Registration and Machine Learning Methods for Brain Imaging of Alzheimer's Disease

Liam Cattell

Brasenose College

University of Oxford

Supervisors

Prof. Julia A. Schnabel

Dr Chloe Hutton

A thesis submitted for the degree of

Doctor of Philosophy in Engineering Science

Hilary Term 2016

Abstract

Alzheimer's disease (AD) is an irreversible neurodegenerative disorder, characterised by memory loss and reduced cognitive function. One of the pathologic biomarkers of AD is the presence of neuritic plaques composed of beta amyloid peptides. Positron emission tomography (PET) imaging is increasingly used to assess the amyloid burden in patients with dementia, and therefore, tools for analysing amyloid PET scans could aid clinicians when diagnosing and treating AD. For this reason, this thesis presents novel methods for quantifying and classifying amyloid in brain PET images. The most important contributions made in this thesis are: 1) The formulation of a dual-modality deformable registration method, based on the demons framework, for simultaneously aligning PET and magnetic resonance (MR) images to a template space, in order to calculate standardised uptake value ratios (SUVRs). 2) The proposal of a novel machine learning method for classifying brain amyloid status, based on histograms of oriented image gradients. 3) An assessment of convolutional deep belief networks for amyloid status classification. 4) An initial investigation towards a complete classification framework for AD using multi-graph learning. The results of the registration experiments suggest that the proposed joint PET-MR registration method can perform as well as similar single-modality methods in terms of registration accuracy, and could provide an improved separation between populations of AD patients and healthy controls when used in the calculation of SUVRs. Although the investigation into convolutional deep belief networks indicated that, at present, they are impractical for amyloid status classification, our novel amyloid status classification method achieved a higher classification accuracy than two other established methods. Moreover, unlike SUVR, our proposed machine learning method does not require specific knowledge of neuroanatomy and can be applied to multiple amyloid PET tracers without substantial recalibration.

I eat my peas with honey;

I've done it all my life.

It makes the peas taste funny,

But it keeps them on the knife.

– Anonymous

Acknowledgements

This work was made possible by an Industrial CASE studentship (11440394) jointly funded by the EPSRC and Siemens Molecular Imaging. However, I would like to use this opportunity to express my gratitude to a number people who have helped me over the last few years.

Firstly, I am indebted to Julia A. Schnabel and Chloe Hutton for their supervision and guidance during this research. Their support has been essential, especially when making the decision to change the focus of my research from image registration to machine learning midway through my project. Julia, I thank you for giving me the opportunity to do this research, and for introducing me to some truly brilliant people, both in Oxford and further afield. Chloe, thank you for your endless enthusiasm and encouragement, and countless discussions about my work.

I must also extend my gratitude to my friends and colleagues at the IBME and Siemens. Not only did they make the offices a more enjoyable place to work, but they dispensed invaluable advice about research, and were always interested to hear about my (lack of) progress. I also thank my friends at Brasenose College (which has been my home for almost eight years) for allowing me to escape the world of biomedical engineering and learn about other fascinating topics, including parasitoid wasps and Kant's critiques of reason. I would also like to thank my Mill St housemates. You graciously allowed me to indulge in many disruptive side-projects, from baking outlandish cakes to renovating a 20-year-old campervan, and you made the tiresome parts of my DPhil infinitely more enjoyable. To my friends elsewhere, I thank you for the late-night Skype calls and pub chats that put this work in perspective.

I owe an enormous amount of thanks to my parents for their everlasting support and love, and to my grandfather for reading this thesis at least once.

Finally, Chelsea, thank you for everything. This is for you.

Contents

List of Tables

List of Figures

List of Algorithms

Chapter 1

Introduction

The overall prevalence of dementia in the Western world rises steeply with increasing age, from 1% in persons aged 60-65 years to 45% in persons aged 90-95 years [Roberts et al. 2008]. Consequently, ageing populations pose a significant public health burden. In the final stages of Alzheimer's disease (AD), the most common type of dementia, sufferers lose their ability to communicate, fail to recognise family and friends, and require around-the-clock care. Ultimately, AD is fatal.

The 2011 criteria and guidelines for diagnosing AD recognise three stages of the disease: preclinical AD, mild cognitive impairment due to AD, and dementia due to AD [McKhann et al. 2011]. Despite attempts to have clear patient categories, AD affects people in different ways and at different rates. As a result, the boundaries between the groups are very blurred. Due to these difficulties in diagnosis, AD is underdiagnosed, and up to half of the estimated people with AD may not know they have it [Alzheimer's Association 2014].

There are two microscopic lesions associated with AD: abnormal fibrous inclusions, called neurofibrillary tangles, within neurons, and neuritic plaques composed of accumulations of amyloid-β (Aβ) peptides [Perl 2010]. In the early 1980s, clinicians were unable to quantify these microscopic lesions in living patients, so the original 1984 criteria for the diagnosis of probable AD required confirmation of dementia by neuropsychological testing, and evidence of cerebral atrophy from computed tomography (CT) imaging [McKhann et al. 1984]. Unfortunately, reductions

in cortical thickness are also observed in elderly individuals with normal cognitive function, making AD diagnosis by that value alone very difficult [Terry 1986].

Due to the development of positron emission tomography (PET) tracers capable of imaging Aβ, the diagnosis guidelines were revised in 2011 to incorporate a positive amyloid scan as an AD biomarker [McKhann et al. 2011]. By combining amyloid PET imaging with other modalities, such as magnetic resonance (MR) imaging, and patient metadata, a more comprehensive picture of the disease could be achieved, leading to more robust diagnosis and patient classification.

For this reason, the underlying aim of this research is to work towards the development of an amyloid PET image analysis tool for aiding the accurate and consistent diagnosis of AD. Although current clinical techniques can adequately distinguish between PET scans that are clearly amyloid positive and clearly amyloid negative, with machine learning methods it may be possible to improve the accuracy of image classification, which may increase the confidence of a clinician's diagnostic decision. To achieve this, this project will combine functional information from amyloid PET with structural information from MR, in a multi-modal setting, by firstly developing a novel spatial normalisation tool, and secondly by using machine learning techniques to derive features which can be used to distinguish between patient groups.

1.1 Thesis Overview

A brief overview of the individual chapters in this thesis, along with their corresponding research contributions, is given below:

Chapter 2 provides essential background information for the later chapters, and is split into three parts. Firstly, we discuss the pathology, biomarkers, and treatment of Alzheimer's disease, as well as current methods for amyloid quantification. Secondly, in anticipation of our research into dual-modality spatial normalisation methods in Chapters 3 and 4, we introduce the concept of image registration. More specifically, we describe the demons registration framework and its extensions in the literature. Finally, we provide an introduction to the machine learning methods,

including deep learning, that will be explored in Chapters 5-8.

Chapter 3 examines the effect of using functional and structural image information during the registration of brain PET volumes to a template. To achieve this, we propose a novel method for combined PET-MR registration. Our framework jointly registers PET-MR brain volume pairs to a template space, by combining the incremental updates from single-modality local correlation coefficient demons registrations. In addition to comparing our PET-MR registration method to single-modality registration methods, we also assess the effect of using a synthetic PET template and a template constructed from real amyloid PET data during the registration process. We show that our combined PET-MR registration method can perform at least as well as single-modality registration methods in terms of registration accuracy. Moreover, we also demonstrate that the amyloid quantification measures obtained using our PET-MR registration method could be better suited to discriminating between populations of AD patients and healthy controls, than those acquired using single-modality registration methods.

Chapter 4 presents an extension to the combined PET-MR registration method proposed in Chapter 3, by posing the registration framework as the minimisation of a dual-modality cost function. We further extend the joint PET-MR registration method to allow for a locally adaptive modality weighting that selects from a set of weighting factors, the best combination of PET and MR to give the maximum local correlation coefficient at each voxel. Our results indicate that, although our PET-MR registration method can outperform a non-linear PET-only registration method in terms of registration accuracy, the ability of the clinical amyloid quantification method to classify AD patients and healthy controls is robust to the registration method used to align the PET volume to the template space.

Chapter 5 explores a template-free approach to the classification of brain amyloid, by proposing a machine learning method that is independent of the amyloid PET tracer, and does not require specific knowledge of neuroanatomy. Our method extracts feature vectors directly from amyloid PET images, which are based on his-

tograms of oriented three-dimensional gradients, and uses them to train a classifier to distinguish between amyloid positive and amyloid negative PET scans. Without recalibrating the hyper-parameters, we apply our method to images from three different amyloid PET tracers, and achieve a classification accuracy greater than an established amyloid PET quantification method.

Chapter 6 assesses the generalisability of the amyloid status classification method proposed in Chapter 5. Using three different parameter combinations, we apply our method to data from three different amyloid tracers, and once again achieve superior classification performance to an established clinical amyloid PET quantification method.

Chapter 7 investigates the potential of using a convolutional deep belief network to classify amyloid status. We examine the effect of the network's depth on classification accuracy, and explore the idea of training the first layer of the network using non-medical image data. For our application, the results indicate that a one-layer convolutional deep belief network can achieve a higher classification accuracy than a two- or three- layer network. Moreover, we also show that a one-layer network trained using non-medical data is able to more accurately classify brain amyloid status than a similar network trained using PET images.

Chapter 8 presents investigatory research towards a complete classification framework for AD that is able to combine data from a range of clinical biomarkers. We compare three existing classification methods from the literature, and propose a novel method based on multi-graph learning. We show that all four methods are able to accurately classify healthy controls versus AD patients, but the classification accuracy decreases considerably for multiclass classification tasks. We identify drawbacks in the design and implementation of each classification framework, and consequently, we propose a set of criteria that any future work towards a complete classification framework for AD should fulfil.

Chapter 9 concludes this thesis by summarising the contributions of our work, highlighting the various accomplishments, and discussing any limitations. We will

also outline directions for future research in this field, in particular our investigation into using a convolutional deep belief network for amyloid status classification highlighted that training deep learning methods using 3D data is computationally very expensive. Therefore, in future, it would be useful to develop novel methods to reduce the training time. Not only would this make deep learning more practical for amyloid status classification, but it could have applications beyond medical image analysis, such as object detection in video sequences.

Chapter 2

Background

This chapter describes the pathology of Alzheimer's disease, how it is diagnosed clinically, and how the disease is imaged in vivo using amyloid PET imaging. The concept of image registration is also introduced, as well as its role in Alzheimer's disease diagnosis and amyloid PET quantification. Finally, several approaches to machine learning are discussed, including the use of image descriptors, support vector machines, and deep learning with restricted Boltzmann machines.

2.1 Alzheimer's Disease

Alzheimer's disease (AD) is an irreversible neurodegenerative disorder, characterised by memory loss and reduced cognitive function due to the progressive impairment of neurons and their connections. The disease invariably leads to death, there is no cure, and there is currently no proven method to stop a patient's decline [Rowe and Villemagne 2013]. AD is by far the most common form of dementia, and accounts for 50-75% of all dementia cases [ADI 2009]; in 2006, it was estimated that 26.6 million people worldwide suffered with the disease. By 2050, the prevalence is expected to quadruple, with 1 in 85 people living with AD [Brookmeyer et al. 2007]. The disease is a social burden to carers and a huge financial burden on healthcare systems because advanced AD patients require round-the-clock care. The cost of AD in 2013 was estimated to be £26.3 billion in the UK alone, with £11.6 billion of that money

being provided in the form of unpaid care by families of people with the disease [Prince et al. 2014].

At present, definitive diagnosis of AD can only be achieved via an autopsy, thus greatly limiting correct diagnosis in living patients. In its early stages, the symptoms of AD are difficult to distinguish from those attributed to normal ageing. In the later stages, AD is often mistaken for other types of dementia, and autopsy studies have shown that many elderly patients have brain abnormalities associated with multiple types of dementia [Clarfield 2003]. Furthermore, there are some medical conditions, such as thyroid problems and depression, that can present with similar symptoms to AD. Unlike dementia, these conditions may be reversible with suitable treatment. Unsurprisingly, incorrect diagnosis of AD is common. One comprehensive study showed that depending on the clinical criteria used to diagnose AD, sensitivity ranged from 70.9% to 87.3%, and specificity ranged from 44.3% to 70.8% [Beach et al. 2012]. For this reason, more accurate diagnosis techniques are urgently required.

Although the exact etiology of AD is still unknown, two fundamental neuropathological lesions have been implicated in the pathogenesis of the disease: neurofibrillary tangles composed of the microtubule-associated protein *tau*, and neuritic plaques consisting of an accumulation of *amyloid-β* (Aβ) peptides [Perl 2010]. These lesions are illustrated in Figure 2.1. The abnormal presence of neurofibrillary tangles was first observed by [Alzheimer 1907] whilst documenting a case of the disease that would eventually bear his name. Tau proteins are responsible for controlling the stability of axonal microtubules. This is achieved, in part, by modifying the tau proteins via phosphorylation. When the tau proteins are hyperphosphorylated, tangles are formed, causing the protein to aggregate in an insoluble form [Rowe and Villemagne 2013]. The insoluble structure interferes with axonal transport, leading to cell death. It has been shown that the extent and distribution of neurofibrillary tangles correlate with the severity of AD [Bierer et al. 1995], however, tangles have also been associated with the pathogenesis of several other diseases, such as progressive supranuclear palsy and frontotemporal dementia [Goedert and Spillantini

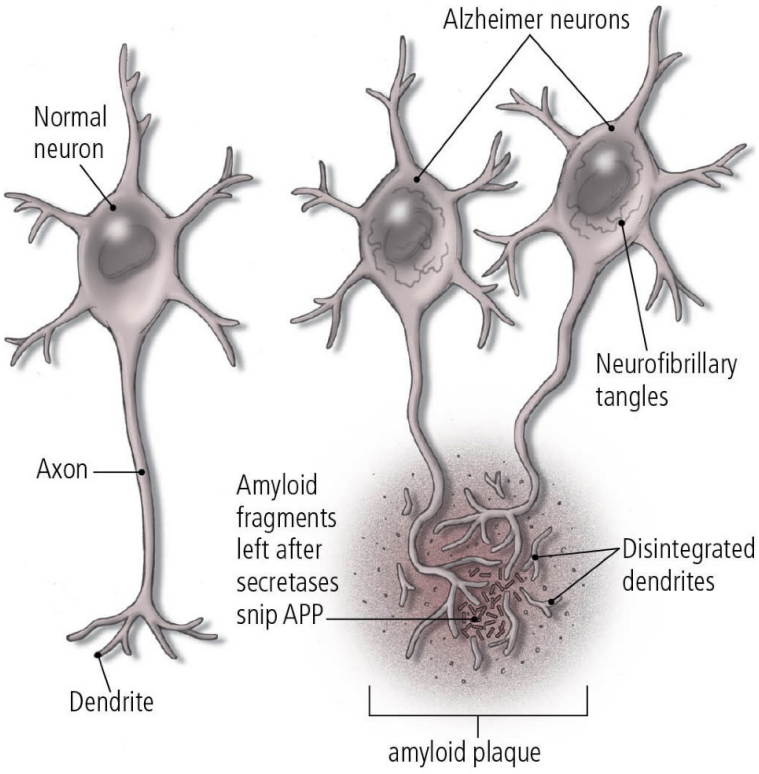


Figure 2.1: An illustration (adapted from [Komaroff 2015]) of a healthy neuron (left) compared to neurons typically found in an Alzheimer's disease patient (right). There are two fundamental neuropathological lesions that have been implicated in the pathogenesis of Alzheimer's disease: neurofibrillary tangles composed of the protein tau, and neuritic plaques consisting of an accumulation of amyloid-β peptides.

2000; Arai et al. 2001].

In contrast, Aβ plaques are not found in frontotemporal dementia [Rowe and Villemagne 2013]. Aβ is derived from a parent protein called the amyloid precursor protein, which fragments after action from two secretases [Perl 2010]. The resulting plaques are neurotoxic, and cause the neurons to die [Mattson 2004]. Aβ plaques are present in all cases of AD, and begin to develop many years before the onset of dementia [Rowe and Villemagne 2013]. As a consequence, accurately imaging amyloid has the potential to confirm the presence of AD, distinguish between AD and other dementias, and diagnose AD at an earlier stage. For this reason, this thesis predominantly focuses on in vivo amyloid imaging.

Amyloid plaques are categorised as either diffuse or dense. Diffuse plaques are an

antecedent to dense plaques, and indicate early plaque formation. Dense plaques[1] are compact and are called *neuritic* when they are associated with local neuronal damage [Mott and Hulette 2011]. The early stages of amyloid deposition suffer from considerable inter-patient variation in the pattern of the deposits, however, [Braak and Braak 1989] attempted to identify three stages of amyloid accumulation. In the first phase, low densities of amyloid deposits are found in the basal portions of the frontal, temporal and occipital lobes. In the second stage, medium densities of amyloid deposits are present in most isocortical regions except the primary sensory areas and the primary motor field. Finally, in stage three, almost all regions of the isocortex (including the primary sensory areas and motor field) contain dense amyloid deposits. In all stages the hippocampus is largely devoid of amyloid.

2.1.1 Disease Biomarkers

In 1984, [McKhann et al. 1984] established a set of criteria for clinical diagnosis of probable AD. The criteria covered medical history, clinical examination, neuropsychological testing, and laboratory assessment. For probable AD to be diagnosed, the patient had to exhibit progressive memory impairment and deficits in at least two other areas of cognition (confirmed by tests such as the Mini-Mental State Examination). Furthermore, other brain disorders had to be excluded using laboratory tests and structural computed tomography (CT) brain imaging. These criteria have been used in many studies in the literature, and have achieved a mean sensitivity and specificity of 81% and 70%, respectively [Knopman et al. 2001].

Twenty-seven years later the diagnosis criteria were updated to reflect the increase in knowledge of the biology of AD [McKhann et al. 2011]. In addition to the original "probable AD" and "possible AD" categories, a third category, "probable or possible AD with evidence of the AD pathophysiological process", was added to the guidelines. This extra category incorporated biomarkers into the diagnosis criteria. For example, disproportionate atrophy in the temporal lobe on structural magnetic

[1] In this thesis, the term "amyloid plaque" is almost invariably used to mean "dense amyloid plaques".

resonance (MR) imaging, low levels of Aβ in cerebrospinal fluid (CSF), or positive positron emission tomography (PET) amyloid imaging [McKhann et al. 2011].

Due to the close proximity of CSF to the brain, it is an obvious source of biomarkers for AD. Biochemical changes in the brain are reflected in the CSF, and the typical levels of CSF total tau (t-tau), phosphorylated tau (p-tau), and Aβ differ between patients with AD and healthy controls [Blennow et al. 2001]. Elevated levels of CSF t-tau have been reported in AD patients compared to cognitively normal controls [Blennow et al. 1995], however, it has been demonstrated that CSF t-tau has a poor specificity when distinguishing between other forms of dementia [Hulstaert et al. 1999]. Similarly, an increase in CSF p-tau has been recorded in AD, vascular dementia, and frontotemporal dementia [Sjögren et al. 2001]. In contrast to the tau biomarkers, a lower level of CSF Aβ is consistently found in AD patients than healthy controls [Blennow et al. 2001]. Nevertheless, low levels of CSF Aβ have also been reported in other dementias [Sjögren et al. 2000].

Another drawback of CSF biomarkers is the lumbar puncture required to extract the CSF sample from the spine. Lumbar punctures are invasive procedures, and although the risk of a severe adverse event is very small [Peskind et al. 2005], the associated headaches and patient anxiety can limit their use clinically [Anoop et al. 2010]. Moreover, there can be inconsistencies with CSF data analysis, due to the sample collection, transportation, and storage [Anoop et al. 2010].

As an alternative to CSF biomarkers, amyloid PET is increasingly used to assess the burden of Aβ in patients with suspected AD. PET is an imaging technique that produces three-dimensional images of functional processes in tissue. The system detects gamma rays emitted from a positron-emitting radionuclide (also called a *radiotracer* or *tracer*) that is injected into the body prior to image acquisition. There is a wide range of different tracers for analysing different processes in the body, and by analysing the areas where tracer uptake does/does not occur, radiologists can assess body functions and identify abnormalities.

There are two major classes of tracer for brain PET imaging. Members of the

first class are designed to enter the brain, often by active transport, and accumulate after a metabolic transformation [Pike 2009]. One example is the tracer ^{18}F-FDG, which is a glucose analogue. ^{18}F-FDG enters the brain by the action of a glucose transporter, where it is phosphorylated by hexokinase [Teune et al. 2013]. The resulting compound accumulates in the brain, and provides a measure of energy metabolism. Amyloid PET tracers fall into the second major class of tracers for brain imaging, which are designed to reversibly bind to a target without being changed structurally [Pike 2009]. Unsurprisingly, amyloid PET imaging uses tracers that bind to Aβ. The quantity and pattern of tracer uptake can be used to designate a scan as amyloid positive or amyloid negative [Landau et al. 2013]. Example axial slices from positive and negative scans for four different amyloid tracers are shown in Figure 2.2. As predicted by the final two stages of amyloid deposition identified by [Braak and Braak 1989], amyloid positive scans are characterised by high specific tracer uptake in grey matter (see red arrows in Figure 2.2). Tracer uptake in white matter is non-specific, and is high in both amyloid positive and amyloid negative scans (see yellow arrows in Figure 2.2).

The reason for the high white matter signal in amyloid PET imaging has yet to be satisfactorily explained in the literature [Kepe et al. 2013]. One explanation is that slower blood flow in white matter compared to grey matter creates a disparity in the clearance rate of the tracer [Fodero-Tavoletti et al. 2009]. However, this does not account for the higher level of non-specific tracer retention witnessed in healthy controls compared to subjects with AD [Barthel et al. 2011]. Another explanation for the high non-specific signal is that the amyloid PET tracers bind to myelin in the white matter with high affinity [Stankoff et al. 2011]. Nevertheless, it has been reported by [Villemagne et al. 2012] that the high non-specific tracer uptake in white matter does not affect the diagnostic accuracy of amyloid PET tracers.

The first tracer to specifically image amyloid plaques in neuronal tissue was ^{11}C-Pittsburgh Compound-B (^{11}C-PiB) [Klunk et al. 2004]. Due to the short half-life of carbon-11 (20 minutes), the compound needs to be prepared on-site and used

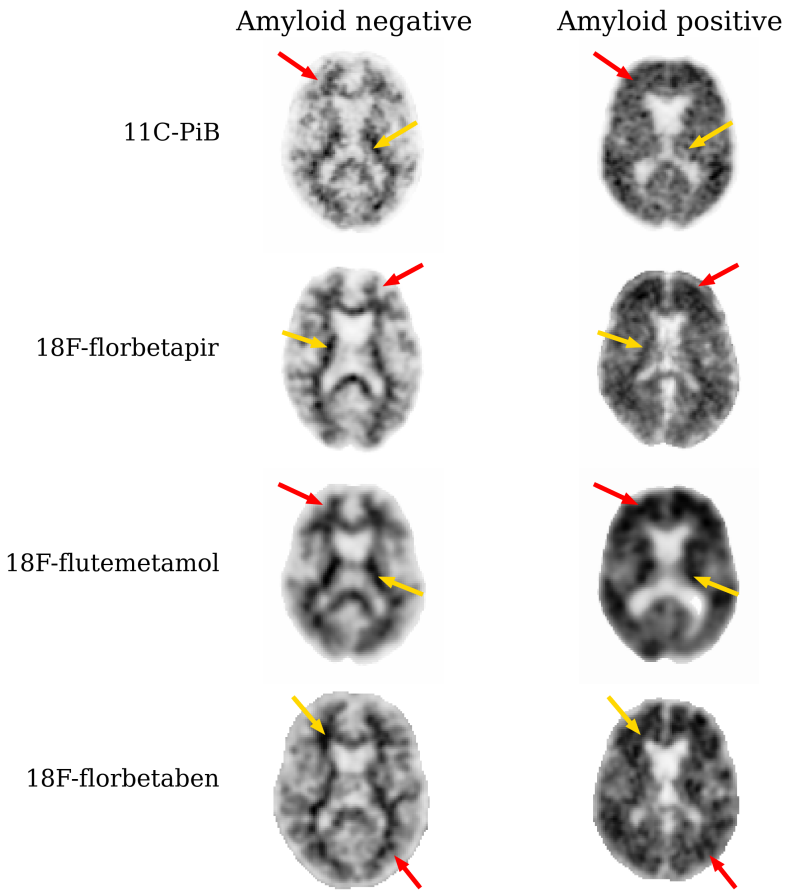


Figure 2.2: Example axial slices from amyloid negative and amyloid positive scans for four different amyloid PET tracers. Amyloid positive scans are characterised by specific tracer uptake in grey matter (red arrows). Tracer uptake in white matter is non-specific, and is high in both amyloid positive and amyloid negative scans (yellow arrows).

immediately. This limits the use of ^{11}C-PiB to centres with an on-site cyclotron, which is uncommon, and hence makes ^{11}C-PiB impractical for routine clinical use. More recently, several other amyloid tracers have been developed using the fluorine-18 isotope, which has a longer half-life of 110 minutes, meaning that it can be prepared off-site and transported to PET imaging centres. Three of these tracers have recently been approved by the U.S. Food and Drug Administration (FDA) for use in clinical diagnosis: ^{18}F-florbetapir, ^{18}F-flutemetamol, and ^{18}F-florbetaben [FDA 2012; FDA 2013; Piramal Enterprises Ltd 2014]. Example axial slices from amyloid PET volumes obtained using these tracers are shown in Figure 2.2. It must be noted that the tracer manufacturers' guidelines state that a positive amyloid scan

alone does not establish a diagnosis of AD [Amyvid 2012; Neuraceq 2014].

2.1.2 Alzheimer's Disease Treatments

Although current medications cannot cure AD, there are medications available to slow the progression of the disease, and lessen symptoms, such as memory loss, for a limited time. Five drugs are approved by the FDA for the treatment of AD [Cummings et al. 2014]. Four of these drugs are classified as acetylcholinesterase inhibitors (AChEIs), which prevent the breakdown of acetylcholine, thereby prolonging the duration of its action as a neurotransmitter [Stella et al. 2015]. The fifth drug, memantine, is an N-methyl-D-aspartate (NMDA) glutamate receptor antagonist. Neuronal cell death is due, in part, to excitotoxicity caused by high levels of the neurotransmitter glutamate. Memantine causes a blockage of the NMDA receptors, which prevents their over-activation, and thus reduces the levels of glutamate [Lipton 2004].

Despite studies showing that AChEIs can improve cognitive function in patients with mild to moderate AD [Davis et al. 1992; Rogers et al. 1998; Rösler et al. 1999; Tariot et al. 2000], there is a debate in the literature about the clinical significance of the improvement [Qizilbash et al. 1998; Steele and Glazier 1999]. A review of the efficacy of AChEIs concluded that there was no difference between them with respect to efficacy [Birks 2006]. Moreover, the same review stated that it is not possible to predict those who will respond to treatment. Common side effects of AChEIs typically include nausea, vomiting, and diarrhoea [Stella et al. 2015]. One AChEI, tacrine, was withdrawn from the U.S. market following continued concerns over its safety [Mehta et al. 2012].

AChEIs are indicated for mild to moderate AD, however, there are studies demonstrating that AChEIs could help patients with severe AD [Feldman et al. 2001; Winblad et al. 2006]. In contrast, memantine is approved for treatment of moderate to severe AD [Reisberg et al. 2003]. Nevertheless, memantine is often prescribed off-label, for mild AD and mild cognitive impairment. A meta-analysis of

clinical trials showed that despite the frequent off-label use of memantine, there is a lack of evidence for a benefit in patients with mild AD [Schneider et al. 2011]. The authors of the meta-analysis also questioned the efficacy of memantine in moderate AD.

Drug development for AD has been difficult. Not only do the benefits of current medications seem marginal, but no new drugs have been approved since 2003 [Cummings et al. 2014]. A study into the trends of AD-drug clinical trials revealed that 99.6% of compounds in a Phase 3 trial failed to advance to the stages necessary for regulatory approval [Cummings et al. 2014]. Despite the difficulties, aducanumab (drug candidate BIIB0307) has shown encouraging early results for the treatment of prodromal AD. Unlike AChEIs and memantine, which only lessen the symptoms of AD, aducanumab is reported to significantly decrease both cognitive decline and Aβ accumulation [Toyn 2015]. Amyloid PET imaging with the ^{18}F-florbetapir tracer was used in the Phase 1B trials to assess the reduction in Aβ plaques [Hang et al. 2015]. Subjects taking a dose of 10mg/kg had a statistically significant decrease in amyloid compared to subjects in the placebo group [Sevigny et al. 2015]. This trial demonstrates that amyloid imaging could play a crucial role in the development and evaluation of drugs to treat AD. Moreover, as more amyloid-modifying drugs are tested, amyloid imaging could become an important tool for monitoring the progression of AD, not just aiding the diagnosis of the disease. It must also be noted that the Phase 3 trials of aducanumab only involve subjects with prodromal AD. As disease-modifying therapies become available, robust detection and quantification of amyloid plaques is essential in order to identify patients who could benefit from novel pharmaceuticals. Consequently, amyloid imaging could become more widespread.

2.1.3 Amyloid PET Quantification

In PET imaging, the standardised uptake value (SUV) is a quantitative tool used to supplement visual assessment of scans, which is the current clinical standard. For example, studies have shown that AD patients exhibit higher SUVs in brain regions

than healthy controls [Lopresti et al. 2005]. The SUV is a dimensionless ratio [Thie 2004]:

$$\mathrm{SUV}_{\mathrm{ROI}} = \frac{Q_{\mathrm{ROI}}}{\mathrm{ID}/m} \tag{2.1}$$

where Q_{ROI} is the radioactivity concentration (in kBq/g) in the region of interest (ROI). ID is the injected dose (in kBq), and m is the body mass (in g).

