prediction of MCI conversion. While they do not report the proportions of MCI-s and MCI-c in their subjects and hence we cannot calculate the overall accuracy, it must be greater than our best result of 74%. However, they define conversion as a subject converting within 18 months rather than three years. Predicting over a short future timespan is an easier problem than over a longer one [Eskildsen et al., 2013] and less clinically useful. Moreover, defining conversion over a shorter time means a smaller proportion of MCI subjects will be converters, reducing the positive predictive value of even a good classification result. Additionally, their work uses CSF data in addition to MRI and FDG-PET, whereas our best performing classifier uses genetic data instead of CSF, which is less invasively obtained. We are able to set our kernel weights by type-II maximum likelihood, avoiding the need for a computationally expensive grid search. The other previously published multikernel method to predict MCI conversion is [Hinrichs et al., 2011]. Although they do define converters with a three year time span, direct comparison of results is again difficult, as they report only an AUC rather than accuracy. The best reported AUC was 0.791, similar to ours but this used longitudinal data, again effectively reducing the time span to predict conversion. They also found the method using only longitudinal image data was more effective than including non-imaging data in their multikernel learning approach. Methods based on features structural imaging alone are also capable of achieving high accuracy. Table 6.10 summarises these results in comparison with our own.

Table 6.10 clearly show the difficulty in making direct comparisons between results. For example, the time within which MCI conversion is defined has a strong effect on results. In [Vounou et al., 2012], tensor based morphometry was used to define a set of voxels that are highly indicative of MCI conversion, and then applied an SVM to these. This method was able to predict conversion with an accuracy of 82%. As this method uses both baseline MRI scans and 24 month follow-up MRI scans to generate Jacobian maps, it is effectively predicting conversion in only a 12 month period rather than three years as we do, and longitudinal data may not be available in all cases.

Parameterisations of the shape of the hippocampus have achieved a greater accu-

Article	Data used	n (MCI-s,MCI-c)	t	acc (%)	AUC
[Young et al., 2013c]	MRI, FDG-PET, ApoE	143 (96, 47)	36	74.10	0.795
[Eskildsen et al., 2013]	MRI	388 (227, 161)	36	73.5	-
[Ye et al., 2012]	MRI, ApoE, cognitive scores	319 (177, 142)	48	-	0.8587
[Wee et al., 2012]	MRI	200 (111, 89)	36	75.05	0.8426
[Zhang et al., 2011]	MRI, FDG-PET, CSF	99 (56, 43)	18	sens 91.5, spec 73.4	-
[Hinrichs et al., 2011]	longitudinal MRI, baseline MRI, longitudinal FDG-PET, baseline FDG-PET, CSF, ApoE, cognitive scores	119	36	-	0.7911
[Coupé et al., 2012]	MRI	405 (238, 167)	36	73.0	-
[Wolz et al., 2011b]	MRI	405 (238, 167)	36	68.00	-
[Nho et al., 2010]	MRI, ApoE, family history	355 (205, 150)	36	71.6	-
[Davatzikos et al., 2008]	MRI, CSF	239 (170, 69)	36	61.7	0.734

Table 6.10: Reported results from a variety of studies for predicting MCI conversion on ADNI data. n = number of subjects, t = number of months over which MCI conversion is defined, acc = accuracy in predicting conversion, if reported, AUC = area under ROC curve of predictions of conversion, if reported.

racy than our approach with conversion defined over three years [Costafreda et al., 2011, Ferrarini et al., 2009], however these used a small number of subjects scanned at a single centre, and also had autopsy confirmed AD subjects available, removing any uncertainty in the training labels. If conversion is defined over a three year period, we believe our method presented here has obtained an accuracy very competitive with the best methods yet published for prediction of conversion to date on ADNI data.

Moreover, our method offers the advantages of probabilistic classification listed in 6.1. The reject option is especially relevant in the case of computer-aided diagnosis. Having a probabilistic classification means that each diagnosis includes an attached degree of confidence rather than a simple binary decision. Clinical decision making is frequently hampered by over-confidence [Berner and Graber, 2008], so an estimate of the certainty of a diagnosis could be of great help, if only as a supplement to decisions made by more traditional methods.

6.5 Conclusion

We have shown that multimodal Gaussian process classifiers can be successfully applied to the prediction of conversion to AD in MCI patients. Prediction of conversion is significantly better for multimodal classification than for any single modality, and also better for GP compared to SVM classification, largely due to the GP's superior ability to exploit multimodal data. Accuracy is state-of-the-art, and to this we can add the advantages of probabilistic classification. A number of extensions to this work are possible. The simplest is to take advantage of new subjects with FDG-PET and CSF data being added to the ADNI database and apply these methods to a larger group of subjects to show greater statistical significance for the advantage of our methods. We perform more sophisticated feature extraction on FDG-PET data and to make use of more complex kernel covariance functions, as described in the following two chapters (7 and 8). We also examine methods to overcome the problem of misdiagnosis leading to noisy training labels in ADNI data in chapter 8 and [Young et al., 2013a].

Chapter 7

Continuous proxies for AD diagnosis and prognosis

7.1 Introduction

In this chapter, we explore an alternative to the classification approach to predicting conversion in MCI subjects, described in the previous two chapters. A classifier to predict conversion in MCI subjects can be trained on labelled examples of MCI-s and MCI-c images, or alternatively on examples of AD patients and healthy controls, under the assumption that MCI-s subjects are more control-like and MCI-c subjects are more AD-like. However, either method ultimately relies on discrete labels designating each subject used for training as being a member of a particular diagnostic group.

This does bring some disadvantages. The labels for training data are, in the cases above, assumed to be always correct. However, a limiting factor in the accuracy of classification studies may be mislabelling of training subjects. The gold standard for diagnosis of AD is autopsy, but most studies use training subjects whose diagnosis has been determined by standard clinical diagnosis, which has been shown to have an error rate of at least 10% [Beach et al., 2012] when compared to retrospective diagnosis when the subjects died and it was consequently possible to confirm (or disconfirm) their earlier diagnosis by autopsy. Furthermore, the same study found that the rate of misdiagnosis varies wildly between AD centres in a multicentre setting very similar to ADNI. This is an issue that has not been widely addressed. The most effective way to do so would be to use only subjects whose diagnosis is confirmed by autopsy; but these are only available in much smaller numbers than those diagnosed in the clinic. An alternative method to estimate the effects of mislabelled data is to use some other classification for which the ground truth is readily available, such as sex, and perform experiments by deliberately changing the labels of some subjects [Young et al., 2013a].

Another problem is that labels for MCI-s and MCI-c subjects are also affected by limited

follow-up time; many subjects deemed as stable may in fact convert after a study has finished. For a study examining conversion, this is not a problem when assessing the accuracy on the test set as it is generally limited to subjects who do or do not convert within a fixed length of time, but may well mean that a training set consisting of MCI-s and MCI-c subjects is suboptimal.

For this reason, MCI-s and MCI-c labels are not used in [Aksu et al., 2011]. Pointing out that training labels for MCI-s and MCI-c are uncertain, they go on to generate their own MCI training labels by following the classification of MCI subjects by an HC versus AD classifier across multiple timepoints. However even this neglects the uncertainty in the HC and AD labels that this scheme ultimately depends on. BrainAge [Gaser et al., 2013] switches the problem to one of regression, with a model being built to predict the chronological age of a large cohort of healthy subjects. This model is then applied to AD and MCI subjects, with the BrainAge defined as the difference between chronological and predicted age. This can then be thresholded to classify subjects into groups such as healthy and AD, or MCI-s and MCI-c with high accuracy.

Our proposed method follows [Gaser et al., 2013] in abandoning discrete disease state labels for training altogether. Like them, it involves performing a regression to predict a continuous proxy for disease status, but instead of age, initially atrophy over a period of one year as measured by the boundary shift integral (BSI) [Leung et al., 2012] is used. This then provides a predicted atrophy rate for each test subject. Gaussian process (GP) regression [Rasmussen and Williams, 2006], with a multiple kernel framework is used to optimally combine MRI, FDG-PET and CSF data. This results in a measure that can predict MCI conversion within three years with a balanced accuracy of 74.6%, as good as state-of-the-art techniques having a much larger training set, including our own previous work using multikernel GPs for classification [Young et al., 2013c]. We refer to this as the BSI experiment, which is previously published in [Young et al., 2013b].

Encouraged by this, we go on to modify the approach and apply it to a much larger group of subjects. In this second experiment instead of age, the target variable for regression is cognitive test scores from the mini-mental state examination (MMSE) [Folstein et al., 1975]. This appears to give better results and is available in a larger number of subjects. This latter point is also due to incorporating data from the ADNI 2 and ADNI-GO databases as well as the original ADNI. We find that we can predict conversion within three years with a balanced accuracy rate of nearly 82% for subjects with FDG-PET scans, and nearly 80% for subjects with structural MRI scans only. This accuracy is amongst the highest yet seen for this problem. We also show that it is heavily dependent on the field strength of the subjects' structural MRI scans, even if only FDG-PET data was used in the regression problem. This is referred to as the MMSE

Disease status	Number	Female	Mean age (sd)
HC	28	19	74.1 (4.5)
MCI-s	38	22	75.3 (7.3)
MCI-c	29	18	75.1 (7.4)
AD	34	23	75.1 (6.8)

Table 7.1: Subject groups and demographics for the BSI experiment.

experiment.

7.2 Materials and methods

7.2.1 Subjects

All ADNI subjects are between 55 and 90 years old, speak English or Spanish, and have a study partner able to provide an independent assessment of functioning. All subjects are willing to undergo neuroimaging and agree to longitudinal follow up, and a subset are willing to undergo lumbar punctures. Subjects with specific psychoactive medication are excluded. Inclusion criteria for normal subjects are MMSE scores between 24 and 30, a CDR of 0, non-depressed and non-demented. Ages are roughly matched to those of AD and MCI subjects. For MCI subjects, the criteria are an MMSE score between 24 and 30, a memory complaint, objective memory loss measured by education adjusted scores on the Wechsler Memory Scale Logical Memory II, a CDR of 0.5, absence of significant levels of impairment in other cognitive domains, essentially preserved activities of daily living, and an absence of dementia. For AD subjects, the criteria are an MMSE scores between 20 and 26, CDR of 0.5 or 1.0, and meeting NINCDS/ADRDA criteria for probable AD. Subjects are designated as HC, AD or MCI at the time of the baseline scan, and for the purposes of this study MCI conversion status is decided by whether subjects who were MCI at baseline were subsequently diagnosed as AD at any stage during the subsequent 36 month follow-up period.

For the BSI experiment, only subjects from the original ADNI study were used. This collected baseline structural MRI scans for all subjects. However FDG-PET scanning and collection of CSF data were only done on partially overlapping subsets of these subjects. Furthermore, calculation of BSI requires a 12-month follow-up structural MRI, which were also missing for some subjects. As our method requires FDG-PET and CSF and a 12-month BSI as well as structural MRI data, only 129 subjects could be included in the study. The details of these are shown in table 7.1.

For the MMSE experiment, we also made use of data from the extensions to the original

Disease status	n (n female, n 1.5T)	Mean age (SD)	Mean MMSE (SD)
Healthy	243 (119, 84)	73.9 (5.9)	29.0 (1.2)
MCI-s	81 (26, 54)	72.0 (7.5)	27.9 (1.6)
MCI-c	91 (37, 57)	72.7 (7.0)	26.9 (1.8)
MCI-other	347 (156, 35)	71.5 (7.4)	28.1 (1.7)
AD	121 (54, 60)	74.7 (7.8)	23.1 (2.0)

Table 7.2: Subject groups and demographics for the PET group in the MMSE experiment.

Disease status	n (n female, n 1.5T)	Mean age (SD)	Mean MMSE (SD)
Healthy	338 (173, 142)	74.4 (5.7)	29.1 (1.1)
MCI-s	133 (46, 84)	72.5 (7.7)	27.8 (1.7)
MCI-c	153 (67, 97)	72.8 (7.3)	26.6 (1.8)
MCI-other	375 (168, 49)	71.8 (7.5)	28.0 (1.8)
AD	185 (87, 101)	74.8 (7.8)	23.1 (2.0)

Table 7.3: Subject groups and demographics for the MRI group in the MMSE experiment.

ADNI project, ADNI 2 and ADNI-GO. ADNI 2 aims to recruit an extra 550 subjects, with a similar proportion of healthy, MCI and AD subjects as the original ADNI. ADNI-GO enables extended follow-up of nearly 500 of the original ADNI participants, as well as recruiting further MCI participants. Imaging protocols for both ADNI 2 and ADNI-GO were designed to ensure compatibility with ADNI data. This mean that the pool of subjects for the MMSE experiment was much larger. Furthermore, we removed the requirement for the subjects to have CSF data to use MRI and FDG-PET only, and as we were using MMSE rather than one year BSI as a regression target, only a small number of subjects for which baseline MMSE scores were missing had to be excluded. As a result there were a total of 883 subjects. This is referred to as the PET group, with details given in table 7.2.

We also explored how well MMSE prediction works using MRI data alone, using an even larger set of 1184 subjects by adding those for which FDG-PET data was not available. This set of subjects is referred to as the MRI group, with details given in tables 7.3.

Both of these sets contained a mixture of subjects scanned at 1.5T and 3T. Subjects enrolled in ADNI-GO and ADNI 2 were all scanned at 3T; whereas the majority of ADNI 1 participants were at 1.5T. As many of the ADNI 2 and GO subjects were recently enrolled at the time of writing, they had less than 36 months of follow-up. It was not therefore possible to reliably label

these subjects as MCI-s or MCI-c. Hence they are not included in the MCI subjects for which accuracy of prediction of conversion is reported. However, unless their corresponding MMSE score was missing they were included in training the regression model. The same applies to the small number of MCI subjects that reverted to healthy status during follow-up. Collectively these subjects are labelled as MCI-other.

7.2.2 MRI data

Images were all T1 weighted structural MRI scans from 1.5T or 3T scanners acquired using the 3D MPRAGE sequence, taken at the baseline time point for each subject. For the subjects in the BSI group, the 12 month follow up scan was also downloaded and used to calculate the atrophy rate. DICOM images were downloaded from the ADNI archive, having been designated the best of back-to-back scans for each subject, and then post-processed to correct for gradient warping, B1 non-uniformity and intensity non-uniformity and undergone phantom based scaling correction. Once downloaded, the images were then converted to NIFTI format for further processing.

7.2.3 MRI image processing

Our MRI features for classification and regression were voxels of whole-brain tissue density maps. To produce these, we followed a similar procedure to [Klöppel et al., 2008] although using slightly different software. Also, the groupwise registration procedure was done initially only for the subjects in the BSI experiment. Once many more subjects' images had been downloaded to make up the group for the MMSE experiment, the groupwise registration procedure was rerun using all the newly obtained subjects.

The first step was to segment the native space, preprocessed scans to produce a GM density map. This was done using NiftySeg [Cardoso et al., 2011] for the subjects in the BSI group and with the 'new segment' module of SPM12, with the maximum cleanup option set, for all subjects in the MMSE experiment. A brain mask produced from the original structural image was then applied to remove any non-brain material. The native space images were also anatomically parcellated into 83 regions with a novel label fusion algorithm [Cardoso et al., 2012] in a multi-atlas label propagation scheme. The resulting parcellations were used to mask out the brainstem and cerebellum from the native space GM segmentations.

The groupwise registration was employed to move all the GM images into a common space. Initially, all T1 weighted images were rigidly registered to a randomly chosen sample, and then averaged together to produce a new target image. This was then repeated with a single round of affine registration of all images to the target, which was then followed by ten rounds

of nonlinear registration, with the updating of the target image also repeated after each round of registrations. This was all done using tools from NiftyReg [Modat et al., 2010]. After this procedure was finished, the transformation from each native T1 weighted image to the final target were applied to the native GM segmentations to resample them into the groupwise space, using trilinear interpolation to maintain the voxel GM densities in the zero to one range. Finally, the GM segmentations in the groupwise space were modulated by the Jacobian determinants of the corresponding final transformations to ensure the total tissue volume remained constant. The result was all GM segmentations in the groupwise space of the BSI experiment subjects or MMSE experiment subjects.

7.2.4 PET image data

Images were again all taken from the baseline scan for each included subject. Images were acquired by scanning 30-60 minutes post injection using scanner-specific protocols. Six five minute frames were acquired for each subject, and then co-registered and averaged. The average images were then rigidly registered to a standard space, and the individual native space frames registered to the standard space average and averaged and intensity normalised in the standard space. Finally, the average images in the standard space were smoothed with a scanner-specific kernel (Joshi et al., 2009) to a uniform isotropic resolution of 8mm FWHM, which is approximately the resolution of the lowest resolution scanners used in ADNI. The postprocessed scans were downloaded as DICOM images.

7.2.5 PET image processing

As previously mentioned, the PET images had already been through substantial processing before being downloaded from the ADNI database. Following this, each native PET image was registered to the corresponding native structural MRI. At this point, the processing used to generate features from the PET images varied between the BSI and MMSE experiments. For the BSI experiments, the native space anatomical parcellations were also transferred to the space of the FDG-PET images for the corresponding subjects. The parcellation was used to normalise each FDG-PET image by its mean cerebellar activity, and then to calculate the mean activity within each anatomical region, generating a set of 83 features for each FDG-PET image. For the MMSE experiment, the PET images were again normalised by the mean activity within the subject's cerebellum. However the PET images were then moved into the groupwise space by again applying the transformations from the native space of each subject to the final groupwise target image, using the same software. Thus for the MMSE experiment the PET features comprised voxel level rather than regional level features.

7.2.6 CSF data

CSF data was only used for subjects in the BSI experiment. CSF samples were obtained from subjects by a lumbar puncture around the time of their baseline scan. Levels of the proteins amyloid- β_{42} ($a\beta_{42}$), tau, and phosphorylated tau were measured and recorded.

7.2.7 Boundary shift integral

The BSI is a method for robustly assessing volume loss of whole brains or brain regions. It calculates a change in volume by integrating across the longitudinal change in position of the boundary between CSF and GM surrounding the region of interest. Preprocessing is needed to extract the region of interest (which in our case is the whole brain) from each image, linearly align the baseline and follow-up images, and correct for intensity inhomogeneity between scans. We use the latest version of BSI [Leung et al., 2012] which uses a symmetric registration scheme to minimise bias and maximise desirable qualities for an atrophy measurement such as inverse consistency and transitivity between multiple timepoints.

We normalise the resulting volume changes by the baseline brain volumes and by the actual interval between baseline and follow-up scans (as the nominal 12 months varies quite widely in practice), and multiply by 100. This produces a normalised brain atrophy rate (BAR) in percentage of original brain volume per year for each subject. These are then used as targets for regression analysis in the BSI experiment. We also experimented with using BSI of the left hippocampus only as a regression target, but found it produced markedly inferior results, largely as a result of reducing the size of the training set due to missing data.

7.2.8 MMSE scores

We used the MMSE scores for each subject at baseline as the targets for our regression problem. MMSE is derived from a questionnaire widely used in screening for dementia and to track cognitive decline, with questions covering a variety of cognitive domains. Scores are given as an integer score up to a maximum of 30. The scores were obtained along with the corresponding images from the ADNI database. Additional experiments were performed using both an alternative neuropsychological test score, ADAS-Cog, and longitudinal measures based on decline in MMSE scores from baseline to one and three year follow-up timepoints as alternative targets for regression. For both of these cases, results were markedly inferior to simply using baseline MMSE and so are not presented here. In the case of longitudinal MMSE, this is because of a much smaller training set due to missing follow-up data and increased noise compared to using baseline MMSE. For baseline ADAS-Cog, the difference is more difficult to explain but possibly it is a noisier measure than MMSE or simply less informative about the underlying disease

severity.

7.2.9 Gaussian processes

The learning portion of the procedure - both classification and regression - was done using GPs. These provide a kernelised, Bayesian framework for both these tasks. For a full explanation of GPs for regression, we refer the reader to [Rasmussen and Williams, 2006] for a much more in depth treatment.

As in chapter 6, for the subjects in the BSI experiment, we combine the MRI, PET and CSF data in a multikernel framework to perform multimodal classification and regression. This means each element of the kernel matrix \mathbf{K} is a linear combination of three subkernels representing the covariance between the MRI, PET and CSF data a pair of subjects, with weights α . A bias term β is also included in the sum. So in the case of multimodal classification using information derived from the MRI, PET and CSF data for each subject the overall kernel is

$$\mathbf{K}_{ij} = \alpha^{MR} \mathbf{x}_i^{MR} \cdot \mathbf{x}_j^{MR} + \alpha^{PET} \mathbf{x}_i^{PET} \cdot \mathbf{x}_j^{PET} + \alpha^{CSF} \mathbf{x}_i^{CSF} \cdot \mathbf{x}_j^{CSF} + \beta \quad (7.1)$$

For a more detailed look at multikernel learning, see section 6.2.9.

7.2.10 Classification and validation in BSI experiment

Predicted BARs for all 129 subjects in the PET experiment were generated regardless of their disease status. This was done in a leave-one-out (LOO) procedure across the entire set. We then used the predicted BAR of MCI subjects to classify them as MCI-s or MCI-c by thresholding. As there was not an a priori reason to threshold at any particular value of the predicted BAR (unlike probabilistic binary classification, where thresholding at 0.5 may be chosen as a starting point) we chose the threshold as the value that best balances sensitivity and specificity. This was done in a second, inner LOO loop nested inside each iteration, to avoid introducing optimistic bias.

We also compared our method to performing direct binary classification on the conversion status, again using GPs. There were three different choices of training group here: train on the MCI-s and MCI-c subjects and labels in an LOO loop, training on AD and control subjects and applying the resulting classifier to the whole MCI population, and a shared label approach. This attempted to increase the amount of training data by having one training group comprise both the control and MCI-s subjects treated as a single class, and another training group comprising both the AD and MCI-c subjects treated as a separate single class. The shared label approach was again done inside an LOOCV loop.

7.2.11 Classification and validation in MMSE experiment

For the regression experiments, predicted MMSE scores were generated for all subjects in a LOO loop in a manner very similar to that used in the BSI experiment. We report the resulting balanced accuracy, as well as area under the receiver operating characteristic curve (AUC). Although this was not the primary focus of our experiments, we also compared predicted to actual MMSE scores as a measure of how well GP regression modelled the data.

To compare the utility of predicted MMSE scores against the more conventional approach to predicting conversion, we also again performed GP classification on the same set of subjects used for the MMSE regression experiment.

To do this, we took the healthy and AD subjects as training data, and then applied the resulting classifier to the MCI-s and MCI-c subjects. This idea relies on the assumption that MCI-s subjects are more healthy-like and MCI-c subjects are more AD-like. In our previous work, we found it produced much better results than training on labelled MCI-s and MCI-c subjects. It also had the advantage of not requiring cross-validation as the training and testing subjects were drawn from different populations. Again we report balanced accuracies (generated with the same LOO thresholding method used with the predicted MMSE scores, applied to the predicted class membership probabilities) and AUC, as well as accuracy, sensitivity and specificity obtained by thresholding at 0.5. Finally, we assess the significance of differences in classification accuracy for the two methods using McNemar's test [McNemar, 1947].

7.3 Results for BSI experiment

The correlation coefficient between predicted and measured BARs for the subjects is 0.38 ($p < 0.0001$) and the root mean squared error is 0.61. However our primary focus is not on the predicted brain atrophy rates themselves, but on whether they can be used to predict conversion in MCI subjects. Figures 7.1 and 7.2 show the spread of both measured and predicted BAR values for all four disease groups (HC, MCI-s, MCI-c, AD).

As shown in figures 7.1 and 7.2, while the mean predicted BARs for each group are similar to the corresponding means for measured BARs, each clinical group occupies a much tighter cluster of values, even allowing for a few outliers (marked as a +). This results in reduced overlap between the clinical groups, which is especially noticeable between the MCI-s and MCI-c groups. To test this, we classify the MCI-s and MCI-c subjects by finding a threshold in predicted BAR that best balances sensitivity and specificity. A nested leave-one-out scheme is used to avoid introducing optimistic bias. The resulting accuracy is 74.6%, which is similar to the best previously reported results. The balanced accuracy and area under the ROC curve

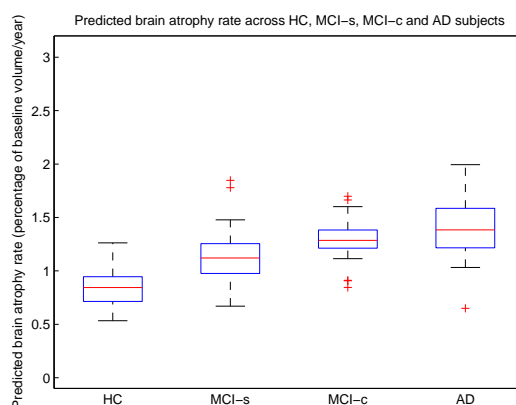
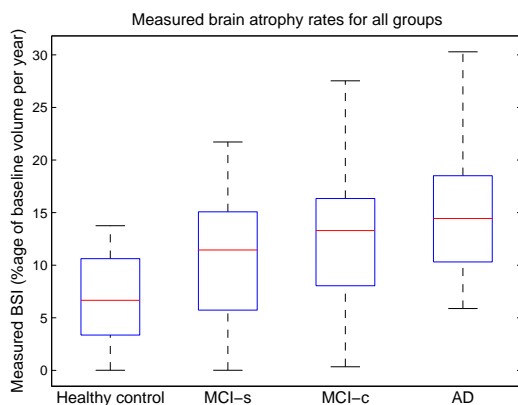


Figure 7.1: Measured BAR across groups

Figure 7.2: Predicted BAR across groups

(AUC) are shown in table 7.4. This also shows results for single modalities, demonstrating the benefit of combining sources of data with multikernel learning.

To illustrate the advantage of our method of atrophy prediction, we also compare it to performing direct binary classification on the conversion status again using GPs. This can be done by training on the MCI subjects only in an LOOCV loop, by training on all subjects, again with an LOOCV loop and grouping HC subjects with MCI-s and MCI-c subjects with AD, and finally by training on the HC and AD subjects, and testing on the MCI subjects. The results are given in table 7.5.

Table 7.4: Accuracy of discrimination between MCI-s and MCI-c with predicted brain atrophy rate

Modalities	Accuracy (%)	AUC
MRI	59.7	0.595
PET	73.1	0.777
CSF	52.2	0.545
MRI, PET	67.2	0.743
MRI, CSF	58.2	0.602
PET, CSF	65.7	0.726
MRI, PET, CSF	74.6	0.725

Table 7.5: Accuracy of discrimination between MCI-s and MCI-c with training on binary diagnostic class labels

Training	Accuracy (%)	AUC
MCI (CV)	40.3	0.401
HC, MCI, AD (CV)	52.2	0.569
HC, AD	55.2	0.661

7.4 Results for MMSE experiment

Results comparing predicted MMSE to binary classification for both FDG-PET and MRI data in the PET group are given below in table 7.6.