An immediately apparent problem with the SUV is that it normalises for body mass. Fat uptake of certain PET tracers (for example, ^{18}F-fluorodeoxyglucose) is very low. Therefore, the SUVs in non-fat tissues for larger subjects may be overestimated relative to the same SUVs in smaller subjects. The higher level of fat will contribute to the overall body mass, but will accumulate less tracer. Although no studies have explicitly looked at the link between body weight and brain tissue SUVs, it is widely accepted that the SUV for many other tissues (including blood) demonstrates a strong positive correlation with weight [Zasadny and Wahl 1993]. A method to correct this problem was proposed by [Zasadny and Wahl 1993], in which lean body mass (mass ignoring fat) is used instead of total body mass. Despite being suggested over two decades ago, the correction is not used in the majority of papers quoting SUVs.

One popular solution to the body habitus issue is to use the standardised uptake value ratio (SUVR), rather than the SUV. The SUVR involves normalising the SUV of a region of interest to a reference region in which the tracer uptake is representative of healthy tissue. Assuming that the same injected dose was used for the ROI SUV and the reference SUV, then the body mass is cancelled out in the calculation:

$$\mathrm{SUVR}_{\mathrm{ROI}} = \frac{\mathrm{SUV}_{\mathrm{ROI}}}{\mathrm{SUV}_{\mathrm{reference}}} = \frac{Q_{\mathrm{ROI}}}{Q_{\mathrm{reference}}} \tag{2.2}$$

The computation of the SUVR can be conducted on static or summed PET frames, negating the need for arterial sampling or dynamic scans. Hence, the SUVR has become the method of choice for normalisation of tracer uptake, due to the

simplicity of the calculation. However, to ensure that the SUVR is only a function of target concentration, and not influenced by blood flow, it should only be calculated once tracer binding has reached steady-state [Pike 2009; Cselényi and Farde 2015]. For fluorine-18 amyloid PET tracers, SUVR is typically calculated at 90 minutes after the injection of the tracer [Villemagne et al. 2011a].

In amyloid PET imaging, the SUVR is used to quantitatively assess brain amyloid burden. Typically, the individual SUVRs for each target region are averaged to form the mean, or *composite*, SUVR [Rowe et al. 2008]. It has been shown that incorporating this ratio as an adjunct to the visual assessment of ^{18}F-florbetapir amyloid scans can decrease inter-reader variability [Pontecorvo et al. 2014; Nayate et al. 2015]. Nevertheless, despite the widespread use of the SUVR in research, there are many ways that its accuracy and robustness could be improved.

One problem affecting meaningful SUVR usage in disease diagnosis and monitoring, is the choice of the ROIs and reference region. Although there is a consensus in the literature about which general brain areas are to be used in the composite SUVR for amyloid imaging, there are differences between tracers and studies [Jack et al. 2008; Jagust et al. 2009; Fleisher et al. 2011; Villemagne et al. 2011a; Barthel et al. 2011; Joshi et al. 2012]. For example, since amyloid deposition does not typically occur in the cerebellum, the reference regions in the aforementioned studies are all cerebellum-based. Nevertheless [Jack et al. 2008] use the cerebellar grey matter, whereas [Joshi et al. 2012] use the whole cerebellum. Moreover, the frontal lobe is a universal target region, but the delineation of the region varies with the study; [Barthel et al. 2011] use the frontal lobe, whereas [Villemagne et al. 2011a] use specific areas within the frontal lobe. In practice, this makes comparison between studies very difficult.

We address this problem in Chapter 5 by proposing a novel method for amyloid status classification that is independent of tracer, and does not require any anatomical regions of interest. Moreover, other methods have been proposed to standardise the SUVR process between tracers. For example, the method proposed

in the Centiloid Project converts SUVRs to a 0-100 scale, where 0 represents definitely amyloid negative, and 100 represents the amount of global amyloid deposition in a typical mild-moderate AD patient [Klunk et al. 2015]. The aim of that project is to provide a "standard" method for quantitative ^{11}C-PiB image analysis, and a method for converting analyses for other tracers to the 100-point scale. Whilst this standardisation is extremely useful, and allows for subjects to be placed on an "amyloid spectrum" (rather than just amyloid positive/negative), each new method and amyloid tracer must be calibrated to the Centiloid method in order for accurate comparisons between studies.

Reliable and reproducible SUVRs also require accurate delineation of the anatomical target and reference regions. Manual delineation is time-consuming and prone to inter-observer variation. Several algorithms exist in the literature for automated brain PET region extraction [Wong et al. 2002; Farinha et al. 2009], however, given the poor spatial resolution and signal-to-noise ratio of PET, boundaries are often indistinguishable and rarely continuous. Moreover, the delineability of regions of interest would be highly dependent on the tracer. If structural magnetic resonance (MR) images are available, they can be used to refine the PET regions, but this could introduce other errors, since the PET and MR scans will need to be aligned using a process called *image registration*. Moreover, [Raniga et al. 2008] showed that SUVRs for all brain regions are significantly lower when normalised using automated cerebellar grey matter segmentation. The large discrepancy makes direct comparison between manual and automated SUVR calculations difficult. One can abrogate the need for delineation by aligning the PET image to a template with known region maps. However, poor alignment could affect the SUVR, as an ROI in the image may not be completely contained within the mask of the ROI in the template. Image registration is introduced in the next section, and the impact of different registration methods on SUVR calculation is explored in Chapters 3 and 4.

2.2 Image Registration

Image registration is one of the fundamental tools in biomedical image analysis. Also known as image matching, registration is the process of determining the spatial transformation s that will align a set of images. The ultimate aim is to map each point x in one image to the corresponding point y in another. In the medical imaging field, arguably three of the most important applications of registration are: 1) multi-modality fusion, where information from different imaging modalities is combined to aid diagnosis; 2) longitudinal analysis, where temporal changes in patients are examined; 3) population modelling to study functional or anatomical variability [Sotiras et al. 2013]. The misalignment of the original images can have many causes, including patient movement and anatomical, physiological or pathological differences.

In this thesis, the applications of image registration broadly correspond to the first and third applications listed by [Sotiras et al. 2013]. Firstly, despite the clinical usefulness of the molecular and functional information provided by PET images on their own, it is widely accepted that registration with anatomical images improves their diagnostic value. For example, registration of MR images with PET volumes would provide a clinician with a more comprehensive picture of function and anatomy together, potentially allowing for accurate determination of the pattern of grey matter tracer uptake compared to non-specific uptake in white matter. Indeed, tracer manufacturer guidelines state that MR or computed tomography (CT) images can be used to help localise tracer uptake in the grey matter [Amyvid 2012; Neuraceq 2014]. The second application of image registration in this thesis relates to population modelling. Prior to SUVR calculation, the PET volume is registered to a template space in which the anatomical ROIs are defined. This allows the SUVRs of different subjects to be compared to one another, and compared to population means.

The choice of registration scheme is dependent on the type of images, the type of spatial transform desired, and the goal of the registration. Rigid registration

can be used to capture global spatial changes between images, such as rotation and translation, but not smaller, local differences. This technique can be extended to fully affine registration by including terms to encompass scaling and shear.

Linear registration methods, such as rigid or affine, are represented by a small number of parameters (6 and 12, respectively, in 3D). In contrast to linear registration, deformable, non-linear registration captures more complex transformations. In non-parametric approaches, the coordinates $s(x)$ of each image voxel in the transformed image need to be specified. Usually, this spatial transformation is described by a displacement field $\mathbf{s}$, which is added to an identity transformation ($\mathrm{Id} : x \mapsto x$):

$$s : x \mapsto x + \mathbf{s}(x) \tag{2.3}$$

Most, if not all, deformable registration methods in the literature can be viewed as an iterative accumulation of updates to the initial estimate of the deformation. The differences between the methods lie in the way in which the deformation updates are estimated, and the way in which they are composed. This variation is a result of the chosen transformation model, cost function, and optimisation framework.

In this thesis we investigate deformable registration methods that use information from MR images to help inform the registration of amyloid PET images to a template for SUVR calculation. In particular, we extend a deformable registration framework known as demons registration, because it has been shown to be able to accurately co-register brain MR images, and is little affected by the choice of subject population [Klein et al. 2009]. The background for demons registration is presented in the next section.

2.2.1 Demons Registration

The demons method is one approach for deformable, non-parametric registration. Originally proposed by [Thirion 1998], the demons algorithm models the registration as a diffusion process, with a parallel to Maxwell's demon (a thought experiment used to solve the Gibb's paradox of thermodynamics). The original "classical"

version of the algorithm has been used extensively by the biomedical image analysis community due to its efficiency and simplicity of implementation.

In its most basic form, the demons algorithm iteratively estimates the transformation s between a fixed (target) image F and a moving (source) image M (see Algorithm 2.1).

Algorithm 2.1 The original demons registration algorithm

1: **while** not converged **do**

2: Given the current deformation field $\mathbf{s}$, compute an update field $\mathbf{u}$

3: $\mathbf{s} \leftarrow \mathbf{s} + \mathbf{u}$

4: Regularise the deformation field by convolving it with a Gaussian kernel G, such that $\mathbf{s} \leftarrow G * \mathbf{s}$

5: **end while**

The update field is a set of forces inspired by the optical flow equations [Horn and Schunck 1981], but renormalised to prevent instabilities for small image gradient values:

$$\mathbf{u} = \frac{(F - M(s)) \cdot \nabla M(s)}{\|\nabla M(s)\|^2 + (F - M(s))^2} \tag{2.4}$$

where $M(s)$ represents the composition of the moving image and the deformation field, i.e. $M(s) = M \circ s$.

From Equation 2.4, it can be seen that the "demons" push according to the local characteristics of the images. If the voxel intensity is lower than the target value, the moving image M is pushed in the direction of the gradient. If the voxel intensity is greater than the target value, the moving image is pushed in the opposite direction to the image gradient.

A major weakness of the classical demons algorithm is the lack of a strong theoretical background. Unlike many other registration methods [Bookstein 1989; Rueckert et al. 1999; Avants et al. 2008], the demons algorithm was not presented as an energy minimisation. Although the idea was influenced by Maxwell's demon, and

the update forces inspired by optical flow, there was no underlying theory that could predict when it would fail and why. Fortunately, since being proposed by [Thirion 1998], several research teams have worked to provide a theoretical framework for the demons algorithm. In addition to advancing the understanding of the algorithm, this work has facilitated the modification of the underlying assumptions. Firstly, [Pennec et al. 1999] showed that the update term is an approximation to the second order gradient descent on the sum of squared differences (SSD) between voxel intensities. A decade later, [Vercauteren et al. 2009] expanded this work by casting the demons algorithm into a minimisation of a well posed global energy $E(s)$. Subsequently, the authors showed that all variants of the demons algorithm can be grouped into one registration framework:

$$E(F, M, s) = \frac{1}{\sigma_i^2}\text{Sim}(F, M \circ s) + \frac{1}{\sigma_T^2}\text{Reg}(s) \tag{2.5}$$

where $\circ$ is the composition operator, $\text{Sim}(F, M \circ s)$ is the image similarity measure (typically SSD), $\text{Reg}(s)$ is the regularisation term, and σ_i and σ_T account for the image noise and amount of regularisation, respectively. In [Thirion 1998], σ_i is the local estimation of the noise $\sigma_i(p) = |F(p) - M \circ s(p)|$. Parameter σ_T could be tuned by evaluating the registration result for a range of values, but usually the value is taken from existing literature.

To avoid computationally expensive iterations during this optimisation, [Cachier et al. 2003] introduced an exact correspondence field c into the energy function. This allows for some errors in s. The resulting energy function is as follows:

$$E(F, M, s) = \frac{1}{\sigma_i^2}\text{Sim}(F, M \circ s) + \frac{1}{\sigma_x^2}\text{Dist}(s, c) + \frac{1}{\sigma_T^2}\text{Reg}(s) \tag{2.6}$$

where $\text{Dist}(s, c)$ forces the displacement field s to be close to the exact correspondences c, and σ_x accounts for correspondence uncertainty.

Due to the introduction of the correspondence field by [Cachier et al. 2003], the optimisation alternates between minimising c and minimising s. The first step

optimises $\frac{1}{\sigma_i^2}\text{Sim}(F, M \circ s) + \frac{1}{\sigma_x^2}\text{Dist}(s, c)$ with respect to c, and the second step optimises $\frac{1}{\sigma_x^2}\text{Dist}(s, c) + \frac{1}{\sigma_T^2}\text{Reg}(s)$ with respect to s.

The demons algorithm owes its efficiency to two factors. Firstly, the Gaussian regularisation of the deformation field is numerically robust and, perhaps more importantly, fast to compute. The second, more profound factor is concealed in Equation 2.4. As previously stated, [Pennec et al. 1999] demonstrated the link between demons and the second order gradient descent on the SSD criterion. This improves the convergence speed of the demons algorithm, so that fewer iterations are required to achieve good results.

Despite the increase in the speed of convergence, the link between the demons method and the gradient descent of the SSD leads to limitations. Since Equation 2.4 does not lend itself to updates for other similarity measures than SSD, the demons algorithm is implicitly restricted to single-modality registration tasks. Fortunately, since its inception, the demons framework has been generalised to allow for other similarity measures [Guimond et al. 2001; Chefd'hotel et al. 2002]. Typically, the generalisation is to compute the update field from the gradient descent optimisation of the similarity measure. For poorly conditioned problems, the steepest descent approach is known to be slow to converge. Consequently, these methods require more iterations to achieve good registration results.

Another problem affecting the demons framework is that all voxels with the same intensity in the moving image could theoretically be mapped to a single point with this intensity in the fixed image. Repeated for all intensity levels, this transformation would perfectly minimise the SSD criterion. However, it would introduce physically and physiologically impossible folding. To avoid this problem, the transformation can be regularised. Regularisation is a critical step in non-rigid registration [Pennec et al. 1999], however the level of regularisation required is dependent on the type and size of the images being used. It has been reported in the literature that "standard values" for Gaussian diffusion-like regularisation (from no regularisation at all to $\sigma = 1$) work well in most cases [Vercauteren et al. 2009]. These parameters are only

reported for MR images, and unfortunately, there appears to be no single value that works in all situations and for all modalities.

In the original demons algorithm, [Thirion 1998] proposed to apply Gaussian smoothing to the overall deformation field $\mathbf{s}$. An alternative method, suggested by [Bro-Nielsen and Gramkow 1996], is to regularise the update field $\mathbf{u}$ instead. The authors also showed that this variant of the demons framework is an approximation to the viscous fluid registration method employed by [Christensen 1994]. For this reason, this type of regularisation, in which only the update is smoothed, is known as *fluid regularisation* (cf. the original *diffusion-like regularisation*). Further to this, [Pennec et al. 1999] suggested combining the two approaches, so that both the update and accumulative deformation fields are regularised.

In addition to extensions of the regularisation, extensions of the demons update scheme have been considered. The additive update scheme (see Algorithm 2.1) is appropriate if one assumes that the displacement fields are elements of a linear vector space. This scheme is similar to the update rules used in Newton-based optimisation methods. However, mapping a point through one spatial transformation and then through another one is equivalent to mapping the point through the composition of the two transformations. Therefore, at each iteration the update $\mathbf{u}$ should be composed with the accumulative spatial transformation s:

$$s \leftarrow s \circ (\mathrm{Id} + \mathbf{u}) \tag{2.7}$$

In the literature, such an update is referred to as a *compositive adjustment*. Although each iteration takes longer to compute, it has been reported that the compositive update scheme converges in fewer iterations than the additive update scheme [Vercauteren et al. 2009].

One caveat of both the additive and compositive demons algorithms is that they do not ensure invertibility of the final spatial transformation. In contrast, diffeomorphic[2] image registration algorithms do ensure that the output transformation

[2]A bijective map ϕ is diffeomorphic iff ϕ and ϕ^{-1} are continuously differentiable.

is invertible. Although an invertible transformation may not always be required, diffeomorphisms preserve the topology and prevent physically implausible folding. Furthermore, diffeomorphic registration is useful if there is no favoured registration direction. However, it must be noted that not all problems benefit from diffeomorphisms. Different applications may have different constraints. For example, clinicians might prefer to see the topology changes caused by ventricular enlargements.

A diffeomorphic approach to demons is presented in [Vercauteren et al. 2007]. The main idea is that the update field $\mathbf{u}$ is assumed to lie in a space tangential to the Riemannian manifold of diffeomorphisms. Thus, the diffeomorphic update (a stationary velocity field in this case) is found by applying an exponential mapping from the tangential space to the manifold of diffeomorphisms (i.e. the diffeomorphic update is $\exp(\mathbf{u})$, not just $\mathbf{u}$). This is in contrast to the compositive demons method in Equation 2.7, where the update field $\mathbf{u}$ is applied directly to the spatial transformation s. At each iteration the diffeomorphic update $\exp(\mathbf{u})$ is composed with the transformation s:

$$s \leftarrow s \circ \exp(\mathbf{u}) \tag{2.8}$$

All three major demons methods (additive, compositive and diffeomorphic) produce registration results that are visually quite similar, but only compositive and diffeomorphic updates are guaranteed to have smooth output transformations [Vercauteren et al. 2009]. Whilst the compositional approach often results in an invertible transform, diffeomorphic demons will always be invertible, albeit at a reported 50% increase in computation time per iteration [Vercauteren et al. 2009].

In the diffeomorphic demons algorithm proposed in [Vercauteren et al. 2007], only the update field at each iteration is encoded using an exponential. Although this keeps the algorithm computationally simple and mostly unchanged from the original demons algorithm, it lacks some of the characteristics of log-domain registration methods that encode the entire transformation with stationary velocity

fields [Ashburner 2007], namely, the ability to obtain the inverse transform at low cost, and a symmetrical registration result. Therefore, the diffeomorphic demons algorithm was extended to cast the complete transformation into the log domain [Vercauteren et al. 2008]. The transformation s is parameterised as a stationary velocity field $\mathbf{v}$, such that $s = \exp(\mathbf{v})$.

The inverse of the transformation s can be obtained easily by a backward computation $s^{-1} = \exp(-\mathbf{v})$. Therefore, a symmetric registration algorithm can be achieved by symmetrising the energy function presented in Equation 2.6 [Vercauteren et al. 2008]:

$$s_{\text{opt}} = \arg\min_{s}(E(F, M, s) + E(F, M, s^{-1}) \tag{2.9}$$

One disadvantage of this method is that it requires the separate optimisation of the two terms, which can be computationally expensive for complicated similarity criteria. Therefore, to reduce the computational cost of using other image similarity measures than SSD, [Lorenzi et al. 2013] proposed optimising in the halfway space, leading to a single energy term by considering $F \circ \exp(-\frac{\mathbf{v}}{2})$ and $M \circ \exp(\frac{\mathbf{v}}{2})$.

It has been shown that registering two MR brain images using a non-symmetric registration method can create a bias in subsequent analyses [Yushkevich et al. 2010]. Therefore, our novel method for joint registration of PET-MR image pairs to a template space (proposed in Chapter 3) extends the symmetric log-domain demons registration method proposed by [Lorenzi et al. 2013].

2.2.2 Deformable Registration and the Brain

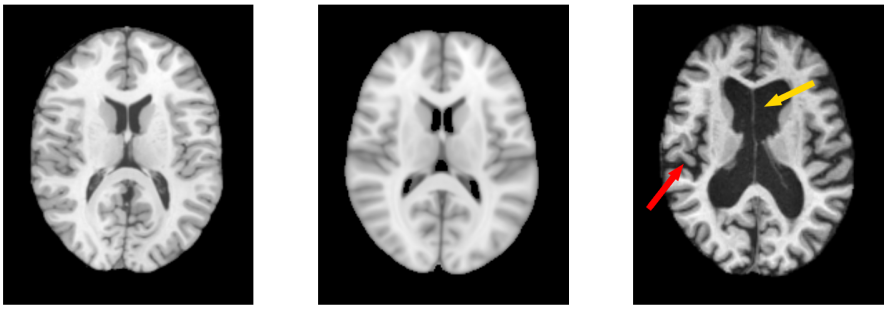
One final problem besetting deformable registration methods, such as demons, is that they seek perfect correspondences between points in the fixed and moving images. This appears to be a useful attribute when estimating large and highly non-linear deformations corresponding to anatomical variability between individuals. Nevertheless, this apparent usefulness has a serious flaw: it assumes a one-to-one mapping between the image pair exists. This almost universally does not hold.

Brains, as with all parts of the human body, are highly variable, such that a point in one brain may not exist in the brain of another subject, or even the same brain at a later time point.

Even if the brains being registered are anatomically identical (i.e. from the same subject), extracting common points and features from different imaging modalities is difficult. For this reason, registration between PET and MR images is often limited to a rigid body transformation [Kiebel et al. 1997; Čížek et al. 2004; Lee et al. 2009b]. Whilst this may be acceptable for current clinical purposes, it might not be sufficient for identifying more subtle brain changes associated with incipient neurological conditions.

To overcome this point-to-point mapping problem, one needs to develop tools for comparing similar brains, without exact correspondences. This in itself leads to further complications, since choosing a suitable common reference space in which to perform comparisons entails determining correspondences between the brains. Established templates such as the Colin 27 [Holmes et al. 1998] and the non-linear MNI-ICBM152 [Grabner et al. 2006] average brains may not adequately represent the majority of the brains to which they are being compared. This is demonstrated in Figure 2.3, where the two templates are compared to an elderly brain. As a result of the ageing process, the elderly brain has larger ventricles and exhibits temporal lobe atrophy (indicated by the arrows in Figure 2.3). Neither of these characteristics are represented in the MR templates. However, whilst one could construct a template for a particular subject group, there is no guarantee that it will match the substructures within all of the brains in question. As a result, only larger scale anatomical features and regions of brain function can be reliably registered.

In this thesis we use and compare two different approaches to creating brain templates that allow matching of macroscopic anatomical regions. These templates allow standardised regions of interest to be used for automated SUVR calculation.



(a) (b) (c)

Figure 2.3: An axial slice through 3D brain volumes at approximately the same anatomical plane. It is immediately apparent that (a) the Colin 27 Average Brain [Holmes et al. 1998], and (b) the 6th generation non-linear MNI-ICBM152 average brain [Grabner et al. 2006] do not adequately represent (c) an example elderly brain. For example, the elderly brain exhibits atrophy in the temporal lobes (red arrow) and enlarged ventricles (yellow arrow).

2.3 Machine Learning in Medical Imaging

As medical imaging has become more widespread, the workload of radiologists has increased rapidly. Visual assessment of medical images is time-consuming and prone to human error, and therefore, many radiologists may welcome an automated approach to support their expert opinion. Computer-aided diagnosis (CAD) tools could assist clinicians with image interpretation by providing a "second opinion". Not only can these tools increase clinician confidence, but they can also reduce the image reading time and offer improved accuracy and consistency of disease diagnosis.

Although SUVR can be used as a CAD tool when combined with an amyloid positivity threshold, its development required an extensive amount of human input. The ROIs, reference region, and amyloid positivity threshold were determined by researchers and radiologists following extensive patient trials. In contrast, a CAD tool in which the diagnostic criteria are acquired via machine learning or data mining could be capable of extracting information imperceptible to physicians.

Machine learning is a field of research concerned with the development of algorithms that can learn from and make predictions about data. Traditionally, machine learning methods have evolved from the study of computer vision and pattern recognition, however, many of these methods are now being applied in radiology for image

segmentation, registration, and CAD [Wang and Summers 2012]. Learning to diagnose a patient's condition from medical images is a two-step process. Firstly, features relevant for discriminating one physical condition or disease state from another must be extracted from appropriate training data. Secondly, a classifier must be trained to separate features belonging to the groups of patients with different conditions. These two steps are described in more detail in the following sections.

2.3.1 Feature Extraction

To be able to distinguish between different disease states, a CAD tool must learn to separate them based on features extracted from biomedical images. These features could be as straightforward as image pixels (or voxels in 3D volumes) [Vandenberghe et al. 2013], or more complicated hierarchies of image filters [Plis et al. 2014]. One of the first feature descriptors was the Scale Invariant Feature Transform (SIFT) [Lowe 1999]. Instead of comparing image patches using pixel intensities alone, SIFT allowed points in an image to be represented robustly by a 128-D vector. In the original paper, the SIFT descriptor was applied to the output of a point-of-interest detector, and used to find corresponding points between images. For object classification problems the keypoint detection step was deemed unnecessary, and thus, SIFT was extended to work on a dense grid of image points [Liu et al. 2008].

Unfortunately, SIFT lacks the ability to encode objects as a set of parts that can deform relative to each other. SIFT is also unable to cope reliably with missing image patches. This could be problematic in medical applications, where certain disease pathologies could move or severely distort image patches (e.g. tumour vs no tumour). Therefore, a more statistical approach to characterise objects was required. This gave rise to the visual bag-of-words [Sivic and Zisserman 2003] and flexible part-based models [Fergus et al. 2005]. Another, more straightforward approach to object detection and classification was proposed by [Dalal and Triggs 2005]. The feature descriptor, known as histogram of oriented gradients (HOG), was successfully used to detect pedestrians in static images, and the authors demonstrated that their

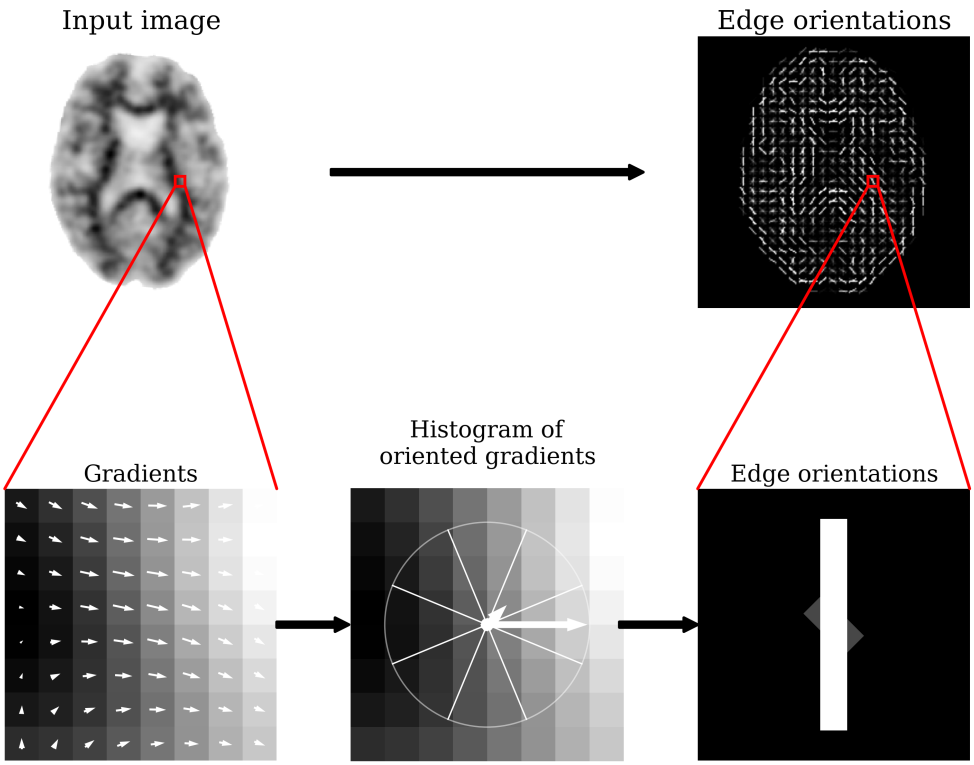
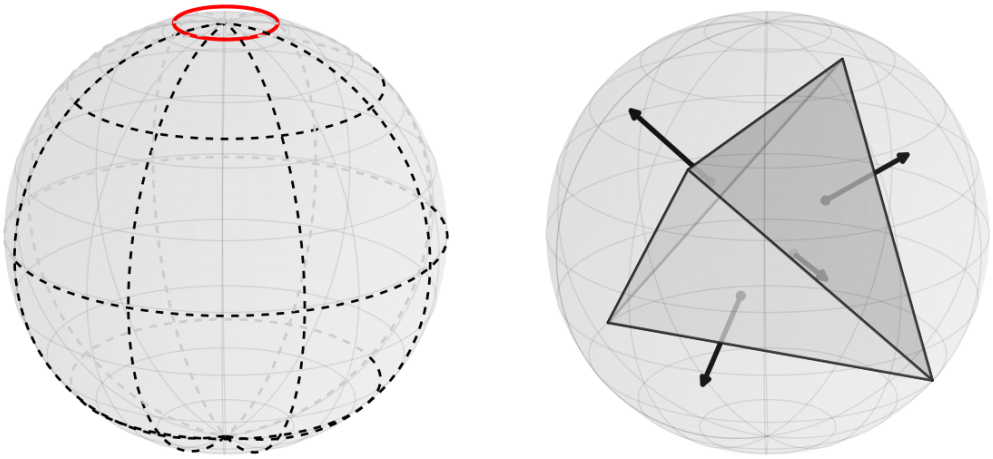


Figure 2.4: An illustration of 2D HOG features for an example ^{18}F-florbetapir PET axial brain slice. The bottom row shows the general steps of the HOG algorithm: image gradients in a single cell (left), quantisation of those gradients (centre), and the edge orientations associated with the histogram of gradients (right). The intensities of edge orientations are determined from the magnitudes of the histogram bins.

detection method resulted in a false positive rate one to two orders of magnitude lower than keypoint-based approaches [Dalal and Triggs 2005]. In the HOG method, the image is partitioned into a grid of uniformly spaced cells, and the normalised histogram of image gradient orientations in each cell forms the set of feature vectors. An illustration of the steps involved in HOG is presented in Figure 2.4.