Data	Method	acc(%)	sens(%)	spec(%)	bal acc(%)	AUC	p
MRI	pMMSE	N/a	N/a	N/a	67.4	0.735	0.737
MRI	classification	58.7	34.1	86.4	66.3	0.714	
FDG-PET	pMMSE	N/a	N/a	N/a	72.7	0.786	0.814
FDG-PET	classification	71.5	55.0	90.1	71.5	0.788	

Table 7.6: Accuracies for predicting conversion to AD in MCI subjects in the MMSE experiment. Method = classification paradigm (predicted MMSE or binary classification), acc, sens, spec = accuracy, sensitivity, specificity of thresholding probabilistic binary classification results at 0.5, bal acc = accuracy at best balance of sensitivity and specificity, p = significance of difference in balanced accuracy between different methods for the same data.

As can be seen, the predicted MMSE method outperforms binary classification using both types of data, although we are unable to show that the small advantage is statistically significant. To explore the limits of classification accuracy using MRI data only, we also report the results for the larger MRI group in table 7.7.

Method	acc(%)	sens(%)	0.5 spec	bal acc(%)	AUC	p-value
pMMSE	N/a	N/a	N/a	68.9	0.761	0.712
classification	65.4	48.4	85.0	68.6	0.743	

Table 7.7: Accuracies for predicting conversion to AD in MCI subjects in the MMSE experiment. Method = classification paradigm (predicted MMSE or binary classification), acc, sens, spec= accuracy, sensitivity, specificity of thresholding probabilistic binary classification results at 0.5, bal acc = accuracy at best balance of sensitivity and specificity, p-value = significance of difference in balanced accuracy between different methods.

7.4.1 Effect of MRI field strength on results

As previously stated, the subjects' structural MRI scans were performed on a mixture of 1.5T and 3T scanners. To see if this variable has an effect on classification accuracy, we break down the results for the predicted MMSE method by magnetic field strength, as shown in table 7.8. The results are produced by taking the previously presented predicted MMSE scores and splitting them into two groups based on the relevant field strength, and then doing the LOO thresh-

olding procedure independently on each group of scores. The results are therefore produced with a training set consisting of all subjects regardless of field strength.

Group	Data	bal acc - all(%)	bal acc - 1.5T(%)	balanced acc - 3T (%)
PET	FDG-PET	72.7	64.9	82.0
PET	MRI	67.4	65.7	80.3
MRI	MRI	68.9	65.2	75.2

Table 7.8: Breakdown of accuracy of predicted MMSE in MCI conversion by MRI field strength

7.4.2 Accuracy of MMSE predictions

The primary purpose of this study was to examine the utility of predicted MMSE scores in forecasting conversion from MCI to AD. The accurate prediction of MMSE scores in individuals was considered secondary and, in fact, completely accurate prediction of MMSE scores would be undesirable for reasons that are explained in the discussion section. Nevertheless, we do assess the ability of our regression model to predict MMSE scores, for the same three groups and types of data as for the prediction of conversion to AD. Results are calculated and presented for all subjects in each group regardless of diagnostic status, rather than for MCI-s and MCI-c subjects only, as was done previously. We report the correlation coefficient r and root mean square error (RMSE) in table 7.9, and present the results as scatter plots in 7.3.

Group	Data	r	RMSE
PET	FDG-PET	0.605	1.97
PET	MRI	0.596	1.99
MRI	MRI	0.574	2.1

Table 7.9: accuracy of predicted MMSE compared to ground truth in, for MRI data in the MRI group, and for PET and MRI data in the PET group.

7.5 Discussion

The results for the BSI experiment show a clear advantage for our method of training on a well-characterised proxy for MCI conversion, rather than the diagnostic status itself. Training on BAR enables us to reach accuracies of up to 74.6%, whereas training on diagnostic labels struggles to perform better than chance. It therefore appears that the use of BAR bypasses the problems caused by binary diagnostic labels. This makes better use of data as subjects can be used for training regardless of diagnostic label, and as parameters are learned automatically

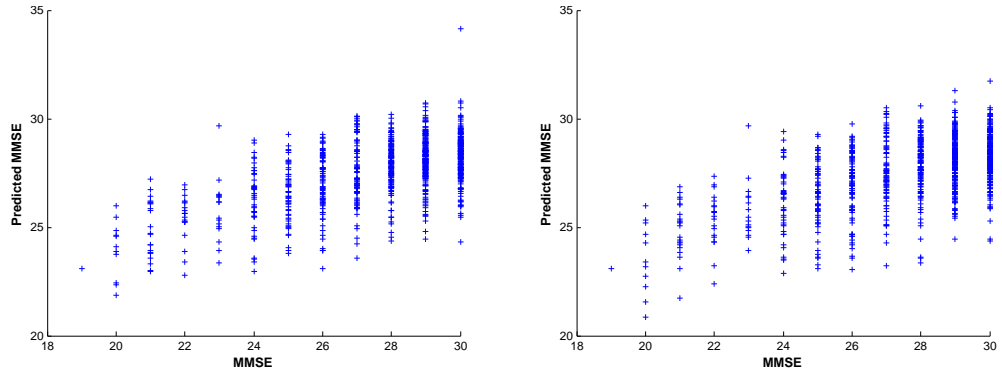


Figure 7.3: Predicted vs actual MMSE for FDG-PET data in PET. group

Figure 7.4: Predicted vs actual MMSE for MRI data in PET. group

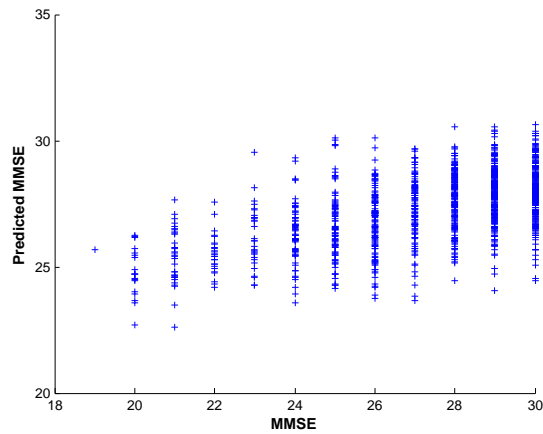


Figure 7.5: Predicted vs actual MMSE for MRI data in MRI. group

there is no need to set subjects aside for tuning. We also show an advantage for multimodal regression. Although direct comparisons between methods are difficult [Young et al., 2013c], the resulting accuracy in forecasting MCI conversion is competitive with the best yet achieved.

Despite this we were motivated to perform the BSI experiments, different in detail but inspired by the same idea, for two reasons. Firstly, the BSI experiment was with a relatively small number of subjects, especially considering MRI, FDG-PET and CSF data were necessary to get the best results (although FDG-PET alone does almost as well). In particular, the binary classification done for the purposes of comparison gave surprisingly poor results when compared to a previous experiment using very similar data. We hypothesised that this was due to the small training set and wanted to determine whether continuous proxy methods would maintain such a large advantage over binary classification with a much larger number of subjects.

The second reason was to see whether a more convenient proxy for conversion than BSI could be used. Although the BSI gave good results, it has the limitation of requiring a follow-up image, which makes an alternative where baseline data only is needed attractive (although it is worth emphasising that, while 12 month follow-up scans are also required to calculate BSI values for *training* data, they are not needed for *testing* data). This also allows us to increase the training set size even further, as MMSE scores at baseline were available for almost all subjects. The utility of BSI as a proxy was spoiled by the fact it was often not listed even for subjects where a follow-up image from the right time was available, and that where it was available, slightly different BSI methods were used on different subjects, meaning the measured BSI values used for training lacked consistency.

The use of MMSE as a target for regression raises the question of why it is necessary at all to do imaging and learning - might measured MMSE itself be able to distinguish MCI-s and MCI-c subjects at baseline? Unfortunately, as for measured BSI, this is not the case. Calculating the balanced accuracy of measured MMSE for discriminating MCI-s and MCI-c subjects in the PET group gives a figure of just 61.5%, compared to the 67.4% and 72.7% for predicted MMSE depending on which type of image data was used. The reason for this is obvious in the diagrams of figures. These show the box plots of measured MMSE scores and both sets of predicted MMSE scores for the subjects in the PET group, broken down by diagnostic group. The 'MCI-other' group contains all baseline MCI subjects that could not be definitely labelled as MCI-s or MCI-c as they lacked sufficient follow-up information or reverted to normal cognition.

The effect of learning to predict MMSE score is very similar to that of learning to predict BAR as shown in figure 7.1 and figure 7.2. As can be seen from figure 7.6, there is a

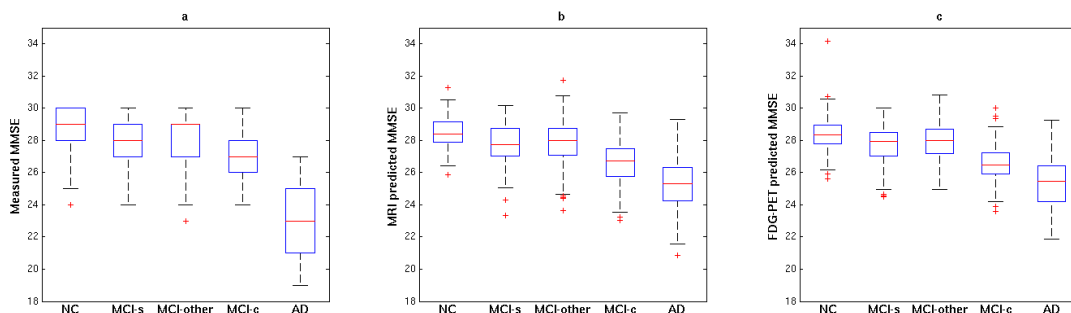


Figure 7.6: Measured and predicted MMSE across clinical groups in the PET group. (a) is measured MMSE, (b) is predicted MMSE from MRI data, (c) is predicted MMSE from FDG-PET data

great deal of overlap between the MCI-s and MCI-c groups in measured MMSE score, which is reflected in its poor classification accuracy. This is the reason why accurately predicting the measured MMSE is undesirable for us in this work. The effect of using predicted rather than measured MMSE is to move all diagnostic groups towards the mean MMSE, but also to tighten the distribution within each diagnostic group. As can be seen, this reduces the overlap between the MCI-s and MCI-c groups, enabling them to be distinguished with greater accuracy. Furthermore, this overlap is smaller in the predictions based on FDG-PET data than those based on MRI data (figure 7.6), reflecting the greater predictive accuracy of FDG-PET data we have already seen. Interestingly, the reverse effect is seen in discrimination between healthy and AD subjects. There is very little overlap between these groups in measured MMSE (largely because MMSE score is one factor used to make the diagnosis) and the movement towards the mean produced by using predicted MMSE scores actually increases the overlap between the two groups.

So, if it does not work by accurately predicting actual cognitive test scores, how does predicted MMSE predict conversion so well? The actual MMSE scores can be seen as coming from a latent variable - some underlying, true disease severity - plus a large level of noise due to individual variability. Fitting a regression model to the MMSE scores may act to partly remove this noise, producing a measure that much more closely reflects the subjects' actual level of cognitive decline. Very similar arguments apply to predicted BAR.

It is therefore unsurprising that either measure would correlate strongly with conversion to AD in MCI subjects. This is still, however, a rather oblique approach compared to binary classification. It does hold some advantages over binary classification as well. The modelling of a continuous measure respects the notion of all subjects being on a spectrum of cognitive

decline. Labels based on diagnosis do not do this - for example, two subjects may be correctly diagnosed as having AD, but with one case much more severe than the other. Diagnostic labels may also be quite unreliable [Beach et al., 2012]. Finally binary classification methods can use either AD and healthy subjects for training, or MCI subjects. But in either case, only about half the subjects can be used for training. Moving to a regression approach means all subjects except a few with missing target variables can be used for training, although this can also apply to some methods using discrete groups such as ordinal regression [Doyle et al., 2013].

We also notice that there is a very large difference in accuracy in predicting conversion between those subjects whose MRI was obtained from a 1.5T scanner and those from a 3T scanner. It is unsurprising that this has some effect, but the magnitude of the resulting difference is surprising: almost 15 percentage points when using MRI data, and more than 17 percentage points when FDG-PET is used. This is most likely due to poorer registration in subjects scanned at the lower field strength. As using PET data introduces a further registration for each subject to the pipeline, between the native PET and native MRI, this could explain why the effect is even stronger in PET data. Although the effects of variation due to different scanners at ADNI centres on classification accuracy has been examined [Abdulkadir et al., 2011], little attention has been paid to this particular variable, probably because significant numbers of 3T scans have only entered the database in large numbers recently.

7.6 Conclusion

This chapter has shown that, while traditional classification techniques reliant on discrete labelled groups can obtain high accuracy for predicting conversion to AD, even higher accuracies can be obtained from a simple regression approach using only one type of image data and a well-chosen continuous proxy. In this way, we can predict conversion to AD within three years in MCI subjects with up to 82% accuracy using PET data, or 80% accuracy using MRI data. These accuracies are, we believe, among the highest yet reported for this well-studied and very clinically important problem (although fair comparison between results is difficult). It is important to note that these results were only obtained for test subjects scanned at 3T, irrespective of whether MRI or FDG-PET data was used. Future plans include studying of the effect of scanner field more rigorously, as it appears to have such strong effects. Additionally the work presented here could be extended, by looking at using target variables other than MMSE, such as rate of change of cognitive scores, and to incorporate multiple targets in a multi-task regression to further improve results.

Chapter 8

Anatomical regional kernels

8.1 Introduction

This chapter introduces a new way to define features derived from brain images that can improve classification accuracy. For Alzheimer’s disease (AD), grey matter (GM) density maps obtained from structural MRI images are typically used as data in the classification. However the actual features derived from the image can take two forms: at the level of the MRI voxel [Klöppel et al., 2008], or as summaries of all GM voxels within different anatomical regions. The regions can be defined by an atlas [Zhang et al., 2011] or can themselves be generated from voxel level data [Fan et al., 2007]. There is a trade-off between these methods. Regional level features reduce the data dimensionality and can introduce prior information relevant to the classification problem, but also eliminate fine detail that may be informative about disease state. Voxel level data can introduce noise by including uninformative brain regions and results in a very high dimensional problem. The different feature extraction methods are compared and discussed in depth in [Cuingnet et al., 2010].

The proposed method combines the strengths of these two approaches. It uses both voxel level features and atlas derived regions, and automatically gives less weight to voxels within less relevant regions. This is done using multiple kernel learning (MKL) as introduced in a previous chapter (6.1). This is usually applied to combine data derived from different imaging modalities [Young et al., 2013c, Zhang et al., 2011] or kernel functions [Hinrichs et al., 2011]. Conversely, in the approach presented here each kernel represents the voxel level data *within a different anatomical region* to produce anatomical regional kernels (ARKs). This takes a similar approach to [Chu et al., 2010], and [Liu et al., 2013] used a related nested region approach. Although the work was developed from our previous use of MKL, and is presented as a specific case of MKL, it is related to other families of methods. Specifically, it can be seen as a way to incorporate explicit spatial regularisation into the classifier. A number of

other methods have been developed to do this specifically for three dimensional medical image data. Spatial smoothness and sparsity can be enforced with a joint ℓ_1 and total variation penalty [Gramfort et al., 2013]. Alternatively a smoothness penalty is derived from the image voxel neighbourhood structure, which can be built into a kernel function for use with an SVM or other kernel method [Cuingnet et al., 2013] or used directly as a term in the objective function [Sabuncu and Leemput, 2012].

Our method and [Sabuncu and Leemput, 2012] can also both be interpreted as a variant of automatic relevance determination (ARD) [Neal, 1996, Rasmussen and Williams, 2006], a Bayesian method of automatic feature selection. Our method, however, operates at the regional level in the kernel space, rather than at the voxel level in the input space. This is enabled by the existence of a brain atlas in a custom groupwise template. Sections 8.3.4 and 8.4.1 explain how this was achieved, and how MKL is performed within a GP framework.

The method is applied to a large population of AD, MCI and control subjects from the ADNI study. In terms of classification accuracy, for classification of AD and control subjects our method outperforms a single kernel with voxel level features by a large margin, and a single kernel with regional features by a smaller amount. It also outperforms both voxel level and regional level features for prediction of conversion in MCI subjects.

We also introduce two new methods to assess the quality of a classifier that exploits the probabilistic predictions made by GPs. Finally, we show that the optimal kernel weights in the MKL formulation are informative about which regions are affected by AD.

8.2 Image and biomarker data

All data were obtained from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database¹. The MRI images were T1 weighted structural scans from a mixture of 1.5T and 3T scanners. All were subjected to quality control and automatically corrected for spatial distortion caused by gradient nonlinearity and B1 field inhomogeneity and downloaded from the ADNI database. Subjects were classified as healthy control (HC), AD or mild cognitive impairment by neuropsychological and clinical testing at the time of the baseline scan, and only HC and AD subjects were used. For the classification experiments, a further quality control step was taken which removed 16 subjects with registration errors, leaving a final total of 627 AD and control subjects plus 346 MCI-s and MCI-c subjects. Their demographics are given in table 1.

¹<http://adni.loni.ucla.edu/>

Table 8.1: Subject groups and demographics

Disease status	n	1.5T	F	Mean age (sd)
HC	376	162	192	74.8 (5.8)
MCI-s	163	109	56	72.5 (7.7)
MCI-c	183	120	78	72.8 (7.3)
AD	251	140	114	75.3 (7.8)

Table 8.2: Demographics of subjects for ARK experiments. n is total number of subjects in the group, 1.5T is the number of subjects in the group whose MRI scan was at 1.5T, F is the number of females in the group. sd is the standard deviation of ages in the group

8.3 Image processing

8.3.1 Groupwise registration

As the method defines features at the voxel level, it was necessary to transfer images into a common space. This was done using the same procedure described in section 7.2.3: all native space images were rigidly and then affinely registered to a randomly chosen image, coalescing the registered images to update the template after each round of registrations. This was then followed by ten rounds of nonrigid registration to produce a final template in the groupwise space. All registrations were performed using the NiftyReg package [Modat et al., 2010].

8.3.2 Image segmentation

All images were segmented into GM, WM, CSF, and non-brain tissues components using the new segment module of SPM12 with the cleanup option set to maximum. A brain mask generated from the original structural image was then applied to the GM segmentations to further exclude any non-brain material.

8.3.3 Image parcellation

The native space images were also anatomically parcellated into 83 regions. This was done with a novel label fusion algorithm [Cardoso et al., 2012] in a multi-atlas label propagation scheme. A library of 30 atlases manually labelled with 83 anatomical regions was used as a basis for the parcellation [Gousias et al., 2008].

8.3.4 Atlas construction

Unlike in other approaches using anatomical regions, features were defined at the level of the voxel rather than regions, requiring that all images share a common space. As kernels were constructed from the voxels within anatomical regions common across subjects, the parcellation defining the region was also required to be in the common space. However, our initial

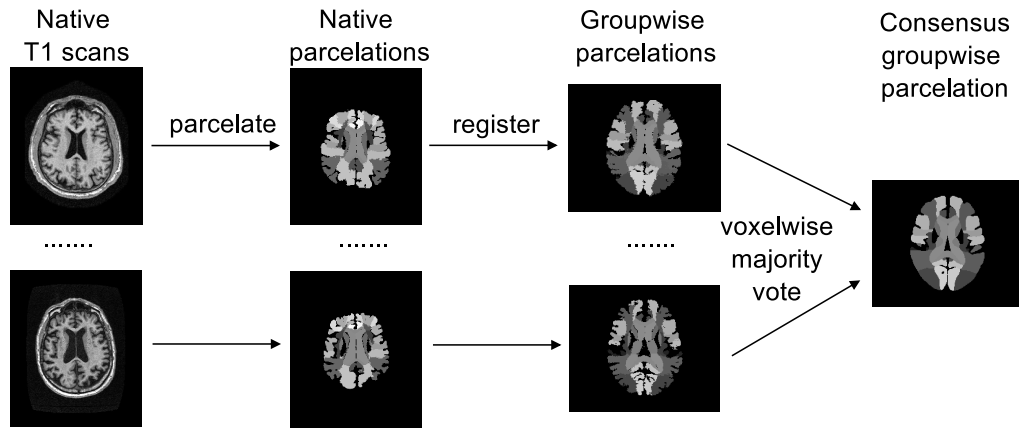


Figure 8.1: Pipeline for constructing atlas in groupwise space

parcellations were in the native spaces of each subject. To combine these initial parcellations in the groupwise space, the following procedure was used. First, all the parcellations were warped into the groupwise space, using the parameters from the native space of each image to the final groupwise template. Care was taken to preserve the integer labels in the parcellations during resampling. Finally to combine the individual parcellations, a consensus atlas was produced by majority voting among the set of N parcellations X to assign a single label l to each voxel v_i of the groupwise space Ω :

$$v_i, i \in \Omega = \arg \max_l \sum_{j=1}^N \begin{cases} 1, & \text{if } X_{i,j} = l \\ 0, & \text{otherwise} \end{cases} \quad (8.1)$$

The pipeline to construct the atlas is summarised graphically in 8.1.

8.4 Gaussian process classification

Gaussian processes (GPs) provide a Bayesian, kernelised framework for solving both regression and classification problems. We refer the reader to previous chapters or [Rasmussen and Williams, 2006] for a more detailed treatment. Briefly, however, a GP (essentially a multivariate Gaussian) forms the prior on the value of a latent function f . For binary classification, the value of the latent function is linked to class membership probability by a sigmoidal function. The GP is parameterised by a mean function $m(\mathbf{x})$ and a covariance kernel function $k(\mathbf{x}, \mathbf{x}')$.

$$p(f(\mathbf{x}), f(\mathbf{x}')) \sim \mathcal{N}(\mathbf{m}, \mathbf{K}), \text{ where } \mathbf{m} = \begin{bmatrix} m(\mathbf{x}) \\ m(\mathbf{x}') \end{bmatrix}, \mathbf{K} = \begin{bmatrix} k(\mathbf{x}, \mathbf{x}) & k(\mathbf{x}, \mathbf{x}') \\ k(\mathbf{x}', \mathbf{x}) & k(\mathbf{x}', \mathbf{x}') \end{bmatrix} \quad (8.2)$$

8.4.1 Gaussian processes as multimodal kernel methods

As equation 8.2 implies, GP classification belongs to the family of kernel methods. Hence a positive sum of valid kernels is a valid kernel, and a valid kernel multiplied by a positive scalar is also a valid kernel. The covariance between the i th and j th subject, \mathbf{K}_{ij} , is a kernel function k of the feature vectors for the i th and j th subject \mathbf{x}_i and \mathbf{x}_j and hyperparameters θ . For ARKs, the final kernel K_{ij} is the weighted sum of 83 linear subkernels, each of which is the dot product between the voxels within a particular anatomical region of the i th and j th image. These regions are defined using masks for each label derived from the groupwise atlas. The covariance hyperparameters are the weights of the subkernels α and bias term β , so the final kernel value \mathbf{K}_{ij} is given by

$$\mathbf{K}_{ij} = k(\mathbf{x}_i, \mathbf{x}_j) = \beta + \sum_{r=1}^{83} \alpha_r (\mathbf{x}_{i,r} \cdot \mathbf{x}_{j,r}) \quad (8.3)$$

where r indexes regions 1 to 83 and β is a bias term. There are thus 84 covariance hyperparameters: $\theta_{cov} = (\alpha_1, \alpha_2, \dots, \alpha_{83}, \beta)$. All the above calculations are carried out within the GPML toolkit², which was modified to take precomputed kernel matrices.

8.5 Results

To generate classification results for the HC and AD subjects, we perform a leave-one-out cross validation (LOOCV) across the entire set of 627 subjects. For the purposes of comparison to existing methods, we also deploy two more conventional methods related to those introduced: using voxel level data for the whole brain, and generating a feature per region, by dividing the total volume of GM by the total intracranial volume to create a normalised amount of GM. Both of these methods are then used with a single kernel linear GP formulation. These are referred to as 'voxels' and 'regions' respectively.

The experimental framework for classifying the MCI subjects as MCI-s or MCI-c is a little different. As we perform transfer learning, by training a classifier on AD and control subjects and then applying it to the MCI-s and MCI-c ones, no cross validation is necessary as the training and test sets are already disjoint.

8.5.1 Binary accuracy

We compare the three methods by thresholding predicted probabilities at 0.5 and comparing to ground truth labels for HC or AD status. The resulting sensitivity, specificity and accuracy are shown in table 2. We also show the area under the ROC curve (AUC), and a p-value for difference in accuracy with McNemar's test. The ARK formulation displays a greater accuracy and

²<http://www.gaussianprocess.org/gpml/code/matlab/doc/>

Method	sens (%)	spec (%)	acc (%)	p vs ARK for acc	AUC	IS (bits)
ARK	80.9	92.6	87.9	-	0.937	0.528
voxels	73.7	93.9	85.8	0.166	0.914	0.395
regions	80.1	91.0	86.6	0.409	0.9275	0.486

Table 8.3: Accuracy of classification between control and AD subjects with ARK, voxels and regions methods. IS = information score. (section 8.5.2)

Method	sens(%)	spec(%)	bal acc(%)	p vs ARK for bal acc	AUC	IS (bits)
ARK	67.8	70.6	68.5	-	0.741	-0.069
voxels	57.9	75.5	66.7	0.374	0.740	0.105
regions	68.9	63.8	66.7	0.555	0.732	-0.071

Table 8.4: Accuracy of classification between MCI-s and MCI-c subjects with ARK, voxels and regions methods. IS = information score. (section 8.5.2)

AUC than both competing methods. While the advantage over the voxels method is substantial, we do not have enough subjects, and thus statistical power, to show that it or the smaller advantage over regions is statistically significant. We can, however, exploit the probabilistic nature of GP classification predictions to show the superior performance of ARKs with other forms of validation.

For classifying MCI-c versus MCI-s, we show the same information, plus balanced accuracy.

8.5.2 Information scoring

The outputs of GP classification are probabilities of a test subject belonging to a particular class (AD, in our case). We can therefore calculate the test log predictive probability ($\log_2 p(y' | (x', \vec{X}, \theta))$) and average this across all test subjects. We then subtract this from the mean log probabilities of a baseline method which does not use the data, but instead simply estimates the class membership probabilities from the prevalences in the training subjects. This tells us how much information, in bits, the classifier was able to extract from the data about the identity of the test subjects [Rasmussen and Williams, 2006]. For perfectly accurate classification with 100% confidence this would be equal to one. Results are also shown in table 8.3 for the HC vs AD classification and table 8.4 for classification of MCI subjects.

As table 8.4 shows, the information scores (IS) can be negative even in cases where accuracy appears high. This is because the score is heavily affected by the more extreme predictions

(that is, probabilities close to one or zero) meaning the score can be negative even if most predictions are correct, if the minority of incorrect ones tend to be extreme. We show a way to visualise the spread of individual predictions in the next section.

8.5.3 Individual predictions

We can also visualise the effects of different methods on individuals for the whole set. Figures 8.2 and 8.3 show the *difference* in predicted $p(AD)$ for all subjects between ARKs and both competing methods. Results are colour-coded so AD subjects are shown in red and NC ones in blue, and sorted by the value of the $p(AD)$ for the competing method. Hence blue (NC) subjects will be represented by a line extending left from the baseline, and red (AD) subjects by a line extending right, if ARKs improve the baseline classification. The plots also show how most subjects are correctly classified: the AD subjects mostly occupied the right hand side of the plots ($p(AD) > 0.5$) and the NC ones the left side of the plots. We can also summarise the differences between individual predictions by again using the information score, using the voxels or regions as a baseline rather than training label prevalences. This gives an advantage for ARK of 0.045 bits over voxels and 0.133 bits over regions.

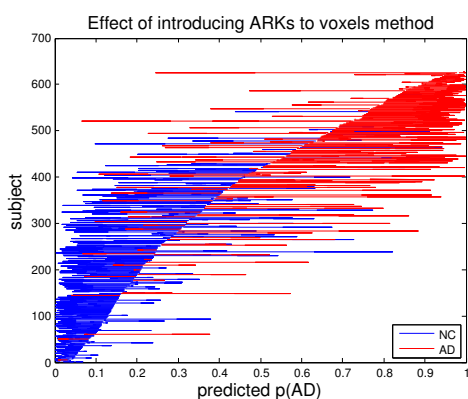


Figure 8.2: Differences between individual predictions of AD versus control status by the ARK and voxel methods

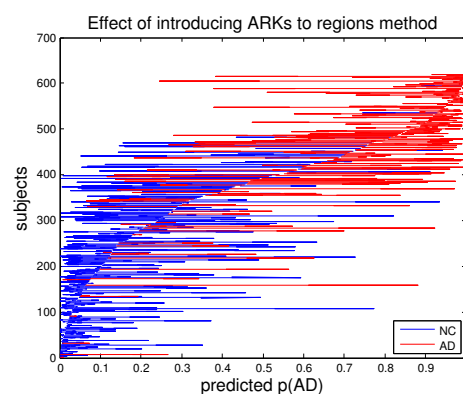


Figure 8.3: Differences between individual predictions of AD versus control status by the ARK and regions methods

The same visualisation can be done for the effects of ARK features on the voxels and regions methods for classifying MCI-s vs MCI-c, as is shown in figures 8.4 and 8.5

These plots reveal why the IS can be so misleading. The ARK formulation has a tendency to make predictions more confident as it allows the classifier greater flexibility to fit the training data than either the voxels or regions method, due to having a much larger number of hyperparameters. This results in improved classification overall; however a number of subjects with predictions that are fairly moderate (close to 0.5) are made into very confident, incorrect

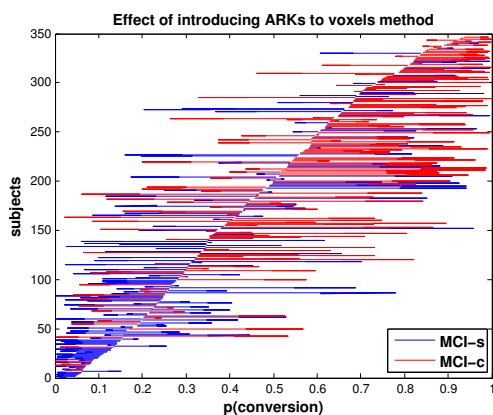


Figure 8.4: Differences between individual predictions of MCI conversion by the ARK and voxel methods

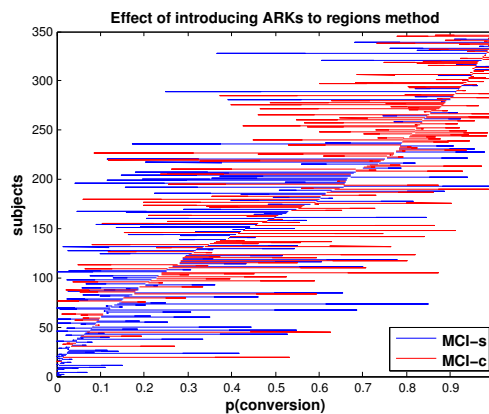


Figure 8.5: Differences between individual predictions of MCI conversion by the ARK and regions methods

predictions. These are not numerous enough to prevent the accuracy of predictions improving in a binary sense, but do nevertheless affect the IS disproportionately.

8.5.4 Effects of scanner field strength

As we showed in a previous chapter (section 7.4.1), the field strength used to acquire the subjects' structural MRI scans has a strong effect on the accuracy of predictions based on the resulting MRI data, or even on PET data for the same subjects if a structural MRI is necessary to process it. Here we split the MCI-s and MCI-c subjects by the field strength at which they were scanned. For both groups (1.5T and 3T) we present the specificity, sensitivity and accuracy by thresholding at 0.5. We also calculate a balanced accuracy in the usual LOO fashion separately for each group. The results are shown in table 8.5. As can be seen, the subjects scanned at 3T have a slightly higher accuracy. More surprisingly perhaps, the balanced accuracies for both groups when calculated separately are *both* better than the balanced accuracy for the pooled group. This is likely because the distributions of predicted probabilities are different for the different field strengths, which results in suboptimal selection of a threshold when the predictions for subjects with the two field strengths are pooled. It should be emphasised that these results were generated by training on all subjects; the splitting by field strength was done after all predictions were computed.

8.6 Discussion

ARKs enable improved classification by combining the strengths of low level (voxel) and high level (regional) features. This results in a classifier that has more flexibility than either voxels or regions do alone. As a result, ARK classification is able to fit the training data better than either;

Field strength	sens(%)	spec(%)	acc(%)	bal acc(%)	AUC
All	67.8	70.6	69.1	68.5	0.741
1.5T	70.0	66.1	68.1	68.6	0.710
3T	63.5	79.6	70.1	72.7	0.788

Table 8.5: Results for ARK classification of MCI-s and MCI-c, broken down by MRI scan field strength

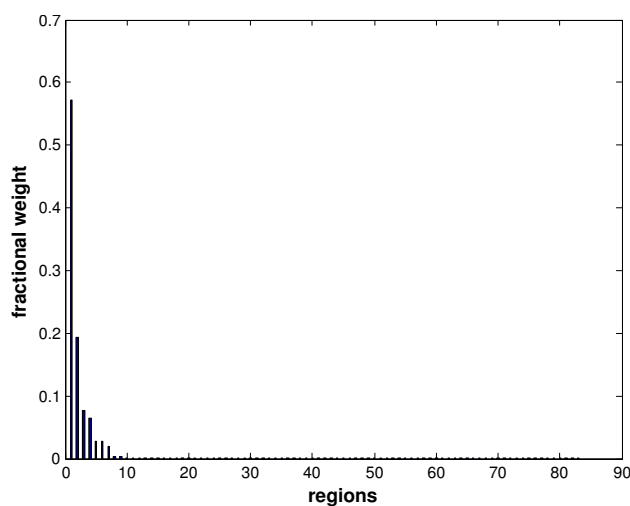


Figure 8.6: Spectrum of regional weights in AD/HC classification. The weights are sparse, with the vast majority of the total weight shared among just a small minority of the regions.

however, the introduction of prior anatomical information means that the hyperparameters are maintained at biologically plausible values which prevents the extra flexibility resulting in too much overfitting.

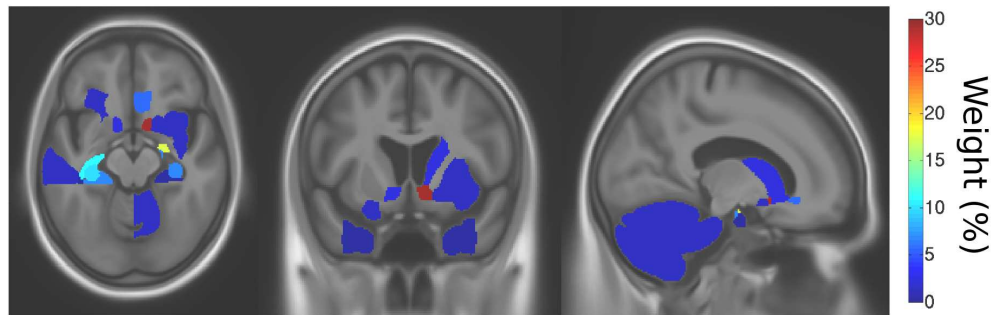
8.6.1 Interpretation of hyperparameters

The optimised weights α tell us about the importance of the corresponding regions in the classification, and hence in AD. For each of the 627 sets of α , we normalise α so they represent a fraction of the total weight, then average each normalised weight across all folds. The spectrum of weights is shown in figure 8.6.

This shows that the weights are fairly sparse, with only 14 regions having weights of more than 1% of the total.

These are shown in figure 8.7. They include temporal lobe regions frequently implicated in AD in studies such as [Braak and Braak, 1995], as well as the GM tissue adjacent to the temporal horn of the left lateral ventricle, which will be very sensitive to expansion of the horn. However, other structures much more widely distributed across the brain are also shown to be

Figure 8.7: Maps of regions with more than 1% of total weight



important in the classification. For example, the largest weight value is given to the right nucleus accumbens, and the left nucleus accumbens and right caudate are also given large weights. This may reflect atrophy due to AD in deep as well as cortical grey matter, as is suggested in [de Jong et al., 2008, Madsen et al., 2010]. One possible alternative explanation is that there may be a number of patients who have dementia due to a cause other than AD, which causes atrophy in a different set of brain regions. It is also possible that the unexpected results may be caused by registration and/or segmentation errors, meaning a small number of highly weighted voxels are assigned to the wrong anatomical region and therefore give it a larger weighting than it would otherwise have.

8.7 Conclusion

Our results show that ARKs successfully combine voxel level data with prior anatomical knowledge, and may offer an accuracy improvement compared to voxel level data alone. The approach also suggests an improvement over features based on predefined regions, although for classification of healthy controls and AD patients this is a smaller advantage than for voxel data. We are also able to visualise both the improvements ARKs bring to individual subjects, and the final weights on the kernels. This may be able to reveal new regions that were previously thought not to be involved in AD. It is interesting to note that the resulting pattern of weights is sparse. This arises naturally from the way the problem is formulated with automatic relevance determination.

The method is quite general, and could be applied both to other imaging modalities, such as PET data, or to other features derived from MRI. All that is required is that low level features, such as voxels, are grouped into regions and can all be transferred into a common space to provide correspondence. This could be, for example, vertexwise cortical thickness data and a labelled cortical atlas.

The chief disadvantage of ARKs is speed of classifier training, due to the high dimension-

ality of the data and the large number of hyperparameters; however this is largely compensated for by the use of modified software that uses precomputed (sub)kernel matrices.

Chapter 9

Conclusions

9.1 Overall conclusions

The work presented in this thesis has significantly advanced the state-of-the art for the problem of predicting progression to AD in MCI subjects. Using the largest cohort of MCI-s and MCI-c subjects available to us, the MRI group in chapters 7 and 8, we have obtained the following balanced accuracies, given in table 9.1.

We report the accuracy for MRI data only here as it is the most widely used imaging modality in dementia, and also the least invasive. However we find that by combining MRI ARK data with ApoE genotype (with can be obtained noninvasively, in comparison with FDG-PET and CSF) we can increase the balanced accuracy still further to 71.7% overall.

It is clear, however, that FDG-PET data does offer superior accuracy to MRI. In every experiment where the same method was applied to the same set of subjects with both PET and MRI, PET always outperforms MRI alone. The best balanced accuracy we obtained with the largest possible subset of MCI subjects – again, the PET group from the MMSE experiments in chapter 7 – gave a balanced accuracy of 75%. Among the PET subjects whose MRI was

Method	Field strength	Bal acc(%)	AUC
ARK	1.5T	68.6	0.710
ARK	3T	72.6	0.788
ARK	all	68.5	0.741
MMSE	1.5T	65.2	0.734
MMSE	3T	0.752	0.811
MMSE	all	68.9	0.761

Table 9.1: Best overall classification with MRI data only. Balanced accuracies for 1.5T or 3T subjects obtained from leave-one-out loop over those subjects only; training was on all subjects.

from a 3T scanner this accuracy rises to 83%, which is the highest yet reported for predicting conversion within three years. However, for the PET subjects whose MRI scan was at 1.5T the accuracy is much lower, suggesting that the scanner field strength affects the processing of PET data even more than it does the MRI itself. This is possibly because processing PET data requires an additional registration step compared to MRI data. This suggests that it may be desirable for the decision about which method should be used to predict conversion to AD in any particular MCI subject to depend on the image quality of the subject's structural MRI. Scanning at 3T if at all possible is desirable, as it will enable prediction of conversion to AD within three years with 75% accuracy if only the MRI is available, and 83% accuracy if FDG-PET is also available.

In the small groups of subjects described in chapter 6 and for the BSI experiment in chapter 7, multimodal prediction using both MRI and FDG-PET was found to give slightly better results than FDG-PET alone. However, when the FDG-PET image processing pipeline was improved, enabling us to use voxel level rather than region level PET features, the advantage of multimodal classification all but disappeared, as shown in table 9.2. The highest accuracy of just over 83% is obtained from PET data only for subjects scanned at 3T, which is not improved by adding MRI data. For subjects scanned at 1.5T, however, adding MRI data does give a very modest improvement. This suggests that multimodal classification may strengthen data that are suboptimal, but if any single modality is very strongly predictive in its own right then combining with other data does not offer any statistically significant improvement. Interestingly, we also found that the best results for subjects scanned at 1.5T or for all subjects together were obtained by training on all subjects, whereas for subjects scanned at 3T the best accuracy came from training only on 3T subjects despite the resulting smaller size of the training set. This is the way the results in table 9.2 were obtained for the 1.5T subjects and all subjects, and for the 3T subjects respectively.

The experiment with continuous proxies, as described in chapter 7, offered slightly higher classification accuracies than the conventional approach involving binary labels if only MRI data is available. Again, however, the advantage appears to be smaller when the number of subjects is larger (unsurprisingly, as the continuous proxy approach makes more efficient use of the data) so in the limit there may be little to choose between the two approaches. It is, however, clear that in the case of binary labels the transfer learning approach of training on AD and healthy subjects, and then applying the resulting classifier to a population of MCI-s and MCI-c subjects, is more effective than training directly on the MCI subjects.

Modality/modalities	Field strength	Bal acc(%)	AUC
PET	1.5T	68.38	0.757
PET	3T	83.08	0.867
PET	all	74.13	0.796
PET+MRI	1.5T	69.12	0.750
PET+MRI	3T	83.08	0.882
PET+MRI	all	73.13	0.794

Table 9.2: Best overall classification with voxel-level PET data only and voxel level PET and MRI data combined with MKL.

9.2 Future research

There are a number of different avenues in which the work presented in this thesis could be extended. As mentioned in the literature review in chapter 4, three primary choices affect the accuracy of diagnosis and prognosis in studies such as the ones introduced here. The first and most important is what is referred to in [Sabuncu and Konukoglu, 2014] as the biological footprint of the condition, which is the magnitude of differences between healthy subjects and patients. However, this is fixed for any specific condition, which means that to improve the prediction of development of AD in MCI patients, efforts must instead be focused on the two other factors affecting accuracy: the choice of classifier or machine learning algorithm, and the choice of data.

Simply utilising more sophisticated or complex classification algorithms does not appear to lead to greater accuracy. This is shown experimentally in [Sabuncu and Konukoglu, 2014], which found that choice of classifier algorithm had a far smaller effect on accuracy than choice of data or biological footprint size. The same observation can be made for the specific problem of separating MCI-s and MCI-c subjects with ADNI data, as accuracies (using 1.5T structural MRI data) appear to have plateaued at around 75% despite a number of researchers employing newer and more sophisticated types of classifiers such as deep learning methods [Suk and Shen, 2013]. However, it is possible to improve results by using existing algorithms in innovative ways to produce novel classification paradigms. This can be done to make better use of existing subjects' data. With binary classification, the training data is normally made up of one particular group - either MCI-s and MCI-c subjects, or in the transfer learning setting control and AD subjects. Either way, a large number of potential training subjects are left unused. However, the unused subjects do contain relevant information and finding a way to make use of all subjects regardless of their disease status would be very desirable. Using contin-

uous proxies is one way to do this, and this waste of data was in fact one of the motivations for introducing them. The continuous proxy approach could itself be improved, by finding better regression techniques and more relevant proxies. It should also be possible to improve results by using multiple proxies (such as MMSE, ADAS-COG and BSI) with multi-task learning or canonical correlation analysis (CCA) [Hotelling, 1936]. This could potentially even be used to relate multivariate patterns of atrophy to baseline GM maps. However, this is not the only way to make use of all the data. Ordinal regression retains discrete labels but be able to use all data regardless of group for training. Initial results for this [Doyle et al., 2014] were unexceptional, however it remains a conceptually attractive approach. Alternatively, semi-supervised learning [Chapelle et al., 2010] may prove useful and has already been applied to AD classification using other techniques [Filipovych and Davatzikos, 2011].

Similarly, we could also make better use of existing data by exploiting multiple timepoints. While longitudinal features have been widely used on ADNI data, MCI-c subjects are generally treated as homogeneous when in fact the time of conversion may vary wildly in terms of when (within the cutoff period) they actually do convert to AD. Survival analysis methods, such as Cox regression [Cox and Oakes, 1984], are perhaps the best way to model such data, and can also be used with GP priors [Barrett and Coolen, 2013, Vanhatalo et al., 2013].

Almost all of these methods would likely be improved by using a sparse method for learning. Sparsity is particularly attractive for AD or MCI classification as it is well known that the pattern of atrophy in AD is well defined and generally restricted to specific areas of the brain [Braak and Braak, 1995]. Therefore enforcing a sparse set of weights for either voxel or region level data should result in a more accurate model as the distribution of weights is likely to be more biologically plausible and hence more directly reflective of the underlying disease processes. Furthermore, promotion of sparsity will help to reduce overfitting which is of particular concern in neuroimaging due to the small number of subjects and frequently very high dimensionality of the data. A sparse approach was used in one of the highest performing methods for predicting MCI conversion without use of PET data [Ye et al., 2012]. Off-the-shelf sparse methods such as the elastic net [Zou and Hastie, 2005] can be applied to neuroimaging data. Alternatively, special methods combining sparsity and spatial smoothness for neuroimaging data have been successfully developed [Gramfort et al., 2013, Cuingnet et al., 2013, Sabuncu and Leemput, 2012]. All of these can be used for direct binary classification or continuous proxy regression, and could also be adapted for more specialised applications such as time-to-event models as in [Sabuncu, 2013].

An intermediate point between using different learning algorithms and different types

of image data would be to find superior features from existing data. In particular, we did not explore the use of cortical thickness measurements in this thesis. Also, scalar momentum features derived from diffeomorphic registrations have been used as features for classification and regression with some success [Marquand et al., 2013, Singh et al., 2014]. Both of these have the advantage that they could be used with the ARK framework to further improve results. Other types of feature are based on extracting statistical information from voxel level data. Hippocampal grading has been shown to yield good results [Coupé et al., 2012], which suggests that features based on MRI texture feature may also be predictive. The long-standing grey level cooccurrence matrix (GLCM) approach to texture has been applied to AD classification [Freeborough and Fox, 1998] with promising results. However the more recent, fully three dimensional and rotation invariant texture features such as local binary patterns [Fehr and Burkhardt, 2008, Banerjee et al., 2013] and statistical geometric features [Chen et al., 1995] may perform better. Finally, it is possible to formulate a kernel in a manner that means the elements of the kernel matrix are not similarities between vectors, but between clouds of points which can have varying cardinalities [Rahimi and Recht, 2007]. This method has already been used successfully with fractional anisotropy (FA) data derived from DTI brain images [Ansari et al., 2014]. It has the advantage of retaining spatial information and thus could potentially be more discriminating as well as not requiring nonlinear registration. An obvious application for this would be to voxel level structural MRI data in intensity normalised, affinely aligned brains or brain ROIs. All these have the advantage of requiring structural MRI data. As mentioned previously, despite the greater accuracy provided by FDG-PET data it is still desirable to improve results based on MRI data alone, due to the lower cost and invasiveness and greater availability of MRI.

The biggest increases in accuracy are likely to come from entirely different types of data, especially if these are informed by greater understanding of the AD process and its causes. It is possible that future biomarkers will be used that are not informed by image data at all, such as those based on lipid levels in blood which have attracted widespread publicity recently [Mapstone et al., 2014]. This method, however, recruited at only two centres, rather than the far greater number used in ADNI, and furthermore it predicted the onset of dementia in healthy subjects rather than predicting conversion to AD in an MCI population. While this difference may appear to be minimal, it is possible that the blood test is actually performing an easier classification due to a greater biological difference between the two groups. Additionally a test for healthy subjects implies screening, which brings its own problems and has different requirements to predicting conversion in the MCI population.

Biomarkers extracted from newer imaging modalities have recently been shown to hold a great deal of promise. This thesis has only employed structural MRI and FDG-PET, to examine brain anatomy and metabolism. However AD is a disease which affects the brain in a variety of ways. In particular, it is known to alter the white matter connectivity between regions, meaning simple measures extracted from DTI such as FA correlate strongly with cognitive measure [Nir et al., 2013], and have been shown to give very good separation between MCI-s and MCI-c subjects in initial small scale studies [van Bruggen et al., 2012]. Similarly, AD also affects brain function, which is detectable in fMRI. This has also shown promise in classifying AD patients and controls on a small scale [Li et al., 2013]. More recently, arterial spin labelling (ASL), a method of using MRI to measure cerebral blood flow, has emerged as a potential biomarker for AD [Wang et al., 2013]. The reason for the low numbers of subjects in all studies involving these advanced MRI modalities is that they have only recently been added to the ADNI protocol. As the relevant data becomes available in larger quantities in the ADNI database, they should enable a great variety of new methods for predicting conversion from MCI to AD to be developed.

Appendix A

Running times and computational complexity

A.1 Experiment and results

In order to give estimates of the likely running times of a typical classification, this section contains the results of a small scale experiment. The experiment was performed on a set of 100 randomly chosen structural MRI scans of subjects from the ADNI dataset, consisting of 50 healthy subjects and 50 with probable AD. The time taken to train a classifier from these data was noted as training time is much greater than the time to make a prediction on an unseen subject. We present results for GP classification using the original GPML toolkit (<http://www.gaussianprocess.org/gpml/code/matlab/doc/>), for our modified version of GPML using a precomputed kernel, and for comparison an SVM, also both with and without a precomputed kernel using LIBSVM (<http://www.csie.ntu.edu.tw/~cjlin/libsvm>). For the GP classification, training here refers to optimisation of the hyperparameters. 200 iterations of the optimiser were used for the GP experiments, the same number as in all previous experiments. SVM training used default parameters. The experiments were performed in Matlab running on a Linux PC with 3.8GB of RAM and two 2.4GHz processors. Results are shown below in table A.1.

A final experiment was performed, using the same task of training a classifier on the structural MRI scans of 50 AD and 50 control subjects. Here, however, the data was in the form of

Classifier	Precomputed kernel	Running time
SVM	no	7.9 seconds
SVM	yes	0.005 seconds
GP	no	17 minutes, 22 seconds
GP	yes	16.8 seconds

Table A.1: Running times for training a model on 50 AD and 50 control subjects

83 multiple regional kernels, as described in chapter 8. This implies setting a much larger number of hyperparameters as one is introduced for each kernel. Correspondingly training time was greater than for the single precomputed kernel GP form, but still reasonable at four minutes, 24 seconds.

A.2 Discussion

The results here are for an experiment using a much smaller number of training subjects than would be used in practice (and is smaller than most of the training sets used in other experiments in this thesis). Estimates of running times in a more realistic setting may be made by extrapolating from these results. GPs do have the disadvantage of scaling relatively poorly. As the algorithm is dominated by a matrix inversion their computational complexity is $\mathcal{O}(n^3)$. For large scale classification, this can, however, be improved on, by the fully independent training conditional (FITC) approximation [Naish-Guzman and Holden, 2007]. This is included in the GPML software, however it was not used for any of the experiments in this thesis. The scaling behaviour of SVMs is slightly better. Although the computational complexity is data dependent and more difficult to characterise, it is generally between $\mathcal{O}(n^2)$ and $\mathcal{O}(n^3)$ [Chang and Lin, 2011]. GP classification does have an advantage for the multikernel formulation, due to its ability to learn a large number of hyperparameters from training data only. The method of [Zhang et al., 2011] could not perform the equivalent; grid search and cross-validation is tractable to set a handful of parameters but not for 83.

The runtimes in table A.1 should be interpreted with caution. Although we would expect GP classification to be slower than an SVM generally due to the greater computational complexity, this is likely exaggerated in the results. This is because GPML is written entirely in native Matlab, whereas LIBSVM is a highly optimised C++ library that can be called from other environments such as Matlab or Python. Unfortunately to my knowledge there are no fast C++ implementations of GP classification with the flexibility of GPML. However, there is considerable scope for making GP classification faster by developing one.

It is quite clear, however, that precomputing a kernel matrix is highly advantageous for both classifiers. This is because, in applications to medical image data, the dimensionality of the data is very high so generating the kernel is a slow operation; furthermore in GPML the kernel is regenerated for each iteration. This is actually necessary as the elements of the kernel matrix depends on the hyperparameters, so the kernel matrix must be updated as the hyperparameters are being optimised. However, for some kernel functions (including linear ones), the kernel matrix is a function of a constant matrix of the same size and the hyperparameters. This constant

matrix can be calculated directly from the data and it is this that is used in our modified version of GPML. Although, as previously stated, generating a precomputed kernel matrix is itself a slow operation, this is not a major problem in practice as it need only be done once. New data can then be added incrementally. If the original data are retained alongside a kernel matrix, then the new training and/or testing data can be incorporated into the matrix by computing the new matrix elements only and inserting them rather than computing the entire matrix from scratch.

Appendix B

Lists of subjects in experiments

B.1 Subjects in experiment in chapter 5

The subjects used in the AD and normal control classification experiment described in chapter 5 are listed in table B.1.

Table B.1: List of subjects used in the control/AD classification experiment in chapter 5. ID = ADNI roster ID number, Status = disease status of subject.