One drawback of the original HOG method is that it was designed to work on 2D images only. Medical images are often 3D (or even 4D, in the case of dynamic scans) volumes, and as a result, a generalised version of HOG for higher dimensions is required. Computing image gradients in higher dimensions is a straightforward task, and in 3D, a logical extension to HOG would involve quantising the gradients using spherical polar coordinates. By dividing the elevation angle and azimuth into equally sized bins, gradients are quantised using a similar system to latitude



(a) (b)

Figure 2.5: (a) Using spherical polar coordinates to quantise the 3D gradients leads to problems at the poles (red circle) because the bins get progressively smaller. (b) Therefore, [Kläser et al. 2008] proposed to approximate a sphere using a regular polyhedron. The 3D gradients are projected on to the vectors from the centre of the polyhedron to the centres of the faces.

and longitude. However, this leads to problems at the poles because the bins get progressively smaller. This is demonstrated by the red circle in Figure 2.5(a). To solve this problem, [Kläser et al. 2008] proposed to use a regular polyhedron as an approximation to a sphere. Rather than have a continuous space of orientations, each side of the polyhedron corresponds to a histogram bin (see Figure 2.5(b)). This method achieved state-of-the-art results when used to classify human actions in video sequences [Kläser et al. 2008]. Although video sequences are 3D in time (x, y, and time t), whereas amyloid PET volumes are 3D in space (x, y, z), we explore this method in Chapter 5 for feature extraction in PET.

2.3.2 Image Classification

Once suitable image features have been extracted from the training data, they are used to train a classifier capable of separating the image categories (e.g. diseased patients vs healthy controls). Many types of classifiers are presented in the literature, including non-parametric approaches such as k-nearest neighbours [Altman 1992], and ensemble learning methods such as random forests [Breiman 2001]. Despite the

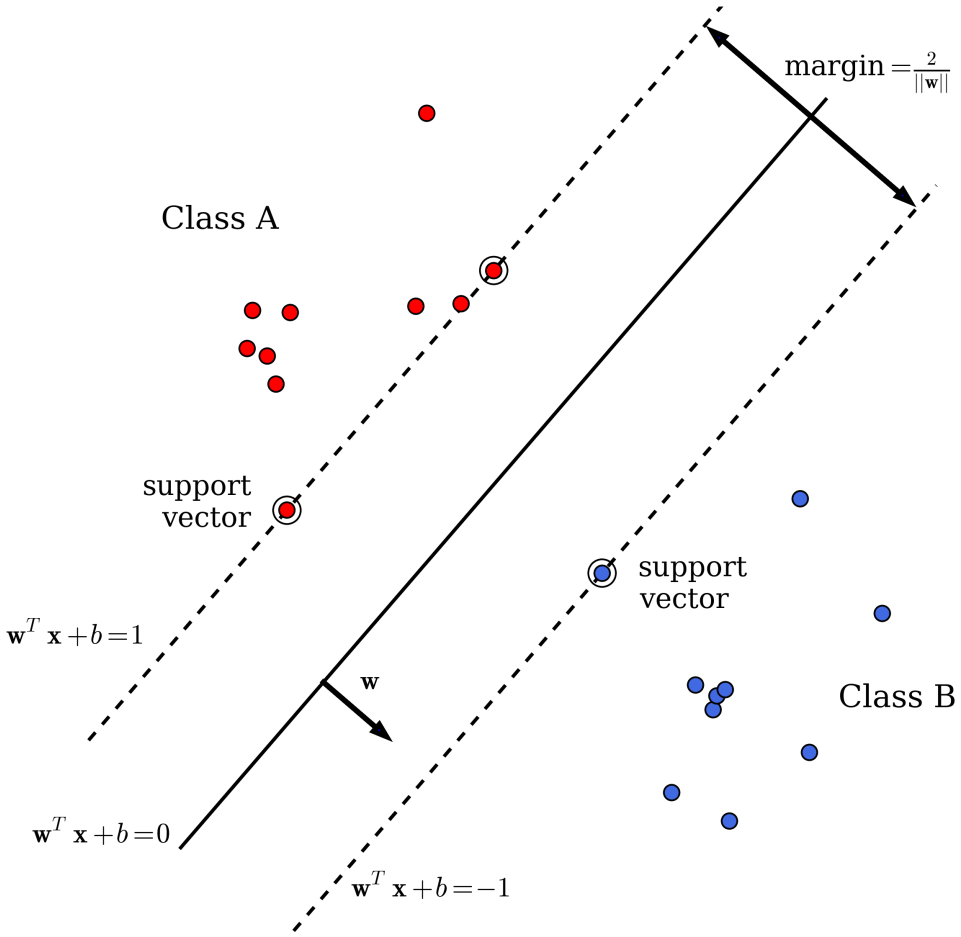


Figure 2.6: An illustration of a linear support vector machine (SVM) classifier using 2D training examples. The red points represent training examples that belong to class A, and the blue points represent training examples that belong to class B. The SVM tries to find the hyperplane (solid black line) that maximises the margin between the two classes. The circled dots correspond to the points that satisfy $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x}+b = \pm 1$ (known as the *support vectors*), where $\mathbf{w}$ is the normal to the separating hyperplane and b is the bias.

wide variety of classifiers available, many papers use a well-understood classification algorithm called a support vector machine (SVM) [Dalal and Triggs 2005; Lee et al. 2009a; Felzenszwalb et al. 2010]. Using a set of correctly labelled data, the SVM tries to find the hyperplane that maximises the margin between the two classes [Cortes and Vapnik 1995]. This hyperplane can then be used to classify previously unseen data (often called *test data*). Points on one side of the hyperplane are classified as class A, and points on the other side of the hyperplane belong class B. This is illustrated by the red and blue points in Figure 2.6.

Given N labelled, linearly separable training examples $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)$, where $\mathbf{x} \in \mathbb{R}^d$ and $y_i \in \{-1, 1\}$, the two classes can be separated by a linear classifier of the form:

$$f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b \tag{2.10}$$

where $\mathbf{w}$ is the normal to the separating hyperplane (also known as the *weight vector*), and b is the bias.

An optimal weight vector $\mathbf{w}$ can be found, such that $\|\mathbf{w}\|^2$ is minimised and the following constraints are satisfied:

$$\begin{aligned} \mathbf{w}^T\mathbf{x}_i + b &\geq 1 \qquad \text{if } y_i = +1 \\ \mathbf{w}^T\mathbf{x}_i + b &\leq -1 \qquad \text{if } y_i = -1 \end{aligned} \tag{2.11}$$

or equivalently:

$$y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 \tag{2.12}$$

The training examples that satisfy this equality (i.e. $y_i(\mathbf{w}^T\mathbf{x}_i + b) = 1$) are called *support vectors*. The support vectors define two hyperplanes: one through the support vectors belonging to class A, and one through the support vectors belonging to class B. In Figure 2.6, the circled dots represent the support vectors, and the dashed lines represent the corresponding hyperplanes (which are lines in the 2D example).

For cases where the training examples are not linearly separable, training errors are permitted, such that the optimisation becomes[3]:

$$\arg\min_{\mathbf{w}} \|\mathbf{w}\|^2 + C\sum_{i}^{N} \xi_i \tag{2.13}$$

subject to

[3]Equation 2.13 is the *primal* form of the SVM, and can be rewritten in terms of the support vectors $\mathbf{x}_j$, negating the need for $\mathbf{w}$. The alternative form of the equation is known as the *dual* form.

$$y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i \tag{2.14}$$

where C is a regularisation parameter, and ξ is a slack variable that allows training examples to exist in the space between the hyperplanes formed by the support vectors. The slack variable also allows training examples to be misclassified.

Equation 2.13 formulates the learning of the SVM as a constrained optimisation. Since $\xi \geq 0$, the slack variable can be written:

$$\xi_i = \max(0, 1 - y_i f(\mathbf{x})) \tag{2.15}$$

Consequently, the learning problem is equivalent to the unconstrained optimisation over $\mathbf{w}$:

$$\arg\min_{\mathbf{w}} \|\mathbf{w}\|^2 + C\sum_{i}^{N} \max(0, 1 - y_i f(\mathbf{x})) \tag{2.16}$$

where $\|\mathbf{w}\|$ is the regularisation term, and $\max(0, 1 - y_i f(\mathbf{x}))$ is the loss function.

One advantage of the linear SVM is that there are no hyper-parameters to optimise other than the constant C. Combined with "black-box" implementations such as LIBSVM [Chang and Lin 2011], this makes SVMs straightforward to apply. Unfortunately, the ease of use of SVMs is also a major drawback; the classifier can be unintuitive, especially when using high dimensional data that cannot be easily visualised. When a test example is classified as one class or the other, the only explanation of the process is that a weighted sum of elements of the test data are higher/lower than some threshold. The ability to convert the output of SVMs into a set of understandable rules is critical for a CAD tool, not only for acceptance by physicians and patients, but also for reducing regulatory barriers to diagnosis support systems based on SVMs [Fung et al. 2005].

Since SVMs try to separate the data in high dimensional space, rather than directly trying to minimise the classification error rate, they are fairly insensitive to the relative numbers of training examples in each class [Drucker et al. 1999]. For

example, adding extra training examples that are far behind the hyperplanes will not change the support vectors or classification boundary. This feature is advantageous in medical applications, where datasets may be biased towards one particular patient group. For this reason, in Chapter 5 we use an SVM to classify brain PET images as amyloid positive or amyloid negative using features based on histograms of 3D gradients.

2.3.3 Deep Learning

In recent years, research in computer vision has moved away from developing image descriptors, and instead, has focused on deep learning. This trend has also been noticeable in medical image analysis, where deep learning methods have recently been used for image segmentation and classification [Plis et al. 2014; Roth et al. 2015]. Figure 2.7 reveals that there were 48 papers related to deep learning in the 2015 MICCAI conference proceedings[4], compared to just three papers in 2011. Deep learning methods are breaking records in classification accuracy [Krizhevsky et al. 2012; Szegedy et al. 2012], and have reduced the classification error on the ImageNet dataset by 21.5% compared to using SIFT features and an SVM [Russakovsky et al. 2015]. Unlike other machine learning approaches, image features are automatically learnt in deep learning. This removes the subjective feature extraction step associated with machine learning, and contributes to their improvements in accuracy.

Despite the increased popularity of deep learning in medical image analysis, and its potential to improve classification accuracy, relatively few studies have explored deep learning methods for PET imaging applications. Rather than using deep learning to learn features directly from PET volumes, the images are often reduced to SUVR-like features and combined with structural MR information prior to the learning process [Suk and Shen 2013]. We address this problem in Chapter 7 by investigating the potential of a deep learning algorithm to derive features capable of classifying brain amyloid status directly from PET images.

[4]These figures were acquired by searching "deep learning" in the online MICCAI conference proceedings.

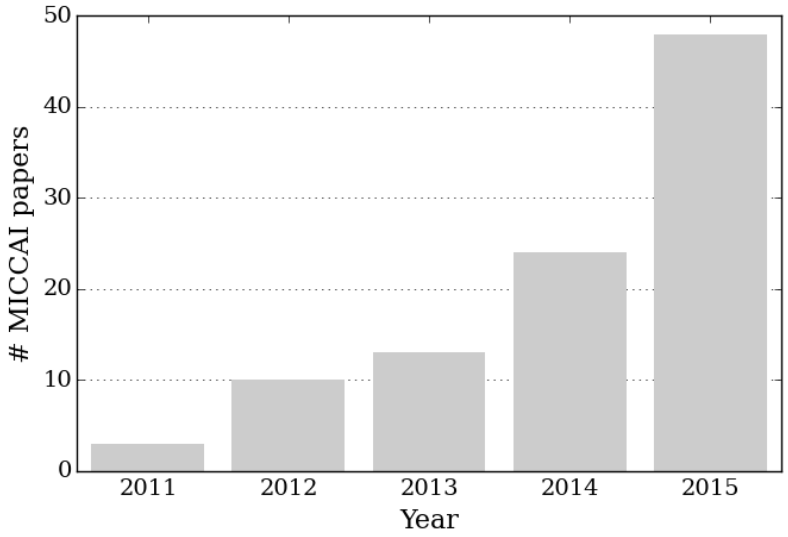


Figure 2.7: The number of papers in the MICCAI conference proceedings related to deep learning over a five year period from 2011 to 2015. Deep learning has clearly become a popular technique in medical image analysis.

Deep learning is characterised by models consisting of multiple layers of non-linear information processing [Deng and Yu 2013]. The depth of the models is another reason for the increased classification accuracy in computer vision tasks. Deep architectures allow the models to learn hierarchical representations of the input data, where the higher-level, more abstract concepts are defined from lower-level, less abstract ones. More abstract features are usually invariant to local changes of the input, and therefore can have greater predictive power [Bengio et al. 2014]. This hierarchy of features is clearly demonstrated in Figure 2.8, which shows the features learnt by a three-layer convolutional deep belief network that was designed to detect faces in photographs [Lee et al. 2009a]. The features in layer 1 correspond to edge detectors, the features in layer 2 correspond to parts of faces, and the features in layer 3 correspond to entire faces. It is not difficult to see why the higher-layer image features are more adept at detecting faces than the features from the first layer.

Many different deep learning methods have been proposed in the literature [LeCun et al. 1998; Hinton et al. 2006; Vincent et al. 2010], and new variants on existing architectures appear constantly. One of the first popular, and successful, approaches to deep learning was the deep belief network (DBN), proposed by [Hin-

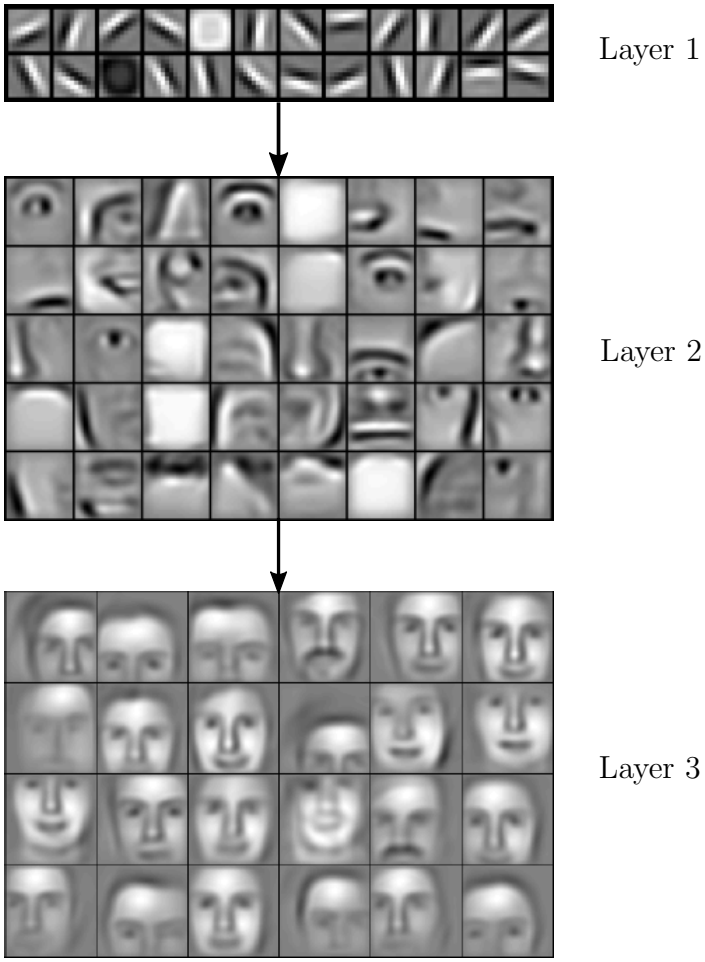


Figure 2.8: The features learnt by a three-layer convolutional deep belief network designed to detect faces in photographs. Higher layers learn higher-level features. The features in layer 1 correspond to edge detectors, the features in layer 2 correspond to parts of faces, and the features in layer 3 correspond to entire faces. This figure was adapted from [Lee et al. 2009a].

ton et al. 2006]. A DBN is composed of layers of restricted Boltzmann machines (RBMs).

An RBM is a two-layer, stochastic neural network, whose neurons form a bipartite graph. Binary visible units $\mathbf{v}$, whose states are known, are connected to binary hidden units $\mathbf{h}$ (the latent factors to be learnt) using symmetrically weighted connections (see Figure 2.9).

The energy of the joint configuration $(\mathbf{v}, \mathbf{h})$ is given by[5]:

[5]This only applies for binary-valued visible and hidden units. Formulations for Gaussian and multinomial units are presented in [Salakhutdinov et al. 2007].

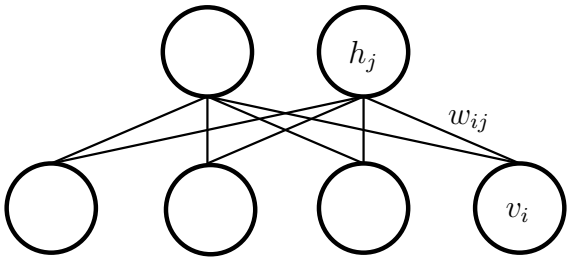


Figure 2.9: A restricted Boltzmann machine (RBM) with two hidden units (top row) and four visible units (bottom row). The units form a bipartite graph, where every visible unit v_i is connected to every hidden unit h_j by a weight w_{ij} and there are no connections within a layer.

$$E(\mathbf{v}, \mathbf{h}) = -\sum_{i,j} v_i h_j w_{ij} - \sum_i a_i v_i - \sum_j b_j h_j \tag{2.17}$$

where w_{ij} is the weight between visible unit v_i and hidden unit h_j, and a_i, b_j are their respective biases.

The probability of a pair of visible and hidden vectors is then defined:

$$p(\mathbf{v}, \mathbf{h}) = \frac{1}{Z} \exp(-E(\mathbf{v}, \mathbf{h})) \tag{2.18}$$

where Z is the partition function:

$$Z = \sum_{\mathbf{v},\mathbf{h}} \exp(-E(\mathbf{v}, \mathbf{h})) \tag{2.19}$$

The probability of visible vector $\mathbf{v}$ is obtained by marginalising over the hidden vectors, such that:

$$p(\mathbf{v}) = \frac{1}{Z} \sum_{\mathbf{h}} \exp(-E(\mathbf{v}, \mathbf{h})) \tag{2.20}$$

Due to the inverse relationship between energy and probability, by lowering the energy of a visible vector, the probability assigned to the visible vector can be increased. The RBM parameters can be optimised by performing stochastic gradient ascent on the log-likelihood of the training data:

$$\frac{\partial p(\mathbf{v})}{\partial w_{ij}} = \langle v_i h_j \rangle_{\text{data}} - \langle v_i h_j \rangle_{\text{model}} \tag{2.21}$$

where angle brackets denote expectations. This leads to an update rule for the weights:

$$\Delta w_{ij} = \epsilon(\langle v_i h_j \rangle_{\text{data}} - \langle v_i h_j \rangle_{\text{model}}) \tag{2.22}$$

where ϵ is the learning rate.

Despite the apparent simplicity of the update rule, the computation of the exact gradient in Equation 2.21 is intractable because it requires running alternating Gibbs sampling for a long time. As a result, a method known as *contrastive divergence* is used to approximate the gradient by running alternating Gibbs sampling for n steps [Hinton 2002]. Since the hidden units are not directly connected to one another, sampling an RBM is very efficient. Given a visible vector $\mathbf{v}$, the probability that the binary state of each hidden unit h_j is set to 1 is:

$$p(h_j = 1|\mathbf{v}) = \sigma(b_j + \sum_i v_i w_{ij}) \tag{2.23}$$

where $\sigma(x) = 1/(1 + \exp(-x))$ is the logistic sigmoid function.

Similarly, to sample the state of a visible unit v_i, given a hidden vector $\mathbf{h}$:

$$p(v_i = 1|\mathbf{h}) = \sigma(a_i + \sum_j h_j w_{ij}) \tag{2.24}$$

The process of alternate sampling can be repeated for any number of iterations, but often values of $n = 1$ are used. Despite this approximation to alternating Gibbs sampling being very crude, it has been shown to be effective for training RBMs [Hinton 2012].

By stacking RBMs, a hierarchical feature representation can be learnt. A deep belief network (DBN) is constructed by using the expected hidden unit values (also called *activations*) from one RBM as the input to another RBM. Each layer is

trained in isolation, and the activations of the second RBM can be connected to a third RBM, and so on.

While DBNs have been successfully used to learn high-level structures in many types of data, including handwritten digits [Hinton et al. 2006], motion capture data [Taylor et al. 2007], and speech [Mohamed et al. 2012], scaling them to work on large images is challenging. A DBN can only use 1D input data, and thus ignores the 2D structure of images. Moreover, as the size of the input image increases, the algorithm becomes increasingly computationally expensive because the vector of weights has the same number of elements as the input. This also means that individual features can only be detected at a single location within the input image.

Similar problems were encountered with deep neural networks, and were solved by using a convolutional architecture inspired by models of the visual cortex [LeCun et al. 1998]. Convolutional deep neural networks (CNN) learn feature detectors that are sensitive to small regions of the input called *receptive fields*. The regions are tiled to cover the entire input image, and the feature detectors are applied over the input space to form outputs known as *feature maps*. Neurons within a feature map have the same weights and biases, constraining them to perform the same operation on different parts of the image [LeCun et al. 1998]. This allows the same feature to be detected at every point in the image, meaning that a large image can be represented using a small number of feature detectors. The same principles were applied to an RBM by [Desjardins and Bengio 2008] and [Lee et al. 2009a], thus forming a convolutional RBM (CRBM). Just as stacked RBMs can form a DBN, stacked CRBMs can form a convolutional DBN (CDBN). CDBNs have been successfully used for face detection in photographs (see Figure 2.8) [Lee et al. 2009a], and have achieved achieved state-of-the-art classification performance on the CIFAR-10 data set, which contains images from ten different classes [Krizhevsky 2010]. Consequently, in this thesis we investigate the ability of CDBNs to classify brain amyloid status from PET images. For clarity, the mathematics of the CDBN are presented prior to that work in Chapter 7.

In this chapter, we have presented information about the pathology and biomarkers of AD, as well as the current method for brain amyloid quantification. We have also introduced the concept of image registration, and discussed approaches to machine learning, including the HOG feature descriptor and SVM classifier. Finally, we described RBMs and their role in the deep learning method known as deep belief networks. In the next chapter we propose a novel deformable registration method for matching amyloid PET images to a template. Our method extends the log-domain demons registration method discussed in Section 2.2.1 by using additional information from MR images to help inform the PET registration process.

Chapter 3

Combined PET-MR Brain Registration

For amyloid PET image quantification, the SUVR is calculated by computing the mean uptake of tracer in a region of interest, divided by the uptake in a reference region (where the uptake is non-specific). Typically, prior to SUVR calculation in brain images, the PET volume is transformed to a template space in which the regions of interest are defined. Current methods for quantitative analysis of amyloid PET imaging use affine or non-linear registration methods to achieve alignment to the PET template [Fleisher et al. 2011; Hutton et al. 2015]. Consequently, SUVRs could be somewhat dependent on the quality of the registration.

For amyloid PET images that are difficult to interpret visually, the tracer manufacturers' guidelines actively recommend the use of structural images to clarify the relationship of the tracer uptake and grey matter anatomy [Amyvid 2012; Neuraceq 2014]. By simultaneously viewing co-registered amyloid PET and MR images from the same individual, the additional anatomical information provided by the MR image can help localise regions of tracer uptake in the PET image, thus improving the diagnostic value of the visual assessment. However, in the SUVR calculation process, the transformation of the amyloid PET image to the template is usually restricted to a single modality. For example, [Fleisher et al. 2011] directly register the

amyloid PET images to an MR template, without using any MR imaging. In contrast, [Jack et al. 2008] register the amyloid PET images to an MR image acquired from the same individual, and then register the MR images to the MR template. The resulting transformations are then applied to the amyloid PET images. As a result, the PET images have no direct influence on the registration to the template space. It has been shown that significantly different SUVRs are obtained when using a PET-driven registration method compared to an MR-driven method with grey matter segmentation [Saint-Aubert et al. 2014]. Despite this discrepancy, it has also been shown that MR-driven and PET-driven approaches to quantitative analysis of amyloid PET images are both adequate for clinical diagnostic purposes [Edison et al. 2013]. Consequently, in this work we wish to remove the discrepancy between the SUVRs obtained from the PET-driven and MR-driven registration methods, whilst improving their potential diagnostic value.

An alternative approach would be to develop a registration method that combines both PET and MR image information. Such a method could exploit all the available data, and would potentially capture changes present in one modality, but not visible in the other. In turn, this could lead to more precise SUVRs. In a clinical setting, not all patients may have an MR scan in addition to their amyloid PET scan. Furthermore, if a patient does have an MR scan, the scan could be of poor quality. In these situations, a purely or mainly PET-driven registration approach is required, however, where a high quality MR image is present, it could very strongly guide the registration of the PET image to the template. As a result, a registration approach which is capable of using a weighted combination of the two modalities could be useful.

Therefore, the aim of this chapter is to assess the performance of a combined modality registration method for registering PET and MR volumes to a dual-modality template space. More accurate registrations might lead to SUVRs that are more representative of tracer uptake in specific brain regions, which in turn could result in SUVRs that are better able to discriminate between groups of AD patients

and healthy controls. To achieve this assessment, we propose a novel approach to combined non-linear PET-MR brain registration. We extend the log-domain demons registration framework [Vercauteren et al. 2008] described in Chapter 2.2.1, by combining the individual update steps in simultaneous single-modality demons registrations. Unlike existing multi-channel demons methods [Peyrat et al. 2010; Forsberg et al. 2011], which only use anatomical images, our proposed method combines functional and anatomical information at each iteration. Furthermore, to account for locally varying intensity inhomogeneity in the images, we employ a local correlation coefficient (LCC) image similarity metric [Cachier et al. 2003; Lorenzi et al. 2013], rather than the standard global criterion, sum of squared differences (SSD). We compared SUVRs following affine registration and five PET-MR LCC-demons registration methods, each with a different weighted combination of PET and MR modalities, including PET-only and MR-only.

The remainder of this chapter is structured as follows: In Section 3.1, the LCC log-domain demons framework is summarised, and the novel registration algorithm is outlined. Section 3.2 introduces the PET-MR data set, and describes the pre-processing steps and template construction. Section 3.3 describes the experiments, and Section 3.4 compares the PET templates and presents the registration results. Finally, Section 3.5 discusses the advantages and limitations of the combined non-linear PET-MR registration method, and concludes this work.

The work in this chapter is based on [Cattell et al. 2014].

3.1 Methods

3.1.1 Local Correlation Coefficient Log-Domain Demons

The log-domain demons algorithm estimates the diffeomorphic transform between a fixed image F and a moving image M [Vercauteren et al. 2008]. As described in Chapter 2.2.1, the transformation s is defined as the Lie group exponential map $s = \exp(\mathbf{v})$, where $\mathbf{v}$ is a stationary velocity field. The algorithm can be cast as the

minimisation of a global energy function, consisting of a similarity criterion Sim, used to measure the resemblance of the aligned images, and a regularisation term Reg:

$$E(F, M, c, s) = \frac{1}{\sigma_i^2}\text{Sim}(F, M, c) + \frac{1}{\sigma_x^2}\|\log(s^{-1} \circ c)\|^2 + \frac{1}{\sigma_T^2}\text{Reg}(s) \tag{3.1}$$

where σ_i accounts for the noise in the image, σ_x accounts for the spatial uncertainty of the correspondences, and σ_T controls the regularisation strength. In order for the optimisation to be well-posed, [Cachier et al. 2003] introduced an auxiliary correspondence field c (parameterised by a stationary velocity field $\mathbf{v}_x$), to decouple the minimisation into two simple steps at each iteration: 1) optimisation of the similarity, given $\mathbf{v}$, and 2) regularisation, given $\mathbf{v}_x$.

Using the Baker-Campbell-Hausdorff (BCH) formula to describe the composition of transformations in the log-domain [Bossa et al. 2007], the coupling term in Equation 3.1 can be written:

$$\begin{aligned}
\log(s^{-1} \circ c) &= \log(\exp(-\mathbf{v}) \circ \exp(\mathbf{v}_x)) \\
&= \text{BCH}(-\mathbf{v}, \mathbf{v}_x) \\
&= -\mathbf{v} + \mathbf{v}_x + \frac{1}{2}[-\mathbf{v}, \mathbf{v}_x] + \frac{1}{12}[-\mathbf{v}[-\mathbf{v}, \mathbf{v}_x]] + O(\|\mathbf{v}_x\|^2) \\
&= \mathbf{u}
\end{aligned} \tag{3.2}$$

where $\mathbf{u}$ is the update field. The Lie bracket $[-\mathbf{v}, \mathbf{v}_x]$ represents a velocity field defined at each point p [Vercauteren et al. 2008]:

$$[-\mathbf{v}, \mathbf{v}_x] = \text{Jac}(-\mathbf{v})(p)\mathbf{v}_x(p) - \text{Jac}(\mathbf{v}_x)(p)\mathbf{v}(p) \tag{3.3}$$

where $\text{Jac}(\mathbf{v})$ represents the Jacobian of the stationary velocity field $\mathbf{v}$.

In practice, the calculations are simplified by using the 0^{th} order approximation of the BCH expansion $\mathbf{u} = \text{BCH}(-\mathbf{v}, \mathbf{v}_x) \simeq -\mathbf{v} + \mathbf{v}_x$ [Lorenzi et al. 2013]. The log-domain demons algorithm is summarised in Algorithm 3.1.

Algorithm 3.1 Log-domain demons registration

1: **while** not converged **do**

2: Given the current transformation $s = \exp(\mathbf{v})$, compute an update field $\mathbf{u}$

3: For fluid-like regularisation, regularise $\mathbf{u}$ by convolving with a Gaussian kernel $\mathbf{u} \leftarrow G * \mathbf{u}$

4: $\mathbf{v} \leftarrow \mathrm{BCH}(\mathbf{v}, \mathbf{u})$

5: For diffusion-like regularisation, regularise $\mathbf{v}$ by convolving with a Gaussian kernel $\mathbf{v} \leftarrow G * \mathbf{v}$

6: **end while**

In the classical form, the choice of single-modality similarity measure is the sum of squared differences (SSD) between the image intensities at each voxel [Thirion 1998; Vercauteren et al. 2009]. Despite its simple implementation, SSD is a global criterion and assumes an identity relationship between the image intensities. This makes it extremely sensitive to the locally varying intensity biases often found in medical images.

By assuming that the bias distribution is locally uniform, [Cachier et al. 2003] proposed a local implementation of the correlation coefficient (LCC). The LCC similarity ρ implicitly estimates the locally linear relationship between the image intensities, thus accounting for additive and multiplicative bias:

$$\rho(F, M) = \int_{\sigma} \frac{\overline{FM}}{\sqrt{\overline{F^2} \cdot \overline{M^2}}} \tag{3.4}$$

where $\overline{F} = \mathbf{G}_{\sigma} * F(x)$ is the local mean image defined by convolution with a Gaussian $\mathbf{G}_{\sigma}$ of kernel size σ.

Unlike SSD, the LCC criterion has a parameter which needs to be chosen based on the application (i.e. the kernel size σ of the Gaussian filter). However, LCC optimisation fits neatly into the demons registration framework, in contrast to mutual information [Studholme et al. 1996], which requires the estimation of the global joint intensity histogram of the images. If $\mathrm{Sim}(F, M \circ s)$ is replaced by $\rho^2(F, M \circ s)$

in Equation 3.1, the correspondence energy can be written as:

$$E(F, M) = -\frac{1}{\sigma_i^2}\rho^2(F, M) + \frac{1}{\sigma_x^2}\|\mathbf{u}\|^2 + \frac{1}{\sigma_T^2}\text{Reg}(s) \tag{3.5}$$

Since the inverse transformation $s^{-1} = \exp(-\mathbf{v})$ can be easily acquired in the log-domain, [Vercauteren et al. 2008] proposed a symmetric registration framework. Unfortunately, the method is computationally expensive for similarity terms more complex than SSD. Therefore, [Lorenzi et al. 2013] proposed to symmetrise the registration framework by optimising in the halfway space, in which the fixed and moving images are resampled in a single energy term. Given two images F' and M', and the symmetric formulation $F = F' \circ \exp(-\frac{\mathbf{v}}{2})$ and $M = M' \circ \exp(\frac{\mathbf{v}}{2})$, [Lorenzi et al. 2013] showed that optimisation of the energy function with respect to the update $\mathbf{u}$ can be computed:

$$\mathbf{u} = -\frac{2\Lambda}{\|\Lambda\|^2 - \frac{4}{\rho^2}\frac{\sigma_i^2}{\sigma_x^2}} \tag{3.6}$$

where

$$\Lambda = \left(\frac{\mathbf{G}_\sigma * (F\nabla M^T)}{\mathbf{G}_\sigma * (FM)} - \frac{\mathbf{G}_\sigma * (M\nabla F^T)}{\mathbf{G}_\sigma * (FM)} + \frac{\mathbf{G}_\sigma * (F\nabla F^T)}{\mathbf{G}_\sigma * (F^2)} - \frac{\mathbf{G}_\sigma * (M\nabla M^T)}{\mathbf{G}_\sigma * (M^2)}\right) \tag{3.7}$$

3.1.2 Combined PET-MR LCC-Demons Registration

In our proposed combined PET-MR registration algorithm, it is assumed that the pairs of PET and MR volumes for each subject are already rigidly registered. This may be as a result of hybrid PET-MR image acquisition, or achieved using a registration algorithm such as FMRIB's Linear Image Registration Tool (FLIRT) [Jenkinson and Smith 2001; Jenkinson et al. 2002]. The non-linear registration of the PET-MR image pairs is then performed by combining the individual update fields from single-modality PET to PET and MR to MR LCC-demons registrations at

each iteration:

$$\mathbf{u} = \alpha \mathbf{u}_{\mathrm{MR}} + (1 - \alpha)\mathbf{u}_{\mathrm{PET}}, \quad 0 \leq \alpha \leq 1 \tag{3.8}$$

where the relative contribution of each modality can be controlled by varying the global weighting α. Hence, we modified the LCC log-domain demons algorithm, as presented in Algorithm 3.2.

Algorithm 3.2 Combined PET-MR LCC log-domain demons registration

1: **while** not converged **do**
2: Use Equation 3.6 to compute an update field $\mathbf{u}_{\mathrm{MR}}$ for the two MR images, and an update field $\mathbf{u}_{\mathrm{PET}}$ for the two PET images
3: Use Equation 3.8 to compute the combined update field $\mathbf{u}$
4: For fluid-like regularisation, regularise $\mathbf{u}$ by convolving with a Gaussian kernel $\mathbf{u} \leftarrow G * \mathbf{u}$
5: $\mathbf{v} \leftarrow \mathrm{BCH}(\mathbf{v}, \mathbf{u})$
6: For diffusion-like regularisation, regularise $\mathbf{v}$ by convolving with a Gaussian kernel $\mathbf{v} \leftarrow G * \mathbf{v}$
7: Apply the same updated transformation to the MR and PET images
8: **end while**

3.2 Materials

3.2.1 Alzheimer's Disease Neuroimaging Initiative Data

Unless otherwise stated, all the data used in this thesis are from the Alzheimer's Disease Neuroimaging Initiative (ADNI) online database[1]. The ADNI was launched in the USA in 2003, and its main objective has been to test whether clinical assessments, MRI, PET, and other biomarkers can be combined to robustly diagnose AD and measure the progression of mild cognitive impairment (MCI). The initiative is

[1] `http://adni.loni.usc.edu/`

the result of a collaboration between numerous academic institutions and private corporations, and subjects have been recruited from 55 sites across the USA and Canada. The initial goal of ADNI was to recruit 800 subjects, but the original ADNI was followed by ADNI-GO and ADNI-2. At the time of this thesis, ADNI have recruited over 1800 participants, aged 50-90, consisting of cognitively normal individuals, people with early or late MCI, and people with early AD. The baseline diagnosis for each subject is determined using neuropsychological tests, such as the Mini-Mental State Exam [Folstein et al. 1975], as well as a general assessment of cognition by a physician. A summary of the ADNI baseline diagnosis criteria is presented in Table 3.1. In addition, all subjects must be in good general health, with no diseases or psychological conditions that could interfere with the study. The duration of the follow-up depends on the ADNI protocol (ADNI-1, ADNI-GO, or ADNI-2). All of the ADNI subjects used in this thesis followed the ADNI-2 protocol [ADNI 2008], which specifies that, in addition to the baseline assessment and MR and amyloid PET imaging, subjects must undergo clinical and neuropsychological assessments every six months, and have a three-dimensional T1-weighted MR scan (on a 3T scanner using a gradient echo sequence) every 12 months. Moreover, ^{18}F-florbetapir amyloid PET imaging is to be performed every two years from the baseline, as funding permits.

3.2.2 Data Acquisition and Pre-Processing

For the experiments in this chapter, 38 subjects (19 AD, 19 healthy controls) were selected from the ADNI database. Each subject had a pair of ^{18}F-florbetapir PET and T1-weighted MR volumes, acquired no more than 12 months apart. Although the scans were acquired at multiple sites, all sites followed the same ADNI protocol (ADNI-2). Prior to the experiments, the PET volumes were rigidly registered to their corresponding MR volumes using Statistical Parametric Mapping version 8 (SPM8)[2]. The whole-head MR and PET volumes were then skull-stripped us-

[2] `http://www.fil.ion.ucl.ac.uk/spm/software/spm8/`

Table 3.1: A summary of the ADNI baseline diagnosis criteria for cognitively normal (CN) controls, subjects with early and late mild cognitive impairment (MCI), and subjects with AD [ADNI 2008].

CN	Early MCI	Late MCI	AD
Subject must be free of memory complaints	Subject must have subjective memory concern	Same as early MCI	Same as early MCI
Normal memory function documented by scoring above education adjusted cutoffs on the Logical Memory II test from the Wechsler Memory Scale – Revised [Wechsler 1987]: • ≥9 for 16+ years of education • ≥5 for 8-15 years of education • ≥3 for 0-7 years of education	Abnormal memory function documented by scoring within the education adjusted ranges on the Logical Memory II test from the Wechsler Memory Scale – Revised [Wechsler 1987]: • 9-11 for 16+ years of education • 5-9 for 8-15 years of education • 3-6 for 0-7 years of education	Abnormal memory function documented by scoring below education adjusted cutoffs on the Logical Memory II test from the Wechsler Memory Scale – Revised [Wechsler 1987]: • ≤8 for 16+ years of education • ≤4 for 8-15 years of education • ≤2 for 0-7 years of education	Same as late MCI
Mini-Mental State Exam score between 24 and 30 (inclusive)	Same as CN	Same as CN	Mini-Mental State Exam score between 20 and 26 (inclusive)
Clinical dementia rating = 0	Clinical dementia rating = 0.5	Same as early MCI	Clinical dementia rating = 0.5 or 1.0
Cognitively normal based on an absence of significant impairment in cognitive function or activities in daily life	General cognition and function sufficiently preserved such that a diagnosis of AD cannot be made by the physician	Same as early MCI	Fulfils the criteria for probable AD in [McKhann et al. 1984]

ing a mask constructed from the tissue segmentations of the MR images (obtained using SPM8 [Ashburner and Friston 2005]). Finally, using FSL's FLIRT software [Jenkinson and Smith 2001; Jenkinson et al. 2002], the MR brain volumes were affinely registered to the Montreal Neurological Institute (MNI) space. The resulting transformations were applied to the ^{18}F-florbetapir brain volumes. The three pre-processing steps are summarised in Figure 3.1.

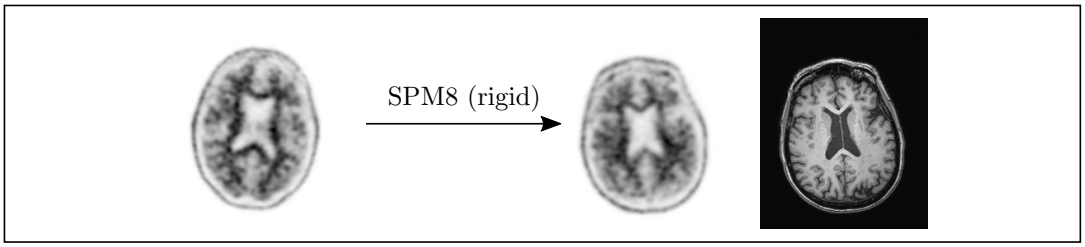
3.2.3 Template Construction

The use of an exemplar subject as the template could introduce a bias towards either the diseased or cognitively normal groups. Therefore, in an attempt to reduce this bias, an MR template was created iteratively from five MR brain volumes from AD patients and five healthy MR brain volumes, selected at random from the set of 38 subjects. The template construction algorithm, originally proposed by [Guimond et al. 2000], is summarised in Algorithm 3.3.

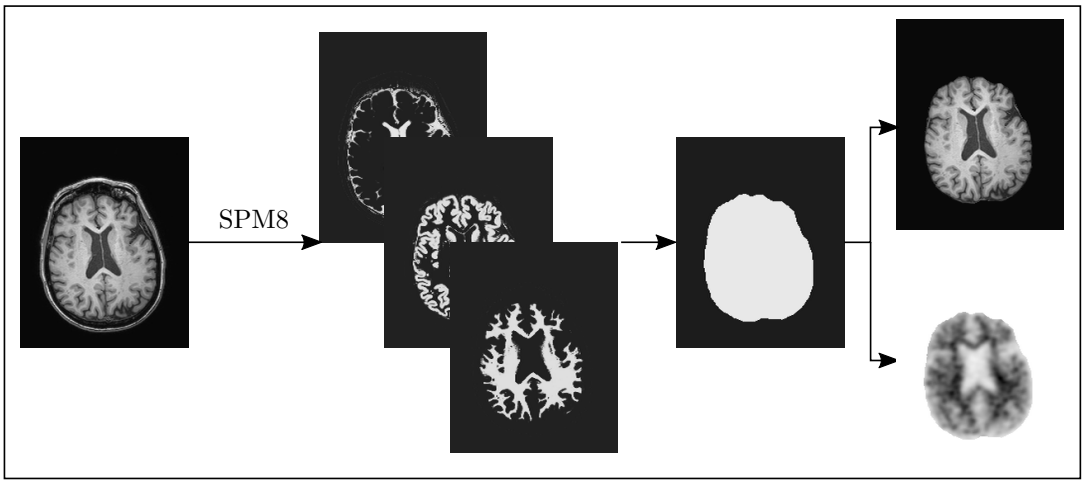
Algorithm 3.3 MR template construction

1: **given** reference image I_R, and set of N images $I_1, \ldots, I_N$
2: Transform the intensities of I_i to be in the intensity range of I_R.
3: Affinely register each I_i to I_R to obtain the affine transformations A_i.
4: **repeat**
5: Non-linearly register each I_i to I_R using A_i as the initial transformation. This provides registered images I_i' and transformations s_i.
6: Compute the mean registered image $\bar{I}' = \frac{1}{N}\sum I_i'$.
7: Compute the mean transformation $\bar{s} = \frac{1}{N}\sum s_i$.
8: Apply $\bar{s}^{-1}$ to $\bar{I}'$ to obtain the anatomical average model I_T.
9: $I_R \leftarrow I_T$
10: **until** convergence

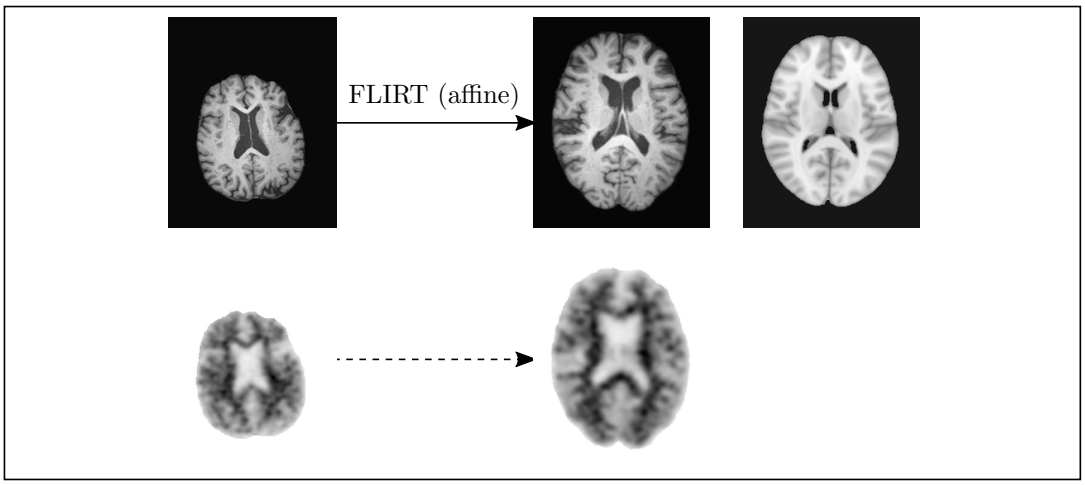
Since the MR brain volumes were affinely registered to MNI space during pre-processing, the initial reference image I_R was chosen to be the non-linear MNI152



(a)



(b)



(c)

Figure 3.1: Prior to the experiments, the PET and MR volumes underwent three pre-processing steps: (a) Using SPM8, the PET volumes were rigidly registered to the corresponding MR volumes in subject space. (b) The MR and PET volumes were skull-stripped using a mask constructed from the MR tissue segmentations (obtained using SPM8). (c) The MR brain volumes were affinely registered to MNI space using FLIRT. The resulting transformations were applied to the corresponding PET brain volumes (dashed line).

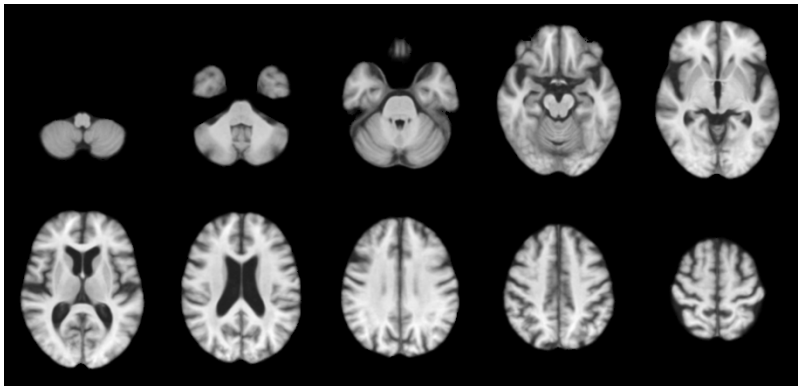
template available in FSL. More specifically, this MR template is known as the 6th generation non-linear MNI-ICBM152 average brain [Grabner et al. 2006]. The original log-domain demons registration algorithm [Vercauteren et al. 2008] was used to non-linearly register the images I_i to the reference (see line 5 in Algorithm 3.3). Axial slices from the final MR template are shown in Figure 3.2(a).

To create the PET template, the spatially normalised PET volumes corresponding to the ten subjects used to generate the MR template were transformed into the template space using the deformation fields s from the final iteration of the MR template construction. The registered PET volumes were then averaged to form the mean registered PET volume. The final amyloid PET template was created by applying the inverse of the mean MR deformation field $\bar{s}$ to the mean registered PET volume. The PET template construction process is summarised in Algorithm 3.4. Axial slices from the PET template are shown in Figure 3.2(b).

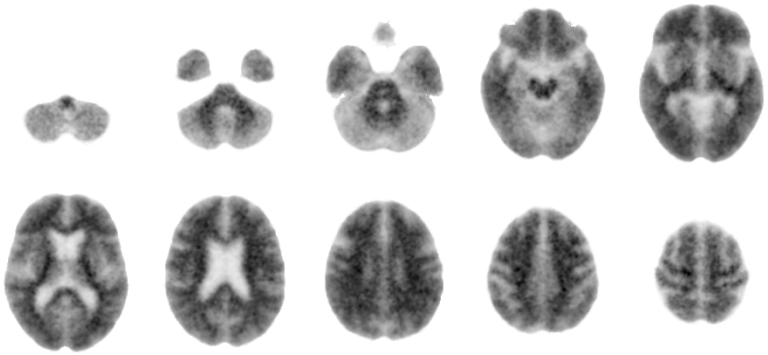
Algorithm 3.4 Real PET template construction

1: **given** set of N PET images $I_1 \dots I_N$ co-registered with the corresponding individuals' MR images, and transformations $s_1 \dots s_N$ used to transform the MR images into template space
2: **for** $i = 1$ to N **do**
3: Apply transformation s_i to I_i to obtain a PET image in template space I'_i.
4: **end for**
5: Compute the mean registered PET image $\bar{I}' = \frac{1}{N} \sum I'_i$.
6: Compute the mean MR transformation $\bar{s} = \frac{1}{N} \sum s_i$.
7: Apply $\bar{s}^{-1}$ to $\bar{I}'$ to obtain the anatomical average real PET template I_T.

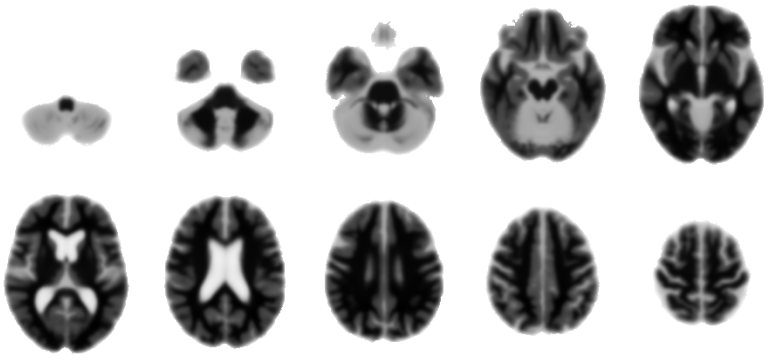
During the pre-processing of the data, the PET volumes were rigidly registered to their corresponding MR volumes. Consequently, any errors in the rigid registrations would be propagated into the PET template construction, which in turn could bias the combined PET-MR registration, since the two templates would not be perfectly aligned. Therefore, to ensure an unbiased starting point for the combined PET-MR registration, we also evaluated a synthetic PET template constructed from the



(a)



(b)



(c)

Figure 3.2: (a) Axial slices from the iterative MR template. (b) The corresponding slices from the PET template constructed from ^{18}F-florbetapir data (real PET template). (c) The corresponding axial slices from the synthetic PET template.

MR template. Similarly to [Peyrat et al. 2012], the synthetic PET template was created by heuristically combining the MR template tissue segmentations (obtained using SPM8) with different weighting factors, such that the synthetic PET template resembled an amyloid PET image. Axial slices from the synthetic PET template are shown in Figure 3.2(c). To avoid confusion between the synthetic PET template and the template constructed from actual ^{18}F-florbetapir data, the latter will henceforth be called the *real* PET template.

3.3 Experiments

3.3.1 Affine and Non-Linear Registration

The MR volumes of the 28 subjects excluded from the template creation were affinely registered to the MR template using FLIRT [Jenkinson and Smith 2001; Jenkinson et al. 2002]. The resulting transformations were then applied to the corresponding PET volumes.

Starting with the affinely registered images, five separate PET-MR LCC log-domain demons registrations were conducted for each of the 28 subjects: $\alpha = 0$ (equivalent to PET to PET registration), $\alpha = 1$ (equivalent to MR to MR registration), and $\alpha = \{0.25, 0.5, 0.75\}$ for the proposed combined PET-MR registration algorithm (with increasing influence of the MR volumes).

In [Lorenzi et al. 2013], the LCC demons registration method was successfully used to register pairs of MR volumes. Since the registration framework proposed in this chapter combines the individual updates of simultaneous single-modality LCC log-domain demons registrations, we used the registration parameters from [Lorenzi et al. 2013] to compute the update $\mathbf{u}_{\text{MR}}$ for the MR images: $\sigma_{\text{LCC}} = 5$ and $\sigma_i/\sigma_x = 0.05$. To date, the LCC demons registration method has not been used to register PET brain volumes. Therefore, in the absence of established parameters for PET registration, we used the same parameters to compute both $\mathbf{u}_{\text{PET}}$ and $\mathbf{u}_{\text{MR}}$. Despite not being optimised for our combined PET-MR registration framework, we

used the regularisation parameters and number of iterations from [Lorenzi et al. 2013]: $\sigma_{\text{fluid}} = 0.5$ and $\sigma_{\text{diffusion}} = 1.5$, with a multi-resolution, multi-scale scheme of $30\times99\times10$ iterations (coarse to fine).

To assess the effect of the real and synthetic PET templates on the registration results, all the PET-MR registration experiments were repeated using both PET templates.

3.3.2 Dice Overlap

To assess the performance of the different registration methods, an overlap agreement measure was used to evaluate how well anatomical regions were registered to one another. The Dice coefficient (also know as the Dice similarity index) quantifies where region labels in the fixed and moving volumes agree or disagree [Zijdenbos et al. 1994]:

$$d = 2\frac{\sum_r |F_r \cap M_r|}{\sum_r (|F_r| + |M_r|)} \tag{3.9}$$

where, in this equation, M_r refers to a transformed region label from the moving image, and F_r refers to the corresponding region label from the fixed image.

The standard cortical regions in native subject space were created automatically using an independent and established registration method. To achieve this, SPM8 was used to perform non-linear registration of each whole-head MR volume (in native subject space) to MNI space. The resulting inverse deformation fields were used to resample the Automatic Anatomic Labelling (AAL) atlas regions [Tzourio-Mazoyer et al. 2002] to native subject space. Following this procedure, any transformations applied to the MR brain volumes (e.g. those described in Sections 3.2.2 and 3.3.1) were also applied to the corresponding atlas. The MR template atlas was constructed by taking the deformed atlases of the subjects comprising the template, and combining them using a majority vote rule at each voxel [Heckemann et al. 2006]. The final MR template atlas is shown in Figure 3.3.

For each registration method, Dice coefficients were calculated between the MR

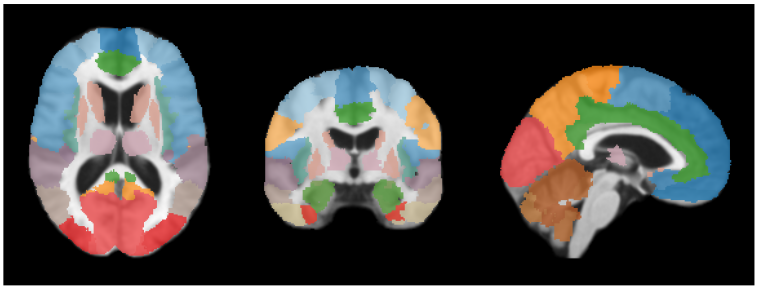


Figure 3.3: Axial, sagittal, and coronal slices of the MR template atlas overlaid on the corresponding slices from the MR template.

template atlas and each of the deformed subject atlases. Paired t-tests were performed between the Dice coefficients of each registration method and every other method. The results were subjected to Bonferroni correction to determine which pairs of registration methods achieved mean Dice coefficients that were significantly different from one another.

3.3.3 Amyloid PET Quantification

Following the registrations, SUVRs were computed from the ratio of ^{18}F-florbetapir retention in regions of interest to the cerebellum. The six cortical regions were shown to have prominent uptake of Aβ in AD versus control [Fleisher et al. 2011]: medial orbital frontal, parietal, temporal, precuneus, anterior cingulate and posterior cingulate. In this work, the target and reference regions were based on the AAL atlas regions in MNI space [Tzourio-Mazoyer et al. 2002], and manually dilated/eroded so they were visually similar to those presented in [Fleisher et al. 2011; Hutton et al. 2015]. The SUVR regions were transformed from their native MNI space to better fit the synthetic PET template, by applying the transformation obtained from the registration of the MNI152 MR average brain to the MR template using SPM8. The final SUVR regions used in this work are presented in Figure 3.4.

In order to best distinguish between patients with AD and healthy individuals using SUVRs, one would expect a large difference in mean SUVR between the two groups, and a small variance for each group. Therefore, using the SUVRs acquired

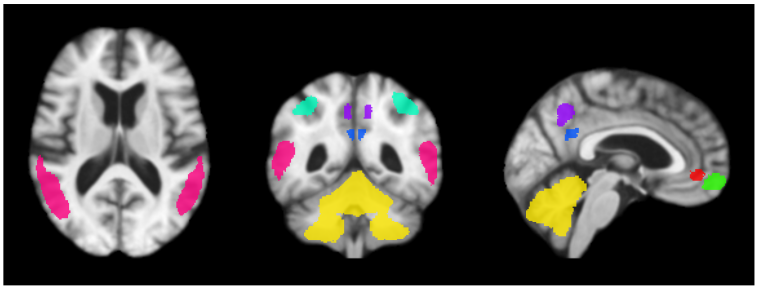


Figure 3.4: The target and reference regions used to calculate the mean SUVR. The six cortical regions shown to have prominent uptake of Aβ in AD versus control are: medial orbital frontal (green), parietal (cyan), temporal (magenta), precuneus (purple), anterior cingulate (red) and posterior cingulate (blue). The reference region is the cerebellum (yellow).

from the registered amyloid PET images, the SUVR means and standard deviations were calculated for each group. Welch's t-test, which allows for different standard deviations for each group, was then used to test whether the population means were different. For this application, the t statistic in Welch's t-test was defined:

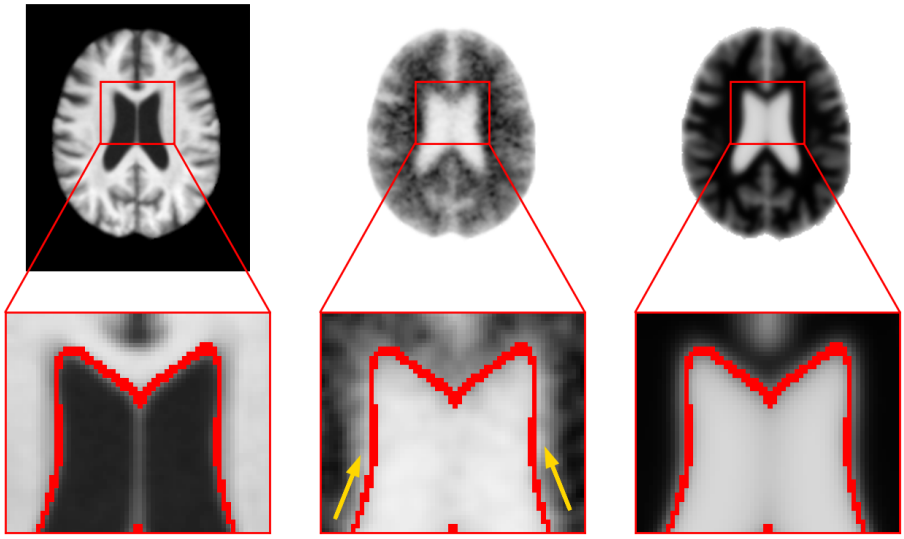
$$t = \frac{\overline{\text{SUVR}}_{\text{AD}} - \overline{\text{SUVR}}_{\text{HC}}}{\sqrt{\frac{\sigma^2_{\text{AD}}}{n_{\text{AD}}} + \frac{\sigma^2_{\text{HC}}}{n_{\text{HC}}}}} \tag{3.10}$$

where the subscript denotes the patient group (AD or healthy control), and $\overline{\text{SUVR}}$, σ^2, and n are the mean SUVR, sample variance, and sample size, respectively. A large t statistic implies that there is a large difference in mean SUVR between the AD and control groups, and a small standard deviation for each group. Consequently, a large t statistic suggests that the SUVRs obtained following registrations by a particular method are well suited to distinguishing between populations of AD patients and cognitively normal controls.