ID	Status	ID	Status	ID	Status	ID	Status	ID	Status	ID	Status
16	NC	272	NC	672	NC	3	AD	400	AD	803	AD
22	NC	295	NC	717	NC	76	AD	426	AD	850	AD
43	NC	312	NC	726	NC	84	AD	431	AD	891	AD
47	NC	327	NC	731	NC	93	AD	457	AD	1082	AD
55	NC	352	NC	751	NC	94	AD	470	AD	1090	AD
56	NC	433	NC	768	NC	129	AD	517	AD	1137	AD
59	NC	454	NC	779	NC	139	AD	535	AD	1144	AD
89	NC	467	NC	843	NC	147	AD	543	AD	1170	AD
95	NC	488	NC	886	NC	149	AD	547	AD	1171	AD
96	NC	498	NC	923	NC	194	AD	577	AD	1209	AD
97	NC	516	NC	926	NC	213	AD	606	AD	1221	AD
123	NC	519	NC	972	NC	221	AD	619	AD	1262	AD
159	NC	520	NC	981	NC	266	AD	690	AD	1285	AD
172	NC	525	NC	984	NC	286	AD	720	AD	1290	AD
177	NC	533	NC	1002	NC	300	AD	733	AD	1337	AD

186	NC	534	NC	1016	NC	310	AD	739	AD	1341	AD
210	NC	555	NC	1035	NC	321	AD	753	AD	1371	AD
232	NC	559	NC	1063	NC	341	AD	754	AD	1373	AD
259	NC	602	NC	1200	NC	366	AD	784	AD	1379	AD
260	NC	648	NC	1250	NC	372	AD	790	AD	1402	AD

B.2 Subjects in experiment in chapter 6

The subjects used in the experiments described in chapter 6 are listed in table B.2. This lists all the subjects in the PET group. For each subject, it is listed whether it was also present in the PET-CSF group.

Table B.2: List of subjects used in the PET group of the MKL experiment in chapter 6. ID = ADNI roster ID number, Status = disease status of subject, CSF = subject has CSF data so is used in PET-CSF group.

ID	Status	CSF	ID	Status	CSF	ID	Status	CSF	ID	Status	CSF
5	NC	Yes	1202	NC	No	200	MCI-s	No	1315	MCI-s	Yes
14	NC	Yes	3	AD	Yes	225	MCI-s	No	1318	MCI-s	No
16	NC	Yes	10	AD	Yes	227	MCI-s	No	1322	MCI-s	No
21	NC	No	53	AD	No	282	MCI-s	No	1346	MCI-s	No
23	NC	Yes	147	AD	Yes	292	MCI-s	Yes	1351	MCI-s	Yes
43	NC	Yes	149	AD	Yes	314	MCI-s	Yes	1378	MCI-s	No
48	NC	No	167	AD	No	354	MCI-s	No	1384	MCI-s	No
55	NC	Yes	183	AD	No	361	MCI-s	Yes	1406	MCI-s	No
67	NC	No	216	AD	No	378	MCI-s	Yes	1407	MCI-s	No
74	NC	No	221	AD	Yes	389	MCI-s	No	1408	MCI-s	No
95	NC	Yes	266	AD	Yes	408	MCI-s	No	1414	MCI-s	Yes
120	NC	Yes	286	AD	Yes	410	MCI-s	Yes	1417	MCI-s	No
123	NC	Yes	316	AD	Yes	414	MCI-s	No	1418	MCI-s	No
130	NC	No	321	AD	Yes	424	MCI-s	Yes	1419	MCI-s	Yes
171	NC	No	341	AD	Yes	446	MCI-s	Yes	1425	MCI-s	No
173	NC	Yes	343	AD	No	461	MCI-s	No	1426	MCI-s	No

223	NC	Yes	370	AD	No	464	MCI-s	No	57	MCI-c	Yes
230	NC	No	374	AD	No	481	MCI-s	Yes	101	MCI-c	Yes
259	NC	Yes	400	AD	Yes	485	MCI-s	No	128	MCI-c	No
262	NC	No	431	AD	Yes	531	MCI-s	Yes	141	MCI-c	No
272	NC	Yes	474	AD	Yes	544	MCI-s	Yes	155	MCI-c	No
283	NC	No	497	AD	No	546	MCI-s	No	204	MCI-c	Yes
301	NC	No	543	AD	Yes	549	MCI-s	No	214	MCI-c	No
311	NC	No	547	AD	Yes	598	MCI-s	Yes	222	MCI-c	Yes
312	NC	Yes	554	AD	No	608	MCI-s	Yes	231	MCI-c	Yes
359	NC	No	565	AD	Yes	621	MCI-s	Yes	240	MCI-c	Yes
386	NC	Yes	577	AD	Yes	626	MCI-s	Yes	256	MCI-c	Yes
416	NC	No	642	AD	No	634	MCI-s	Yes	258	MCI-c	Yes
419	NC	No	682	AD	No	656	MCI-s	No	293	MCI-c	Yes
459	NC	Yes	712	AD	No	669	MCI-s	No	325	MCI-c	No
498	NC	Yes	720	AD	Yes	673	MCI-s	Yes	326	MCI-c	Yes
500	NC	No	740	AD	No	679	MCI-s	No	344	MCI-c	Yes
502	NC	No	754	AD	Yes	698	MCI-s	No	362	MCI-c	Yes
522	NC	No	760	AD	No	709	MCI-s	No	394	MCI-c	Yes
526	NC	No	786	AD	No	715	MCI-s	No	511	MCI-c	Yes
555	NC	Yes	836	AD	Yes	718	MCI-s	Yes	513	MCI-c	No
575	NC	No	850	AD	Yes	746	MCI-s	Yes	567	MCI-c	Yes
576	NC	No	889	AD	No	748	MCI-s	Yes	675	MCI-c	No
610	NC	Yes	891	AD	Yes	770	MCI-s	No	695	MCI-c	No
618	NC	Yes	979	AD	No	800	MCI-s	Yes	708	MCI-c	No
637	NC	Yes	1001	AD	No	865	MCI-s	No	723	MCI-c	Yes
647	NC	No	1041	AD	Yes	909	MCI-s	No	860	MCI-c	No
648	NC	Yes	1044	AD	Yes	914	MCI-s	No	861	MCI-c	Yes
657	NC	Yes	1056	AD	No	919	MCI-s	No	904	MCI-c	Yes
672	NC	Yes	1090	AD	Yes	925	MCI-s	Yes	906	MCI-c	Yes
680	NC	Yes	1109	AD	Yes	932	MCI-s	Yes	941	MCI-c	Yes
686	NC	Yes	1157	AD	No	945	MCI-s	No	978	MCI-c	Yes
731	NC	Yes	1164	AD	No	947	MCI-s	No	997	MCI-c	Yes
734	NC	No	1171	AD	Yes	950	MCI-s	Yes	1007	MCI-c	No
741	NC	No	1205	AD	No	961	MCI-s	Yes	1010	MCI-c	Yes

751	NC	Yes	1221	AD	Yes	973	MCI-s	Yes	1033	MCI-c	Yes
778	NC	Yes	1254	AD	No	976	MCI-s	No	1130	MCI-c	Yes
779	NC	Yes	1281	AD	Yes	994	MCI-s	Yes	1135	MCI-c	No
813	NC	No	1283	AD	No	1030	MCI-s	Yes	1217	MCI-c	Yes
818	NC	Yes	1285	AD	Yes	1032	MCI-s	No	1240	MCI-c	No
842	NC	No	1307	AD	No	1043	MCI-s	Yes	1282	MCI-c	No
843	NC	Yes	1339	AD	No	1073	MCI-s	Yes	1311	MCI-c	No
845	NC	No	1341	AD	Yes	1103	MCI-s	No	1393	MCI-c	Yes
862	NC	No	1368	AD	No	1106	MCI-s	No	1394	MCI-c	Yes
863	NC	No	1371	AD	Yes	1114	MCI-s	No	1398	MCI-c	Yes
866	NC	Yes	1373	AD	Yes	1118	MCI-s	No	1412	MCI-c	No
898	NC	No	1379	AD	Yes	1120	MCI-s	Yes	1423	MCI-c	Yes
934	NC	No	1382	AD	No	1165	MCI-s	No	1427	MCI-c	No
967	NC	No	1402	AD	Yes	1186	MCI-s	No			
972	NC	Yes	33	MCI-s	Yes	1215	MCI-s	No			
985	NC	No	80	MCI-s	No	1218	MCI-s	No			
1002	NC	Yes	112	MCI-s	Yes	1224	MCI-s	Yes			
1023	NC	No	135	MCI-s	Yes	1246	MCI-s	No			
1063	NC	Yes	142	MCI-s	No	1260	MCI-s	Yes			
1133	NC	No	158	MCI-s	Yes	1265	MCI-s	Yes			
1194	NC	No	160	MCI-s	No	1275	MCI-s	No			
1197	NC	No	188	MCI-s	Yes	1314	MCI-s	No			

B.3 Subjects in experiments in chapter 7

The subjects used in the BSI experiment described in chapter 7 are listed in table B.3.

Table B.3: List of subjects used in the BSI experiment in chapter 7. ID = ADNI roster ID number, Status = disease status of subject.

ID	Status	ID	Status	ID	Status	ID	Status	ID	Status	ID	Status
5	NC	866	NC	850	AD	424	MCI-s	1265	MCI-s	997	MCI-c
16	NC	972	NC	891	AD	446	MCI-s	1315	MCI-s	1010	MCI-c
23	NC	1002	NC	1041	AD	481	MCI-s	1351	MCI-s	1033	MCI-c

43	NC	1063	NC	1044	AD	531	MCI-s	1419	MCI-s	1130	MCI-c
55	NC	3	AD	1090	AD	544	MCI-s	57	MCI-c	1217	MCI-c
120	NC	10	AD	1109	AD	598	MCI-s	101	MCI-c	1393	MCI-c
173	NC	147	AD	1171	AD	608	MCI-s	204	MCI-c	1394	MCI-c
272	NC	149	AD	1221	AD	621	MCI-s	222	MCI-c	1398	MCI-c
312	NC	221	AD	1281	AD	626	MCI-s	231	MCI-c	1423	MCI-c
386	NC	266	AD	1341	AD	673	MCI-s	256	MCI-c		
459	NC	286	AD	1371	AD	718	MCI-s	258	MCI-c		
498	NC	316	AD	1373	AD	746	MCI-s	293	MCI-c		
555	NC	321	AD	1379	AD	748	MCI-s	326	MCI-c		
610	NC	341	AD	1402	AD	932	MCI-s	344	MCI-c		
648	NC	400	AD	33	MCI-s	950	MCI-s	362	MCI-c		
657	NC	431	AD	112	MCI-s	961	MCI-s	394	MCI-c		
672	NC	474	AD	135	MCI-s	973	MCI-s	511	MCI-c		
680	NC	543	AD	158	MCI-s	994	MCI-s	567	MCI-c		
686	NC	547	AD	188	MCI-s	1030	MCI-s	723	MCI-c		
731	NC	565	AD	292	MCI-s	1043	MCI-s	861	MCI-c		
751	NC	577	AD	314	MCI-s	1073	MCI-s	904	MCI-c		
779	NC	720	AD	361	MCI-s	1120	MCI-s	906	MCI-c		
818	NC	754	AD	378	MCI-s	1224	MCI-s	941	MCI-c		
843	NC	836	AD	410	MCI-s	1260	MCI-s	978	MCI-c		

The subjects used in the MMSE experiment described in chapter 7 are listed in table B.4. This table lists all subjects in the MRI group used for this experiment. For each subject, as well as the disease group, the scanner field strength and whether the subject had an FDG-PET scan and was therefore also in the PET group is listed. The small number of scans done in 2.9T scanners are grouped with the 3T scans when examining the effects of different scanner field strengths. The MCI-u status is for subjects which are MCI at baseline, but due to having been only recently enrolled in ADNI cannot be assigned to MCI-s or MCI-c. In continuous proxy experiments these subjects can be used for training as long as they have the appropriate proxy data, but cannot be used for testing.

Table B.4: List of subjects used in the MRI group of the MMSE experiment in chapter 7. ID = ADNI roster ID number, Status = disease status of subject, T = field strength of structural MRI in Tesla, PET = subject has PET data so is used in PET group.

ID	Status	T	PET	ID	Status	T	PET	ID	Status	T	PET
2	NC	1.5	Yes	760	AD	1.5	Yes	4414	MCI-c	3	Yes
5	NC	1.5	Yes	777	AD	1.5	Yes	4432	MCI-c	3	No
8	NC	1.5	Yes	790	AD	1.5	No	4502	MCI-c	3	Yes
14	NC	1.5	Yes	793	AD	1.5	No	4515	MCI-c	3	Yes
16	NC	1.5	Yes	796	AD	1.5	No	4530	MCI-c	3	Yes
19	NC	1.5	No	812	AD	3	No	4595	MCI-c	3	Yes
21	NC	1.5	Yes	828	AD	3	No	4661	MCI-c	3	Yes
22	NC	1.5	No	836	AD	1.5	Yes	4680	MCI-c	3	Yes
23	NC	1.5	Yes	850	AD	1.5	Yes	4689	MCI-c	3	Yes
31	NC	2.9	No	852	AD	1.5	No	4706	MCI-c	3	Yes
40	NC	1.5	No	853	AD	1.5	No	4712	MCI-c	3	Yes
56	NC	1.5	No	889	AD	1.5	Yes	4796	MCI-c	3	Yes
58	NC	2.9	No	891	AD	1.5	Yes	4857	MCI-c	3	Yes
59	NC	1.5	No	916	AD	3	No	4888	MCI-c	3	Yes
61	NC	2.9	No	929	AD	1.5	Yes	4899	MCI-c	3	Yes
67	NC	1.5	Yes	955	AD	1.5	No	4918	MCI-c	3	Yes
68	NC	1.5	No	979	AD	1.5	Yes	4928	MCI-c	3	Yes
70	NC	1.5	No	991	AD	1.5	Yes	38	MCI-u	1.5	No
72	NC	1.5	No	996	AD	2.9	No	60	MCI-u	1.5	No
74	NC	1.5	Yes	1001	AD	1.5	Yes	112	MCI-u	1.5	Yes
81	NC	1.5	No	1018	AD	1.5	No	138	MCI-u	1.5	Yes
86	NC	1.5	No	1024	AD	1.5	No	188	MCI-u	1.5	Yes
89	NC	1.5	No	1041	AD	1.5	Yes	282	MCI-u	1.5	Yes
90	NC	1.5	Yes	1044	AD	1.5	Yes	284	MCI-u	1.5	No
96	NC	1.5	Yes	1055	AD	2.9	No	377	MCI-u	1.5	Yes
97	NC	1.5	Yes	1056	AD	1.5	Yes	384	MCI-u	3	No
106	NC	1.5	No	1059	AD	1.5	Yes	393	MCI-u	3	No

113	NC	1.5	No	1079	AD	1.5	No	397	MCI-u	3	No
120	NC	1.5	Yes	1082	AD	2.9	No	410	MCI-u	1.5	Yes
123	NC	1.5	Yes	1101	AD	3	No	414	MCI-u	1.5	Yes
125	NC	1.5	No	1102	AD	1.5	No	417	MCI-u	2.9	No
156	NC	1.5	No	1109	AD	1.5	Yes	429	MCI-u	1.5	No
166	NC	1.5	No	1137	AD	1.5	No	443	MCI-u	1.5	Yes
171	NC	1.5	Yes	1144	AD	1.5	Yes	458	MCI-u	1.5	No
172	NC	1.5	No	1152	AD	1.5	No	485	MCI-u	1.5	Yes
173	NC	1.5	Yes	1157	AD	1.5	Yes	544	MCI-u	1.5	Yes
177	NC	1.5	No	1161	AD	1.5	Yes	551	MCI-u	1.5	Yes
184	NC	1.5	No	1164	AD	1.5	Yes	566	MCI-u	1.5	Yes
186	NC	1.5	No	1170	AD	2.9	No	579	MCI-u	1.5	No
223	NC	1.5	Yes	1171	AD	1.5	Yes	669	MCI-u	1.5	Yes
229	NC	1.5	No	1185	AD	3	No	702	MCI-u	1.5	No
230	NC	1.5	Yes	1192	AD	1.5	No	721	MCI-u	1.5	Yes
232	NC	1.5	Yes	1205	AD	1.5	Yes	739	MCI-u	1.5	No
245	NC	1.5	Yes	1221	AD	1.5	Yes	748	MCI-u	1.5	Yes
257	NC	1.5	No	1253	AD	3	No	783	MCI-u	1.5	Yes
259	NC	1.5	Yes	1254	AD	1.5	Yes	821	MCI-u	1.5	No
260	NC	3	No	1257	AD	1.5	Yes	832	MCI-u	1.5	No
262	NC	1.5	Yes	1262	AD	3	No	890	MCI-u	1.5	No
272	NC	1.5	Yes	1263	AD	1.5	Yes	924	MCI-u	1.5	Yes
283	NC	1.5	Yes	1281	AD	1.5	Yes	928	MCI-u	3	No
295	NC	1.5	No	1285	AD	1.5	Yes	957	MCI-u	1.5	Yes
298	NC	1.5	No	1289	AD	3	No	958	MCI-u	1.5	Yes
301	NC	1.5	Yes	1290	AD	1.5	Yes	1028	MCI-u	1.5	Yes
303	NC	2.9	No	1296	AD	1.5	No	1038	MCI-u	1.5	Yes
312	NC	1.5	Yes	1307	AD	1.5	Yes	1051	MCI-u	1.5	No
319	NC	1.5	Yes	1308	AD	1.5	No	1074	MCI-u	1.5	Yes
327	NC	1.5	Yes	1337	AD	1.5	No	1092	MCI-u	1.5	Yes
337	NC	1.5	No	1354	AD	1.5	Yes	1103	MCI-u	1.5	Yes
352	NC	1.5	Yes	1368	AD	1.5	Yes	1104	MCI-u	3	No
359	NC	1.5	Yes	1373	AD	1.5	Yes	1204	MCI-u	1.5	Yes
360	NC	1.5	Yes	1377	AD	1.5	No	1215	MCI-u	1.5	Yes

363	NC	1.5	Yes	1382	AD	1.5	Yes	1231	MCI-u	1.5	No
382	NC	2.9	No	1385	AD	2.9	No	1245	MCI-u	1.5	Yes
386	NC	1.5	Yes	1397	AD	1.5	Yes	1275	MCI-u	1.5	Yes
403	NC	2.9	No	1435	AD	1.5	No	1277	MCI-u	2.9	No
405	NC	3	No	4024	AD	3	Yes	1279	MCI-u	2.9	No
413	NC	1.5	No	4039	AD	3	Yes	1293	MCI-u	3	No
416	NC	1.5	Yes	4136	AD	3	Yes	1294	MCI-u	1.5	Yes
419	NC	1.5	Yes	4152	AD	3	Yes	1309	MCI-u	2.9	No
436	NC	1.5	No	4153	AD	3	Yes	1343	MCI-u	1.5	Yes
454	NC	1.5	Yes	4172	AD	3	Yes	1366	MCI-u	1.5	No
459	NC	1.5	Yes	4195	AD	3	Yes	1400	MCI-u	1.5	Yes
467	NC	1.5	Yes	4201	AD	3	Yes	1408	MCI-u	1.5	Yes
484	NC	1.5	Yes	4209	AD	3	Yes	1411	MCI-u	1.5	Yes
488	NC	2.9	No	4211	AD	3	Yes	1420	MCI-u	1.5	Yes
489	NC	1.5	Yes	4223	AD	3	Yes	1426	MCI-u	1.5	Yes
493	NC	2.9	No	4252	AD	3	Yes	2007	MCI-u	3	Yes
498	NC	1.5	Yes	4258	AD	3	Yes	2010	MCI-u	3	Yes
500	NC	1.5	Yes	4280	AD	3	Yes	2018	MCI-u	3	Yes
502	NC	1.5	Yes	4282	AD	3	Yes	2022	MCI-u	3	Yes
506	NC	1.5	Yes	4307	AD	3	Yes	2027	MCI-u	3	Yes
516	NC	1.5	No	4353	AD	3	Yes	2036	MCI-u	3	Yes
519	NC	1.5	No	4373	AD	3	Yes	2037	MCI-u	3	Yes
522	NC	1.5	Yes	4494	AD	3	Yes	2042	MCI-u	3	Yes
526	NC	1.5	Yes	4500	AD	3	Yes	2043	MCI-u	3	Yes
534	NC	1.5	Yes	4526	AD	3	Yes	2045	MCI-u	3	Yes
538	NC	1.5	No	4546	AD	3	Yes	2055	MCI-u	3	Yes
545	NC	1.5	No	4549	AD	3	Yes	2058	MCI-u	3	Yes
548	NC	1.5	No	4589	AD	3	Yes	2060	MCI-u	3	Yes
555	NC	1.5	Yes	4591	AD	3	Yes	2061	MCI-u	3	Yes
558	NC	1.5	No	4641	AD	3	Yes	2063	MCI-u	3	Yes
559	NC	1.5	No	4657	AD	3	Yes	2068	MCI-u	3	Yes
575	NC	1.5	Yes	4660	AD	3	Yes	2072	MCI-u	3	Yes
576	NC	1.5	Yes	4672	AD	3	Yes	2073	MCI-u	3	Yes
578	NC	1.5	No	4686	AD	3	Yes	2074	MCI-u	3	Yes

601	NC	1.5	No	4692	AD	3	Yes	2077	MCI-u	3	No
602	NC	3	No	4696	AD	3	Yes	2079	MCI-u	3	Yes
605	NC	3	No	4707	AD	3	Yes	2083	MCI-u	3	Yes
610	NC	1.5	Yes	4728	AD	3	Yes	2087	MCI-u	3	Yes
618	NC	1.5	Yes	4730	AD	3	Yes	2093	MCI-u	3	Yes
622	NC	3	No	4733	AD	3	Yes	2099	MCI-u	3	Yes
640	NC	2.9	No	4755	AD	3	Yes	2100	MCI-u	3	Yes
643	NC	1.5	No	4756	AD	3	Yes	2106	MCI-u	3	Yes
647	NC	1.5	Yes	4770	AD	3	Yes	2109	MCI-u	3	Yes
657	NC	1.5	Yes	4774	AD	3	Yes	2116	MCI-u	3	Yes
672	NC	1.5	Yes	4802	AD	3	Yes	2119	MCI-u	3	Yes
677	NC	3	No	4827	AD	3	Yes	2121	MCI-u	3	Yes
680	NC	1.5	Yes	4845	AD	3	Yes	2123	MCI-u	3	Yes
681	NC	1.5	No	4853	AD	3	Yes	2125	MCI-u	3	Yes
684	NC	1.5	No	4863	AD	3	Yes	2133	MCI-u	3	Yes
685	NC	1.5	No	4867	AD	3	Yes	2138	MCI-u	3	Yes
686	NC	1.5	Yes	4892	AD	3	Yes	2142	MCI-u	3	Yes
692	NC	1.5	No	4894	AD	3	Yes	2146	MCI-u	3	Yes
717	NC	1.5	No	4905	AD	3	Yes	2148	MCI-u	3	Yes
726	NC	1.5	No	4906	AD	3	Yes	2150	MCI-u	3	Yes
731	NC	1.5	Yes	4910	AD	3	Yes	2151	MCI-u	3	Yes
734	NC	1.5	Yes	4938	AD	3	No	2153	MCI-u	3	Yes
741	NC	1.5	Yes	4940	AD	3	Yes	2155	MCI-u	3	Yes
751	NC	1.5	Yes	4949	AD	3	Yes	2164	MCI-u	3	Yes
761	NC	1.5	No	4954	AD	3	Yes	2168	MCI-u	3	Yes
767	NC	1.5	No	4962	AD	3	Yes	2171	MCI-u	3	Yes
768	NC	1.5	Yes	4971	AD	3	Yes	2180	MCI-u	3	Yes
779	NC	1.5	Yes	5012	AD	3	Yes	2182	MCI-u	3	Yes
810	NC	1.5	No	5015	AD	3	Yes	2183	MCI-u	3	Yes
813	NC	1.5	Yes	5018	AD	3	Yes	2184	MCI-u	3	Yes
818	NC	1.5	Yes	5019	AD	3	Yes	2185	MCI-u	3	Yes
842	NC	1.5	Yes	5028	AD	3	Yes	2187	MCI-u	3	Yes
843	NC	1.5	Yes	4	MCI-s	1.5	No	2190	MCI-u	3	Yes
862	NC	1.5	Yes	51	MCI-s	1.5	Yes	2191	MCI-u	3	Yes

863	NC	1.5	Yes	107	MCI-s	1.5	No	2193	MCI-u	3	Yes
866	NC	1.5	Yes	116	MCI-s	1.5	No	2194	MCI-u	3	Yes
876	NC	1.5	No	150	MCI-s	1.5	Yes	2195	MCI-u	3	Yes
883	NC	1.5	Yes	160	MCI-s	1.5	Yes	2196	MCI-u	3	Yes
886	NC	1.5	No	176	MCI-s	1.5	No	2200	MCI-u	3	Yes
896	NC	1.5	No	178	MCI-s	1.5	Yes	2205	MCI-u	3	Yes
898	NC	1.5	Yes	200	MCI-s	1.5	Yes	2208	MCI-u	3	Yes
899	NC	1.5	No	225	MCI-s	1.5	Yes	2210	MCI-u	3	Yes
920	NC	2.9	No	273	MCI-s	1.5	No	2213	MCI-u	3	Yes
923	NC	1.5	No	276	MCI-s	1.5	No	2219	MCI-u	3	Yes
926	NC	3	No	285	MCI-s	1.5	Yes	2220	MCI-u	3	Yes
931	NC	1.5	No	291	MCI-s	1.5	Yes	2225	MCI-u	3	Yes
934	NC	1.5	Yes	292	MCI-s	1.5	Yes	2233	MCI-u	3	Yes
951	NC	1.5	No	307	MCI-s	2.9	No	2234	MCI-u	3	Yes
963	NC	3	No	324	MCI-s	3	No	2238	MCI-u	3	Yes
967	NC	1.5	Yes	351	MCI-s	1.5	No	2239	MCI-u	3	Yes
969	NC	1.5	No	361	MCI-s	1.5	Yes	2240	MCI-u	3	Yes
972	NC	1.5	Yes	376	MCI-s	2.9	No	2245	MCI-u	3	Yes
1002	NC	1.5	Yes	378	MCI-s	1.5	Yes	2247	MCI-u	3	Yes
1013	NC	1.5	Yes	389	MCI-s	1.5	Yes	2248	MCI-u	3	Yes
1014	NC	1.5	No	407	MCI-s	1.5	Yes	2264	MCI-u	3	Yes
1016	NC	2.9	No	408	MCI-s	1.5	Yes	2274	MCI-u	3	Yes
1023	NC	1.5	Yes	445	MCI-s	1.5	No	2284	MCI-u	3	Yes
1063	NC	1.5	Yes	448	MCI-s	3	No	2301	MCI-u	3	Yes
1086	NC	2.9	No	449	MCI-s	1.5	No	2304	MCI-u	3	Yes
1094	NC	1.5	No	464	MCI-s	1.5	Yes	2307	MCI-u	3	Yes
1098	NC	2.9	No	469	MCI-s	2.9	No	2315	MCI-u	3	Yes
1123	NC	3	No	481	MCI-s	1.5	Yes	2316	MCI-u	3	Yes
1169	NC	3	No	501	MCI-s	2.9	No	2324	MCI-u	3	Yes
1190	NC	3	No	505	MCI-s	1.5	No	2332	MCI-u	3	Yes
1194	NC	1.5	Yes	546	MCI-s	1.5	Yes	2333	MCI-u	3	Yes
1195	NC	1.5	Yes	552	MCI-s	1.5	Yes	2336	MCI-u	3	Yes
1197	NC	1.5	Yes	557	MCI-s	1.5	No	2347	MCI-u	3	Yes
1202	NC	1.5	Yes	590	MCI-s	1.5	Yes	2357	MCI-u	3	Yes