3.4 Results

3.4.1 Visual Assessment of the Templates

In this chapter we have proposed a method for simultaneously registering PET and MR volumes to a template space for SUVR calculation. In addition to comparing



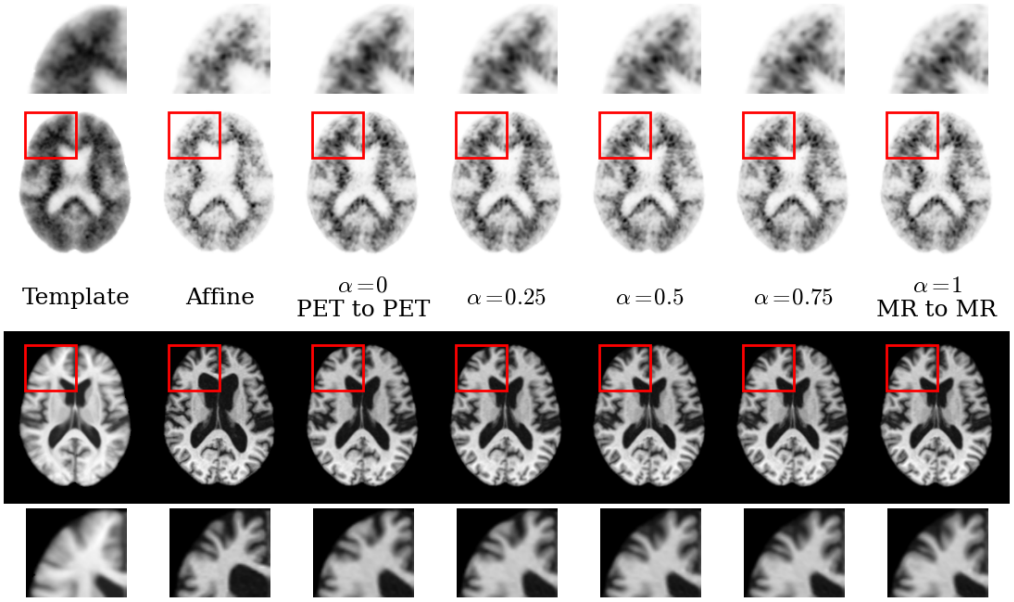
(a) (b) (c)

Figure 3.5: Example axial slice from the (a) MR template, (b) the real PET template, and (c) the synthetic PET template. In the zoomed-in regions, the outline of the ventricles from the MR template is superimposed (shown in red) on the ventricles of all three templates. The ventricles of the real PET template do not exactly match the ventricles of the MR template, as low tracer uptake is visible outside of the red boundary (yellow arrows).

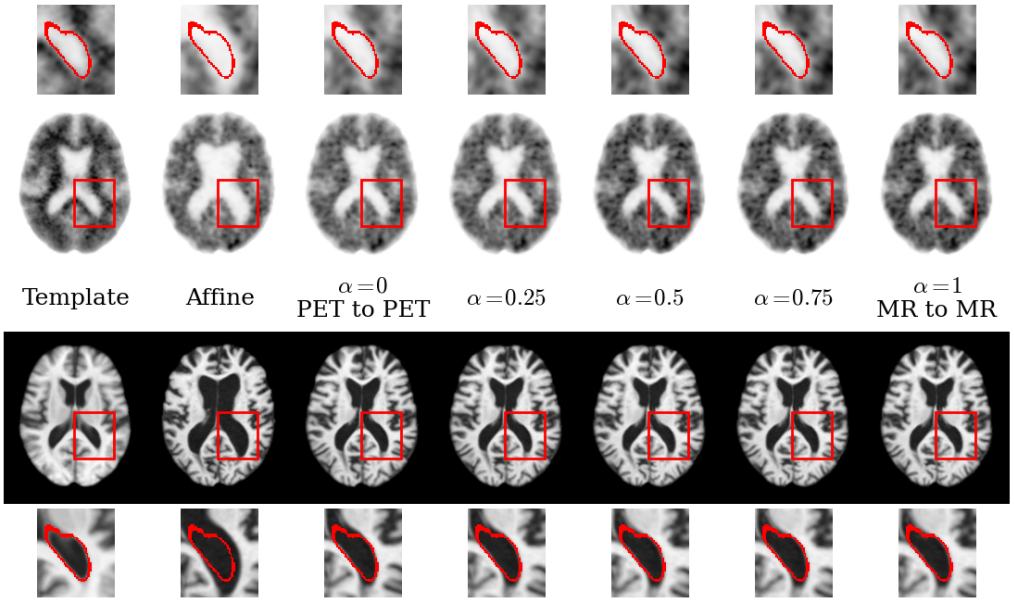
our combined PET-MR registration method to non-linear single-modality registration methods, we also compared the effect of using a real PET template and a synthetic PET template on the PET-MR registration accuracy and subsequent SUVRs. Figure 3.5 compares a corresponding axial slice of the MR template, real PET template, and synthetic PET template. The outline of the ventricles in the MR template is superimposed (shown in red) on the ventricles of all three templates in the zoomed-in region. The ventricles of the real PET template do not exactly match the ventricles of the MR template, as areas of low tracer uptake are visible outside of the red ventricle boundary. This is highlighted by the yellow arrows in Figure 3.5(b). In contrast, the MR-derived synthetic PET template is, by construction, perfectly registered to the MR template.

3.4.2 Registration Results Using the Real PET Template

Figure 3.6 shows the registration results for the AD patients that achieved the smallest and largest range of Dice coefficients across the six registration methods

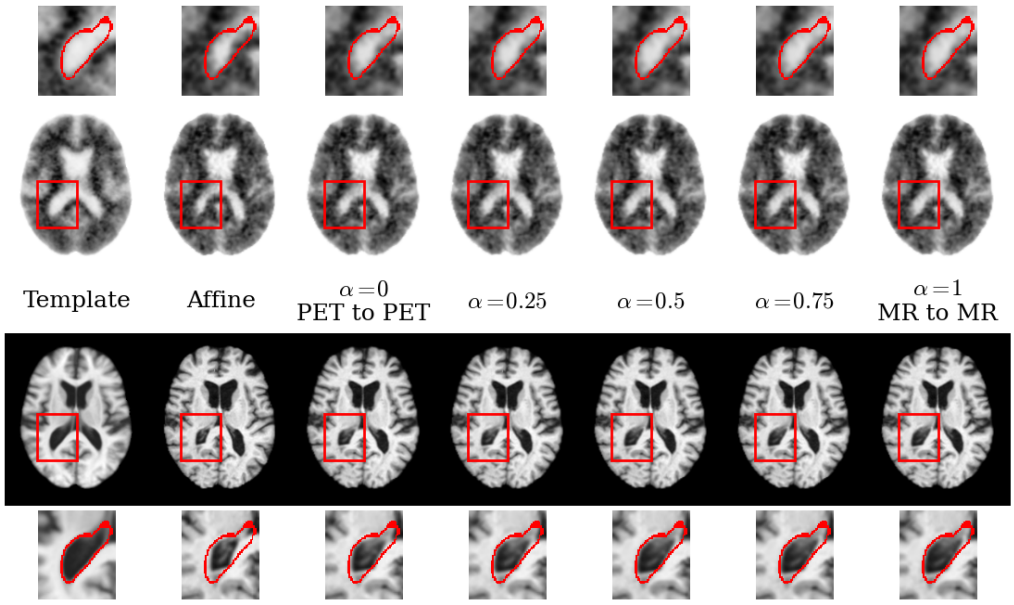


(a)

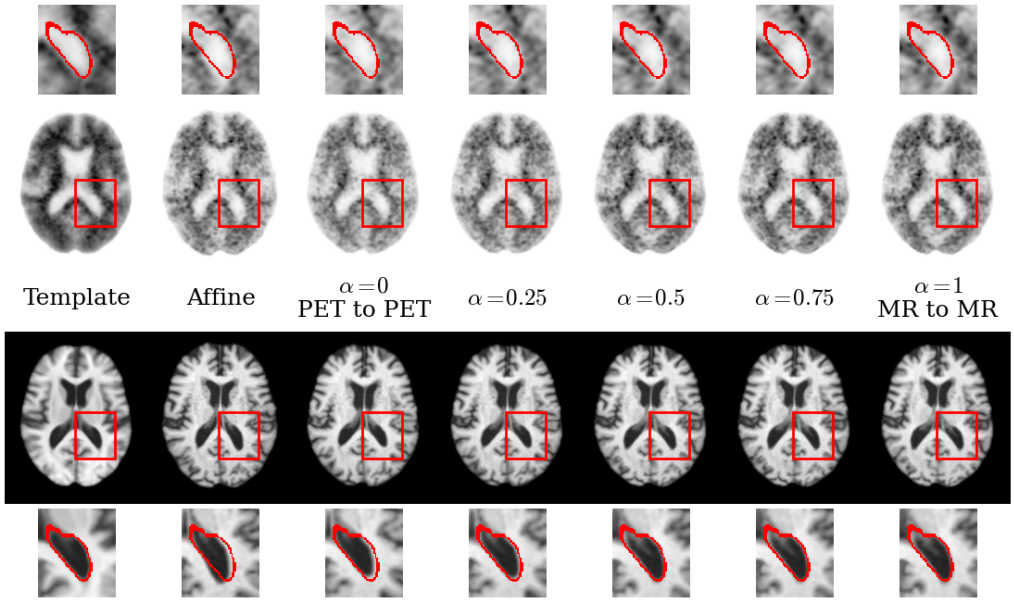


(b)

Figure 3.6: Sample axial slices from the volumetric registration results for the AD subjects that achieved (a) the smallest and (b) the largest range of Dice coefficients across the six registration methods using the real PET template. The middle two rows of each sub-figure show the corresponding axial slice from the deformed PET and MR volumes, respectively. The top and bottom rows of each sub-figure show a close-up of the region highlighted by the red box, with the outline of the MR template ventricles superimposed in red. In (a) the $\alpha = 0$ registration method exhibits poor registration in the frontal lobe compared to the other methods, since the gyri are unnaturally stretched. In (b), the non-linear registration methods show improved registration results compared to the affine method, since the registered ventricles better match the MR template ventricle outline.



(a)



(b)

Figure 3.7: Sample axial slices from the volumetric registration results for the healthy controls that achieved (a) the smallest and (b) the largest range of Dice coefficients across the six registration methods using the real PET template. The middle two rows of each sub-figure show the same axial slice from the deformed PET and MR volumes, respectively. The top and bottom rows of each sub-figure show a close-up of the region highlighted by the red box, with the outline of the MR template ventricles superimposed in red. In both sub-figures, the non-linear registration methods show improved registration results compared to the affine method, since the registered ventricles better match the MR template ventricle outline.

using the real PET template. The equivalent results for the healthy controls are presented in Figure 3.7. In Figures 3.6(b) and 3.7 the non-linear registration methods ($\alpha = \{0, 0.25, 0.5, 0.75, 1\}$) show improved registration compared to the affine method. This is highlighted by the close-up regions in the figures, where the ventricles in the non-linearly registered MR images more closely match the ventricles of the MR template (superimposed in red) than the ventricles in the affinely registered MR images. In Figure 3.6(a), the non-linear PET to PET ($\alpha = 0$) registration method exhibits poor registration in the frontal lobe compared to the other non-linear methods. The $\alpha = 0$ registered PET image shows that the tracer uptake in the white matter has been over-stretched, leading to a distortion visible in the MR image. This effect decreases as the contribution of the PET modality decreases (i.e. as α increases).

3.4.3 Registration Results Using the Synthetic PET Template

Figure 3.8 presents the registration results for the AD patients that achieved the smallest and largest range of Dice coefficients across the six registration methods using the synthetic PET template. These are the same AD patients as shown in Figure 3.6. Once again, the non-linear registration methods ($\alpha = \{0, 0.25, 0.5, 0.75, 1\}$) demonstrate improved registration in the ventricles compared to the affine registration method, since the registered ventricles better match the MR template ventricle outline (overlaid in red in the zoomed-in regions).

Figure 3.9 shows the registration results for the healthy controls that achieved the smallest and largest range of Dice coefficients across the different registration methods using the synthetic PET template. The same healthy control is shown in Figures 3.7(a) and 3.9(a), but the subject presented in Figure 3.9(b) is different to the subject in Figure 3.7(b). Identically to Figures 3.6(b), 3.7, and 3.8, the zoomed-in regions in Figure 3.9(a) indicate that the non-linear registration methods show improved registration compared to the affine registration method. Once again, the

affinely registered ventricles do no match the outline of the ventricles from the MR template.

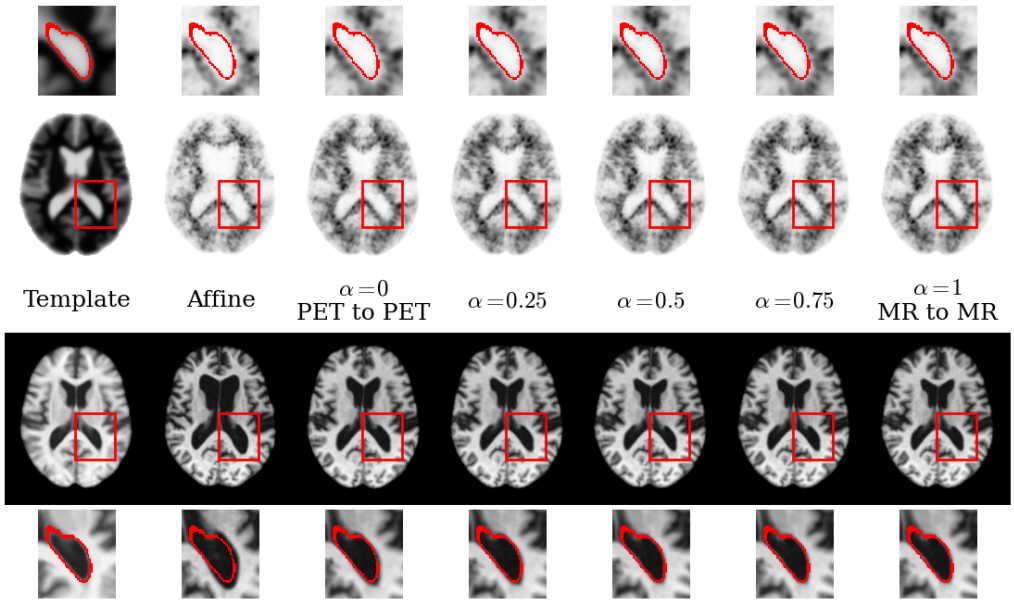
Similarly to Figure 3.6(a), the PET to PET ($\alpha = 0$) registration method in Figure 3.9(b) exhibits poor registration in the frontal brain area compared to the other methods. This is demonstrated in the close-up regions, where it is clear that the gyri are unnaturally stretched.

3.4.4 Registration Accuracy

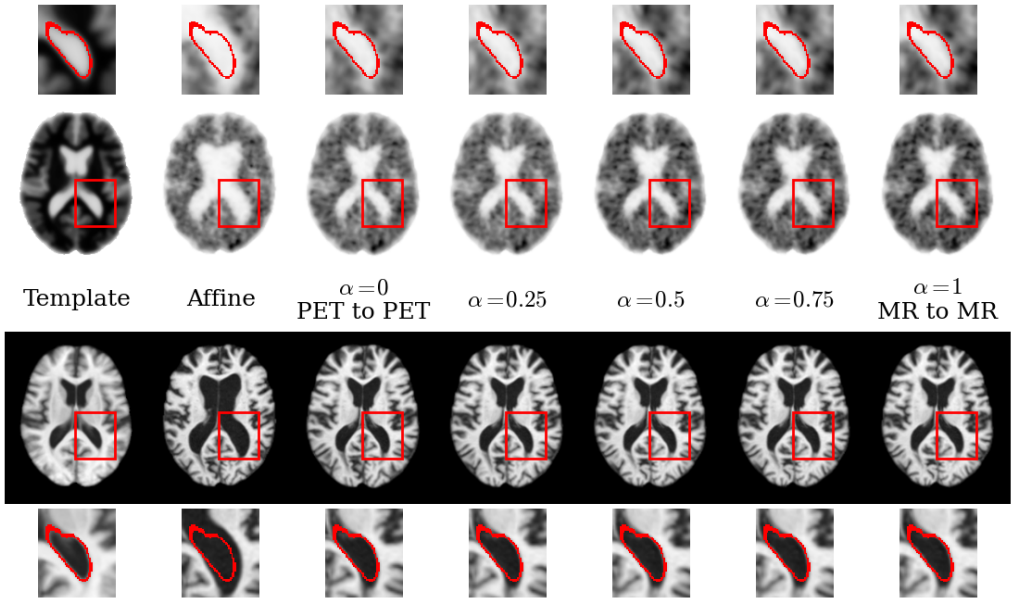
Figure 3.10 shows the distributions of Dice coefficients for each registration method. For comparison, the results acquired using the synthetic PET template are shown in blue, and the results from the real PET template are shown in red. As expected, the median Dice coefficient for the affine registration method (0.72) was smaller than the median Dice coefficients of the non-linear methods. Moreover, the paired t-tests on the overlap measures showed that all of the non-linear registration methods – except the $\alpha = 0$ method using synthetic template – resulted in significantly higher mean Dice coefficients than the affine method ($p < 0.01$, corrected).

For both the real and synthetic PET templates, the median Dice coefficient for the non-linear registration methods increased with increasing α. The $\alpha = 0$ registration method using the synthetic PET template resulted in the lowest median Dice coefficient of all the non-linear methods (0.74). Furthermore, the mean Dice coefficients for the $\alpha = 0$ and $\alpha = 0.25$ registration methods were significantly lower using the synthetic PET template compared to the real PET template ($p < 0.01$). This is indicated by the $\star$ symbol in Figure 3.10. In contrast, the $\alpha = 1$ registration method and both $\alpha = 0.75$ registration methods achieved the same, highest median Dice coefficient (0.84) and the same interquartile range (0.02).

The affine registration method resulted in the largest interquartile range (0.05), while the $\alpha = 0.5$ PET-MR registration method using the synthetic template had the smallest interquartile range (0.01). Although the interquartile range of the $\alpha = 0.5$ registration method was smaller for the synthetic PET template than the real PET

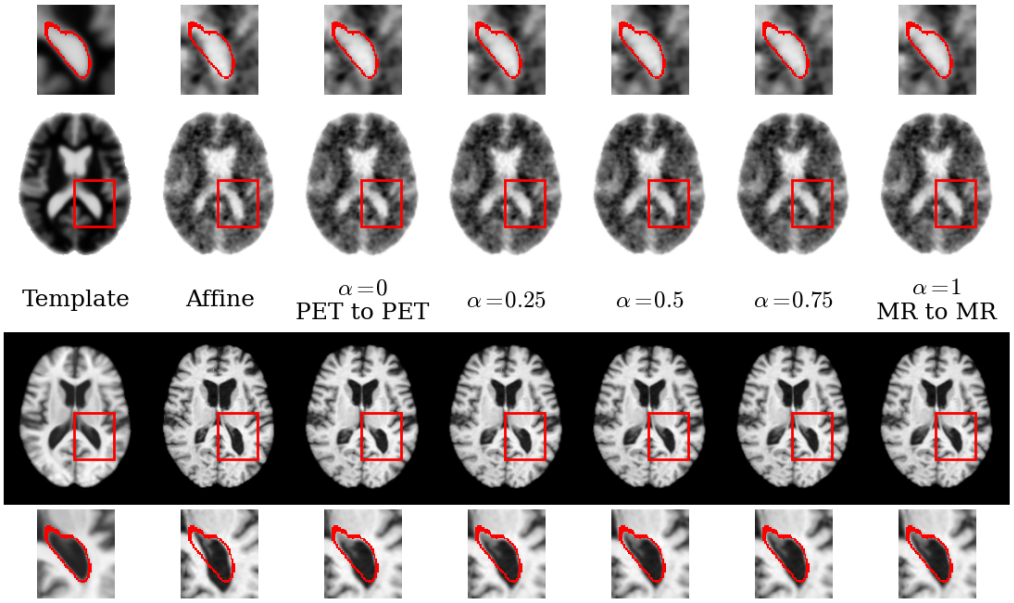


(a)

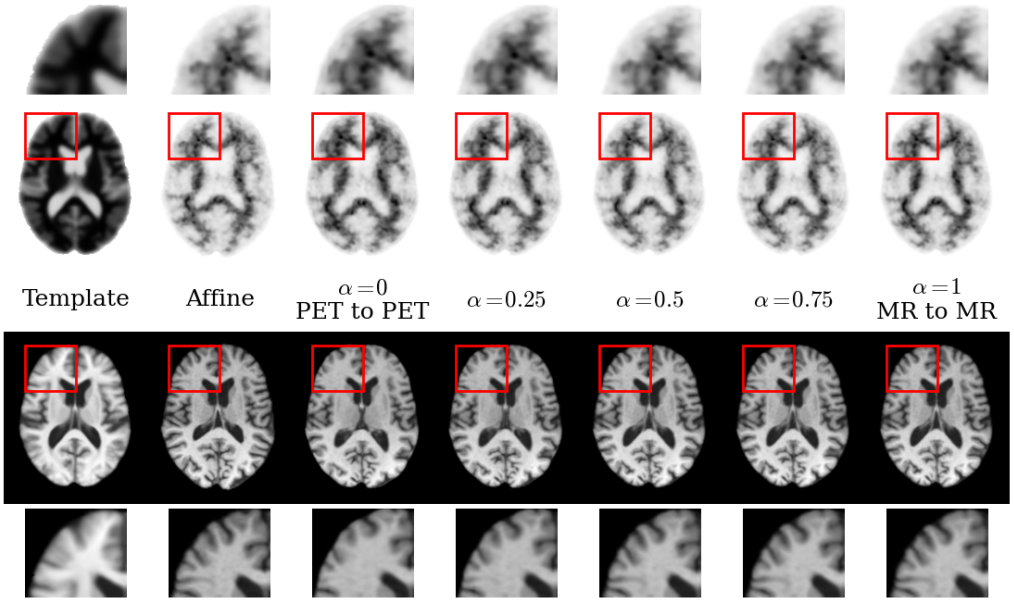


(b)

Figure 3.8: Sample axial slices from the volumetric registration results for the AD subjects that achieved (a) the smallest and (b) the largest range of Dice coefficients across the six registration methods using the synthetic PET template. The middle two rows of each sub-figure show the same axial slice from the deformed PET and MR volumes, respectively. The top and bottom rows of each sub-figure show a close-up of the region highlighted by the red box, with the outline of the MR template ventricles superimposed in red. In both sub-figures, the non-linear methods show improved registration in the ventricles compared to the affine registration method, since the registered ventricles better match the MR template ventricle boundary.



(a)



(b)

Figure 3.9: Sample axial slices from the volumetric registration results for the healthy controls that achieved (a) the smallest and (b) the largest range of Dice coefficients across the six registration methods using the synthetic PET template. The middle two rows of each sub-figure show the same axial slice from the deformed PET and MR volumes, respectively. The top and bottom rows of each sub-figure show a close-up of the region highlighted by the red box, with the outline of the MR template ventricles superimposed in red. In (a), the non-linear registration methods show improved registration results compared to the affine registration method, since the registered ventricles better match the MR template ventricle boundary. In (b), the $\alpha = 0$ registration method exhibits poor registration in the frontal lobe compared to the other methods, since the white matter is over-stretched.

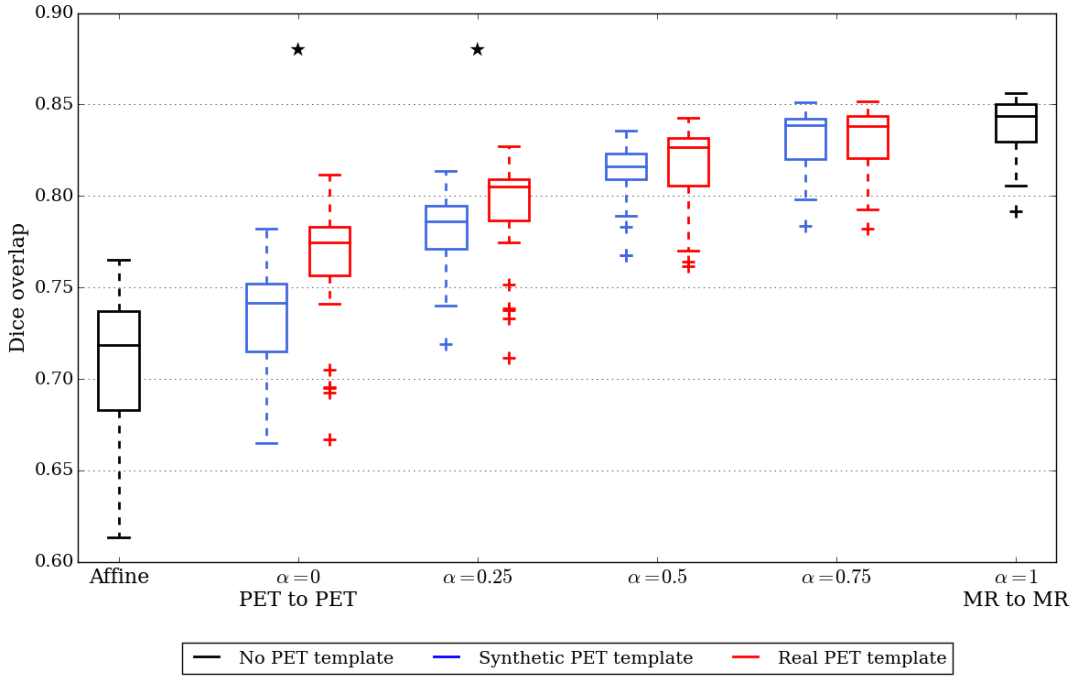


Figure 3.10: Distributions of overlap measure by registration method. For comparison, the Dice coefficients from the synthetic PET template are shown in blue, and the overlaps from the real PET template are shown in red. The $\star$ symbols denote methods for which there was a significant difference in Dice coefficients between the synthetic and real PET templates ($p < 0.01$). The box and whisker plots show the Dice coefficients between the deformed image labels and the MR template labels, averaged over all regions in the label set, and over all subjects. Each box has lines at the lower quartile, median, and upper quartile. The whiskers extend from each edge of the box to the most extreme values within 1.5 times the interquartile range from the box. Outliers (+) have values beyond the ends of the whiskers.

template (0.01 and 0.03, respectively), the median Dice coefficient was higher for the real PET template (0.83 and 0.82, respectively).

The paired t-tests on the Dice coefficients of the non-linear registration methods using the synthetic PET template showed that the mean Dice coefficients were all significantly different from one another ($p < 0.01$, corrected). Similarly, for the real PET template, all pairs of mean Dice coefficients were significantly different from one another ($p < 0.01$, corrected).

3.4.5 Standardised Uptake Value Ratio

Table 3.2 presents the mean SUVR of each subject group (AD or cognitively normal controls) for each registration method, as well as the corresponding t statistic from

Welch's t-test, which was used to assess the ability of the SUVRs to discriminate between the subject populations. A large t statistic suggests that the SUVRs acquired following registrations by a particular method are well suited to distinguishing between groups of AD patients and healthy controls. The results for the affine and $\alpha = 1$ registration methods are identical across both PET templates because they are not influenced by the PET template.

As expected, for both PET templates, the AD group had a significantly higher mean SUVR than the control group across all registration methods ($p < 0.01$). The standard deviation of the mean SUVRs was also higher for the AD group across all of the registration methods tested. The non-linear PET to PET ($\alpha = 0$) registration method consistently achieved the smallest standard deviation of mean SUVRs. However, for both PET templates, it resulted in the smallest absolute difference between the subject groups (0.31 and 0.32, respectively). Despite achieving the largest absolute difference in mean SUVR between the subject groups (0.38), the MR to MR ($\alpha = 1$) registration resulted in the largest standard deviation for the AD patients (0.29), and the second largest standard deviation for the healthy controls (0.20).

Three registration methods, all using the synthetic PET template, obtained the same, highest t statistic (4.08): PET to PET ($\alpha = 0$), and the combined PET-MR registration methods with $\alpha = 0.25$ and $\alpha = 0.5$. Although the $\alpha = 0$ and $\alpha = 0.25$ registration methods achieved a lower t statistic using the real PET template (4.02 and 4.05, respectively), the $\alpha = 0.5$ registration method also achieved the highest t statistic using the real PET template (4.07). Furthermore, the combined PET-MR registration method with $\alpha = 0.75$ resulted in the same t statistic for both PET templates (4.07).

Table 3.2: Mean (standard deviation) SUVR for both subject groups (AD and healthy controls), and the t statistic by registration method: affine, PET to PET ($\alpha = 0$), proposed combined PET-MR ($\alpha = \{0.25, 0.5, 0.75\}$), and MR to MR ($\alpha = 1$). The results for the affine and $\alpha = 1$ registration methods are identical across both PET templates because they are not influenced by the PET template. The best results for each PET template are highlighted in bold.

	PET template	Affine	$\alpha = 0$	$\alpha = 0.25$	$\alpha = 0.5$	$\alpha = 0.75$	$\alpha = 1$
SUVR AD	real	1.49 (0.28)	1.45 **(0.23)**	1.45 (0.24)	1.46 (0.25)	1.47 (0.27)	1.48 (0.29)
	synthetic	1.49 (0.28)	1.45 **(0.24)**	1.46 **(0.24)**	1.47 (0.26)	1.47 (0.27)	1.48 (0.29)
SUVR HC	real	1.12 (0.21)	1.14 **(0.18)**	1.13 **(0.18)**	1.12 (0.19)	1.11 (0.20)	1.10 (0.20)
	synthetic	1.12 (0.21)	1.13 **(0.17)**	1.13 (0.18)	1.12 (0.19)	1.11 (0.20)	1.10 (0.20)
t statistic	real	4.00	4.02	4.05	**4.07**	**4.07**	4.04
	synthetic	4.00	**4.08**	**4.08**	**4.08**	4.07	4.04

3.5 Discussion

3.5.1 Visual Comparison of PET Templates

Figure 3.5 shows that the real PET template does not exactly match the MR template. This result was expected, and is one of the main arguments in favour of using an MR-derived synthetic PET template. The reason for the differences between the real PET template and the MR template can probably be attributed to the pre-processing of the data. Small misregistrations in the PET to MR rigid registration during the pre-processing steps would have been propagated into the MR and real PET template construction (yellow arrows in 3.5). These differences do not exist in the MR-derived synthetic PET template. Moreover, the nature of amyloid PET imaging and the noise in the images mean that the images lack sharp edge-type features. As a result, the boundary between the ventricles and white matter in the MR image might not exactly match the same boundary in the amyloid PET image, even if the two images were accurately registered. This effect would be compounded by the number of images used to create the MR and real PET templates, leading to the observed mismatch between them.

There are numerous factors and parameters involved in the creation of both PET

templates that were not explored in this work. For example, the relative weightings of the MR tissues used to create the synthetic PET template could have been varied, or the number of subjects used to construct the real PET template could have been altered. The development of an optimal PET template is an area of research in its own right, and beyond the scope of this work. For that reason, we chose to only compare a single instance of a synthetic and real PET template. Moreover, the real and synthetic PET templates used in this chapter achieved adequate registration results, such that variations were not considered.

3.5.2 Registration Quality

Visual assessment of the results presented in Figures 3.6-3.9 suggests that non-linear registration methods show improved registration compared to the affine method. This assessment is confirmed by the distributions of overlap measures in Figure 3.10, because the median Dice coefficients of the non-linear registration methods are all higher than the median Dice overlap of the affine method. A further visual assessment of the results presented in Figures 3.6(a) and 3.9(b) also suggests that the PET to PET ($\alpha = 0$) registration methods resulted in poorer registrations than the other non-linear methods, since the white matter was over-stretched. The distributions of the Dice coefficients, presented in Figure 3.10, further support this assessment, because for both the real and synthetic PET templates, the PET to PET ($\alpha = 0$) registration method resulted in the lowest median Dice coefficient (0.77 and 0.74, respectively), and largest interquartile range (0.03 and 0.04, respectively), of all the non-linear methods. Nevertheless, for $\alpha = 0$ and $\alpha = 0.25$, registrations using the real PET template resulted in significantly higher Dice coefficients ($p < 0.01$) and smaller interquartile ranges than the registrations involving the synthetic PET template. These results suggest that if MR images are unavailable, the real PET template may provide more accurate registrations than the synthetic PET template.

Figure 3.10 also shows that the median Dice coefficient increases with increasing contribution of the MR volume during registration, before plateauing at $\alpha = 0.75$.

The combined PET-MR registration method with $\alpha = 0.75$ achieved the same, highest median Dice coefficient for both PET templates (0.84, equal with the $\alpha = 1$ method). Furthermore, the $\alpha = 0.5$ method using the synthetic PET template resulted in the lowest interquartile range (0.01). This may be due to the functional information acting as a spatially varying regulariser on the log-domain demons updates.

One reason for the $\alpha = 0$ method resulting in relatively poor Dice coefficients compared to the other non-linear registration methods is highlighted by the zoomed-in region in Figures 3.6(a) and 3.9(b). For some subjects, the $\alpha = 0$ method exhibited poor registration in the frontal lobe. This could be due to strong contrast between the grey and white matter in the amyloid PET scans of those subjects. If the PET template does not have clear grey-white matter contrast, the white matter of ^{18}F-florbetapir PET brain volumes will be deformed to match the strong contrast at the brain edge. This is plainly visible in the axial PET slice for the $\alpha = 0$ registration method in Figures 3.6(a) and 3.9(b); strong contrast between the grey and white matter can be observed in the frontal lobe of the affinely registered image (highlighted by the zoomed-in regions), but the high tracer uptake in the white matter is stretched to fit the edge of the brain in the PET to PET ($\alpha = 0$) case. This problem decreases as the influence of the MR modality increases, suggesting that the MR images influence the registration such that unnatural deformations cannot occur.

An amyloid negative PET scan exhibits more tracer uptake in white matter than in grey matter, creating clear grey-white matter contrast. Therefore, since the amyloid PET scans of cognitively normal controls are more likely to be classified as amyloid negative, unnatural deformations of the white matter, such as those in Figure 3.9(b), are less surprising. However, stretched white matter is surprising for AD patients, which are more likely to have an amyloid positive scan that lacks grey-white matter contrast. A visual assessment of the amyloid PET scan in Figure 3.6(a) suggests that the subject is amyloid negative, despite having an AD diagnosis

in the ADNI database. Discord between clinical diagnoses and ^{18}F-florbetapir interpretations has been recorded in the literature [Frey 2015], and this could explain why the gyri in Figure 3.6(a) are distorted.

Other than using a combined PET-MR registration approach to reduce unnatural deformations during the registration process, another possible solution to this problem is to increase the regularisation during the $\alpha = 0$ registration. This would prevent the brain volume from being deformed unrealistically, but could result in poor registration if large deformations are required. Alternatively, a spatially varying regularisation approach could be employed, in order to specifically limit the deformation in particular brain regions. The registration parameters used in this chapter were same as those used in [Lorenzi et al. 2013]. In that work, the LCC log-domain demons framework was used to register MR brain volumes to other MR brain volumes. Therefore, there is no guarantee that parameters suited to MR to MR registration are appropriate for PET registration.

Another potential solution would be to use an adaptive PET template, as suggested by [Lundqvist et al. 2013]. The authors recognised that registering amyloid scans with different patterns of tracer uptake to the same template could produce poor registration results. To overcome this problem, [Lundqvist et al. 2013] proposed to select an appropriate template on a continuous scale from an extreme amyloid negative scan to an extreme amyloid positive scan. To avoid having to select the most appropriate template prior to registration, the template is automatically generated during the registration process.

3.5.3 SUVR Separation

Due to the large variation in elderly brains, and differences in the progression of AD, a range of SUVRs is expected in both groups. Nevertheless, the mean SUVRs (and standard deviations) in this work fall within the ranges reported in [Fleisher et al. 2011] and [Hutton et al. 2015]. Furthermore, the non-linear registration methods exhibited smaller standard deviations than the affine method for the healthy

controls, suggesting more precise SUVRs. A similar trend was observed for the $\alpha = \{0, 0.25, 0.5, 0.75\}$ registration methods for the diseased population.

To best distinguish between populations of AD patients and healthy controls, one would expect a large difference in mean SUVR between the groups, and a small standard deviation for each group. Despite the relatively poor registration results compared to the other non-linear methods, the PET to PET ($\alpha = 0$) registration method achieved the smallest standard deviation for the diseased subjects and the healthy controls for both PET templates. However, as reported in Section 3.4.5, the $\alpha = 0$ registration method resulted in the smallest difference in the mean SUVRs of the two subject groups. Conversely, the MR to MR ($\alpha = 1$) method achieved the largest standard deviations of all the non-linear registration methods (AD: 0.29, and healthy controls: 0.20), but exhibited the largest difference in mean SUVR between the two populations (0.38). This suggests that a combined PET-MR registration method could be a good compromise between a large difference in mean SUVR and a small standard deviation. This is reflected in Table 3.2, where the combined PET-MR registration methods with $\alpha = \{0.25, 0.5, 0.75\}$ exhibit consistently high t-values. Moreover, the $\alpha = 0.5$ registration method achieved the highest t statistic for both the real and synthetic PET templates. This indicates that a combined PET-MR registration method could be well suited to discriminate between populations of healthy individuals and AD patients.

The mean SUVRs and standard deviations presented in Table 3.2 are very similar for both PET templates, indicating that SUVR is robust to the choice of template. Nevertheless, when using the real PET template, the $\alpha = \{0, 0.25, 0.5\}$ registration methods resulted in slightly different mean SUVRs than when using the synthetic PET template. Consequently, these methods have lower t statistics for the real PET template compared to the synthetic PET template. This suggests that for small values of the modality weighting parameter α, using the synthetic PET template to drive the registration may result in mean SUVRs that are more capable of distinguishing between groups of AD patients and healthy controls.

3.5.4 Methodological Considerations

In an attempt to sample close to the optimum global modality weighting, we evaluated our proposed PET-MR registration method with five different weightings $\alpha = \{0, 0.25, 0.5, 0.75, 1\}$. However, it is possible that the optimal modality weighting is spatially varying. Therefore, in Chapter 4, we will extend our current method to include a locally adaptive modality weighting.

The pre-processing methods applied in this work could also have had an effect on the non-linear registration results, especially the methods with little or no influence from the MR modality. In the first stage of pre-processing, the PET images were rigidly registered to their corresponding MR volumes (see Figure 3.1(a)). Secondly, as described in Section 3.3.1, the affinely registered PET volumes were created by affinely registering the MR volumes to the MR template, and applying the resulting transformations to the corresponding PET volumes. These two steps potentially introduce errors into the affinely registered PET images used as the starting points for the non-linear registrations. Affinely registering the PET volumes to the template space directly, rather than using an MR-based transformation, would result in different starting points for the non-linear PET to PET ($\alpha = 0$) registrations. In turn, this could improve the quality of the non-linear PET to PET registration results.

Closer inspection of the Dice coefficients of individual brain regions revealed that the relatively poor PET to PET ($\alpha = 0$) registration was not limited to the frontal lobe. The largest differences in overlap measure between the $\alpha = 0$ registration methods and every other non-linear method were found in the cerebellum, specifically in the "cerebellum_7b_L" and "cerebellum_7b_R" AAL regions (labels 9051 and 9052, respectively). These regions are small (see Figure 3.11); combined, they occupy just 6.8cm^3 of the total brain volume, which measures over 1900cm^3. Consequently, a small misregistration could have a profound effect on the Dice coefficient in these regions. Since all of the regions in the brain atlas carry the same weight in the mean Dice coefficient, poor overlap in small regions, due to slight misregistration,

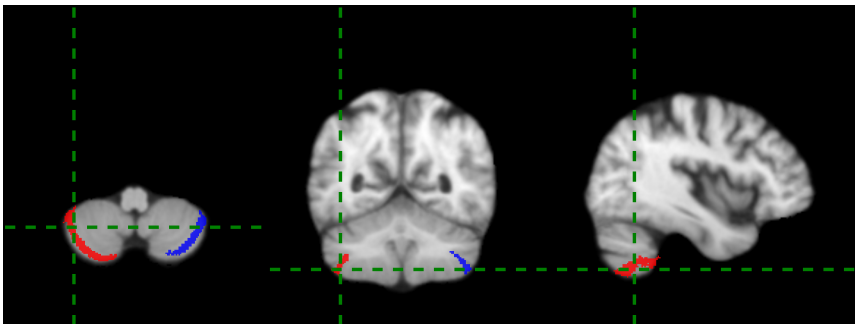


Figure 3.11: The largest differences in Dice coefficient between the $\alpha = 0$ registration method and every other non-linear method were found in the "cerebellum_7b_L" (red) and "cerebellum_7b_R" (blue) regions (AAL labels 9051 and 9052, respectively). To help the reader localise the regions within the brain, the dashed green lines on each image plane show the positions of the other two image planes.

could have a substantial impact on the overall overlap measurement.

In this work, the cortical region atlases used to compute the Dice coefficients were created by registering the AAL template (in MNI space) to subject space. Due to errors in these registrations, it is unlikely that the subject atlases will be as accurate as expertly delineated cortical regions. As a consequence, there may be a systematic error in the Dice coefficients presented in Figure 3.10. However, although the true Dice coefficients may differ from the values in this chapter, the trend in the results is still valid, since the same atlases were used to evaluate the quality of all of the registration methods.

Finally, in this chapter we used SUVR to provide a quantitative measure of brain amyloid burden. However, since amyloid plaque pathology can be found in cognitively normal controls, using SUVRs to discriminate based on clinical status of AD should be viewed with caution [Wolk et al. 2012]. Moreover, in a clinical setting, amyloid PET imaging should not be used to establish a diagnosis of AD, and must only be used as an adjunct to other diagnostic evaluations [Amyvid 2012]. Although we used the mean SUVR to distinguish between populations of AD patients and healthy controls, the values were used to assess the differences between the registration methods, and not to provide a clinical diagnosis of individual subjects.

3.5.5 Conclusion

In this chapter, we proposed a novel combined PET-MR registration method, using a weighting between single-modality updates within the LCC log-domain demons framework. Our proposed method is able to exploit all of the available data, and the results suggest that our method can perform at least as well as the single-modality registration methods with regards to registration accuracy. Furthermore, our results also indicate that the mean SUVRs acquired using our combined PET-MR registration method may be more suitable for distinguishing between populations of AD patients and healthy controls than SUVRs obtained using single-modality registration methods. We also assessed the effect of using a synthetic PET template, as well as a PET template constructed from real ^{18}F-florbetapir data, during registrations of amyloid PET volumes to a template space. We demonstrated that for small values of α, where PET is the dominant modality in the PET-MR registration process, the real PET template can achieve significantly higher registration accuracy than the synthetic PET template. However, the SUVRs obtained using the PET-MR registration method with the synthetic PET template may be better suited to discriminating between groups of healthy controls and AD subjects than those acquired using the real PET template.

In the next chapter we will extend our proposed PET-MR registration method by posing the registration framework as the minimisation of a single energy function. Moreover, the choice of the weighting parameter α is further assessed, by examining a spatially varying modality combination on a larger validation data set. The aim of the next chapter is to assess the robustness of SUVR calculation to the registration method used to align the amyloid PET volumes to the template. To evaluate this, we will compare the classification accuracy of SUVRs obtained using different registration methods when used to classify subjects as AD patients or healthy controls.

Chapter 4

The Effect of Registration on SUVR

The primary aim of Chapter 3 was to investigate the potential of using joint-modality image information during registration of PET volumes to a template. The results in Chapter 3 demonstrated that, for the purpose of separating populations of AD patients and healthy controls using SUVR, it may be beneficial to register the PET volume to a template space using a joint-modality approach that combines information from both PET and MR images. Nevertheless, the improvements of the combined modality registration method over single-modality methods (in terms of population separation) were small; SUVRs appeared to be robust to the type of PET template (synthetic vs real) and the choice of registration method. Nevertheless, the previous work was conducted on a small data set (only 28 subjects), and therefore, we revisit the experiments in this chapter using a larger collection of subjects. Moreover, the combined PET-MR registration method used in Chapter 3 had no rigorous mathematical foundation; the framework relied on a weighted combination of updates from two, separate single-modality demons registrations. This might have affected the convergence of the registration method, and ultimately, the composite SUVR.

The aim of this chapter is to assess the robustness of SUVR to the registra-

tion method used to align the PET volume to the template space. We compare PET-based and MR-based non-linear registration methods for amyloid PET quantification, and investigate the influence of joint-modality PET-MR registration on the final SUVR. To achieve this, we extend the combined PET-MR LCC-demons registration algorithm from Chapter 3 by introducing a dual-modality cost function. Similarly to Chapter 3, the contribution of each modality to the registration process is controlled by a weighting parameter. We further extend the method to allow for a locally adaptive modality weighting (LAMW) that selects from a set of weighting factors the best combination of PET and MR to yield the maximum LCC at each voxel. Following affine and non-linear registrations, cortex-to-cerebellar SUVRs were calculated in 100 subjects (50 AD, 50 healthy controls) from the ADNI database, and compared across the different registration methods. In addition to testing the LAMW method, five separate weightings of modality contributions (ranging from PET-only to MR-only) were investigated.

The remainder of this chapter is structured as follows: In Section 4.1 the update field of the joint-modality registration is derived from a dual-modality cost function, and the LAMW method is introduced. Section 4.2 gives details about the data set and the template selection. Sections 4.3 and 4.4 present the experiments and results, respectively. Finally, Section 4.5 discusses the effect of PET-based versus MR-based versus joint-modality registration for amyloid PET quantification, and concludes this work.

The work in this chapter is based on [Cattell et al. 2015].

4.1 Methods

4.1.1 Dual-Modality Cost Function

Unlike the PET-MR registration method in Chapter 3, which combined separate single-modality LCC demons registration updates at each iteration, the following registration method is framed more mathematically. Given a pair of PET and MR

volumes, the energy function from [Lorenzi et al. 2013] (see Equation 3.5) is extended to allow for two modalities:

$$E = -\frac{\alpha}{\sigma_{i_M}^2}\rho^2(\mathbf{u}, F'_M, M'_M) - \frac{1-\alpha}{\sigma_{i_P}^2}\rho^2(\mathbf{u}, F'_P, M'_P) + \frac{1}{\sigma_x^2}\|\mathbf{u}\|^2 \tag{4.1}$$

where α controls the relative contribution of each modality, σ_i accounts for the noise in the image, σ_x accounts for the spatial uncertainty of the correspondences, and ρ is the LCC. Subscript M denotes the MR images, subscript P denotes the PET images, and F'_k and M'_k are given by symmetric resampling of images F_k and M_k, respectively [Lorenzi et al. 2013]:

$$\begin{aligned} F'_k &= F_k \circ \exp\left(-\frac{\mathbf{u}}{2}\right) = F_k - \nabla F_k^T \frac{\mathbf{u}}{2} + O(\|\mathbf{u}\|^2) \\ M'_k &= M_k \circ \exp\left(\frac{\mathbf{u}}{2}\right) = M_k + \nabla M_k^T \frac{\mathbf{u}}{2} + O(\|\mathbf{u}\|^2) \end{aligned} \tag{4.2}$$

where $O(\|\mathbf{u}\|^2)$ denotes the higher-order terms in the Taylor expansion. Dropping the subscripts for clarity, the LCC term ρ for a single image modality can be written:

$$\rho(\mathbf{u}, F', M') = \frac{\mathbf{G}_\sigma * (F'M')}{(\mathbf{G}_\sigma * (F'^2))^{\frac{1}{2}}(\mathbf{G}_\sigma * (M'^2))^{\frac{1}{2}}} \tag{4.3}$$

where $\mathbf{G}_\sigma$ is a Gaussian of kernel size σ, and $*$ denotes the convolution operation. Analogously to [Lorenzi et al. 2013], the factor in the denominator can be approximated using a Taylor series:

$$\begin{aligned} (\mathbf{G}_\sigma * (F'^2))^{-\frac{1}{2}} &\simeq \left(\mathbf{G}_\sigma * \left(F - \nabla F^T \frac{\mathbf{u}}{2}\right)^2\right)^{-\frac{1}{2}} + O(\|\mathbf{u}\|^2) \\ &\simeq \left(\mathbf{G}_\sigma * \left(F^2 - F\nabla F^T \mathbf{u}\right)^2\right)^{-\frac{1}{2}} + O(\|\mathbf{u}\|^2) \\ &\simeq (\mathbf{G}_\sigma * (F^2))^{-\frac{1}{2}} \left(1 - \frac{\mathbf{G}_\sigma * (F\nabla F^T \mathbf{u})}{\mathbf{G}_\sigma * (F^2)}\right)^{-\frac{1}{2}} + O(\|\mathbf{u}\|^2) \\ &\simeq (\mathbf{G}_\sigma * (F^2))^{-\frac{1}{2}} \left(1 + \frac{\mathbf{G}_\sigma * (F\nabla F^T \mathbf{u})}{2\mathbf{G}_\sigma * (F^2)}\right) + O(\|\mathbf{u}\|^2) \\ &\simeq \frac{1}{(\mathbf{G}_\sigma * (F^2))^{\frac{1}{2}}} + \frac{\mathbf{G}_\sigma * (F\nabla F^T \mathbf{u})}{2(\mathbf{G}_\sigma * (F^2))^{\frac{3}{2}}} + O(\|\mathbf{u}\|^2) \end{aligned} \tag{4.4}$$

Similarly:

$$(\mathbf{G}_\sigma * (M'^2))^{-\frac{1}{2}} \simeq \frac{1}{(\mathbf{G}_\sigma * (M^2))^{\frac{1}{2}}} - \frac{\mathbf{G}_\sigma * (M\nabla M^T\mathbf{u})}{2(\mathbf{G}_\sigma * (M^2))^{\frac{3}{2}}} + O(\|\mathbf{u}\|^2) \tag{4.5}$$

Disregarding higher-order terms, the LCC term can be expanded:

$$\begin{aligned}\rho(\mathbf{u}, F', M') &\simeq \frac{\mathbf{G}_\sigma * \left(\left(F - \nabla F^T\frac{\mathbf{u}}{2}\right)\left(M + \nabla M^T\frac{\mathbf{u}}{2}\right)\right)}{(\mathbf{G}_\sigma * (F'^2))^{\frac{1}{2}}(\mathbf{G}_\sigma * (M'^2))^{\frac{1}{2}}} + O(\|\mathbf{u}\|^2) \\ &\simeq \frac{\rho(F,M)}{2}\Bigg(2 + \frac{\mathbf{G}_\sigma * (F\nabla M^T\mathbf{u})}{\mathbf{G}_\sigma * (FM)} - \frac{\mathbf{G}_\sigma * (M\nabla F^T\mathbf{u})}{\mathbf{G}_\sigma * (FM)} \\ &\quad + \frac{\mathbf{G}_\sigma * (F\nabla F^T\mathbf{u})}{\mathbf{G}_\sigma * (F^2)} - \frac{\mathbf{G}_\sigma * (M\nabla M^T\mathbf{u})}{\mathbf{G}_\sigma * (M^2)}\Bigg) + O(\|\mathbf{u}\|^2)\end{aligned} \tag{4.6}$$

By letting $\rho = \rho(F, M)$, and assuming a sufficiently smooth update field, such that $\mathbf{G}_\sigma * (\nabla I^T \cdot \mathbf{u}) \simeq \mathbf{G}_\sigma * (\nabla I^T) \cdot \mathbf{u}$, the LCC term can be simplified:

$$\rho(\mathbf{u}, F', M') \simeq \rho + \frac{\rho}{2}\Lambda\mathbf{u} + O(\|\mathbf{u}\|^2) \tag{4.7}$$

where, identically to Equation 3.7:

$$\Lambda = \left(\frac{\mathbf{G}_\sigma * (F\nabla M^T)}{\mathbf{G}_\sigma * (FM)} - \frac{\mathbf{G}_\sigma * (M\nabla F^T)}{\mathbf{G}_\sigma * (FM)} + \frac{\mathbf{G}_\sigma * (F\nabla F^T)}{\mathbf{G}_\sigma * (F^2)} - \frac{\mathbf{G}_\sigma * (M\nabla M^T)}{\mathbf{G}_\sigma * (M^2)}\right) \tag{4.8}$$

The squared LCC is therefore:

$$\rho^2(\mathbf{u}, F', M') \simeq \left(\rho + \frac{\rho}{2}\Lambda\mathbf{u}\right)^2 = \rho^2\left(1 + \Lambda\mathbf{u} + \frac{1}{4}\mathbf{u}^T\Lambda^T\Lambda\mathbf{u}\right) \tag{4.9}$$

The energy in Equation 4.1 is minimised with respect to the update field $\mathbf{u}$ at $\nabla E = 0$. Using the result from Equation 4.9, and reintroducing the subscripts M and P to denote MR and PET images respectively, ∇E can be written:

$$\nabla E = -\frac{\alpha}{\sigma_{i_M}^2}\rho_M^2\left(\Lambda_M + \frac{1}{2}\|\Lambda_M\|^2\mathbf{u}\right) - \frac{1-\alpha}{\sigma_{i_P}^2}\rho_P^2\left(\Lambda_P + \frac{1}{2}\|\Lambda_P\|^2\mathbf{u}\right) + \frac{2}{\sigma_x^2}\mathbf{u} \tag{4.10}$$

Setting $\nabla E = 0$, the update field is given by:

$$\mathbf{u} = -\frac{2\left(\frac{\alpha}{\sigma_{i_M}^2}\rho_M^2\Lambda_M + \frac{(1-\alpha)}{\sigma_{i_P}^2}\rho_P^2\Lambda_P\right)}{\frac{\alpha}{\sigma_{i_M}^2}\rho_M^2\|\Lambda_M\|^2 + \frac{(1-\alpha)}{\sigma_{i_P}^2}\rho_P^2\|\Lambda_P\|^2 - \frac{4}{\sigma_x^2}} \tag{4.11}$$

where, for $\alpha = \{0, 1\}$ this is equivalent to the update field for single-modality LCC log-domain demons (see Equation 3.6), as presented in [Lorenzi et al. 2013].

Assuming that $\sigma_{i_M}^2 = \sigma_{i_P}^2 = \sigma_i^2$, Equation 4.11 can be simplified:

$$\mathbf{u} = -\frac{2\left(\alpha\rho_M^2\Lambda_M + (1-\alpha)\rho_P^2\Lambda_P\right)}{\alpha\rho_M^2\|\Lambda_M\|^2 + (1-\alpha)\rho_P^2\|\Lambda_P\|^2 - 4\frac{\sigma_i^2}{\sigma_x^2}} \tag{4.12}$$

Hence, the PET-MR LCC-demons algorithm can be described by the iterations in Algorithm 4.1.

Algorithm 4.1 Joint PET-MR LCC log-domain demons registration

1: **while** not converged **do**
2: Given the current transformation, compute an update field $\mathbf{u}$ using Equation 4.12
3: For fluid-like regularisation, regularise $\mathbf{u}$ by convolving with a Gaussian kernel $\mathbf{u} \leftarrow G * \mathbf{u}$
4: $\mathbf{v} \leftarrow \mathrm{BCH}(\mathbf{v}, \mathbf{u})$, where BCH is the Baker-Campbell-Hausdorff formula (see Equation 3.2)
5: For diffusion-like regularisation, regularise $\mathbf{v}$ by convolving with a Gaussian kernel $\mathbf{v} \leftarrow G * \mathbf{v}$
6: Apply the same updated transformation to the MR and PET images
7: **end while**

4.1.2 Locally Adaptive Modality Weighting

Fixing the modality weighting α for all subjects and all brain regions could be improved upon by using a personalised, locally adaptive weighting. Therefore, the

proposed PET-MR registration method is extended by introducing an adaptive α-map. Given a set of α values $\{\alpha_1, \ldots, \alpha_n\}$, the algorithm selects the α value that yields the largest mean squared LCC at each voxel, resulting in a locally adaptive α-map for each subject. This is summarised in Algorithm 4.2.

Algorithm 4.2 Joint PET-MR registration with LAMW

1: **while** not converged **do**
2: **for** $i = \{1, \ldots, n\}$ **do**
3: Compute an update field **u** using Equation 4.12 with α_i.
4: Run lines 3-6 from Algorithm 4.1. This generates temporary images F_{Mi}, M_{Mi}, F_{Pi}, and M_{Pi}.
5: Compute the mean squared correlation (MSC)
$\text{MSC}_i = \frac{1}{2}(\rho^2(F_{Mi}, M_{Mi}) + \rho^2(F_{Pi}, M_{Pi}))$
6: **end for**
7: Create a locally adaptive modality combination map α_{MSC} by selecting the α value which yields the largest MSC at each voxel
8: Compute an update field **u** using Equation 4.12 with α_{MSC}
9: Run lines 3-6 from Algorithm 4.1 to update the PET and MR images.
10: **end while**

4.2 Materials

4.2.1 Data Acquisition and Pre-Processing

In addition to the 38 subjects selected for Chapter 3, a further 72 subjects (36 AD, 36 healthy controls) were obtained from ADNI, bringing the total to 110 subjects (55 AD, 55 healthy controls). The same selection criteria were used, such that each new subject had an ^{18}F-florbetapir PET volume and a T1-weighted MR volume acquired no more than 12 months apart. Prior to the experiments, the PET and MR volumes underwent the pre-processing steps described in Chapter 3.2.2.