1203	NC	1.5	Yes	607	MCI-s	3	No	2360	MCI-u	3	Yes
1206	NC	3	No	608	MCI-s	1.5	Yes	2363	MCI-u	3	Yes
1222	NC	3	No	613	MCI-s	3	No	2367	MCI-u	3	Yes
1232	NC	2.9	No	621	MCI-s	1.5	Yes	2373	MCI-u	3	Yes
1242	NC	2.9	No	626	MCI-s	1.5	Yes	2374	MCI-u	3	Yes
1249	NC	2.9	No	641	MCI-s	1.5	Yes	2378	MCI-u	3	Yes
1250	NC	3	No	644	MCI-s	1.5	No	2379	MCI-u	3	Yes
1251	NC	3	No	656	MCI-s	1.5	Yes	2380	MCI-u	3	Yes
1256	NC	3	No	671	MCI-s	1.5	No	2381	MCI-u	3	Yes
1261	NC	1.5	No	673	MCI-s	1.5	Yes	2389	MCI-u	3	Yes
1267	NC	3	No	679	MCI-s	1.5	Yes	2392	MCI-u	3	No
1276	NC	2.9	No	698	MCI-s	1.5	Yes	2394	MCI-u	3	Yes
1280	NC	1.5	No	709	MCI-s	1.5	Yes	2395	MCI-u	3	Yes
1288	NC	3	No	715	MCI-s	1.5	Yes	2405	MCI-u	3	Yes
1301	NC	3	No	746	MCI-s	1.5	Yes	2407	MCI-u	3	Yes
2201	NC	3	Yes	771	MCI-s	2.9	No	4007	MCI-u	3	Yes
4010	NC	3	Yes	782	MCI-s	1.5	No	4029	MCI-u	3	Yes
4018	NC	3	Yes	792	MCI-s	2.9	No	4030	MCI-u	3	Yes
4020	NC	3	Yes	851	MCI-s	1.5	No	4034	MCI-u	3	Yes
4021	NC	3	Yes	867	MCI-s	1.5	No	4036	MCI-u	3	Yes
4026	NC	3	Yes	871	MCI-s	1.5	No	4051	MCI-u	3	Yes
4028	NC	3	Yes	908	MCI-s	1.5	No	4053	MCI-u	3	Yes
4032	NC	3	Yes	912	MCI-s	2.9	No	4054	MCI-u	3	Yes
4037	NC	3	Yes	919	MCI-s	1.5	Yes	4058	MCI-u	3	Yes
4041	NC	3	Yes	950	MCI-s	1.5	Yes	4059	MCI-u	3	Yes
4043	NC	3	Yes	989	MCI-s	1.5	No	4061	MCI-u	3	Yes
4050	NC	3	Yes	994	MCI-s	1.5	Yes	4063	MCI-u	3	Yes
4060	NC	3	Yes	1030	MCI-s	1.5	Yes	4072	MCI-u	3	Yes
4066	NC	3	Yes	1031	MCI-s	2.9	No	4073	MCI-u	3	Yes
4075	NC	3	Yes	1032	MCI-s	1.5	Yes	4077	MCI-u	3	Yes
4076	NC	3	Yes	1034	MCI-s	1.5	Yes	4079	MCI-u	3	Yes
4080	NC	3	Yes	1040	MCI-s	1.5	No	4115	MCI-u	3	Yes
4081	NC	3	Yes	1046	MCI-s	3	No	4122	MCI-u	3	Yes
4082	NC	3	Yes	1052	MCI-s	1.5	No	4127	MCI-u	3	Yes

4084	NC	3	Yes	1072	MCI-s	3	No	4128	MCI-u	3	Yes
4090	NC	3	Yes	1078	MCI-s	1.5	Yes	4133	MCI-u	3	Yes
4092	NC	3	Yes	1088	MCI-s	2.9	No	4143	MCI-u	3	Yes
4093	NC	3	Yes	1097	MCI-s	1.5	No	4146	MCI-u	3	Yes
4100	NC	3	Yes	1106	MCI-s	1.5	Yes	4149	MCI-u	3	Yes
4104	NC	3	Yes	1118	MCI-s	1.5	Yes	4157	MCI-u	3	Yes
4119	NC	3	Yes	1122	MCI-s	1.5	Yes	4159	MCI-u	3	Yes
4120	NC	3	Yes	1131	MCI-s	3	No	4160	MCI-u	3	Yes
4121	NC	3	Yes	1140	MCI-s	1.5	No	4162	MCI-u	3	Yes
4125	NC	3	Yes	1149	MCI-s	3	No	4168	MCI-u	3	Yes
4139	NC	3	Yes	1182	MCI-s	1.5	No	4169	MCI-u	3	Yes
4148	NC	3	Yes	1186	MCI-s	1.5	Yes	4170	MCI-u	3	Yes
4150	NC	3	Yes	1187	MCI-s	1.5	No	4171	MCI-u	3	Yes
4151	NC	3	Yes	1199	MCI-s	1.5	Yes	4175	MCI-u	3	Yes
4155	NC	3	Yes	1210	MCI-s	1.5	Yes	4184	MCI-u	3	Yes
4158	NC	3	Yes	1227	MCI-s	1.5	No	4187	MCI-u	3	Yes
4164	NC	3	Yes	1246	MCI-s	1.5	Yes	4188	MCI-u	3	Yes
4173	NC	3	Yes	1255	MCI-s	1.5	No	4194	MCI-u	3	Yes
4174	NC	3	Yes	1260	MCI-s	1.5	Yes	4197	MCI-u	3	Yes
4177	NC	3	Yes	1269	MCI-s	1.5	No	4199	MCI-u	3	Yes
4179	NC	3	Yes	1284	MCI-s	1.5	No	4205	MCI-u	3	Yes
4198	NC	3	Yes	1300	MCI-s	1.5	No	4206	MCI-u	3	Yes
4200	NC	3	Yes	1314	MCI-s	1.5	Yes	4210	MCI-u	3	Yes
4208	NC	3	Yes	1338	MCI-s	3	No	4212	MCI-u	3	Yes
4213	NC	3	Yes	1346	MCI-s	1.5	Yes	4214	MCI-u	3	Yes
4218	NC	3	Yes	1357	MCI-s	1.5	Yes	4216	MCI-u	3	Yes
4222	NC	3	Yes	1378	MCI-s	1.5	Yes	4219	MCI-u	3	Yes
4225	NC	3	Yes	1406	MCI-s	1.5	Yes	4220	MCI-u	3	Yes
4255	NC	3	Yes	1414	MCI-s	1.5	Yes	4226	MCI-u	3	Yes
4257	NC	3	No	1418	MCI-s	1.5	Yes	4229	MCI-u	3	Yes
4262	NC	3	Yes	1419	MCI-s	1.5	Yes	4232	MCI-u	3	Yes
4266	NC	3	Yes	1421	MCI-s	1.5	Yes	4235	MCI-u	3	Yes
4269	NC	3	Yes	2002	MCI-s	3	Yes	4237	MCI-u	3	Yes
4270	NC	3	Yes	2003	MCI-s	3	Yes	4241	MCI-u	3	Yes

4275	NC	3	Yes	2011	MCI-s	3	Yes	4250	MCI-u	3	Yes
4276	NC	3	Yes	2031	MCI-s	3	Yes	4251	MCI-u	3	Yes
4277	NC	3	Yes	2057	MCI-s	3	Yes	4256	MCI-u	3	Yes
4278	NC	3	Yes	2070	MCI-s	3	Yes	4259	MCI-u	3	Yes
4279	NC	3	Yes	2199	MCI-s	3	Yes	4263	MCI-u	3	Yes
4288	NC	3	Yes	2237	MCI-s	3	Yes	4268	MCI-u	3	Yes
4291	NC	3	Yes	2278	MCI-s	3	Yes	4271	MCI-u	3	Yes
4292	NC	3	Yes	4067	MCI-s	3	Yes	4281	MCI-u	3	Yes
4308	NC	3	Yes	4134	MCI-s	3	Yes	4285	MCI-u	3	Yes
4313	NC	3	Yes	4186	MCI-s	3	Yes	4287	MCI-u	3	Yes
4320	NC	3	Yes	4217	MCI-s	3	Yes	4293	MCI-u	3	Yes
4335	NC	3	Yes	4260	MCI-s	3	Yes	4297	MCI-u	3	Yes
4337	NC	3	Yes	4274	MCI-s	3	Yes	4300	MCI-u	3	Yes
4339	NC	3	Yes	4332	MCI-s	3	Yes	4301	MCI-u	3	Yes
4340	NC	3	Yes	4403	MCI-s	3	Yes	4302	MCI-u	3	Yes
4343	NC	3	Yes	4408	MCI-s	3	Yes	4303	MCI-u	3	Yes
4345	NC	3	Yes	4431	MCI-s	3	Yes	4309	MCI-u	3	Yes
4348	NC	3	Yes	4517	MCI-s	3	Yes	4310	MCI-u	3	Yes
4349	NC	3	Yes	4524	MCI-s	3	Yes	4311	MCI-u	3	Yes
4350	NC	3	Yes	4556	MCI-s	3	Yes	4312	MCI-u	3	Yes
4352	NC	3	Yes	4594	MCI-s	3	Yes	4324	MCI-u	3	Yes
4357	NC	3	Yes	4601	MCI-s	3	Yes	4328	MCI-u	3	Yes
4367	NC	3	Yes	4798	MCI-s	3	No	4331	MCI-u	3	Yes
4369	NC	3	Yes	4871	MCI-s	3	Yes	4346	MCI-u	3	Yes
4371	NC	3	Yes	4883	MCI-s	3	Yes	4351	MCI-u	3	Yes
4372	NC	3	Yes	4907	MCI-s	3	Yes	4354	MCI-u	3	Yes
4382	NC	3	Yes	4944	MCI-s	3	No	4356	MCI-u	3	Yes
4384	NC	3	Yes	4945	MCI-s	3	No	4359	MCI-u	3	Yes
4385	NC	3	Yes	5004	MCI-s	3	No	4360	MCI-u	3	Yes
4386	NC	3	Yes	42	MCI-c	1.5	No	4363	MCI-u	3	Yes
4387	NC	3	Yes	45	MCI-c	1.5	No	4377	MCI-u	3	Yes
4388	NC	3	Yes	54	MCI-c	1.5	Yes	4381	MCI-u	3	Yes
4389	NC	3	Yes	77	MCI-c	1.5	No	4383	MCI-u	3	Yes
4391	NC	3	Yes	98	MCI-c	1.5	No	4390	MCI-u	3	Yes

4393	NC	3	Yes	101	MCI-c	1.5	Yes	4392	MCI-u	3	Yes
4396	NC	3	Yes	108	MCI-c	1.5	No	4394	MCI-u	3	Yes
4399	NC	3	Yes	111	MCI-c	1.5	No	4395	MCI-u	3	Yes
4400	NC	3	Yes	126	MCI-c	1.5	No	4404	MCI-u	3	Yes
4401	NC	3	Yes	128	MCI-c	1.5	Yes	4405	MCI-u	3	Yes
4410	NC	3	Yes	141	MCI-c	1.5	Yes	4406	MCI-u	3	Yes
4421	NC	3	Yes	161	MCI-c	1.5	Yes	4415	MCI-u	3	Yes
4422	NC	3	Yes	179	MCI-c	1.5	No	4417	MCI-u	3	Yes
4424	NC	3	Yes	182	MCI-c	1.5	No	4419	MCI-u	3	Yes
4427	NC	3	Yes	195	MCI-c	1.5	No	4420	MCI-u	3	Yes
4428	NC	3	Yes	204	MCI-c	1.5	Yes	4423	MCI-u	3	Yes
4429	NC	3	Yes	227	MCI-c	1.5	Yes	4426	MCI-u	3	Yes
4441	NC	3	Yes	231	MCI-c	1.5	Yes	4430	MCI-u	3	Yes
4442	NC	3	Yes	240	MCI-c	1.5	Yes	4434	MCI-u	3	Yes
4446	NC	3	Yes	243	MCI-c	1.5	No	4438	MCI-u	3	Yes
4448	NC	3	Yes	249	MCI-c	1.5	No	4443	MCI-u	3	Yes
4449	NC	3	Yes	256	MCI-c	1.5	Yes	4444	MCI-u	3	Yes
4453	NC	3	Yes	258	MCI-c	1.5	Yes	4447	MCI-u	3	Yes
4464	NC	3	Yes	269	MCI-c	1.5	No	4455	MCI-u	3	Yes
4466	NC	3	Yes	289	MCI-c	1.5	Yes	4456	MCI-u	3	Yes
4469	NC	3	Yes	294	MCI-c	1.5	Yes	4462	MCI-u	3	Yes
4474	NC	3	Yes	314	MCI-c	1.5	Yes	4463	MCI-u	3	Yes
4482	NC	3	Yes	325	MCI-c	1.5	Yes	4465	MCI-u	3	Yes
4483	NC	3	Yes	331	MCI-c	2.9	No	4467	MCI-u	3	Yes
4485	NC	3	Yes	336	MCI-c	1.5	No	4468	MCI-u	3	Yes
4488	NC	3	Yes	344	MCI-c	1.5	Yes	4473	MCI-u	3	Yes
4491	NC	3	Yes	362	MCI-c	1.5	Yes	4475	MCI-u	3	Yes
4496	NC	3	Yes	388	MCI-c	2.9	No	4476	MCI-u	3	Yes
4499	NC	3	Yes	434	MCI-c	1.5	No	4480	MCI-u	3	Yes
4503	NC	3	Yes	461	MCI-c	1.5	Yes	4489	MCI-u	3	Yes
4505	NC	3	Yes	507	MCI-c	1.5	No	4498	MCI-u	3	Yes
4516	NC	3	Yes	511	MCI-c	1.5	Yes	4510	MCI-u	3	Yes
4545	NC	3	Yes	513	MCI-c	1.5	Yes	4514	MCI-u	3	Yes
4552	NC	3	Yes	518	MCI-c	1.5	No	4521	MCI-u	3	Yes

4555	NC	3	Yes	539	MCI-c	1.5	No	4531	MCI-u	3	Yes
4558	NC	3	Yes	549	MCI-c	1.5	Yes	4536	MCI-u	3	Yes
4559	NC	3	Yes	567	MCI-c	1.5	Yes	4538	MCI-u	3	Yes
4560	NC	3	Yes	568	MCI-c	3	No	4539	MCI-u	3	Yes
4566	NC	3	Yes	604	MCI-c	2.9	No	4547	MCI-u	3	Yes
4576	NC	3	Yes	611	MCI-c	1.5	No	4548	MCI-u	3	Yes
4577	NC	3	Yes	625	MCI-c	3	No	4553	MCI-u	3	Yes
4578	NC	3	Yes	631	MCI-c	1.5	No	4557	MCI-u	3	Yes
4579	NC	3	Yes	638	MCI-c	2.9	No	4562	MCI-u	3	Yes
4580	NC	3	Yes	649	MCI-c	2.9	No	4565	MCI-u	3	Yes
4585	NC	3	Yes	675	MCI-c	1.5	Yes	4571	MCI-u	3	Yes
4587	NC	3	Yes	695	MCI-c	1.5	Yes	4582	MCI-u	3	Yes
4604	NC	3	Yes	708	MCI-c	1.5	Yes	4584	MCI-u	3	Yes
4607	NC	3	Yes	723	MCI-c	1.5	Yes	4590	MCI-u	3	Yes
4609	NC	3	Yes	725	MCI-c	2.9	No	4596	MCI-u	3	Yes
4612	NC	3	Yes	727	MCI-c	2.9	No	4597	MCI-u	3	Yes
4620	NC	3	Yes	729	MCI-c	1.5	No	4605	MCI-u	3	Yes
4632	NC	3	Yes	750	MCI-c	1.5	No	4611	MCI-u	3	Yes
4637	NC	3	Yes	769	MCI-c	3	No	4613	MCI-u	3	Yes
4638	NC	3	Yes	834	MCI-c	1.5	No	4614	MCI-u	3	Yes
4643	NC	3	Yes	835	MCI-c	2.9	No	4621	MCI-u	3	Yes
4644	NC	3	Yes	839	MCI-c	1.5	No	4623	MCI-u	3	Yes
4645	NC	3	Yes	856	MCI-c	1.5	No	4624	MCI-u	3	Yes
4649	NC	3	Yes	860	MCI-c	1.5	Yes	4626	MCI-u	3	Yes
4652	NC	3	Yes	865	MCI-c	1.5	Yes	4636	MCI-u	3	Yes
4739	NC	3	Yes	869	MCI-c	1.5	No	4646	MCI-u	3	Yes
4762	NC	3	Yes	873	MCI-c	1.5	No	4654	MCI-u	3	Yes
4795	NC	3	Yes	874	MCI-c	1.5	No	4659	MCI-u	3	Yes
4832	NC	3	Yes	887	MCI-c	1.5	No	4668	MCI-u	3	Yes
4835	NC	3	Yes	906	MCI-c	1.5	Yes	4674	MCI-u	3	Yes
4843	NC	3	Yes	909	MCI-c	1.5	Yes	4675	MCI-u	3	Yes
4855	NC	3	Yes	913	MCI-c	2.9	No	4678	MCI-u	3	Yes
4872	NC	3	Yes	915	MCI-c	1.5	No	4679	MCI-u	3	Yes
4878	NC	3	Yes	922	MCI-c	2.9	No	4711	MCI-u	3	Yes

4900	NC	3	Yes	941	MCI-c	1.5	Yes	4713	MCI-u	3	Yes
4921	NC	3	Yes	947	MCI-c	1.5	Yes	4715	MCI-u	3	Yes
4951	NC	3	Yes	952	MCI-c	1.5	No	4722	MCI-u	3	Yes
4952	NC	3	Yes	954	MCI-c	1.5	No	4723	MCI-u	3	Yes
5040	NC	3	Yes	973	MCI-c	1.5	Yes	4736	MCI-u	3	Yes
3	AD	1.5	Yes	976	MCI-c	1.5	Yes	4742	MCI-u	3	Yes
7	AD	1.5	No	982	MCI-c	1.5	No	4743	MCI-u	3	Yes
10	AD	1.5	Yes	987	MCI-c	1.5	Yes	4746	MCI-u	3	Yes
29	AD	1.5	No	1004	MCI-c	1.5	No	4750	MCI-u	3	Yes
53	AD	1.5	Yes	1007	MCI-c	1.5	Yes	4757	MCI-u	3	Yes
76	AD	1.5	No	1010	MCI-c	1.5	Yes	4764	MCI-u	3	Yes
83	AD	1.5	No	1015	MCI-c	1.5	No	4765	MCI-u	3	Yes
84	AD	1.5	No	1043	MCI-c	1.5	Yes	4769	MCI-u	3	Yes
88	AD	1.5	No	1054	MCI-c	1.5	No	4777	MCI-u	3	Yes
91	AD	1.5	No	1057	MCI-c	1.5	Yes	4780	MCI-u	3	Yes
93	AD	1.5	No	1070	MCI-c	1.5	No	4782	MCI-u	3	Yes
94	AD	1.5	No	1077	MCI-c	1.5	Yes	4791	MCI-u	3	Yes
110	AD	1.5	No	1117	MCI-c	3	No	4799	MCI-u	3	Yes
129	AD	1.5	No	1121	MCI-c	3	No	4803	MCI-u	3	Yes
139	AD	2.9	No	1126	MCI-c	3	No	4804	MCI-u	3	Yes
162	AD	1.5	No	1130	MCI-c	1.5	Yes	4805	MCI-u	3	Yes
183	AD	1.5	Yes	1135	MCI-c	1.5	Yes	4806	MCI-u	3	Yes
194	AD	1.5	No	1138	MCI-c	3	No	4813	MCI-u	3	Yes
213	AD	1.5	Yes	1148	MCI-c	2.9	No	4814	MCI-u	3	Yes
221	AD	1.5	Yes	1217	MCI-c	1.5	Yes	4815	MCI-u	3	Yes
228	AD	1.5	Yes	1240	MCI-c	1.5	Yes	4816	MCI-u	3	Yes
266	AD	1.5	Yes	1243	MCI-c	1.5	Yes	4817	MCI-u	3	Yes
299	AD	1.5	No	1244	MCI-c	1.5	No	4823	MCI-u	3	Yes
300	AD	1.5	No	1247	MCI-c	3	No	4825	MCI-u	3	Yes
310	AD	1.5	No	1265	MCI-c	1.5	Yes	4838	MCI-u	3	No
321	AD	1.5	Yes	1271	MCI-c	1.5	No	4842	MCI-u	3	Yes
332	AD	2.9	No	1282	MCI-c	1.5	Yes	4849	MCI-u	3	Yes
341	AD	1.5	Yes	1299	MCI-c	1.5	Yes	4852	MCI-u	3	Yes
343	AD	1.5	Yes	1311	MCI-c	1.5	Yes	4869	MCI-u	3	Yes

366	AD	1.5	No	1315	MCI-c	1.5	Yes	4873	MCI-u	3	Yes
370	AD	1.5	Yes	1331	MCI-c	3	No	4876	MCI-u	3	Yes
374	AD	1.5	Yes	1351	MCI-c	1.5	Yes	4885	MCI-u	3	Yes
392	AD	2.9	No	1363	MCI-c	1.5	No	4889	MCI-u	3	Yes
404	AD	2.9	No	1389	MCI-c	3	No	4891	MCI-u	3	Yes
426	AD	1.5	No	1394	MCI-c	1.5	Yes	4893	MCI-u	3	Yes
431	AD	1.5	Yes	1398	MCI-c	1.5	Yes	4898	MCI-u	3	Yes
438	AD	1.5	Yes	1407	MCI-c	1.5	Yes	4902	MCI-u	3	Yes
457	AD	3	No	1412	MCI-c	1.5	Yes	4904	MCI-u	3	Yes
470	AD	1.5	Yes	1423	MCI-c	1.5	Yes	4909	MCI-u	3	Yes
487	AD	2.9	No	1427	MCI-c	1.5	Yes	4917	MCI-u	3	Yes
492	AD	1.5	Yes	2047	MCI-c	3	Yes	4919	MCI-u	3	Yes
517	AD	1.5	No	2216	MCI-c	3	Yes	4920	MCI-u	3	No
528	AD	1.5	No	2398	MCI-c	3	Yes	4925	MCI-u	3	Yes
535	AD	1.5	Yes	4005	MCI-c	3	Yes	4926	MCI-u	3	Yes
543	AD	1.5	Yes	4035	MCI-c	3	Yes	4929	MCI-u	3	Yes
547	AD	1.5	Yes	4042	MCI-c	3	Yes	4936	MCI-u	3	Yes
554	AD	1.5	Yes	4057	MCI-c	3	Yes	4941	MCI-u	3	Yes
577	AD	1.5	Yes	4094	MCI-c	3	Yes	4955	MCI-u	3	Yes
592	AD	1.5	No	4096	MCI-c	3	Yes	4960	MCI-u	3	Yes
606	AD	3	No	4102	MCI-c	3	Yes	4974	MCI-u	3	Yes
619	AD	1.5	No	4114	MCI-c	3	Yes	4976	MCI-u	3	Yes
642	AD	1.5	Yes	4131	MCI-c	3	Yes	4985	MCI-u	3	Yes
653	AD	1.5	Yes	4167	MCI-c	3	Yes	4989	MCI-u	3	Yes
733	AD	2.9	No	4189	MCI-c	3	Yes	5014	MCI-u	3	Yes
740	AD	1.5	Yes	4203	MCI-c	3	Yes				
753	AD	2.9	No	4240	MCI-c	3	Yes				
754	AD	1.5	Yes	4366	MCI-c	3	Yes				
759	AD	1.5	No	4402	MCI-c	3	Yes				

B.4 Subjects in experiment in chapter 8

The subjects used in the experiments described in chapter 8 are listed in table B.5. For each subject, as well as the disease group, the scanner field strength is listed. The small number

of scans done in 2.9T scanners are grouped with the 3T scans when examining the effects of different scanner field strengths.

Table B.5: List of subjects used in the ARK experiments in chapter

8. ID = ADNI roster ID number, Status = disease status of subject,

T = field strength of structural MRI in Tesla.