4.2.2 Template Selection

To reduce the potential bias associated with selecting a single subject as the template, the MR template from Chapter 3 was used in the experiments in this chapter. The template was constructed iteratively from five AD MR brains and five healthy MR brains, chosen at random from the initial set of 38 subjects, using the method proposed in [Guimond et al. 2000] (see Chapter 3.2.3).

The aim of this chapter is to assess the effect of registration on SUVR calculation. Therefore, we used the synthetic PET template from Chapter 3, since our results in that chapter indicated that SUVRs obtained using a synthetic PET template are marginally better suited to separating groups of AD patients and healthy controls than SUVRs acquired using a PET template constructed from real ^{18}F-florbetapir data.

4.3 Experiments

4.3.1 Affine and Non-Linear Registration

The MR volumes of the 100 subjects (50 AD, 50 healthy controls) excluded from the template creation were affinely registered to the MR template using FSL's FLIRT [Jenkinson and Smith 2001; Jenkinson et al. 2002]. The resulting transformations were applied to the corresponding PET volumes. The 100 subjects then underwent PET-MR LCC log-domain demons registrations for a range of fixed modality weightings: $\alpha = \{0, 0.25, 0.5, 0.75, 1\}$. As in Chapter 3, the values $\alpha = 0$ and $\alpha = 1$ correspond to single modality PET to PET and MR to MR LCC-demons registrations, respectively. Furthermore, the same registration parameters were used: $\sigma_{\text{fluid}} = 0.5$, $\sigma_{\text{diffusion}} = 1.5$, $\sigma_{\text{LCC}} = 5$, and $\sigma_i/\sigma_x = 0.05$, with a multi-resolution, multi-scale scheme of 30×99×10 iterations (coarse to fine).

Following this experiment, the subjects were registered to the templates using the LAMW method described in Section 4.1.2. The set of combination weightings used to determine the locally optimal one was $\alpha = \{0, 0.25, 0.5, 0.75, 1\}$.

4.3.2 Dice Overlap

Similarly to Chapter 3, to quantitatively evaluate the accuracy of the registration results, the Dice coefficient was computed between cortical regions in each subject's registered MR volume and the MR template [Zijdenbos et al. 1994]. The subject cortical region atlases for the 72 new subjects were constructed using the method outlined in Chapter 3.3.2.

Paired t-tests between each registration method and every other method were used to assess whether any differences in Dice coefficient were statistically significant.

4.3.3 Amyloid PET Quantification

Following the registrations, SUVRs were computed using the same method as described in Chapter 3.3.3. A t-test was used to determine if there was a significant separation between the mean SUVRs of the two subject groups (AD and healthy control). Receiver operating characteristic (ROC) analyses were also performed for each registration method, by using a varying discrimination threshold on the composite SUVRs and calculating the true and false positive rates using the subject groups. The areas under the ROC curve (AUC) were calculated, as well as the classification accuracy, sensitivity and specificity at the SUVR threshold proposed in [Fleisher et al. 2011] (amyloid positive for SUVR > 1.08). Subjects with a composite SUVR greater than 1.08 were classified as belonging to the AD population, and subjects with a composite SUVR equal to or less than the threshold were classified as belonging to the group of healthy controls. The clinical diagnoses from ADNI were used as the gold standard for the classification of SUVR.

Although there is a strong correlation between SUVR and clinical diagnosis, as explained in Chapter 3.5.4, amyloid plaque pathology can be found in cognitively normal controls, and therefore patients should not be clinically diagnosed with AD using SUVR alone. However, in the experiments in this chapter, the classification results provide a useful measure to compare the SUVRs obtained using the different registration methods, and assess the robustness of the SUVR calculation to the

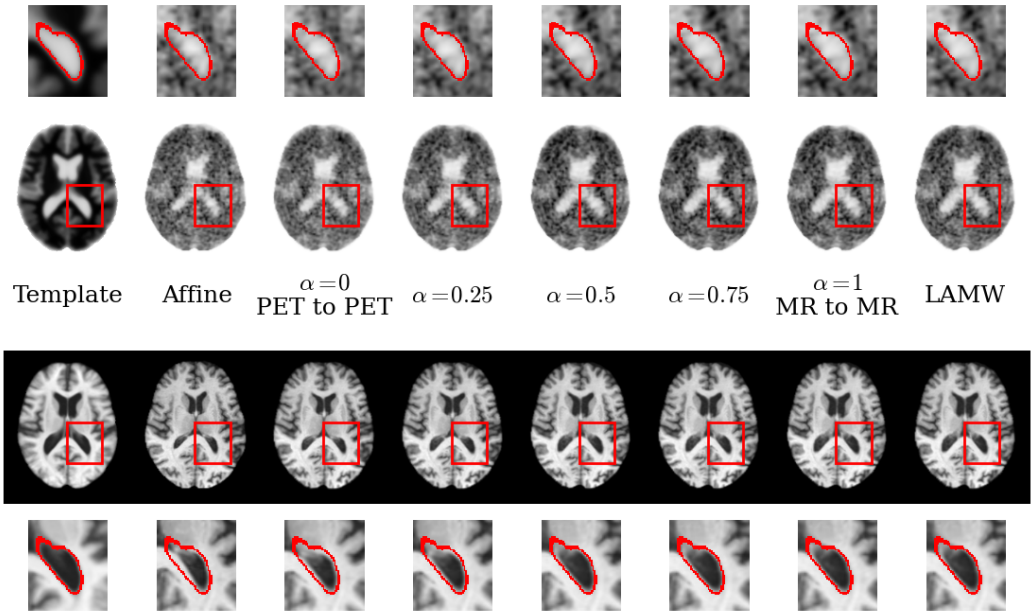
registration method used to align the amyloid PET volume to the template space.

4.4 Results

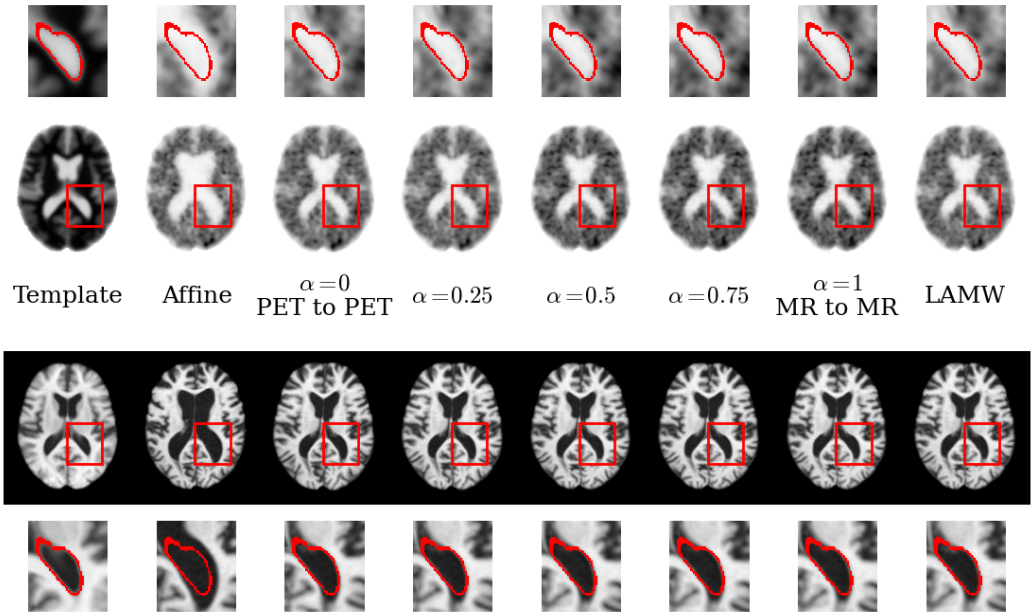
4.4.1 Registration Results

Figure 4.1 shows an axial slice from the registration results of the two AD subjects that achieved the smallest and largest range of Dice coefficients across the different registration methods. The non-linear registration methods ($\alpha = \{0, 0.25, 0.5, 0.75, 1\}$ and LAMW) in Figure 4.1(b) exhibit improved registration compared to the affine method. This is most evident in the ventricles, which are highlighted in the zoomed-in regions. The non-linearly registered ventricles more closely match the MR template ventricle boundary (shown in red) than the affinely registered ventricles. A similar trend is observed in Figure 4.1(a), but the improvement in registration is less pronounced in the PET to PET ($\alpha = 0$) case. Figure 4.2 shows the registration results for the healthy controls that obtained the smallest and largest range of Dice coefficients, respectively. Once again, the non-linear registration methods ($\alpha = \{0, 0.25, 0.5, 0.75, 1\}$ and LAMW) demonstrate improved registration in the ventricles, compared to the affine registration method.

The distributions of Dice coefficients by registration method are shown in Figure 4.3. Analogously to Chapter 3, the median Dice coefficients of the non-linear registration methods increased with increasing α, from 0.74 for $\alpha = 0$ (PET to PET registration), to 0.83 for ($\alpha = 1$) (MR to MR registration). Unsurprisingly, the affine method resulted in the lowest median Dice coefficient (0.71), and highest interquartile range (0.05), of all the registration methods tested. The smallest interquartile range (0.02) was achieved by the joint PET-MR registration method with $\alpha = 0.75$. The LAMW method resulted in the same median Dice coefficient as the joint PET-MR method with $\alpha = 0.5$ (0.81). Despite this, the paired t-tests on the overlap measures showed that the mean Dice coefficients of the LAMW and $\alpha = 0.5$ registration methods were significantly different ($p < 0.01$, corrected for 21

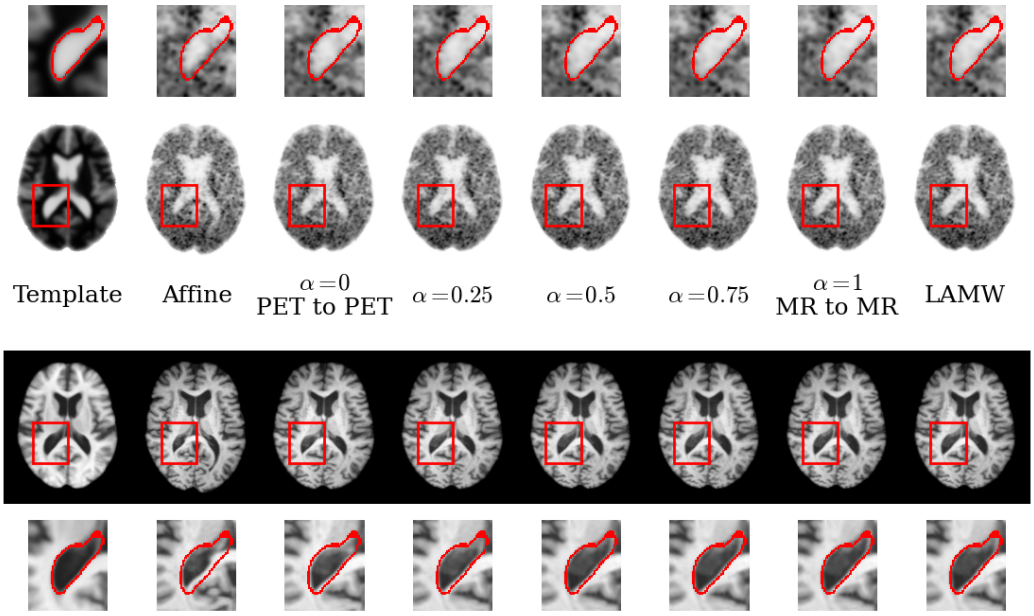


(a)

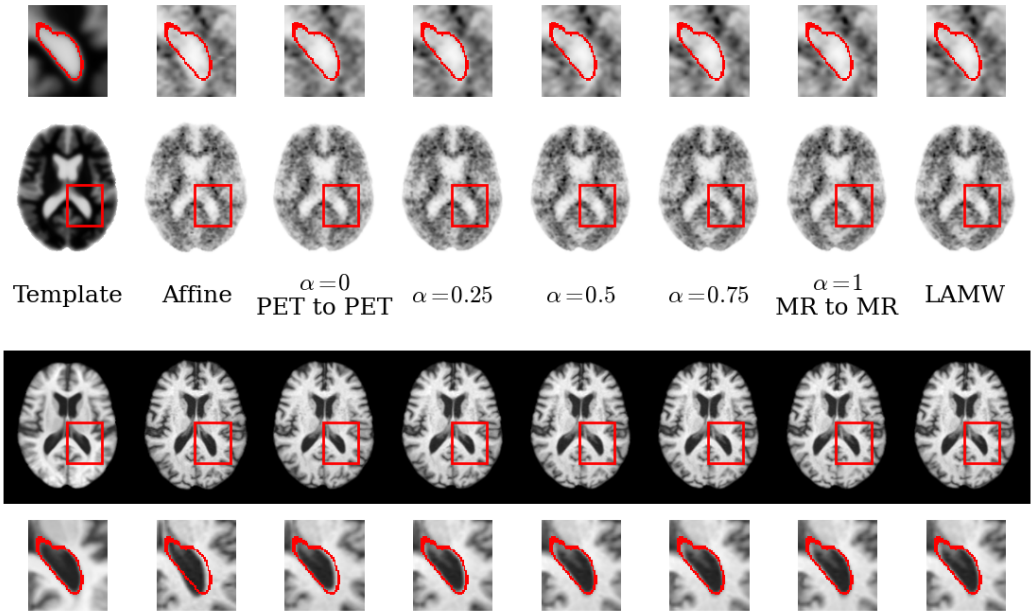


(b)

Figure 4.1: Sample axial slices from the volumetric registration results for the AD subjects that achieved (a) the smallest and (b) the largest range of Dice coefficients across the seven registration methods. The middle two rows of each sub-figure show the same axial slice from the deformed PET and MR volumes, respectively. The top and bottom rows of each sub-figure show a close-up of the region highlighted by the red box, with the outline of the MR template ventricles superimposed in red. In (b), the non-linear methods show improved registration in the ventricles compared to the affine registration method, since the registered ventricles better match the MR template ventricle boundary. A similar trend in observed in (a) but the improvement in registration is less noticeable in the PET to PET ($\alpha = 0$) case.



(a)



(b)

Figure 4.2: Sample axial slices from the volumetric registration results for the healthy controls that achieved (a) the smallest and (b) the largest range of Dice coefficients across the seven registration methods. The middle two rows of each sub-figure show the same axial slice from the deformed PET and MR volumes, respectively. The top and bottom rows of each sub-figure show a close-up of the region highlighted by the red box, with the outline of the MR template ventricles superimposed in red. In both sub-figures, the non-linear registration methods show improved registration in the ventricles compared to the affine method, since the registered ventricles better match the MR template ventricle boundary.

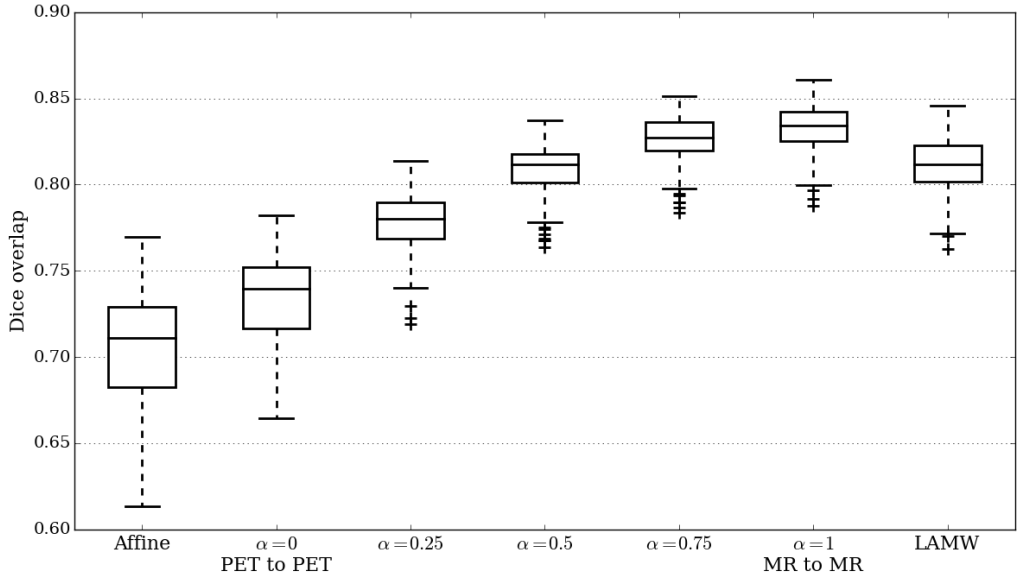


Figure 4.3: Distributions of overlap measure by registration method. The box and whisker plots show the Dice coefficients between the deformed image labels and the MR template labels, averaged over all regions in the label set, and over all subjects. Each box has lines at the lower quartile, median, and upper quartile. The whiskers extend from each edge of the box to the most extreme values within 1.5 times the interquartile range from the box. Outliers (+) have values beyond the ends of the whiskers.

comparisons). In fact, all pairs of mean Dice coefficients were significantly different from one another ($p < 0.01$, corrected).

For ease of comparison with the registration results from Chapter 3, Figure 4.4 shows the distributions of Dice coefficients for the same subset of 28 subjects (14 AD, 14 healthy controls) used in the experiments in Chapter 3. The Dice coefficients from the PET-MR registration method proposed in this chapter are shown in red, and the Dice coefficients from the combined PET-MR registration method with averaged single-modality updates (Chapter 3) are displayed in blue. Only the Dice coefficients for the $\alpha = \{0.25, 0.5, 0.75\}$ methods are compared in Figure 4.4, since the two registration methods are equivalent for $\alpha = 0$ and $\alpha = 1$. Although the Dice coefficients of the two PET-MR registration methods are very similar, the Dice coefficients were significantly different ($p < 0.01$) for α values of $\alpha = 0.25$ and $\alpha = 0.75$ (indicated by the $\star$ symbols in Figure 4.4).

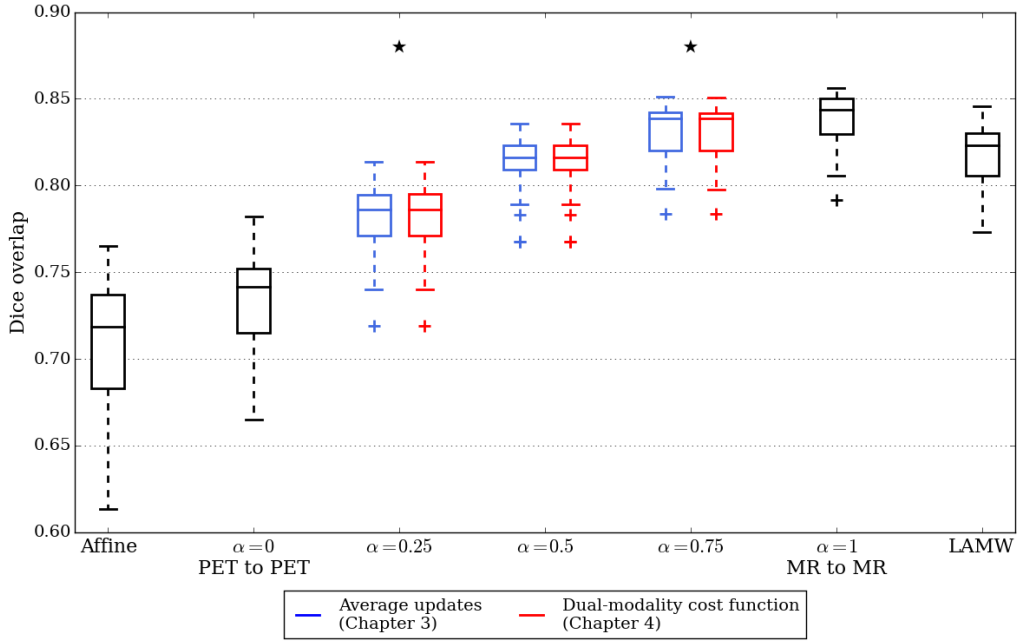


Figure 4.4: Distributions of overlap measure by registration method. For comparison, the Dice coefficients from the combined PET-MR registration method with averaged single-modality updates (Chapter 3) are shown in blue, and the overlaps from the joint PET-MR registration method proposed in this chapter are shown in red. The $\star$ symbols denote α values for which there was a significant difference in Dice coefficients between the two registration methods ($p < 0.01$). The box and whisker plots show the Dice coefficients between the deformed image labels and the MR template labels, averaged over all regions in the label set, and over all subjects. Each box has lines at the lower quartile, median, and upper quartile. The whiskers extend from each edge of the box to the most extreme values within 1.5 times the interquartile range from the box. Outliers (+) have values beyond the ends of the whiskers.

4.4.2 Standardised Uptake Value Ratio

Table 4.1 presents the mean SUVR of the diseased and healthy subject groups, and the classification accuracy, sensitivity, and specificity at an SUVR threshold of 1.08. Across all of the registration methods tested, a significantly higher ($p < 0.01$) mean SUVR was calculated for the AD group compared to the group of healthy controls. This is consistent with our results in Chapters 3, as well as results in the literature [Fleisher et al. 2011]. Similarly to Chapter 3, the non-linear MR to MR ($\alpha = 1$) registration method gave the largest absolute difference in mean SUVR between the two subject populations (0.33), and the PET to PET ($\alpha = 0$) method resulted in the smallest difference between the two groups (0.29). However, the PET to

Table 4.1: Mean (standard deviation) SUVR, classification accuracy (ACC), sensitivity (SEN) and specificity (SPE) at a SUVR threshold of 1.08, by registration method: affine, PET-only ($\alpha = 0$), proposed joint PET-MR ($\alpha = \{0.25, 0.5, 0.75\}$), MR-only ($\alpha = 1$), and joint modality with a locally adaptive modality weighting (LAMW). The best results are highlighted in bold.

	Affine	$\alpha = 0$	$\alpha = 0.25$	$\alpha = 0.5$	$\alpha = 0.75$	$\alpha = 1$	LAMW
SUVR AD	1.44 (0.28)	1.43 **(0.23)**	1.43 (0.24)	1.43 (0.26)	1.44 (0.27)	1.44 (0.28)	1.43 (0.25)
SUVR HC	1.12 (0.25)	1.14 **(0.20)**	1.13 (0.21)	1.12 (0.22)	1.12 (0.23)	1.11 (0.24)	1.13 (0.21)
$ACC_{1.08}$	0.74	0.74	0.74	0.74	**0.76**	0.75	0.75
$SEN_{1.08}$	0.84	**0.88**	0.86	0.86	0.86	0.84	**0.88**
$SPE_{1.08}$	0.64	0.60	0.62	0.62	**0.66**	**0.66**	0.62

PET ($\alpha = 0$) registration method resulted in the lowest standard deviation of mean SUVRs for both subject groups (0.23 and 0.20, respectively).

The PET to PET ($\alpha = 0$) registration method achieved the joint highest sensitivity (0.88), but the lowest specificity (0.60). The opposite trend was achieved by the MR to MR ($\alpha = 1$) method, which obtained the joint highest specificity (0.66), but the lowest sensitivity (0.84). The joint-modality registration method with $\alpha = 0.75$ gave the highest classification accuracy (0.76) at an SUVR threshold of 1.08, although the accuracies for each registration method were within 0.02 of one another. This lack of variation is also evident in Figure 4.5. Although the joint-modality registration method with $\alpha = 0.75$ resulted in the highest AUC (0.76), the DeLong method [DeLong et al. 1988] showed that there was no statistically significant difference between the AUCs of each method ($p < 0.01$, corrected). In addition, the ROC curves for each registration method are almost identical.

4.4.3 LAMW Analysis

Figure 4.6 shows the mean locally adaptive α-maps from the final iteration of the LAMW registration method, across all the diseased subjects and all the healthy controls. In both population groups, the optimal α value in the white matter is predominantly $\alpha = 1$ (i.e. MR modality only). In the ventricles and grey matter

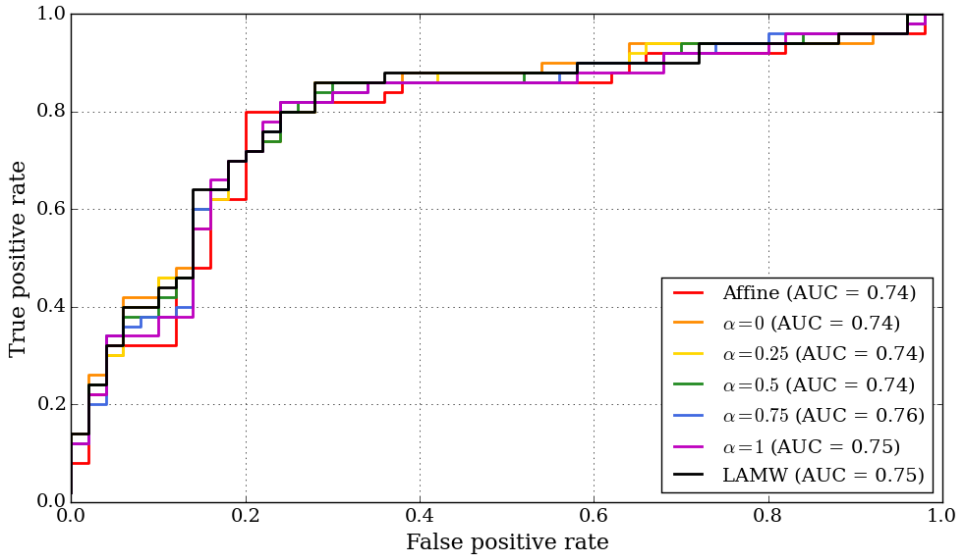


Figure 4.5: ROC curves for subject classification based on the composite SUVRs obtained from each registration method: affine, PET-only ($\alpha = 0$), proposed joint PET-MR ($\alpha = \{0.25, 0.5, 0.75\}$), MR-only ($\alpha = 1$), and joint modality with a locally adaptive modality weighting (LAMW). The area under each ROC curve (AUC) is presented in the figure legend. The ROC curves are almost identical for all methods, and there is no statistically significant difference between the AUCs ($p < 0.01$, corrected).

regions, the optimal α values are typically lower than in the white matter. For healthy controls, there are regions of $\alpha \simeq 0.25$ in the grey matter, which are not visible in the mean AD α-map. These areas are indicated by the black arrows in Figure 4.6.

4.5 Discussion

4.5.1 Registration Quality

Visual assessment of the results in Figures 4.1 and 4.2 indicates that the non-linear registration methods exhibit improved registration compared to the affine method. Nevertheless, Figure 4.1(a) suggests that the improvement in registration is less noticeable in the PET to PET ($\alpha = 0$) case than the other non-linear methods, since the registered ventricles do not match the boundary of the MR template ventricles as closely as the methods with $\alpha > 0$. The overlap measures presented in Figure 4.3 support this assessment, as the median Dice coefficients increase with increasing

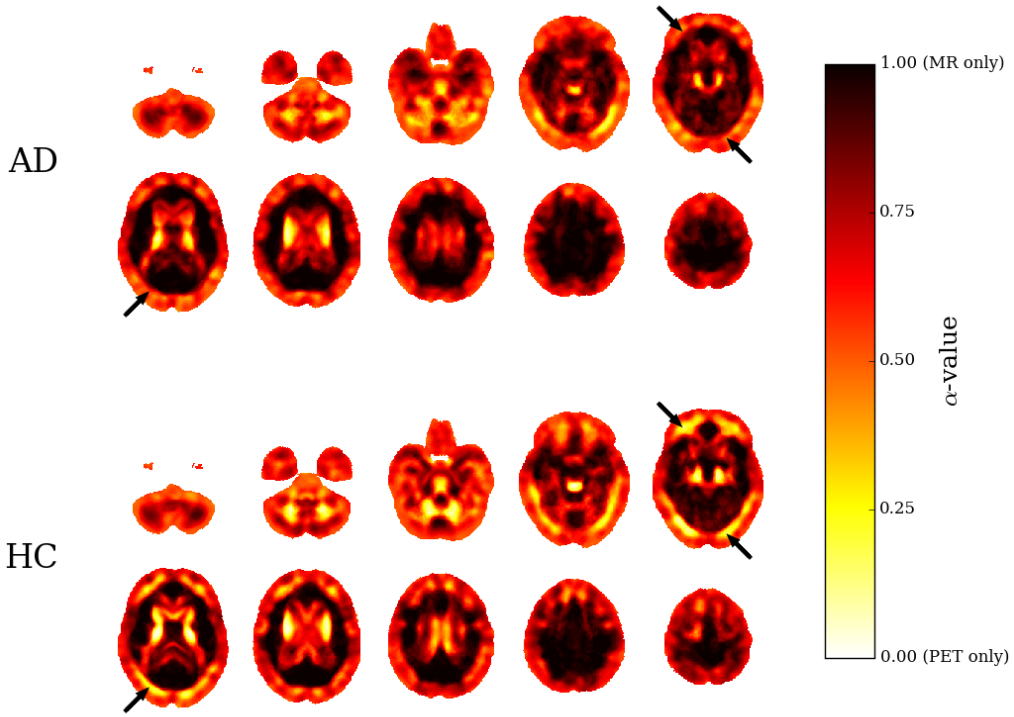


Figure 4.6: Top: Axial slices from the mean locally adaptive α-map from the final iteration of the LAMW registration method across all AD subjects. Bottom: Axial slices from the mean locally adaptive α-map across all healthy controls. The α-map values range from $\alpha = 0$ (PET only) to $\alpha = 1$ (MR only). In grey matter regions, the optimal α values are typically lower than in the white matter. For healthy controls there are regions of $\alpha \simeq 0.25$ in the grey matter, which are not visible in the mean AD α-map. These areas are indicated by the black arrows.

influence from the MR modality. This is similar to the trend observed in Chapter 3.