ID	Status	T	ID	Status	T	ID	Status	T	ID	Status	T
2	NC	1.5	4179	NC	3	1056	AD	1.5	1187	MCI-s	1.5
5	NC	1.5	4198	NC	3	1059	AD	1.5	1199	MCI-s	1.5
8	NC	1.5	4200	NC	3	1079	AD	1.5	1210	MCI-s	1.5
14	NC	1.5	4208	NC	3	1081	AD	2.9	1227	MCI-s	1.5
16	NC	1.5	4213	NC	3	1082	AD	2.9	1246	MCI-s	1.5
19	NC	1.5	4218	NC	3	1083	AD	2.9	1255	MCI-s	1.5
21	NC	1.5	4222	NC	3	1090	AD	1.5	1260	MCI-s	1.5
22	NC	1.5	4225	NC	3	1095	AD	1.5	1268	MCI-s	1.5
23	NC	1.5	4234	NC	3	1101	AD	3	1269	MCI-s	1.5
31	NC	2.9	4254	NC	3	1102	AD	1.5	1284	MCI-s	1.5
40	NC	1.5	4255	NC	3	1109	AD	1.5	1300	MCI-s	1.5
48	NC	1.5	4257	NC	3	1137	AD	1.5	1314	MCI-s	1.5
56	NC	1.5	4262	NC	3	1144	AD	1.5	1318	MCI-s	1.5
58	NC	2.9	4266	NC	3	1152	AD	1.5	1338	MCI-s	3
59	NC	1.5	4269	NC	3	1157	AD	1.5	1340	MCI-s	3
61	NC	2.9	4270	NC	3	1161	AD	1.5	1346	MCI-s	1.5
66	NC	1.5	4275	NC	3	1164	AD	1.5	1357	MCI-s	1.5
67	NC	1.5	4276	NC	3	1170	AD	2.9	1378	MCI-s	1.5
68	NC	1.5	4277	NC	3	1171	AD	1.5	1384	MCI-s	1.5
70	NC	1.5	4278	NC	3	1184	AD	1.5	1406	MCI-s	1.5
72	NC	1.5	4279	NC	3	1185	AD	3	1414	MCI-s	1.5
74	NC	1.5	4288	NC	3	1192	AD	1.5	1417	MCI-s	1.5
81	NC	1.5	4290	NC	3	1201	AD	1.5	1418	MCI-s	1.5
86	NC	1.5	4291	NC	3	1205	AD	1.5	1419	MCI-s	1.5
89	NC	1.5	4292	NC	3	1209	AD	3	1421	MCI-s	1.5
90	NC	1.5	4308	NC	3	1221	AD	1.5	2002	MCI-s	3

96	NC	1.5	4313	NC	3	1248	AD	1.5	2003	MCI-s	3
97	NC	1.5	4320	NC	3	1253	AD	3	2011	MCI-s	3
106	NC	1.5	4335	NC	3	1254	AD	1.5	2031	MCI-s	3
113	NC	1.5	4337	NC	3	1257	AD	1.5	2057	MCI-s	3
118	NC	1.5	4339	NC	3	1262	AD	3	2070	MCI-s	3
120	NC	1.5	4340	NC	3	1263	AD	1.5	2199	MCI-s	3
123	NC	1.5	4343	NC	3	1281	AD	1.5	2237	MCI-s	3
125	NC	1.5	4345	NC	3	1283	AD	1.5	2278	MCI-s	3
130	NC	1.5	4348	NC	3	1285	AD	1.5	4067	MCI-s	3
156	NC	1.5	4349	NC	3	1289	AD	3	4134	MCI-s	3
166	NC	1.5	4350	NC	3	1290	AD	1.5	4186	MCI-s	3
171	NC	1.5	4352	NC	3	1296	AD	1.5	4217	MCI-s	3
172	NC	1.5	4357	NC	3	1304	AD	3	4260	MCI-s	3
173	NC	1.5	4367	NC	3	1307	AD	1.5	4274	MCI-s	3
177	NC	1.5	4369	NC	3	1308	AD	1.5	4332	MCI-s	3
184	NC	1.5	4371	NC	3	1334	AD	1.5	4403	MCI-s	3
186	NC	1.5	4372	NC	3	1337	AD	1.5	4408	MCI-s	3
196	NC	1.5	4376	NC	3	1339	AD	1.5	4431	MCI-s	3
223	NC	1.5	4382	NC	3	1341	AD	1.5	4517	MCI-s	3
229	NC	1.5	4384	NC	3	1354	AD	1.5	4524	MCI-s	3
230	NC	1.5	4385	NC	3	1368	AD	1.5	4556	MCI-s	3
232	NC	1.5	4386	NC	3	1371	AD	1.5	4594	MCI-s	3
245	NC	1.5	4387	NC	3	1373	AD	1.5	4601	MCI-s	3
257	NC	1.5	4388	NC	3	1377	AD	1.5	4694	MCI-s	3
259	NC	1.5	4389	NC	3	1379	AD	1.5	4745	MCI-s	3
260	NC	3	4391	NC	3	1382	AD	1.5	4798	MCI-s	3
262	NC	1.5	4393	NC	3	1385	AD	2.9	4871	MCI-s	3
272	NC	1.5	4396	NC	3	1391	AD	1.5	4883	MCI-s	3
283	NC	1.5	4399	NC	3	1397	AD	1.5	4907	MCI-s	3
295	NC	1.5	4400	NC	3	1402	AD	1.5	4944	MCI-s	3
298	NC	1.5	4401	NC	3	1409	AD	1.5	4945	MCI-s	3
301	NC	1.5	4410	NC	3	1430	AD	1.5	5004	MCI-s	3
303	NC	2.9	4421	NC	3	1435	AD	1.5	30	MCI-c	2.9
311	NC	1.5	4422	NC	3	4009	AD	3	41	MCI-c	1.5

312	NC	1.5	4424	NC	3	4024	AD	3	42	MCI-c	1.5
319	NC	1.5	4427	NC	3	4039	AD	3	45	MCI-c	1.5
327	NC	1.5	4428	NC	3	4136	AD	3	50	MCI-c	1.5
337	NC	1.5	4429	NC	3	4152	AD	3	54	MCI-c	1.5
352	NC	1.5	4433	NC	3	4153	AD	3	77	MCI-c	1.5
359	NC	1.5	4441	NC	3	4172	AD	3	98	MCI-c	1.5
360	NC	1.5	4442	NC	3	4192	AD	3	101	MCI-c	1.5
363	NC	1.5	4446	NC	3	4195	AD	3	108	MCI-c	1.5
382	NC	2.9	4448	NC	3	4201	AD	3	111	MCI-c	1.5
386	NC	1.5	4449	NC	3	4209	AD	3	126	MCI-c	1.5
403	NC	2.9	4453	NC	3	4211	AD	3	128	MCI-c	1.5
405	NC	3	4464	NC	3	4215	AD	3	141	MCI-c	1.5
413	NC	1.5	4466	NC	3	4223	AD	3	161	MCI-c	1.5
416	NC	1.5	4469	NC	3	4252	AD	3	179	MCI-c	1.5
419	NC	1.5	4474	NC	3	4258	AD	3	182	MCI-c	1.5
436	NC	1.5	4482	NC	3	4280	AD	3	195	MCI-c	1.5
441	NC	3	4483	NC	3	4282	AD	3	204	MCI-c	1.5
454	NC	1.5	4485	NC	3	4307	AD	3	217	MCI-c	1.5
459	NC	1.5	4488	NC	3	4338	AD	3	222	MCI-c	1.5
467	NC	1.5	4491	NC	3	4353	AD	3	227	MCI-c	1.5
484	NC	1.5	4496	NC	3	4373	AD	3	231	MCI-c	1.5
488	NC	2.9	4499	NC	3	4379	AD	3	240	MCI-c	1.5
489	NC	1.5	4503	NC	3	4477	AD	3	241	MCI-c	1.5
493	NC	2.9	4505	NC	3	4494	AD	3	243	MCI-c	1.5
498	NC	1.5	4508	NC	3	4500	AD	3	249	MCI-c	1.5
500	NC	1.5	4512	NC	3	4526	AD	3	256	MCI-c	1.5
502	NC	1.5	4516	NC	3	4546	AD	3	258	MCI-c	1.5
506	NC	1.5	4545	NC	3	4549	AD	3	269	MCI-c	1.5
516	NC	1.5	4552	NC	3	4583	AD	3	289	MCI-c	1.5
519	NC	1.5	4555	NC	3	4589	AD	3	293	MCI-c	1.5
520	NC	1.5	4558	NC	3	4591	AD	3	294	MCI-c	1.5
522	NC	1.5	4559	NC	3	4615	AD	3	314	MCI-c	1.5
526	NC	1.5	4560	NC	3	4641	AD	3	325	MCI-c	1.5
533	NC	1.5	4566	NC	3	4657	AD	3	326	MCI-c	1.5

534	NC	1.5	4576	NC	3	4660	AD	3	331	MCI-c	2.9
538	NC	1.5	4577	NC	3	4672	AD	3	336	MCI-c	1.5
545	NC	1.5	4578	NC	3	4686	AD	3	344	MCI-c	1.5
548	NC	1.5	4579	NC	3	4692	AD	3	362	MCI-c	1.5
553	NC	3	4580	NC	3	4696	AD	3	388	MCI-c	2.9
555	NC	1.5	4585	NC	3	4707	AD	3	390	MCI-c	1.5
558	NC	1.5	4586	NC	3	4728	AD	3	394	MCI-c	1.5
559	NC	1.5	4587	NC	3	4730	AD	3	423	MCI-c	1.5
575	NC	1.5	4599	NC	3	4732	AD	3	434	MCI-c	1.5
576	NC	1.5	4604	NC	3	4733	AD	3	461	MCI-c	1.5
578	NC	1.5	4607	NC	3	4755	AD	3	507	MCI-c	1.5
601	NC	1.5	4609	NC	3	4756	AD	3	511	MCI-c	1.5
602	NC	3	4612	NC	3	4770	AD	3	513	MCI-c	1.5
605	NC	3	4616	NC	3	4774	AD	3	518	MCI-c	1.5
610	NC	1.5	4620	NC	3	4783	AD	3	539	MCI-c	1.5
618	NC	1.5	4632	NC	3	4801	AD	3	549	MCI-c	1.5
622	NC	3	4637	NC	3	4802	AD	3	563	MCI-c	1.5
640	NC	2.9	4638	NC	3	4820	AD	3	567	MCI-c	1.5
643	NC	1.5	4643	NC	3	4827	AD	3	568	MCI-c	3
647	NC	1.5	4644	NC	3	4845	AD	3	572	MCI-c	3
648	NC	1.5	4645	NC	3	4853	AD	3	604	MCI-c	2.9
657	NC	1.5	4649	NC	3	4863	AD	3	611	MCI-c	1.5
672	NC	1.5	4652	NC	3	4867	AD	3	625	MCI-c	3
677	NC	3	4688	NC	3	4887	AD	3	631	MCI-c	1.5
680	NC	1.5	4739	NC	3	4892	AD	3	638	MCI-c	2.9
681	NC	1.5	4762	NC	3	4894	AD	3	649	MCI-c	2.9
684	NC	1.5	4795	NC	3	4905	AD	3	658	MCI-c	1.5
685	NC	1.5	4832	NC	3	4906	AD	3	667	MCI-c	1.5
686	NC	1.5	4835	NC	3	4910	AD	3	675	MCI-c	1.5
692	NC	1.5	4843	NC	3	4911	AD	3	695	MCI-c	1.5
711	NC	1.5	4855	NC	3	4924	AD	3	697	MCI-c	1.5
717	NC	1.5	4872	NC	3	4938	AD	3	708	MCI-c	1.5
726	NC	1.5	4878	NC	3	4940	AD	3	723	MCI-c	1.5
731	NC	1.5	4900	NC	3	4949	AD	3	725	MCI-c	2.9

734	NC	1.5	4921	NC	3	4954	AD	3	727	MCI-c	2.9
741	NC	1.5	4951	NC	3	4962	AD	3	729	MCI-c	1.5
751	NC	1.5	4952	NC	3	4971	AD	3	750	MCI-c	1.5
761	NC	1.5	5040	NC	3	4992	AD	3	752	MCI-c	2.9
767	NC	1.5	3	AD	1.5	5012	AD	3	769	MCI-c	3
768	NC	1.5	7	AD	1.5	5013	AD	3	834	MCI-c	1.5
778	NC	1.5	10	AD	1.5	5015	AD	3	835	MCI-c	2.9
779	NC	1.5	29	AD	1.5	5017	AD	3	839	MCI-c	1.5
810	NC	1.5	53	AD	1.5	5018	AD	3	856	MCI-c	1.5
813	NC	1.5	76	AD	1.5	5019	AD	3	860	MCI-c	1.5
818	NC	1.5	78	AD	2.9	5028	AD	3	861	MCI-c	1.5
824	NC	1.5	83	AD	1.5	4	MCI-s	1.5	865	MCI-c	1.5
842	NC	1.5	84	AD	1.5	33	MCI-s	1.5	869	MCI-c	1.5
843	NC	1.5	88	AD	1.5	51	MCI-s	1.5	873	MCI-c	1.5
862	NC	1.5	91	AD	1.5	102	MCI-s	1.5	874	MCI-c	1.5
863	NC	1.5	93	AD	1.5	107	MCI-s	1.5	878	MCI-c	1.5
866	NC	1.5	94	AD	1.5	116	MCI-s	1.5	887	MCI-c	1.5
876	NC	1.5	110	AD	1.5	135	MCI-s	1.5	906	MCI-c	1.5
883	NC	1.5	129	AD	1.5	150	MCI-s	1.5	909	MCI-c	1.5
886	NC	1.5	139	AD	2.9	158	MCI-s	1.5	913	MCI-c	2.9
896	NC	1.5	149	AD	1.5	160	MCI-s	1.5	915	MCI-c	1.5
898	NC	1.5	162	AD	1.5	169	MCI-s	1.5	922	MCI-c	2.9
899	NC	1.5	167	AD	1.5	176	MCI-s	1.5	941	MCI-c	1.5
907	NC	1.5	183	AD	1.5	178	MCI-s	1.5	947	MCI-c	1.5
920	NC	2.9	194	AD	1.5	200	MCI-s	1.5	952	MCI-c	1.5
923	NC	1.5	213	AD	1.5	225	MCI-s	1.5	954	MCI-c	1.5
926	NC	3	216	AD	1.5	273	MCI-s	1.5	973	MCI-c	1.5
931	NC	1.5	219	AD	1.5	276	MCI-s	1.5	976	MCI-c	1.5
934	NC	1.5	221	AD	1.5	285	MCI-s	1.5	982	MCI-c	1.5
951	NC	1.5	228	AD	1.5	288	MCI-s	1.5	987	MCI-c	1.5
963	NC	3	266	AD	1.5	290	MCI-s	3	997	MCI-c	1.5
967	NC	1.5	299	AD	1.5	291	MCI-s	1.5	1004	MCI-c	1.5
969	NC	1.5	300	AD	1.5	292	MCI-s	1.5	1007	MCI-c	1.5
972	NC	1.5	310	AD	1.5	307	MCI-s	2.9	1010	MCI-c	1.5

981	NC	1.5	316	AD	1.5	324	MCI-s	3	1015	MCI-c	1.5
984	NC	1.5	321	AD	1.5	339	MCI-s	1.5	1043	MCI-c	1.5
985	NC	1.5	328	AD	1.5	351	MCI-s	1.5	1054	MCI-c	1.5
1002	NC	1.5	332	AD	2.9	361	MCI-s	1.5	1057	MCI-c	1.5
1013	NC	1.5	341	AD	1.5	376	MCI-s	2.9	1066	MCI-c	3
1014	NC	1.5	343	AD	1.5	378	MCI-s	1.5	1070	MCI-c	1.5
1016	NC	2.9	356	AD	1.5	389	MCI-s	1.5	1073	MCI-c	1.5
1021	NC	1.5	366	AD	1.5	407	MCI-s	1.5	1077	MCI-c	1.5
1023	NC	1.5	370	AD	1.5	408	MCI-s	1.5	1117	MCI-c	3
1035	NC	2.9	372	AD	1.5	424	MCI-s	1.5	1121	MCI-c	3
1063	NC	1.5	374	AD	1.5	445	MCI-s	1.5	1126	MCI-c	3
1086	NC	2.9	392	AD	2.9	448	MCI-s	3	1130	MCI-c	1.5
1094	NC	1.5	404	AD	2.9	449	MCI-s	1.5	1135	MCI-c	1.5
1098	NC	2.9	426	AD	1.5	464	MCI-s	1.5	1138	MCI-c	3
1099	NC	1.5	431	AD	1.5	469	MCI-s	2.9	1148	MCI-c	2.9
1123	NC	3	438	AD	1.5	481	MCI-s	1.5	1213	MCI-c	1.5
1169	NC	3	457	AD	3	501	MCI-s	2.9	1217	MCI-c	1.5
1190	NC	3	470	AD	1.5	505	MCI-s	1.5	1224	MCI-c	1.5
1194	NC	1.5	474	AD	1.5	546	MCI-s	1.5	1240	MCI-c	1.5
1195	NC	1.5	487	AD	2.9	552	MCI-s	1.5	1243	MCI-c	1.5
1197	NC	1.5	492	AD	1.5	557	MCI-s	1.5	1244	MCI-c	1.5
1200	NC	1.5	497	AD	1.5	588	MCI-s	1.5	1247	MCI-c	3
1202	NC	1.5	517	AD	1.5	590	MCI-s	1.5	1265	MCI-c	1.5
1203	NC	1.5	528	AD	1.5	607	MCI-s	3	1271	MCI-c	1.5
1206	NC	3	535	AD	1.5	608	MCI-s	1.5	1282	MCI-c	1.5
1222	NC	3	543	AD	1.5	613	MCI-s	3	1295	MCI-c	1.5
1232	NC	2.9	547	AD	1.5	621	MCI-s	1.5	1299	MCI-c	1.5
1242	NC	2.9	554	AD	1.5	626	MCI-s	1.5	1311	MCI-c	1.5
1249	NC	2.9	565	AD	1.5	641	MCI-s	1.5	1315	MCI-c	1.5
1250	NC	3	577	AD	1.5	644	MCI-s	1.5	1331	MCI-c	3
1251	NC	3	592	AD	1.5	656	MCI-s	1.5	1351	MCI-c	1.5
1256	NC	3	606	AD	3	671	MCI-s	1.5	1363	MCI-c	1.5
1261	NC	1.5	619	AD	1.5	673	MCI-s	1.5	1387	MCI-c	2.9
1267	NC	3	627	AD	1.5	679	MCI-s	1.5	1389	MCI-c	3

1276	NC	2.9	642	AD	1.5	698	MCI-s	1.5	1393	MCI-c	1.5
1280	NC	1.5	653	AD	1.5	709	MCI-s	1.5	1394	MCI-c	1.5
1288	NC	3	690	AD	1.5	715	MCI-s	1.5	1398	MCI-c	1.5
1301	NC	3	691	AD	3	746	MCI-s	1.5	1407	MCI-c	1.5
1306	NC	1.5	696	AD	1.5	770	MCI-s	1.5	1412	MCI-c	1.5
2201	NC	3	699	AD	1.5	771	MCI-s	2.9	1423	MCI-c	1.5
4003	NC	3	724	AD	2.9	782	MCI-s	1.5	1425	MCI-c	1.5
4010	NC	3	730	AD	1.5	792	MCI-s	2.9	1427	MCI-c	1.5
4014	NC	3	733	AD	2.9	800	MCI-s	1.5	2047	MCI-c	3
4018	NC	3	740	AD	1.5	830	MCI-s	3	2216	MCI-c	3
4020	NC	3	753	AD	2.9	851	MCI-s	1.5	2398	MCI-c	3
4021	NC	3	754	AD	1.5	867	MCI-s	1.5	4005	MCI-c	3
4026	NC	3	759	AD	1.5	871	MCI-s	1.5	4015	MCI-c	3
4028	NC	3	760	AD	1.5	908	MCI-s	1.5	4035	MCI-c	3
4032	NC	3	777	AD	1.5	912	MCI-s	2.9	4042	MCI-c	3
4037	NC	3	784	AD	1.5	914	MCI-s	1.5	4057	MCI-c	3
4041	NC	3	790	AD	1.5	919	MCI-s	1.5	4094	MCI-c	3
4043	NC	3	793	AD	1.5	921	MCI-s	1.5	4096	MCI-c	3
4050	NC	3	796	AD	1.5	925	MCI-s	1.5	4102	MCI-c	3
4060	NC	3	812	AD	3	945	MCI-s	1.5	4114	MCI-c	3
4066	NC	3	814	AD	3	950	MCI-s	1.5	4131	MCI-c	3
4075	NC	3	816	AD	1.5	961	MCI-s	1.5	4167	MCI-c	3
4076	NC	3	828	AD	3	989	MCI-s	1.5	4189	MCI-c	3
4080	NC	3	836	AD	1.5	994	MCI-s	1.5	4203	MCI-c	3
4081	NC	3	841	AD	1.5	1030	MCI-s	1.5	4240	MCI-c	3
4082	NC	3	844	AD	3	1031	MCI-s	2.9	4366	MCI-c	3
4084	NC	3	850	AD	1.5	1032	MCI-s	1.5	4402	MCI-c	3
4086	NC	3	852	AD	1.5	1034	MCI-s	1.5	4414	MCI-c	3
4090	NC	3	853	AD	1.5	1040	MCI-s	1.5	4432	MCI-c	3
4092	NC	3	884	AD	1.5	1045	MCI-s	1.5	4502	MCI-c	3
4093	NC	3	889	AD	1.5	1046	MCI-s	3	4515	MCI-c	3
4100	NC	3	891	AD	1.5	1052	MCI-s	1.5	4530	MCI-c	3
4104	NC	3	916	AD	3	1072	MCI-s	3	4595	MCI-c	3
4119	NC	3	929	AD	1.5	1078	MCI-s	1.5	4661	MCI-c	3

4120	NC	3	938	AD	1.5	1080	MCI-s	1.5	4680	MCI-c	3
4121	NC	3	955	AD	1.5	1088	MCI-s	2.9	4689	MCI-c	3
4125	NC	3	956	AD	1.5	1097	MCI-s	1.5	4706	MCI-c	3
4139	NC	3	979	AD	1.5	1106	MCI-s	1.5	4712	MCI-c	3
4148	NC	3	991	AD	1.5	1114	MCI-s	1.5	4784	MCI-c	3
4150	NC	3	996	AD	2.9	1118	MCI-s	1.5	4796	MCI-c	3
4151	NC	3	999	AD	1.5	1122	MCI-s	1.5	4857	MCI-c	3
4155	NC	3	1001	AD	1.5	1131	MCI-s	3	4888	MCI-c	3
4158	NC	3	1018	AD	1.5	1140	MCI-s	1.5	4899	MCI-c	3
4164	NC	3	1024	AD	1.5	1149	MCI-s	3	4918	MCI-c	3
4173	NC	3	1027	AD	1.5	1155	MCI-s	1.5	4928	MCI-c	3
4174	NC	3	1041	AD	1.5	1182	MCI-s	1.5			
4176	NC	3	1044	AD	1.5	1183	MCI-s	1.5			
4177	NC	3	1055	AD	2.9	1186	MCI-s	1.5			

References

- [Abdulkadir et al., 2011] Abdulkadir, A., Mortamet, B., Vemuri, P., Jack Jr., C. R., Krueger, G., and Klöppel, S. (2011). Effects of hardware heterogeneity on the performance of SVM Alzheimer’s disease classifier. *NeuroImage*, 58(3):785–792.
- [Aizerman et al., 1964] Aizerman, A., Braverman, E., and Rozoner, L. (1964). Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, 25:821–837.
- [Aksu et al., 2011] Aksu, Y., Miller, D. J., Kesidis, G., Bigler, D. C., and Yang, Q. X. (2011). An MRI-derived definition of MCI-to-AD conversion for long-term, automatic prognosis of MCI patients. *PLoS ONE*, 6(10):e25074.
- [Albert et al., 2011] Albert, M. S., DeKosky, S. T., Dickson, D., Dubois, B., Feldman, H. H., Fox, N. C., Gamst, A., Holtzman, D. M., Jagust, W. J., Petersen, R. C., Snyder, P. J., Carrillo, M. C., Thies, B., and Phelps, C. H. (2011). The diagnosis of mild cognitive impairment due to alzheimers disease: Recommendations from the national institute on aging-alzheimers association workgroups on diagnostic guidelines for Alzheimer’s disease. *Alzheimer’s Dementia*, 7(3):270 – 279.
- [Aljabar et al., 2009] Aljabar, P., Heckemann, R. A., Hammers, A., Hajnal, J. V., and Rueckert, D. (2009). Multi-atlas based segmentation of brain images: Atlas selection and its effect on accuracy. *NeuroImage*, 46(3):726–738.
- [Ansari et al., 2014] Ansari, M. H., Coen, M. H., Bendlin, B. B., Sager, M. A., and Johnson, S. C. (2014). A spatially sensitive kernel to predict cognitive performance from short-term changes in neural structure. In *AAAI*, pages 1157–1163.
- [Ashburner, 2007] Ashburner, J. (2007). A fast diffeomorphic image registration algorithm. *NeuroImage*, 38(1):95–113.

- [Ashburner and Friston, 2000] Ashburner, J. and Friston, K. J. (2000). Voxel-based Morphometry-the methods. *NeuroImage*, 11(6):805–821.
- [Ashburner and Friston, 2005] Ashburner, J. and Friston, K. J. (2005). Unified segmentation. *NeuroImage*, 26(3):839–851.
- [Ashby and Smith, 2000] Ashby, D. and Smith, A. F. (2000). Evidence-based medicine as Bayesian decision-making. *Statistics in Medicine*, 19(23):3291–3305.
- [Bach and Lanckriet, 2004] Bach, F. R. and Lanckriet, G. R. G. (2004). Multiple kernel learning, conic duality, and the SMO algorithm. In *In Proceedings of the 21st International Conference on Machine Learning (ICML)*.
- [Banerjee et al., 2013] Banerjee, J., Moelker, A., Niessen, W., and van Walsum, T. (2013). 3D lbp-based rotationally invariant region description. In Park, J.-I. and Kim, J., editors, *Computer Vision - ACCV 2012 Workshops*, volume 7728 of *Lecture Notes in Computer Science*, pages 26–37. Springer Berlin Heidelberg.
- [Barnes et al., 2004] Barnes, J., Scahill, R. I., Boyes, R. G., Frost, C., Lewis, E. B., Rossor, C. L., Rossor, M. N., and Fox, N. C. (2004). Differentiating AD from aging using semiautomated measurement of hippocampal atrophy rates. *NeuroImage*, 23(2):574–581.
- [Barrett and Coolen, 2013] Barrett, J. and Coolen, A. (2013). Gaussian process regression for survival analysis with interval censored data. (1312.1591).
- [Beach et al., 2012] Beach, T. G., Monsell, S. E., Phillips, L. E., and Kukull, W. (2012). Accuracy of the clinical diagnosis of alzheimer disease at national institute on aging alzheimer disease centers, 20052010. *Journal of Neuropathology & Experimental Neurology*, 71(4):266–273.
- [Belkin and Niyogi, 2003] Belkin, M. and Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396.
- [Berchtold and Cotman, 1998] Berchtold, N. C. and Cotman, C. W. (1998). Evolution in the conceptualization of dementia and Alzheimer’s disease: Greco-roman period to the 1960s. *Neurobiology of Aging*, 19(3):173–189.
- [Berner and Graber, 2008] Berner, E. S. and Graber, M. L. (2008). Overconfidence as a cause of diagnostic error in medicine. *The American journal of medicine*, 121:S2–23.

- [Birks and Harvey, 1996] Birks, J. and Harvey, R. J. (1996). Donepezil for dementia due to Alzheimer's disease. In *Cochrane Database of Systematic Reviews*. John Wiley & Sons, Ltd.
- [Bishop, 2007] Bishop, C. M. (2007). *Pattern Recognition and Machine Learning*. Springer.
- [Blennow, 2004] Blennow, K. (2004). Cerebrospinal fluid protein biomarkers for Alzheimer's disease. *NeuroRx*, 1(2):213–225.
- [Boser et al., 1992] Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on computational learning theory, COLT '92*, pages 144–152, New York, NY, USA. ACM.
- [Bouras et al., 1994] Bouras, C., Hof, P. R., Giannakopoulos, P., Michel, J.-P., and Morrison, J. H. (1994). Regional distribution of neurofibrillary tangles and senile plaques in the cerebral cortex of elderly patients: A quantitative evaluation of a one-year autopsy population from a geriatric hospital. *Cerebral Cortex*, 4(2):138–150.
- [Braak and Braak, 1995] Braak, H. and Braak, E. (1995). Staging of Alzheimer's disease-related neurofibrillary changes. *Neurobiology of Aging*, 16(3):271–278.
- [Brier, 1950] Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3.
- [Bro-Nielsen and Gramkow, 1996] Bro-Nielsen, M. and Gramkow, C. (1996). Fast fluid registration of medical images. In Höhne, K. and Kikinis, R., editors, *Visualization in Biomedical Computing*, volume 1131 of *Lecture Notes in Computer Science*, pages 265–276. Springer Berlin Heidelberg.
- [Brookmeyer et al., 2007] Brookmeyer, R., Johnson, E., Ziegler-Graham, K., and Arrighi, H. M. (2007). Forecasting the global burden of Alzheimer's disease. *Alzheimer's & Dementia*, 3(3):186–191.
- [Bylesjö et al., 2006] Bylesjö, M., Rantalainen, M., Cloarec, O., Nicholson, J. K., Holmes, E., and Trygg, J. (2006). OPLS discriminant analysis: combining the strengths of PLS-DA and SIMCA classification. *Journal of Chemometrics*, 20(8-10):341–351.
- [Cardoso et al., 2012] Cardoso, M., Modat, M., Ourselin, S., Keihaninejad, S., and Cash, D. (2012). Multi-STEPS: multi-label similarity and truth estimation for propagated segmen-

- tations. In *2012 IEEE Workshop on Mathematical Methods in Biomedical Image Analysis (MMBIA)*, pages 153–158.
- [Cardoso et al., 2011] Cardoso, M. J., Clarkson, M. J., Ridgway, G. R., Modat, M., Fox, N. C., and Ourselin, S. (2011). LoAd: a locally adaptive cortical segmentation algorithm. *NeuroImage*, 56:1386–1397.
- [Cawley and Talbot, 2004] Cawley, G. C. and Talbot, N. L. C. (2004). Fast exact leave-one-out cross-validation of sparse least-squares support vector machines. *Neural Networks*, 17:1467–1475.
- [Cedazo-Minguez and Winblad, 2010] Cedazo-Minguez, A. and Winblad, B. (2010). Biomarkers for alzheimers disease and other forms of dementia: Clinical needs, limitations and future aspects. *Experimental Gerontology*, 45(1):5–14.
- [Chang and Lin, 2011] Chang, C.-C. and Lin, C.-J. (2011). Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27.
- [Chapelle et al., 2010] Chapelle, O., Schölkopf, B., and Zien, A. (2010). *Semi-supervised learning*. MIT, Cambridge, Mass.; London.
- [Chen et al., 1995] Chen, Y. Q., Nixon, M. S., and Thomas, D. W. (1995). Statistical geometrical features for texture classification. *Pattern Recognition*, 28(4):537–552.
- [Cho et al., 2012] Cho, Y., Seong, J.-K., Jeong, Y., and Shin, S. Y. (2012). Individual subject classification for Alzheimer’s disease based on incremental learning using a spatial frequency representation of cortical thickness data. *NeuroImage*, 59(3):2217–2230.
- [Chong and Sahadevan, 2005] Chong, M. S. and Sahadevan, S. (2005). Preclinical Alzheimer’s disease: diagnosis and prediction of progression. *Lancet neurology*, 4(9):576–579.
- [Chu et al., 2010] Chu, C., Bandettini, P., Ashburner, J., Marquand, A., and Kloeppel, S. (2010). Classification of neurodegenerative diseases using Gaussian process classification with automatic feature determination. In *2010 First Workshop on Brain Decoding: Pattern Recognition Challenges in Neuroimaging (WBD)*, pages 17–20. IEEE.
- [Chu et al., 20] Chu, C., Hsu, A.-L., Chou, K.-H., Bandettini, P., and Lin, C. (20). Does feature selection improve classification accuracy? Impact of sample size and feature selection on classification using anatomical magnetic resonance images. *NeuroImage*, 60(1):59–70.

- [Chupin et al., 2009] Chupin, M., Gérardin, E., Cuingnet, R., Boutet, C., Lemieux, L., Lehéricy, S., Benali, H., Garnero, L., and Colliot, O. (2009). Fully automatic hippocampus segmentation and classification in Alzheimer's disease and mild cognitive impairment applied on data from ADNI. *Hippocampus*, 19(6):579–587.
- [Chupin et al., 2007] Chupin, M., Mukuna-Bantumbakulu, A. R., Hasboun, D., Bardinnet, E., Baillet, S., Kinkingnéhun, S., Lemieux, L., Dubois, B., and Garnero, L. (2007). Anatomically constrained region deformation for the automated segmentation of the hippocampus and the amygdala: Method and validation on controls and patients with Alzheimer's disease. *NeuroImage*, 34(3):996–1019.
- [Corder et al., 1994] Corder, E. H., Saunders, A. M., Risch, N. J., Strittmatter, W. J., Schmechel, D. E., Gaskell, Jr, P. C., Rimmler, J. B., Locke, P. A., Conneally, P. M., and Schmechel, K. E. (1994). Protective effect of apolipoprotein e type 2 allele for late onset alzheimer disease. *Nature genetics*, 7(2):180–184.
- [Corder et al., 1993] Corder, E. H., Saunders, A. M., Strittmatter, W. J., Schmechel, D. E., Gaskell, P. C., Small, G. W., Roses, A. D., Haines, J. L., and Pericak-Vance, M. A. (1993). Gene dose of apolipoprotein e type 4 allele and the risk of Alzheimer's disease in late onset families. *Science*, 261(5123):921–923.
- [Cortes and Vapnik, 1995] Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.
- [Costafreda et al., 2011] Costafreda, S. G., Dinov, I. D., Tu, Z., Shi, Y., Liu, C.-Y., Kloszewska, I., Mecocci, P., Soinen, H., Tsolaki, M., Vellas, B., Wahlund, L.-O., Spenger, C., Toga, A. W., Lovestone, S., and Simmons, A. (2011). Automated hippocampal shape analysis predicts the onset of dementia in mild cognitive impairment. *NeuroImage*, 56(1):212–219.
- [Coupé et al., 2012] Coupé, P., Eskildsen, S. F., Manón, J. V., Fonov, V. S., Pruessner, J. C., Allard, M., and Collins, D. L. (2012). Scoring by nonlocal image patch estimator for early detection of Alzheimer's disease. *NeuroImage: Clinical*, 1(1):141–152.
- [Cox and Oakes, 1984] Cox, D. R. and Oakes, D. (1984). *Analysis of Survival Data*. CRC Press.
- [Csernansky et al., 2000] Csernansky, J. G., Wang, L., Joshi, S., Miller, J. P., Gado, M., Kido, D., McKeel, D., Morris, J. C., and Miller, M. I. (2000). Early DAT is distinguished from aging by high-dimensional mapping of the hippocampus. *Neurology*, 55(11):1636–1643.

- [Cui et al., 2011] Cui, Y., Liu, B., Luo, S., Zhen, X., Fan, M., Liu, T., Zhu, W., Park, M., Jiang, T., Jin, J. S., and the Alzheimer's Disease Neuroimaging Initiative (2011). Identification of conversion from mild cognitive impairment to Alzheimer's disease using multivariate predictors. *PLoS ONE*, 6(7):e21896.
- [Cuingnet et al., 2010] Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehéricy, S., Habert, M.-O., Chupin, M., Benali, H., and Colliot, O. (2010). 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.
- [Cuingnet et al., 2013] Cuingnet, R., Glaunès, J. A., Chupin, M., Benali, H., and Colliot, O. (2013). Spatial and anatomical regularization of SVM: A general framework for neuroimaging data. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(3):682–696.
- [Dale et al., 1999] Dale, A. M., Fischl, B., and Sereno, M. I. (1999). Cortical surface-based analysis: I. segmentation and surface reconstruction. *NeuroImage*, 9(2):179–194.
- [Davatzikos et al., 2008] Davatzikos, C., Fan, Y., Wu, X., Shen, D., and Resnick, S. M. (2008). Detection of prodromal Alzheimer's disease via pattern classification of magnetic resonance imaging. *Neurobiology of Aging*, 29(4):514–523.
- [de Jong et al., 2008] de Jong, L. W., van der Hiele, K., Veer, I. M., Houwing, J. J., Westendorp, R. G. J., Bollen, E. L. E. M., de Bruin, P. W., Middelkoop, H. A. M., van Buchem, M. A., and van der Grond, J. (2008). Strongly reduced volumes of putamen and thalamus in alzheimer's disease: an MRI study. *Brain*, 131(12):3277–3285.
- [Dempster et al., 1977] Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):38, 1.
- [Desikan et al., 2009] Desikan, R. S., Cabral, H. J., Hess, C. P., Dillon, W. P., Glastonbury, C. M., Weiner, M. W., Schmansky, N. J., Greve, D. N., Salat, D. H., Buckner, R. L., Fischl, B., and Alzheimer's Disease Neuroimaging Initiative (2009). Automated MRI measures identify individuals with mild cognitive impairment and Alzheimer's disease. *Brain*, 132(8):2048–2057.
- [Dickerson et al., 2008] Dickerson, B. C., Bakkour, A., Salat, D. H., Feczko, E., Pacheco, J., Greve, D. N., Grodstein, F., Wright, C. I., Blacker, D., Rosas, H. D., Sperling, R. A., Atri, A., Growdon, J. H., Hyman, B. T., Morris, J. C., Fischl, B., and Buckner, R. L. (2008).

- The cortical signature of Alzheimer's disease: Regionally specific cortical thinning relates to symptom severity in very mild to mild AD dementia and is detectable in asymptomatic amyloid-positive individuals. *Cerebral Cortex*, 19(3):497–510.
- [Domingos, 2012] Domingos, P. (2012). A few useful things to know about machine learning. *Commun. ACM*, 55(10):78–87.
- [Doyle et al., 2013] Doyle, O., Ashburner, J., Zelaya, F., Williams, S., Mehta, M., and Marquand, A. (2013). Multivariate decoding of brain images using ordinal regression. *NeuroImage*, 81(0):347–357.
- [Doyle et al., 2014] Doyle, O. M., Westman, E., Marquand, A. F., Mecocci, P., Vellas, B., Tsolaki, M., Koszewska, I., Soininen, H., Lovestone, S., Williams, S. C. R., and Simmons, A. (2014). Predicting progression of Alzheimer's disease using ordinal regression. *PLoS ONE*, 9(8):e105542.
- [Drzezga et al., 2003] Drzezga, A., Lautenschlager, N., Siebner, H., Riemenschneider, M., Willoch, F., Minoshima, S., Schwaiger, M., and Kurz, A. (2003). Cerebral metabolic changes accompanying conversion of mild cognitive impairment into Alzheimer's disease: a PET follow-up study. *European journal of nuclear medicine and molecular imaging*, 30(8):1104–1113.
- [Dubois and Albert, 2004] Dubois, B. and Albert, M. L. (2004). Amnesic MCI or prodromal Alzheimer's disease? *The Lancet Neurology*, 3(4):246 – 248.
- [Eskildsen et al., 2013] Eskildsen, S. F., Coupé, P., García-Lorenzo, D., Fonov, V., Pruessner, J. C., and Collins, D. L. (2013). Prediction of Alzheimer's disease in subjects with mild cognitive impairment from the ADNI cohort using patterns of cortical thinning. *NeuroImage*, 65(0):511–521.
- [Eskildsen and Ostergaard, 2006] Eskildsen, S. F. and Ostergaard, L. R. (2006). Active surface approach for extraction of the human cerebral cortex from MRI. In Larsen, R., Nielsen, M., and Sporring, J., editors, *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2006*, volume 4191 of *Lecture Notes in Computer Science*, pages 823–830. Springer Berlin Heidelberg.
- [Fan et al., 2007] Fan, Y., Shen, D., Gur, R. C., Gur, R. E., and Davatzikos, C. (2007). COM-PARE: classification of morphological patterns using adaptive regional elements. *IEEE Transactions on Medical Imaging*, 26(1):93–105.

- [Fawcett, 2006] Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recogn. Lett.*, 27(8):861–874.
- [Fehr and Burkhardt, 2008] Fehr, J. and Burkhardt, H. (2008). 3D rotation invariant local binary patterns. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pages 1–4.
- [Ferrarini et al., 2009] Ferrarini, L., Frisoni, G. B., Pievani, M., Reiber, J. H., Ganzola, R., and Milles, J. (2009). Morphological hippocampal markers for automated detection of Alzheimer’s disease and mild cognitive impairment converters in magnetic resonance images. *Journal of Alzheimer’s Disease*, 17(3):643–659.
- [Ferrarini et al., 2007] Ferrarini, L., Olofsen, H., Palm, W. M., van Buchem, M. A., Reiber, J. H., and Admiraal-Behloul, F. (2007). GAMEs: growing and adaptive meshes for fully automatic shape modeling and analysis. *Medical Image Analysis*, 11(3):302–314.
- [Filipovych and Davatzikos, 2011] Filipovych, R. and Davatzikos, C. (2011). Semi-supervised pattern classification of medical images: Application to mild cognitive impairment (mci). *NeuroImage*, 55(3):1109–1119.
- [Fischl and Dale, 2000] Fischl, B. and Dale, A. M. (2000). Measuring the thickness of the human cerebral cortex from magnetic resonance images. *Proceedings of the National Academy of Sciences of the United States of America*, 97(20):11050–11055.
- [Fischl et al., 1999a] Fischl, B., Sereno, M. I., and Dale, A. M. (1999a). Cortical surface-based analysis: II: inflation, flattening, and a surface-based coordinate system. *NeuroImage*, 9(2):195–207.
- [Fischl et al., 1999b] Fischl, B., Sereno, M. I., Tootell, R. B., and Dale, A. M. (1999b). High-resolution intersubject averaging and a coordinate system for the cortical surface. *Human brain mapping*, 8(4):272–284.
- [Fisher, 1936] Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188.
- [Fjell et al., 2010a] Fjell, A. M., Walhovd, K. B., Fennema-Notestine, C., McEvoy, L. K., Hagler, D. J., Holland, D., Blennow, K., Brewer, J. B., and Dale, A. M. (2010a). Brain atrophy in healthy aging is related to CSF levels of amyloid-beta-42. *Cerebral Cortex*, 20(9):2069–2079.

- [Fjell et al., 2010b] Fjell, A. M., Walhovd, K. B., Fennema-Notestine, C., McEvoy, L. K., Hager, D. J., Holland, D., Brewer, J. B., and Dale, A. M. (2010b). CSF biomarkers in prediction of cerebral and clinical change in mild cognitive impairment and Alzheimer’s disease. *The Journal of Neuroscience*, 30(6):2088–2101.
- [Folstein et al., 1975] Folstein, M. F., Folstein, S. E., and McHugh, P. R. (1975). ”Mini-mental state” : A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, 12(3):189–198.
- [Fonteijn et al., 2012] Fonteijn, H. M., Modat, M., Clarkson, M. J., Barnes, J., Lehmann, M., Hobbs, N. Z., Scahill, R. I., Tabrizi, S. J., Ourselin, S., Fox, N. C., and Alexander, D. C. (2012). An event-based model for disease progression and its application in familial Alzheimer’s disease and Huntington’s disease. *NeuroImage*, 60(3):1880–1889.
- [Forsberg et al., 2008] Forsberg, A., Engler, H., Almkvist, O., Blomquist, G., Hagman, G., Wall, A., Ringheim, A., Lngstrm, B., and Nordberg, A. (2008). PET imaging of amyloid deposition in patients with mild cognitive impairment. *Neurobiology of Aging*, 29(10):1456–1465.
- [Franke and Gaser, 2012] Franke, K. and Gaser, C. (2012). Longitudinal changes in individual *BrainAGE* in healthy aging, mild cognitive impairment, and Alzheimer’s disease. *GeroPsych: The Journal of Gerontopsychology and Geriatric Psychiatry*, 25(4):235–245.
- [Franke et al., 2010] Franke, K., Ziegler, G., Klöppel, S., and Gaser, C. (2010). Estimating the age of healthy subjects from t1-weighted MRI scans using kernel methods: Exploring the influence of various parameters. *NeuroImage*, 50(3):883–892.
- [Freeborough and Fox, 1998] Freeborough, P. and Fox, N. (1998). Mr image texture analysis applied to the diagnosis and tracking of Alzheimer’s disease. *Medical Imaging, IEEE Transactions on*, 17(3):475–478.
- [Gaser et al., 2013] Gaser, C., Franke, K., Klöppel, S., Koutsouleris, N., Sauer, H., and Alzheimer’s Disease Neuroimaging Initiative (2013). BrainAGE in mild cognitive impaired patients: Predicting the conversion to Alzheimer’s disease. *PLoS ONE*, 8(6):e67346.
- [Gerardin et al., 2009] Gerardin, E., Chételat, G., Chupin, M., Cuingnet, R., Desgranges, B., Kim, H.-S., Niethammer, M., Dubois, B., Lehericy, S., Garnero, L., Eustache, F., and Colliot, O. (2009). Multidimensional classification of hippocampal shape features discrimi-

- nates Alzheimer's disease and mild cognitive impairment from normal aging. *NeuroImage*, 47(4):1476–1486.
- [Gerig et al., 2001] Gerig, G., Styner, M., Jones, D., Weinberger, D., and Lieberman, J. (2001). Shape analysis of brain ventricles using SPHARM. In *Proceedings of the IEEE Workshop on Mathematical Methods in Biomedical Image Analysis (MMBIA'01)*, MMBIA '01, pages 171–178, Washington, DC, USA. IEEE Computer Society.
- [Geula and Mesulam, 1995] Geula, C. and Mesulam, M. M. (1995). Cholinesterases and the pathology of alzheimer disease. *Alzheimer disease and associated disorders*, 9 Suppl 2:23–28.
- [Ghiso and Frangione, 2002] Ghiso, J. and Frangione, B. (2002). Amyloidosis and Alzheimer's disease. *Advanced Drug Delivery Reviews*, 54(12):1539–1551.
- [Gibbs and MacKay, 2000] Gibbs, M. and MacKay, D. J. C. (2000). Variational Gaussian process classifiers. *IEEE Transactions on Neural Networks*, 11(6):1458–1464.
- [Gousias et al., 2008] Gousias, I. S., Rueckert, D., Heckemann, R. A., Dyet, L. E., Boardman, J. P., Edwards, A. D., and Hammers, A. (2008). Automatic segmentation of brain MRIs of 2-year-olds into 83 regions of interest. *NeuroImage*, 40(2):672–684.
- [Gramfort et al., 2013] Gramfort, A., Thirion, B., and Varoquaux, G. (2013). Identifying predictive regions from fMRI with TV-L1 prior. In *Proceedings of the 2013 International Workshop on Pattern Recognition in Neuroimaging*, PRNI '13, pages 17–20, Washington, DC, USA. IEEE Computer Society.
- [Gray et al., 2012] Gray, K. R., Wolz, R., Heckemann, R. A., Aljabar, P., Hammers, A., and Rueckert, D. (2012). Multi-region analysis of longitudinal FDG-PET for the classification of Alzheimer's disease. *NeuroImage*, 60(1):221–229.
- [Guyon and Elisseeff, 2003] Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 3:1157–1182.
- [Guyon et al., 2002] Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Mach. Learn.*, 46(1-3):389–422.
- [Hammers et al., 2003] Hammers, A., Allom, R., Koepp, M. J., Free, S. L., Myers, R., Lemieux, L., Mitchell, T. N., Brooks, D. J., and Duncan, J. S. (2003). Three-dimensional

- maximum probability atlas of the human brain, with particular reference to the temporal lobe. *Human Brain Mapping*, 19(4):224–247.
- [Hebert et al., 2003] Hebert, L. E., Scherr, P. A., Bienias, J. L., Bennett, D. A., and Evans, D. A. (2003). Alzheimer disease in the US population: prevalence estimates using the 2000 census. *Archives of neurology*, 60(8):1119–1122.
- [Heckemann et al., 2006] Heckemann, R. A., Hajnal, J. V., Aljabar, P., Rueckert, D., and Hammers, A. (2006). Automatic anatomical brain MRI segmentation combining label propagation and decision fusion. *NeuroImage*, 33(1):115–126.
- [Herholz et al., 2002] Herholz, K., Salmon, E., Perani, D., Baron, J.-C., Holthoff, V., Frölich, L., Schönknecht, P., Ito, K., Mielke, R., Kalbe, E., Zündorf, G., Delbeuck, X., Pelati, O., Anchisi, D., Fazio, F., Kerrouche, N., Desgranges, B., Eustache, F., Beuthien-Baumann, B., Menzel, C., Schröder, J., Kato, T., Arahata, Y., Henze, M., and Heiss, W.-D. (2002). Discrimination between alzheimer dementia and controls by automated analysis of multicenter FDG PET. *NeuroImage*, 17(1):302–316.
- [Hinrichs et al., 2011] Hinrichs, C., Singh, V., Xu, G., and Johnson, S. C. (2011). Predictive markers for AD in a multi-modality framework: an analysis of MCI progression in the ADNI population. *NeuroImage*, 55(2):574–589.
- [Hoang Duc et al., 2013] Hoang Duc, A. K., Modat, M., Leung, K. K., Cardoso, M. J., Barnes, J., Kadir, T., Ourselin, S., and Alzheimer’s Disease Neuroimaging Initiative (2013). Using manifold learning for atlas selection in multi-atlas segmentation. *PloS one*, 8(8):e70059.
- [Holtzman, 2011] Holtzman, D. M. (2011). CSF biomarkers for Alzheimer’s disease: current utility and potential future use. *Neurobiology of Aging*, 32, Supplement 1(0):S4–S9.
- [Hotelling, 1936] Hotelling, H. (1936). Relations Between Two Sets of Variates. *Biometrika*, 28(3/4):321–377.
- [Jack et al., 2013] Jack, C. R., Knopman, D. S., Jagust, W. J., Petersen, R. C., Weiner, M. W., Aisen, P. S., Shaw, L. M., Vemuri, P., Wiste, H. J., Weigand, S. D., Lesnick, T. G., Pankratz, V. S., Donohue, M. C., and Trojanowski, J. Q. (2013). Update on hypothetical model of Alzheimer’s disease biomarkers. *Lancet neurology*, 12(2):207–216.
- [Jack et al., 2010] Jack, C. R., Knopman, D. S., Jagust, W. J., Shaw, L. M., Aisen, P. S., Weiner, M. W., Petersen, R. C., and Trojanowski, J. Q. (2010). Hypothetical model of dynamic biomarkers of the Alzheimer’s pathological cascade. *Lancet neurology*, 9(1):119.

- [Jack et al., 1999] Jack, C. R., Petersen, R. C., Xu, Y. C., O'Brien, P. C., Smith, G. E., Ivnik, R. J., Boeve, B. F., Waring, S. C., Tangalos, E. G., and Kokmen, E. (1999). Prediction of AD with MRI-Based hippocampal volume in mild cognitive impairment. *Neurology*, 52(7):1397–1403.
- [Jolliffe, 2005] Jolliffe, I. (2005). Principal component analysis. In *Encyclopedia of Statistics in Behavioral Science*. John Wiley & Sons, Ltd.
- [Jones et al., 2000] Jones, S. E., Buchbinder, B. R., and Aharon, I. (2000). Three-dimensional mapping of cortical thickness using laplace's equation. *Human Brain Mapping*, 11(1):12–32.
- [Joshi et al., 2009] Joshi, A., Koeppe, R. A., and Fessler, J. A. (2009). Reducing between scanner differences in multi-center PET studies. *NeuroImage*, 46(1).
- [Kim et al., 2005] Kim, J. S., Singh, V., Lee, J. K., Lerch, J., Ad-Dab'bagh, Y., MacDonald, D., Lee, J. M., Kim, S. I., and Evans, A. C. (2005). Automated 3-D extraction and evaluation of the inner and outer cortical surfaces using a laplacian map and partial volume effect classification. *NeuroImage*, 27(1):210–221.
- [Kist and Hastie, 1995] Kist, P. and Hastie, I. R. (1995). Alzheimer's disease. *Postgraduate medical journal*, 71(834):204–205.
- [Klöppel et al., 2008] Klöppel, S., Stonnington, C. M., Chu, C., Draganski, B., Scahill, R. I., Rohrer, J. D., Fox, N. C., Jack, C. R., Ashburner, J., and Frackowiak, R. S. J. (2008). Automatic classification of MR scans in Alzheimer's disease. *Brain*, 131(3):681–689.
- [Kohavi, 1995] Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2, IJCAI'95*, pages 1137–1143, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.
- [Kohonen, 1990] Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, 78(9):1464–1480.
- [Kriegeskorte et al., 2009] Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. F., and Baker, C. I. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. *Nature Neuroscience*, 12(5):535–540.
- [L J P Van Der Maaten, 2007] L J P Van Der Maaten, E. O. P. (2007). Dimensionality reduction: A comparative review. *Journal of Machine Learning Research - JMLR*.

- [Landau et al., 2010] Landau, S. M., Harvey, D., Madison, C. M., Reiman, E. M., Foster, N. L., Aisen, P. S., Petersen, R. C., Shaw, L. M., Trojanowski, J. Q., Jack, C. R., Weiner, M. W., and Jagust, W. J. (2010). Comparing predictors of conversion and decline in mild cognitive impairment. *Neurology*, 75(3):230–238.
- [Lashuel et al., 2002] Lashuel, H. A., Hartley, D. M., Balakhaneh, D., Aggarwal, A., Teichberg, S., and Callaway, D. J. E. (2002). New class of inhibitors of amyloid-beta fibril formation IMPLICATIONS FOR THE MECHANISM OF PATHOGENESIS IN ALZHEIMER'S DISEASE ., *Journal of Biological Chemistry*, 277(45):42881–42890.
- [Lee et al., 2005] Lee, H.-g., Perry, G., Moreira, P. I., Garrett, M. R., Liu, Q., Zhu, X., Takeda, A., Nunomura, A., and Smith, M. A. (2005). Tau phosphorylation in Alzheimer's disease: pathogen or protector? *Trends in Molecular Medicine*, 11(4):164–169.
- [Lerch et al., 2008] Lerch, J. P., Pruessner, J., Zijdenbos, A. P., Collins, D. L., Teipel, S. J., Hampel, H., and Evans, A. C. (2008). Automated cortical thickness measurements from MRI can accurately separate Alzheimer's patients from normal elderly controls. *Neurobiology of Aging*, 29(1):23–30.
- [Leung et al., 2011] Leung, K. K., Barnes, J., Modat, M., Ridgway, G. R., Bartlett, J. W., Fox, N. C., and Ourselin, S. (2011). Brain maps: An automated, accurate and robust brain extraction technique using a template library. *NeuroImage*, 55(3):1091 – 1108.
- [Leung et al., 2012] Leung, K. K., Ridgway, G. R., Ourselin, S., and Fox, N. C. (2012). Consistent multi-time-point brain atrophy estimation from the boundary shift integral. *NeuroImage*, 59(4):3995–4005.
- [Li et al., 2013] Li, Y., Qin, Y., Chen, X., and Li, W. (2013). Exploring the functional brain network of Alzheimer's disease: Based on the computational experiment. *PLoS ONE*, 8(9):e73186.
- [Lipton, 2006] Lipton, S. A. (2006). Paradigm shift in neuroprotection by NMDA receptor blockade: Memantine and beyond. *Nature Reviews Drug Discovery*, 5(2):160–170.
- [Liu et al., 2014] Liu, F., Wee, C.-Y., Chen, H., and Shen, D. (2014). Inter-modality relationship constrained multi-modality multi-task feature selection for Alzheimer's disease and mild cognitive impairment identification. *NeuroImage*, 84(0):466 – 475.
- [Liu et al., 2013] Liu, F., Zhou, L., Shen, C., and Yin, J. (2013). Multiple kernel learning in the primal for multi-modal Alzheimer's disease classification. arXiv e-print 1310.0890.

- [Lötjönen et al., 2010] Lötjönen, J. M., Wolz, R., Koikkalainen, J. R., Thurfjell, L., Waldemar, G., Soininen, H., Rueckert, D., and Alzheimer's Disease Neuroimaging Initiative (2010). Fast and robust multi-atlas segmentation of brain magnetic resonance images. *NeuroImage*, 49(3):2352–2365.
- [Lovestone et al., 2009] Lovestone, S., Francis, P., Kloszewska, I., Mecocci, P., Simmons, A., Soininen, H., Spenger, C., Tsolaki, M., Vellas, B., Wahlund, L.-O., Ward, M., and AddNeuroMed Consortium (2009). AddNeuroMed—the european collaboration for the discovery of novel biomarkers for Alzheimer's disease. *Annals of the New York Academy of Sciences*, 1180:36–46.
- [Madsen et al., 2010] Madsen, S., Ho, A., Hua, X., Saharan, P., Toga, A., Jr, C. J., Weiner, M., and Thompson, P. (2010). 3D maps localize caudate nucleus atrophy in 400 Alzheimer's disease, mild cognitive impairment, and healthy elderly subjects. *Neurobiology of Aging*, 31(8):1312 – 1325. Alzheimer's Disease Neuroimaging Initiative (ADNI) Studies.
- [Magnin et al., 2009] Magnin, B., Mesrob, L., Kinkingnehun, S., Péligrini-Issac, M., Colliot, O., Sarazin, M., Dubois, B., Lehericy, S., and Benali, H. (2009). Support vector machine-based classification of Alzheimer's disease from whole-brain anatomical MRI. *Neuroradiology*, 51(2):73–83.
- [Mapstone et al., 2014] Mapstone, M., Cheema, A. K., Fiandaca, M. S., Zhong, X., Mhyre, T. R., MacArthur, L. H., Hall, W. J., Fisher, S. G., Peterson, D. R., Haley, J. M., Nazar, M. D., Rich, S. A., Berlau, D. J., Peltz, C. B., Tan, M. T., Kawas, C. H., and Federoff, H. J. (2014). Plasma phospholipids identify antecedent memory impairment in older adults. *Nat Med*, 20(4):415–418.
- [Marquand et al., 2010] Marquand, A., Howard, M., Brammer, M., Chu, C., Coen, S., and Mourão-Miranda, J. (2010). Quantitative prediction of subjective pain intensity from whole-brain fMRI data using Gaussian processes. *NeuroImage*, 49(3):2178–2189.
- [Marquand et al., 2013] Marquand, A. F., Filippone, M., Ashburner, J., Girolami, M., Mourao-Miranda, J., Barker, G. J., Williams, S. C. R., Leigh, P. N., and Blain, C. R. V. (2013). Automated, high accuracy classification of Parkinsonian disorders: A pattern recognition approach. *PLoS ONE*, 8(7):e69237.
- [Mazziotta et al., 2001] Mazziotta, J., Toga, A., Evans, A., Fox, P., Lancaster, J., Zilles, K., Woods, R., Paus, T., Simpson, G., Pike, B., Holmes, C., Collins, L., Thompson, P., MacDon-

- ald, D., Iaconi, M., Schormann, T., Amunts, K., Palomero-Gallagher, N., Geyer, S., Parsons, L., Narr, K., Kabani, N., Le Goualher, G., Boomsma, D., Cannon, T., Kawashima, R., and Mazoyer, B. (2001). A probabilistic atlas and reference system for the human brain: International consortium for brain mapping (ICBM). *Philosophical Transactions of the Royal Society of London. Series B*, 356(1412):1293–1322.
- [McKhann et al., 2011] McKhann, G. M., Knopman, D. S., Chertkow, H., Hyman, B. T., Jack, C. R., Kawas, C. H., Klunk, W. E., Koroshetz, W. J., Manly, J. J., Mayeux, R., Mohs, R. C., Morris, J. C., Rossor, M. N., Scheltens, P., Carrillo, M. C., Thies, B., Weintraub, S., and Phelps, C. H. (2011). The diagnosis of dementia due to Alzheimer’s disease: Recommendations from the national institute on aging-Alzheimer’s association workgroups on diagnostic guidelines for Alzheimer’s disease. *Alzheimer’s & Dementia: The Journal of the Alzheimer’s Association*, 7(3):263–269.
- [McNemar, 1947] McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.
- [Meek et al., 1998] Meek, P. D., McKeithan, K., and Schumock, G. T. (1998). Economic considerations in Alzheimer’s disease. *Pharmacotherapy*, 18(2 Pt 2):68–73; discussion 79–82.
- [Minka, 2001] Minka, T. P. (2001). Expectation propagation for approximate Bayesian inference. In *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, UAI ’01, pages 362–369, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.
- [Misra et al., 2009] Misra, C., Fan, Y., and Davatzikos, C. (2009). Baseline and longitudinal patterns of brain atrophy in MCI patients, and their use in prediction of short-term conversion to AD: results from ADNI. *NeuroImage*, 44(4):1415–1422.
- [Mitchell, 1997] Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill, Inc., New York, NY, USA, 1 edition.
- [Modat et al., 2012] Modat, M., Daga, P., Cardoso, M., Ourselin, S., Ridgway, G., and Ashburner, J. (2012). Parametric non-rigid registration using a stationary velocity field. In *2012 IEEE Workshop on Mathematical Methods in Biomedical Image Analysis (MMBIA)*, pages 145–150.
- [Modat et al., 2010] Modat, M., Ridgway, G. R., Taylor, Z. A., Lehmann, M., Barnes, J., Hawkes, D. J., Fox, N. C., and Ourselin, S. (2010). Fast free-form deformation using graphics processing units. *Computer Methods and Programs in Biomedicine*, 98(3):278–284.

- [Morra et al., 2008] Morra, J. H., Tu, Z., Apostolova, L. G., Green, A. E., Avedissian, C., Madsen, S. K., Parikshak, N., Hua, X., Toga, A. W., Jack, Jr, C. R., Weiner, M. W., Thompson, P. M., and Alzheimer’s Disease Neuroimaging Initiative (2008). Validation of a fully automated 3D hippocampal segmentation method using subjects with Alzheimer’s disease mild cognitive impairment, and elderly controls. *NeuroImage*, 43(1):59–68.
- [Mosconi et al., 2010] Mosconi, L., Berti, V., Glodzik, L., Pupi, A., De Santi, S., and de Leon, M. J. (2010). Pre-clinical detection of Alzheimer’s disease using FDG-PET, with or without amyloid imaging. *Journal of Alzheimer’s disease: JAD*, 20(3):843–854.
- [Naish-Guzman and Holden, 2007] Naish-Guzman, A. and Holden, S. (2007). The generalized fitc approximation. In *Advances in Neural Information Processing Systems*, pages 1057–1064.
- [Natterer, 1986] Natterer, F. (1986). Computerized tomography. In *The Mathematics of Computerized Tomography*, pages 1–8. Vieweg+Teubner Verlag.
- [Neal, 1996] Neal, R. M. (1996). *Bayesian Learning for Neural Networks*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.
- [Newcombe, 1998] Newcombe, R. G. (1998). Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine*, 17(8):857–872.
- [Nho et al., 2010] Nho, K., Shen, L., Kim, S., Risacher, S. L., West, J. D., Foroud, T., Jack, C. R., Weiner, M. W., and Saykin, A. J. (2010). Automatic prediction of conversion from mild cognitive impairment to probable Alzheimer’s disease using structural magnetic resonance imaging. *AMIA Annual Symposium Proceedings*, 2010:542–546.
- [Nickisch and Rasmussen, 2008] Nickisch, H. and Rasmussen, C. E. (2008). Approximations for binary Gaussian process classification. *Journal of Machine Learning Research*, 9:2035–2078.
- [Nir et al., 2013] Nir, T. M., Jahanshad, N., Villalon-Reina, J. E., Toga, A. W., Jack, C. R., Weiner, M. W., and Thompson, P. M. (2013). Effectiveness of regional DTI measures in distinguishing Alzheimer’s disease, MCI, and normal aging. *NeuroImage: Clinical*, 3.
- [Nordberg et al., 2010] Nordberg, A., Rinne, J. O., Kadir, A., and Långström, B. (2010). The use of PET in alzheimer disease. *Nature Reviews Neurology*, 6(2):78–87.

- [Peng et al., 2005] Peng, H., Long, F., and Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238.
- [Petersen et al., 1999] Petersen, R. C., Smith, G. E., Waring, S. C., Ivnik, R. J., Tangalos, E. G., and Kokmen, E. (1999). Mild cognitive impairment: clinical characterization and outcome. *Archives of Neurology*, 56(3):303–308.
- [Platt, 2000] Platt, J. (2000). Probabilities for SV machines. In *Advances in Large-Margin Classifiers*, pages 61–74. MIT Press, Cambridge, MA.
- [Querbes et al., 2009] Querbes, O., Aubry, F., Pariente, J., Lotterie, J.-A., Demonet, J.-F., Duret, V., Puel, M., Berry, I., Fort, J.-C., Celsis, P., and The Alzheimer’s Disease Neuroimaging Initiative (2009). Early diagnosis of Alzheimer’s disease using cortical thickness: impact of cognitive reserve. *Brain*, 132(8):2036–2047.
- [Rahimi and Recht, 2007] Rahimi, A. and Recht, B. (2007). Random features for large-scale kernel machines. In *In Neural Information Processing Systems*.
- [Ranginwala et al., 2008] Ranginwala, N. A., Hynan, L. S., Weiner, M. F., and White, 3rd, C. L. (2008). Clinical criteria for the diagnosis of Alzheimer’s disease: still good after all these years. *The American journal of geriatric psychiatry: official journal of the American Association for Geriatric Psychiatry*, 16(5):384–388.
- [Rasmussen and Williams, 2006] Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. MIT Press.
- [Raz et al., 2005] Raz, N., Lindenberger, U., Rodrigue, K. M., Kennedy, K. M., Head, D., Williamson, A., Dahle, C., Gerstorf, D., and Acker, J. D. (2005). Regional brain changes in aging healthy adults: General trends, individual differences and modifiers. *Cerebral Cortex*, 15(11):1676–1689.
- [Reisberg et al., 2003] Reisberg, B., Doody, R., Stöffler, A., Schmitt, F., Ferris, S., and Möbius, H. J. (2003). Memantine in moderate-to-severe Alzheimer’s disease. *New England Journal of Medicine*, 348(14):1333–1341.
- [Rohr, 2000] Rohr, K. (2000). Elastic registration of multimodal medical images: A survey. *KI*, 14:11–17.

- [Rosen et al., 1984] Rosen, W. G., Mohs, R. C., and Davis, K. L. (1984). A new rating scale for Alzheimer’s disease. *The American journal of psychiatry*, 141(11):1356–1364.
- [Rowe et al., 2010] Rowe, C. C., Ellis, K. A., Rimajova, M., Bourgeat, P., Pike, K. E., Jones, G., Frupp, J., Tochon-Danguy, H., Morandau, L., O’Keefe, G., Price, R., Raniga, P., Robins, P., Acosta, O., Lenzo, N., Szoeki, C., Salvado, O., Head, R., Martins, R., Masters, C. L., Ames, D., and Villemagne, V. L. (2010). Amyloid imaging results from the australian imaging, biomarkers and lifestyle (AIBL) study of aging. *Neurobiology of Aging*, 31(8):1275–1283.
- [Rueckert et al., 1999] Rueckert, D., Sonoda, L. I., Hayes, C., Hill, D. L., Leach, M. O., and Hawkes, D. J. (1999). Nonrigid registration using free-form deformations: application to breast MR images. *IEEE transactions on medical imaging*, 18(8):712–721.
- [Sabuncu, 2013] Sabuncu, M. (2013). A Bayesian algorithm for image-based time-to-event prediction. In *Machine Learning in Medical Imaging*, volume 8184 of *Lecture Notes in Computer Science*, pages 74–81. Springer International Publishing.
- [Sabuncu and Konukoglu, 2014] Sabuncu, M. R. and Konukoglu, E. (2014). Clinical prediction from structural brain MRI scans: A large-scale empirical study. *Neuroinformatics*, pages 1–16.
- [Sabuncu and Leemput, 2012] Sabuncu, M. R. and Leemput, K. V. (2012). The relevance voxel machine (rvoxm): A self-tuning Bayesian model for informative image-based prediction. *IEEE Trans. Med. Imaging*, 31(12):2290–2306.
- [Selkoe, 2004] Selkoe, D. J. (2004). Cell biology of protein misfolding: The examples of Alzheimer’s and Parkinson’s diseases. *Nature Cell Biology*, 6(11):1054–1061.
- [Shannon, 1948] Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal, The*, 27(3):379–423.
- [Shen and Davatzikos, 2002] Shen, D. and Davatzikos, C. (2002). HAMMER: hierarchical attribute matching mechanism for elastic registration. *IEEE Transactions on Medical Imaging*, 21(11):1421–1439.
- [Shi et al., 2007] Shi, Y., Thompson, P. M., de Zubicaray, G. I., Rose, S. E., Tu, Z., Dinov, I., and Toga, A. W. (2007). Direct mapping of hippocampal surfaces with intrinsic shape context. *NeuroImage*, 37(3):792–807.

- [Singh et al., 2014] Singh, N., Thomas Fletcher, P., Samuel Preston, J., King, R. D., Marron, J. S., Weiner, M. W., and Joshi, S. (2014). Quantifying anatomical shape variations in neurological disorders. *Medical Image Analysis*, 18(3):616–633.
- [Singh et al., 2012] Singh, N., Wang, A., Sankaranarayanan, P., Fletcher, P., and Joshi, S. (2012). Genetic, structural and functional imaging biomarkers for early detection of conversion from MCI to AD. In Ayache, N., Delingette, H., Golland, P., and Mori, K., editors, *Medical Image Computing and Computer-Assisted Intervention MICCAI 2012*, volume 7510 of *Lecture Notes in Computer Science*, pages 132–140. Springer Berlin / Heidelberg.
- [Stahl, 2000] Stahl, S. M. (2000). The new cholinesterase inhibitors for Alzheimer’s disease, part 2: illustrating their mechanisms of action. *The Journal of clinical psychiatry*, 61(11):813–814.
- [Studholme et al., 1999] Studholme, C., Hill, D. L. G., and Hawkes, D. J. (1999). An overlap invariant entropy measure of 3D medical image alignment. *Pattern Recognition*, 32(1):71–86.
- [Suk et al., 2013] Suk, H.-I., Lee, S.-W., and Shen, D. (2013). Latent feature representation with stacked auto-encoder for AD/MCI diagnosis. *Brain Structure and Function*, pages 1–19.
- [Suk and Shen, 2013] Suk, H.-I. and Shen, D. (2013). Deep learning-based feature representation for AD/MCI classification. In Mori, K., Sakuma, I., Sato, Y., Barillot, C., and Navab, N., editors, *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2013*, volume 8150 of *Lecture Notes in Computer Science*, pages 583–590. Springer Berlin Heidelberg.
- [Sunderland T et al., 2003] Sunderland T, Linker G, Mirza N, and et al (2003). Decreased β -amyloid_{1–42} and increased tau levels in cerebrospinal fluid of patients with alzheimer disease. *JAMA*, 289(16):2094–2103.
- [Tibshirani, 1994] Tibshirani, R. (1994). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288.
- [Tipping, 2001] Tipping, M. E. (2001). Sparse Bayesian learning and the relevance vector machine. *J. Mach. Learn. Res.*, 1:211–244.
- [Trygg and Wold, 2002] Trygg, J. and Wold, S. (2002). Orthogonal projections to latent structures (o-PLS). *Journal of Chemometrics*, 16(3):119–128.

- [Turing, 1950] Turing, A. M. (1950). I.computing machinery and intelligence. *Mind*, LIX(236):433–460.
- [Tzourio-Mazoyer et al., 2002] Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., and Joliot, M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage*, 15(1):273–289.
- [V. Vapnik, 1963] V. Vapnik, A. L. (1963). Pattern recognition using generalized portrait method. *Automation and Remote Control*, 24:774–780.
- [van Bruggen et al., 2012] van Bruggen, T., Stieltjes, B., Thomann, P. A., Parzer, P., Meinzer, H.-P., and Fritzsche, K. H. (2012). Do Alzheimer-specific microstructural changes in mild cognitive impairment predict conversion? *Psychiatry Research: Neuroimaging*, 203(2-3):184–193.
- [Van Leemput et al., 1999a] Van Leemput, K., Maes, F., Vandermeulen, D., and Suetens, P. (1999a). Automated model-based bias field correction of MR images of the brain. *IEEE TRANSACTIONS ON MEDICAL IMAGING*, 18:885–896.
- [Van Leemput et al., 1999b] Van Leemput, K., Maes, F., Vandermeulen, D., and Suetens, P. (1999b). Automated model-based tissue classification of MR images of the brain. *IEEE transactions on medical imaging*, 18(10):897–908.
- [Vanhatalo et al., 2013] Vanhatalo, J., Riihimäki, J., Hartikainen, J., Jylänki, P., Tolvanen, V., and Vehtari, A. (2013). GPstuff: Bayesian modeling with Gaussian processes. *J. Mach. Learn. Res.*, 14(1):1175–1179.
- [Vardi et al., 1985] Vardi, Y., Shepp, L. A., and Kaufman, L. (1985). A statistical model for positron emission tomography. *Journal of the American Statistical Association*, 80(389):8–20.
- [Varma and Simon, 2006] Varma, S. and Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7(1):91.
- [Vemuri et al., 2008] Vemuri, P., Gunter, J. L., Senjem, M. L., Whitwell, J. L., Kantarci, K., Knopman, D. S., Boeve, B. F., Petersen, R. C., and Jack Jr., C. R. (2008). Alzheimer’s disease diagnosis in individual subjects using structural MR images: Validation studies. *NeuroImage*, 39(3):1186–1197.

- [Vemuri et al., 2009] Vemuri, P., Wiste, H. J., Weigand, S. D., Shaw, L. M., Trojanowski, J. Q., Weiner, M. W., Knopman, D. S., Petersen, R. C., and Jack, Jr, C. R. (2009). MRI and CSF biomarkers in normal, MCI, and AD subjects: predicting future clinical change. *Neurology*, 73(4):294–301.
- [Vounou et al., 2012] Vounou, M., Janousova, E., Wolz, R., Stein, J. L., Thompson, P. M., Rueckert, D., and Montana, G. (2012). Sparse reduced-rank regression detects genetic associations with voxel-wise longitudinal phenotypes in Alzheimer’s disease. *NeuroImage*, 60(1):700–716.
- [Walhovd et al., 2010] Walhovd, K. B., Fjell, A. M., Brewer, J., McEvoy, L. K., Fennema-Notestine, C., Hagler, D. J., Jennings, R. G., Karow, D., Dale, A. M., and Alzheimer’s Disease Neuroimaging Initiative (2010). Combining MR imaging, positron-emission tomography, and CSF biomarkers in the diagnosis and prognosis of Alzheimer’s disease. *AJNR. American journal of neuroradiology*, 31(2):347–354.
- [Wang et al., 2013] Wang, Z., Das, S. R., Xie, S. X., Arnold, S. E., Detre, J. A., and Wolk, D. A. (2013). Arterial spin labeled MRI in prodromal Alzheimer’s disease: A multi-site study. *NeuroImage: Clinical*, 2(0):630–636.
- [Warfield et al., 2004] Warfield, S. K., Zou, K. H., and Wells, W. M. (2004). Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. *IEEE Transactions on Medical Imaging*, 23(7):903–921.
- [Wee et al., 2012] Wee, C.-Y., Yap, P.-T., Zhang, D., Denny, K., Browndyke, J. N., Potter, G. G., Welsh-Bohmer, K. A., Wang, L., and Shen, D. (2012). Identification of MCI individuals using structural and functional connectivity networks. *NeuroImage*, 59(3):2045–2056.
- [Wells et al., 1996] Wells, 3rd, W. M., Viola, P., Atsumi, H., Nakajima, S., and Kikinis, R. (1996). Multi-modal volume registration by maximization of mutual information. *Medical image analysis*, 1(1):35–51.
- [Westman et al., 2012] Westman, E., Aguilar, C., Muehlboeck, J.-S., and Simmons, A. (2012). Regional magnetic resonance imaging measures for multivariate analysis in Alzheimer’s disease and mild cognitive impairment. *Brain Topography*, 26(1):9–23.
- [Williams and Barber, 1998] Williams, C. K. I. and Barber, D. (1998). Bayesian classification with Gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1342–1351.

- [Wischik et al., 2008] Wischik, C. M., Bentham, P., Wischik, D. J., and Seng, K. M. (2008). Tau aggregation inhibitor (tai) therapy with remberTM arrests disease progression in mild and moderate Alzheimer’s disease over 50 weeks. *Alzheimer’s & Dementia: The Journal of the Alzheimer’s Association*, 4(4):T167–T167.
- [Wolpert, 1996] Wolpert, D. H. (1996). The lack of a priori distinctions between learning algorithms. *Neural Comput.*, 8(7):1341–1390.
- [Wolz et al., 2011a] Wolz, R., Aljabar, P., Hajnal, J., Lotjonen, J., and Rueckert, D. (2011a). Manifold learning combining imaging with non-imaging information. In *2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 1637–1640.
- [Wolz et al., 2010] Wolz, R., Aljabar, P., Hajnal, J. V., Hammers, A., Rueckert, D., and Alzheimer’s Disease Neuroimaging Initiative (2010). LEAP: learning embeddings for atlas propagation. *NeuroImage*, 49(2):1316–1325.
- [Wolz et al., 2011b] Wolz, R., Julkunen, V., Koikkalainen, J., Niskanen, E., Zhang, D. P., Rueckert, D., Soininen, H., Lötjönen, J., and the Alzheimer’s Disease Neuroimaging Initiative (2011b). Multi-method analysis of MRI images in early diagnostics of Alzheimer’s disease. *PLoS ONE*, 6(10):e25446.
- [Xu et al., 2004] Xu, Q.-S., Liang, Y.-Z., and Du, Y.-P. (2004). Monte Carlo cross-validation for selecting a model and estimating the prediction error in multivariate calibration. *Journal of Chemometrics*, 18(2):112–120.
- [Yakushev et al., 2009] Yakushev, I., Hammers, A., Fellgiebel, A., Schmidtman, I., Scheurich, A., Buchholz, H.-G., Peters, J., Bartenstein, P., Lieb, K., and Schreckenberger, M. (2009). SPM-based count normalization provides excellent discrimination of mild Alzheimer’s disease and amnesic mild cognitive impairment from healthy aging. *NeuroImage*, 44(1):43–50.
- [Yankner et al., 1990] Yankner, B. A., Duffy, L. K., and Kirschner, D. A. (1990). Neurotrophic and neurotoxic effects of amyloid beta protein: reversal by tachykinin neuropeptides. *Science*, 250(4978):279–282.
- [Ye et al., 2012] Ye, J., Farnum, M., Yang, E., Verbeeck, R., Lobanov, V., Raghavan, N., Novak, G., DiBernardo, A., and Narayan, V. (2012). Sparse learning and stability selection for predicting MCI to AD conversion using baseline ADNI data. *BMC Neurology*, 12(1).

- [Yin et al., 2007] Yin, Y. I., Bassit, B., Zhu, L., Yang, X., Wang, C., and Li, Y.-M. (2007). Gamma-secretase substrate concentration modulates the Amyloid-beta42/Amyloid-beta40 ratio: implications for Alzheimer's disease. *Journal of Biological Chemistry*, 282(32):23639–23644.
- [Young et al., 2013a] Young, J., Ashburner, J., and Ourselin, S. (2013a). Wrapper methods to correct mislabelled training data. In *2013 International Workshop on Pattern Recognition in Neuroimaging (PRNI)*, pages 170–173.
- [Young et al., 2012] Young, J., Modat, M., Cardoso, M., Ashburner, J., and Ourselin, S. (2012). Classification of Alzheimer's disease patients and controls with Gaussian processes. In *2012 9th IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1523–1526.
- [Young et al., 2013b] Young, J., Modat, M., Cardoso, M., Ashburner, J., and Ourselin, S. (2013b). An oblique approach to prediction of conversion to Alzheimer's disease with multikernel Gaussian processes. In *3rd NIPS Workshop on Machine Learning and Interpretation in NeuroImaging*.
- [Young et al., 2013c] Young, J., Modat, M., Cardoso, M. J., Mendelson, A., Cash, D., and Ourselin, S. (2013c). Accurate multimodal probabilistic prediction of conversion to Alzheimer's disease in patients with mild cognitive impairment. *NeuroImage: Clinical*, 2(0):735–745.
- [Yuan et al., 2012] Yuan, L., Wang, Y., Thompson, P. M., Narayan, V. A., and Ye, J. (2012). Multi-source feature learning for joint analysis of incomplete multiple heterogeneous neuroimaging data. *NeuroImage*, 61(3):622–632.
- [Zasadny and Wahl, 1993] Zasadny, K. R. and Wahl, R. L. (1993). Standardized uptake values of normal tissues at PET with 2-[fluorine-18]-fluoro-2-deoxy-d-glucose: variations with body weight and a method for correction. *Radiology*, 189(3):847–850.
- [Zekry et al., 2002] Zekry, D., Hauw, J.-J., and Gold, G. (2002). Mixed dementia: Epidemiology, diagnosis, and treatment. *Journal of the American Geriatrics Society*, 50(8):1431–1438.
- [Zhang et al., 2011] Zhang, D., Wang, Y., Zhou, L., Yuan, H., and Shen, D. (2011). Multi-modal classification of Alzheimer's disease and mild cognitive impairment. *NeuroImage*, 55(3):856–867.
- [Zou and Hastie, 2005] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320.