Although the LAMW registration method selects the α value (from a predefined set of values) that maximises the LCC at each voxel, the LAMW method achieved a lower median Dice coefficient (0.81) than those achieved by choosing $\alpha = 0.75$ and $\alpha = 1$, which had a median Dice coefficient of 0.83. As suggested in Chapter 3, this result could indicate that the PET volume acts as a regulariser on the log-domain demons updates. Another reason for the relatively lower median Dice overlap for the LAMW method could be that the locally adaptive α-maps had not fully converged.

The results presented in Figure 4.4 suggest that, despite the more rigorous mathematical basis of the PET-MR registration method proposed in this chapter compared to the registration method in Chapter 3, the two methods behave very similarly. The

median Dice coefficients are similar for both methods, and the two update equations (Equations 3.8 and 4.12) are equivalent in the single-modality cases (i.e. $\alpha = 0$ and $\alpha = 1$). Moreover, the registration parameters used in this chapter were identical to those used in Chapter 3, which could also account for the similarity of the two registration methods.

To simplify the update equation in Equation 4.11, it was assumed that noise in the PET volume is equal to the noise in the MR volume. In reality, the noise will probably be different between the two modalities. Furthermore, the registration parameters used in this work were taken from [Lorenzi et al. 2013], in which the authors proposed a registration method for MR images. These parameters may be suboptimal for PET data, and therefore, it may be possible to improve the registration results of the PET-dominant ($\alpha < 0.5$) PET-MR registration method with more suitable parameters.

4.5.2 Robustness of SUVR Calculation

Although the new PET-MR registration method produced very similar registration results to the method proposed in Chapter 3, improvements in registration quality were not the focus of this chapter. Instead, the aim was to assess the robustness of SUVR to the registration method, on a larger data set than Chapter 3. The results in Table 4.1 suggest that the registration method used to align the amyloid PET volume to the template has little effect on the SUVR. Despite a 0.12 increase in median Dice coefficient between the affine and $\alpha = 1$ registration methods, the mean SUVRs remained almost constant. Consequently, as reported in Table 4.1, the classification accuracy of SUVR, when used to classify AD patients or healthy controls, does not vary much between the different registration methods.

In this work, we did not optimise the SUVR classification threshold (1.08). Instead, we adopted an independent amyloid positivity threshold from published ^{18}F-florbetapir PET literature [Fleisher et al. 2011]. Therefore, it is possible that the threshold does not produce the maximum possible classification accuracy for our

data set. However, although it might be possible to improve the classification accuracies using a threshold derived from the ROC curves, the relative differences between the registration methods will still be minimal. This is because the there is little variation in the composite SUVRs across the different registration methods.

4.5.3 Optimal Modality Combination

The α-maps in Figure 4.6 show that the optimum combination of modalities is spatially varying. In the white matter, the optimum α value was primarily $\alpha = 1$, suggesting that the optimum modality combination in this region was to use MR only, because the correlation between the PET images and the PET template was lower in the white matter than the correlation between the MR images and the MR template. This could be due to the homogeneity of the white matter in the PET volumes and the synthetic PET template. In comparison, in areas where there is clear contrast in the PET images, such as the brain and ventricle boundaries, the LAMW algorithm selected a combination of both the PET and MR modalities ($\alpha < 1$). Moreover, the mean α-map for the healthy controls contains regions of $\alpha \simeq 0.25$, which are not visible in the mean α-map of the diseased population (as indicated by the black arrows in Figure 4.6). These areas of low α correspond to the regions where one would expect clear grey-white matter contrast in an amyloid negative PET scan, such as the frontal lobe. Consequently, the optimal modality combination in these regions favours an increased amount of PET relative to MR. Interestingly, there are no regions of $\alpha = 0$ in the α-maps in Figure 4.6, suggesting that useful information for image registration contained within the PET volumes is also available to some degree in the MR volumes.

4.5.4 Conclusion

The aim of this chapter was to assess the robustness of SUVR to the registration method used to register the amyloid PET image to the template space. To achieve this, we extended the combined PET-MR registration method proposed in Chapter

3 by posing the registration framework as the minimisation of a dual-modality cost function. We also introduced a locally adaptive modality weighting, which selects the optimum combination of PET and MR modalities at each voxel. Regardless of the registration methods presented here, there is a significant separation in SUVRs between populations of AD patients and healthy controls. Although the absolute difference in mean SUVR differs with the registration methods used, classification of AD patients versus cognitively normal controls based purely on composite SUVR is not affected by the registration method.

In this chapter, we evaluated the classification accuracy of SUVR when used to classify subjects as AD patients or healthy controls as defined by their clinical diagnoses in ADNI. Due to the robustness of SUVR to the registration method used to align the PET volume to the template, we showed that the classification accuracy did not vary considerably between the different registration methods tested. However, our experiments highlighted a drawback of using SUVR for classification, namely, the optimum SUVR classification threshold will potentially vary depending on the amyloid PET data set and the cortical regions used to calculate the SUVRs. Therefore, in the next chapter we will explore SUVR-free approaches to amyloid status classification, and propose a machine learning method that does not require a predefined classification threshold or cortical regions of interest.

Chapter 5

Amyloid Status Classification Using 3D HOG

As discussed in Chapters 3 and 4, brain amyloid burden can be evaluated quantitatively by calculating the ratio of tracer uptake in a set of target brain regions to non-specific tracer uptake in a reference region [Jack et al. 2008; Jagust et al. 2009; Fleisher et al. 2011; Villemagne et al. 2011a; Barthel et al. 2011; Joshi et al. 2012]. This ratio is known as the standardised uptake value ratio (SUVR). Typically, the individual SUVRs for each target region are averaged to form the mean, or *composite*, SUVR [Rowe et al. 2008]. It has been shown that incorporating this ratio as an adjunct to the visual assessment of ^{18}F-florbetapir scans can decrease inter-reader variability [Pontecorvo et al. 2014; Nayate et al. 2015].

The composite SUVR is usually used in a discrete fashion, where subjects above a particular threshold are designated as amyloid positive, and subjects below a certain threshold are designated as amyloid negative [Landau et al. 2013]. The thresholds are dependent on the type of tracer, the brain regions used to calculate the composite SUVR, and the delineation quality of those regions. Examples of amyloid positivity thresholds and SUVR target and reference regions are shown in Table 5.1. Although there is a consensus in the literature about which general brain areas are to be used in the composite SUVR, there are differences in the details [Jagust et al. 2009;

Jack et al. 2008]. In practice, for a specific tracer the correct set of regions and thresholds must be known and applied. For example, since amyloid deposition does not typically occur in the cerebellum, the reference regions presented in Table 5.1 are all cerebellum-based. Nevertheless [Jack et al. 2008] use the cerebellar grey matter, whereas [Joshi et al. 2012] use the entire cerebellum. Moreover, the frontal lobe is a universal target region, but the delineation of the region varies with the study; [Barthel et al. 2011] use the frontal lobe, whereas [Villemagne et al. 2011a] use specific areas within the frontal lobe.

A method for amyloid status classification, independent of SUVR, was proposed by [Vandenberghe et al. 2013]. The authors classified ^{18}F-flutemetamol scans as amyloid positive or amyloid negative using a support vector machine (SVM, see Chapter 2.3) trained directly on the voxel intensities. Using leave-one-out testing, the method achieved 100% agreement with the expert visual image assessments. Despite this remarkable result, the method was only tested with data from a single amyloid PET tracer. Moreover, by using all the voxels as the feature vector, it is possible that the performance of the classification method could be affected by noise in the training data.

For this reason, we propose an alternative machine learning method for amyloid status classification, which could serve as an adjunct to visual image interpretation, like composite SUVR, but without the need for defining tracer-specific regions of interest and amyloid positivity thresholds. Unlike the method proposed by [Vandenberghe et al. 2013], our method trains an SVM using features based on histograms of oriented 3D gradients (3D HOG). By using feature descriptors computed over image patches, rather than the voxel intensities directly, our proposed method could be more robust to noise and spatially varying intensity levels.

We show that our 3D HOG + SVM method can be used across a range of amyloid tracers, without the need to define different brain regions or positivity thresholds, and we compare the accuracy of amyloid status classification obtained using our method with the intensity-based SVM, and with the standard approach based on

Table 5.1: Examples of the different regions used to calculate composite SUVRs for ^{11}C-PiB, ^{18}F-florbetapir, and ^{18}F-florbetaben. The regions and amyloid positive/negative threshold are specific to each study.

Tracer	Study	Target regions	Reference region	Threshold
^{11}C-PiB	Jagust et al. (2009) [Jagust et al. 2009]	anterior cingulate, posterior cingulate/precuneus, prefrontal, lateral temporal, parietal cortex	cerebellar grey matter	1.465
	Jack et al. (2008) [Jack et al. 2008]	anterior cingulate, prefrontal, orbitofrontal, parietal, posterior cingulate/precuneus, temporal	cerebellar grey matter	1.5
^{18}F-florbetapir	Fleisher et al. (2011) [Fleisher et al. 2011]	medial orbital frontal, temporal, anterior cingulate, posterior cingulate, parietal lobe, precuneus	cerebellum	1.08
	Joshi et al. (2012) [Joshi et al. 2012]	frontal, temporal, parietal, anterior cingulate, posterior cingulate, precuneus	whole cerebellum	1.10
^{18}F-florbetaben	Villemagne et al. (2011) [Villemagne et al. 2011a]	dorsolateral prefrontal, ventrolateral prefrontal, orbitofrontal, superior parietal, lateral temporal, lateral occipital, anterior cingulate, posterior cingulate	cerebellar cortex	1.4
	Barthel et al. (2011) [Barthel et al. 2011]	frontal, parietal, lateral temporal, anterior cingulate, posterior cingulate, occipital	cerebellar cortex	1.39

SUVR.

The remainder of this chapter is structured as follows: Section 5.1 gives an overview of the data and the visual assessment procedure used to designate images as amyloid positive or amyloid negative. In addition to reviewing both the intensity-based SVM and SUVR classification methods, Section 5.2 introduces 3D HOG and outlines our proposed 3D HOG + SVM method. In Section 5.3, the optimal 3D HOG hyper-parameters are presented, and the classification results of the three methods are compared. Finally, in Section 5.4, the advantages of the proposed 3D HOG + SVM method over the other two classification methods are discussed.

5.1 Materials

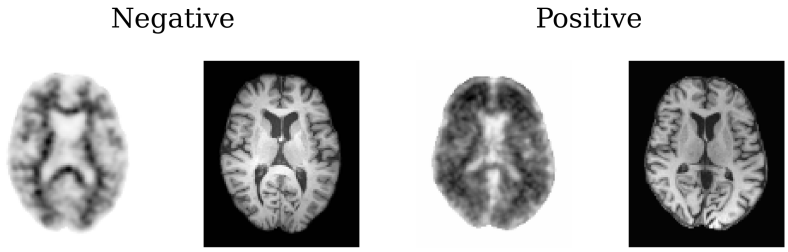
5.1.1 Data Acquisition and Pre-Processing

In addition to the 100 ^{18}F-florbetapir PET and T1-weighted MR volumes used in Chapter 4, ^{18}F-florbetapir PET and MR images from a further 194 subjects were gathered from the ADNI database. The two volumes for each subject selected for this study were acquired no more than 12 months apart. The images were pre-processed using the method detailed in Chapter 3.2.2. In brief, the PET and MR volumes were rigidly co-registered, skull-stripped, and transformed to MNI space. Prior to the classification experiments in this chapter, the transformed ^{18}F-florbetapir PET brain volumes were downsampled to 2×2×2mm resolution, in order to match the resolution of the PET volumes used by [Vandenberghe et al. 2013]. Axial slices from examples of amyloid negative and amyloid positive ^{18}F-florbetapir brain volumes, along with their corresponding MR slices, are shown in Figure 5.1(a).

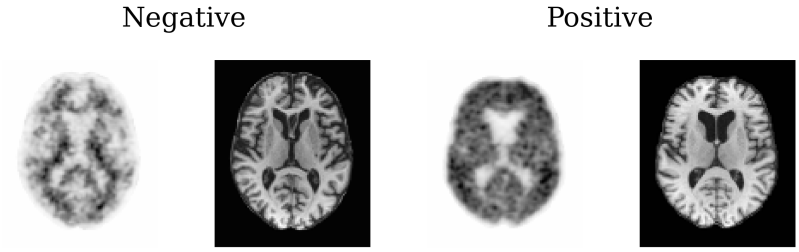
Unlike Chapters 3 and 4, which only used ^{18}F-florbetapir PET data, in this work, 214 ^{11}C-PiB PET and corresponding T1-weighted MR volumes were also downloaded from the ADNI database. The data were acquired from 102 subjects, but even though some subjects had multiple scans (26 subjects had one scan, 42 subjects had two scans, 32 subjects had three scans, and two subjects had four scans), each

^{11}C-PiB/MR pair was treated independently. Ideally, in order to avoid introducing a bias, only images from a single time point should be used to train and test the classification methods. However, to increase the size of the ^{11}C-PiB data set, we used multiple scans from individual subjects. Each ^{11}C-PiB volume and its corresponding MR volume were acquired within 12 months of one another. Furthermore, the ^{11}C-PiB scans underwent the same pre-processing as the ^{18}F-florbetapir volumes. Example amyloid positive and amyloid negative axial slices from the ^{11}C-PiB data set are shown in Figure 5.1(b).

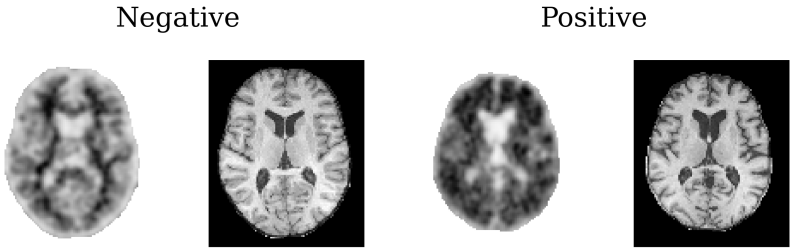
In addition to ^{18}F-florbetapir and ^{11}C-PiB PET data, 150 ^{18}F-florbetaben volumes were used in this chapter. The ^{18}F-florbetaben PET volumes and corresponding T1-weighted MR volumes were provided from the phase 2A clinical trial of ^{18}F-florbetaben in 150 participants. The participant details and imaging protocols are all provided in [Barthel et al. 2011]. Seventeen subjects were excluded due to severe image artefacts in the PET or MR images. Since the ^{18}F-florbetaben data were obtained from a clinical trial, rather than the ADNI database (like the ^{18}F-florbetapir and ^{11}C-PiB data), a different pre-processing method was used. Firstly, the MR and ^{18}F-florbetaben PET volumes were co-registered using an in-house rigid registration algorithm that is briefly described later. The ^{18}F-florbetaben images were then registered to a synthetic PET template in MNI space using an affine version of the in-house registration software, and downsampled to 2×2×2mm resolution in order to match the resolution of the PET volumes used by [Vandenberghe et al. 2013]. The in-house rigid and affine registration algorithms were implemented using routines customised from Siemens syngo.PET Amyloid Plaque (sPAP) quantification software. The resulting transformations were applied to the MR volumes. Analogously to the ^{18}F-florbetapir and ^{11}C-PiB pre-processing, a brain mask was constructed using tissue segmentations obtained from SPM8. Both the ^{18}F-florbetaben PET volumes and their corresponding MR volumes were skull-stripped. Figure 5.1(c) shows an example axial slice for a ^{18}F-florbetaben amyloid negative and ^{18}F-florbetaben amyloid positive brain volume. The corresponding axial MR slices are also shown



(a) ^{18}F-florbetapir



(b) ^{11}C-PiB



(c) ^{18}F-florbetaben

Figure 5.1: Axial slices from example amyloid negative (left) and amyloid positive (right) PET volumes for (a) ^{18}F-florbetapir, (b) ^{11}C-PiB, and (c) ^{18}F-florbetaben. The corresponding axial MR slices are shown to the right of each PET image. These slices were selected after the PET and MR volumes had been preprocessed.

in Figure 5.1(c).

5.1.2 Visual Assessment

In Chapters 3 and 4, the clinical diagnoses from ADNI were used as the gold standard for classification. However, in this chapter, the gold standard amyloid status was determined for each subject using criteria based on visual assessments from three image readers. Data were excluded from the study if the median rating was neither amyloid positive nor amyloid negative. To interpret the PET volumes as either amyloid positive or amyloid negative, the three image readers (one clinical

expert, one senior neuro-PET researcher, and the author of this thesis[1]) interpreted the 641 images a total of six times. The junior researcher assessed all of the images three times, the senior neuro-PET researcher interpreted all the images twice, and the clinical expert read all of the images once. The tracers were assessed one at a time (i.e all ^{18}F-florbetapir scans were interpreted before the ^{11}C-PiB scans), but for each tracer the images were presented in a random order to prevent observer memory affecting the assessments. Each reader was given instructions on how to display and interpret the images on a set of prearranged slices, without access to the corresponding MR image. For the ^{18}F-florbetapir and ^{18}F-florbetaben scans, the instructions were based on those provided by the tracer manufacturers [Amyvid 2012; Neuraceq 2014]. The only difference was the addition of an "equivocal" image class, for images that did not clearly fulfil the definitions of positive or negative scans[2]. The instructions for visual assessment of ^{11}C-PiB were based on a combination of those by [Suotunen et al. 2010] and [Cohen et al. 2013]. Again, an equivocal image class was included for images that could not be designated as either amyloid positive or amyloid negative.

Using Fleiss' kappa to assess the inter-reader reliability, the six image interpretations showed good agreement for all tracers (^{18}F-florbetapir: $\kappa = 0.71$, ^{11}C-PiB: $\kappa = 0.81$, ^{18}F-florbetaben: $\kappa = 0.84$). The gold standard amyloid status was determined from the median of the six image interpretations. Any images for which the median interpretation was not amyloid positive or amyloid negative were discarded. In total, 30 ^{18}F-florbetapir images, five ^{11}C-PiB images, and five ^{18}F-florbetaben images were discarded. The final ^{18}F-florbetapir dataset comprised 264 scans, the final ^{11}C-PiB dataset consisted of 209 scans, and the final ^{18}F-florbetaben dataset contained 128 scans. The demographics of these are summarised in Table 5.2. It should be noted that although the inter-rater agreement and number of equivocal scans varies by tracer, this is not a reflection of the tracers themselves. The discrep-

[1]Denoted as the junior researcher in this work.

[2]Note that since the reading instructions were very thorough, "equivocal" was typically only selected when image quality was particularly poor.

Table 5.2: Demographics of the amyloid positive (P) and amyloid negative (N) subjects used in this chapter.

	^{18}F-florbetapir		^{11}C-PiB		^{18}F-florbetaben	
	P	N	P	N	P	N
# subjects	149	115	167	42	53	75
Age ± std dev.	75.6±7.6	74.7±8.5	76.7±7.6	76.1±7.4	72.0±7.9	69.0±7.0
Sex (male/female)	84/65	60/55	103/64	27/15	27/26	35/40

ancies are predominantly due to differences in image quality and the distributions of amyloid burden.

5.2 Methods

In this chapter we compare three separate amyloid classification methods: the proposed method, which uses histograms of oriented three-dimensional gradients as inputs to an SVM (3D HOG + SVM), another SVM-based method using image intensity directly, and SUVR. The three methods are outlined in Sections 5.2.1-5.2.3.

5.2.1 Histogram of Oriented 3D Gradients

5.2.1.1 Derivation of Feature Vectors

Image descriptors have been widely used in computer vision to describe characteristics such as texture, motion, and shapes in 2D images and video sequences [Lowe 1999; Belongie and Malik 2000; Dalal and Triggs 2005]. Given a region of interest, a descriptor represents the region as a feature vector. By applying machine learning techniques to these feature vectors, they can be used to detect objects in images. As mentioned in Chapter 2.3, one such method, histogram of oriented gradients (HOG), has been used successfully to detect pedestrians in static images [Dalal and Triggs 2005]. The image is partitioned into a grid of uniformly spaced cells, and the normalised histogram of image gradient orientations in each cell forms the set of feature vectors (see Figure 2.4). The key concepts of this two-dimensional method were generalised to three dimensions by [Kläser et al. 2008]. Although originally

used for action recognition in video volumes, this chapter proposes to apply a similar technique to PET volumes to classify brain amyloid status.

In order to compute histograms of oriented gradients across an isotropically resampled PET volume, the volume is partitioned into a uniform grid of cells $\mathbf{c}_i$ of size $k \times k \times k$ voxels, where k is set by the user. Each cell is divided into $S \times S \times S$ sub-blocks $\mathbf{b}_j$, where S is user-defined, and for each sub-block the mean gradient is computed. In the same manner as [Kläser et al. 2008], the mean gradient is calculated using a 3D extension of the integral image (also known as a summed area table), popularised by [Viola and Jones 2001]. Given a volume v and its gradient $\nabla v = (\frac{\partial v}{\partial x}, \frac{\partial v}{\partial y}, \frac{\partial v}{\partial z})^T$, the integral volume can be written as:

$$I(x, y, z) = \sum_{x' \leq x, y' \leq y, z' \leq z} \nabla v(x', y', z') \tag{5.1}$$

The mean 3D gradient $\bar{\mathbf{g}} = (\bar{g}_x, \bar{g}_y, \bar{g}_z)^T$ within a cuboid of size $w \times h \times d$ at position $(x, y, z)^T$ is then given by:

$$\begin{aligned} \bar{\mathbf{g}} = (&I(x + w, y + h, z + d) - I(x, y + h, z + d) \\ &- I(x + w, y, z + d) + I(x, y, z + d)) - (I(x + w, y + h, z) \\ &- I(x, y + h, z) - I(x + w, y, z) + I(x, y, z)) \end{aligned} \tag{5.2}$$

Following computation of the 3D gradient $\bar{\mathbf{g}}$, its orientation is quantised into a histogram with n discrete bins. As explained in Chapter 2, using spherical polar coordinates to quantise the 3D gradient orientations leads to problems at the "poles", because the bins get progressively smaller. Therefore, we adopted the solution employed by [Kläser et al. 2008], which uses a regular polyhedron as an approximation to a sphere. Rather than have a continuous space of orientations, each side of the polyhedron corresponds to a histogram bin. In 3D space, there are only five polyhedra constructed from congruent regular polygons with the same number of faces meeting at each vertex. These polyhedra are known as Platonic solids: tetrahedron, hexahedron (cube), octahedron, dodecahedron, and icosahedron. They have 4, 6, 8,

12, and 20 faces, respectively.

To quantise a 3D gradient $\bar{\mathbf{g}}$ with respect to its orientation, $\bar{\mathbf{g}}$ is projected on to the axes going through the origin of the coordinate system and the centre of all faces of the polyhedron. Letting P be the matrix of face centre coordinates $\mathbf{p}_1, \ldots, \mathbf{p}_n$, the projection $\hat{\mathbf{q}}$ of $\bar{\mathbf{g}}$ is:

$$\hat{\mathbf{q}} = \frac{P \cdot \bar{\mathbf{g}}}{\|\bar{\mathbf{g}}\|_2} \tag{5.3}$$

Opposite gradient directions can be quantised into the same histogram bin by halving the set of face centre coordinates and taking the absolute value of $\hat{\mathbf{q}}$. Histograms organised in this manner are said to have "half-orientation".

Since $\bar{\mathbf{g}}$ should only vote in one histogram bin, the projection $\hat{\mathbf{q}}$ is thresholded. The threshold $t = \mathbf{p}_i^T \cdot \mathbf{p}_j$ is subtracted from $\hat{\mathbf{q}}$ and all negative elements are set to zero. The magnitude of the gradient is distributed according to the thresholded projection $\hat{\mathbf{q}}'$:

$$\mathbf{q} = \frac{\|\bar{\mathbf{g}}\|_2 \cdot \hat{\mathbf{q}}'}{\|\hat{\mathbf{q}}'\|_2} \tag{5.4}$$

The histogram $\mathbf{h}_{\mathbf{c}_i}$ for a given cell $\mathbf{c}_i$ is the sum of the quantised mean gradients of the sub-blocks $\mathbf{q}_{\mathbf{b}_j}$ in that cell:

$$\mathbf{h}_{\mathbf{c}_i} = \sum_{j=1}^{S^3} \mathbf{q}_{\mathbf{b}_j} \tag{5.5}$$

The histograms $\mathbf{h}_{\mathbf{c}_i}$ for each cell are concatenated to form the final feature vector for the entire volume.

5.2.1.2 Classification

A classifier is required to separate the feature vectors associated with different image classes (e.g. amyloid positive and amyloid negative). Typically a support vector machine (SVM) is used to classify HOG features [Dalal and Triggs 2005; Kläser et al. 2008]. In this work, the SVM implementation in the scikit-learn package for

Python [Pedregosa et al. 2011] is used.

5.2.1.3 Parameter Optimisation

In order to determine the optimum cell size k, number of sub-blocks S, and number of histogram bins for the 3D HOG method, the ^{18}F-florbetapir data set was split into a training and test set. The training set comprised 133 subjects (75 positive, 58 negative), and the test set consisted of 131 subjects (74 positive, 57 negative). The 3D HOG and SVM parameters were optimised on the ^{18}F-florbetapir training data set only. With the exception of Figure 5.2, all of the ^{18}F-florbetapir results reported in Section 5.3 were generated using the test set only. To assess the generalisability of the method, no 3D HOG parameter optimisation was conducted using the ^{11}C-PiB and ^{18}F-florbetaben data, so the entire data sets were used for testing.

To optimise the parameters for the 3D HOG feature descriptors, feature vectors from the ^{18}F-florbetapir training set volumes were computed for a range of parameter values. Cell size ranged from $k = 4$ voxels to $k = 32$ voxels in increments of 4 voxels, and the number of sub-blocks S were in the set $S = \{1, 2, 4\}$. The number of histograms bins (dodecahedron and icosahedron) was also assessed, as well as the effect of full- and half-orientation. A comprehensive grid search of parameters was conducted, resulting in 96 different parameter combinations.

For each of the 3D HOG parameter combinations, an SVM classifier was trained using the corresponding feature vectors of the ^{18}F-florbetapir training set. In order to ascertain the optimum SVM parameters to use on the test data, ten-fold stratified cross-validation was performed on the training set. The training set was randomly divided in to 10 subsets, each with the same proportion of amyloid positive/negative subjects. Nine subsets were used to train the SVM, and the remaining subset was used as the validation test data set. This was repeated, such that each subset was used as the test set. An SVM with a Gaussian radial basis function (RBF) kernel was used, and the slackness variable C (where $C = 10^i$ for $i = \{-2, \ldots, 3\}$) and free parameter of the RBF γ (where $\gamma = 10^i$ for $i = \{-5, \ldots, 2\}$) were optimised using

a grid search of parameters [Boser et al. 1992; Cortes and Vapnik 1995].

5.2.1.4 Testing

The 3D HOG parameters and SVM parameters that gave the highest classification accuracy, sensitivity, and specificity on the ^{18}F-florbetapir training set were applied to the ^{18}F-florbetapir test set. These parameter values were also applied to the ^{11}C-PiB and ^{18}F-florbetaben data. In order to assess the performance of the 3D HOG features for amyloid status classification, leave-one-out testing was used on each PET tracer separately. For each fold, the SVM was trained using all of the images except one. The remaining image was then used as the test image. This process was repeated until all of the images had been used as the test image. Following leave-one-out testing for all three tracers, the mean classification accuracy, sensitivity, and specificity were calculated. By adjusting the SVM classifier's decision boundary, receiver operating characteristic (ROC) analysis was performed on each of the tracers.

5.2.2 Standardised Uptake Value Ratio

5.2.2.1 Quantification Software

The ratio of tracer uptake in a set of target brain regions to non-specific tracer uptake in a reference region, also known as SUVR, was computed using the commercially available Siemens syngo.PET Amyloid Plaque (sPAP) quantification software. Prior to SUVR calculation, the software automatically registers the subject's PET volume to a synthetic PET template, in Montreal Neurological Institute (MNI) stereotactic space, in which the cortical regions of interest are defined [Peyrat et al. 2012; Hutton et al. 2015]. Analogously to Chapters 3 and 4, the predefined set of six target regions for ^{18}F-florbetapir were: the frontal, parietal, anterior cingulate, posterior cingulate, precuneus, and temporal lobes [Hutton et al. 2015]. The reference region was the whole cerebellum. For ^{18}F-florbetaben, slightly different predefined target and reference regions were used (target regions: frontal, parietal, anterior cingulate,

posterior cingulate, temporal, occipital lobes, reference: cerebellar cortex [Barthel et al. 2011]). Note that the different tracers used different sets of regions according to the published literature ([Hutton et al. 2015] and [Barthel et al. 2011], respectively).

The sPAP quantification method has been validated for use with ^{18}F-florbetapir and ^{18}F-florbetaben [Peyrat et al. 2012; Hutton et al. 2014; Hutton et al. 2015], but not for ^{11}C-PiB because it is not an FDA-approved tracer. However, it was still possible to use sPAP for quantification of the ^{11}C-PiB data. Based on the literature [Jagust et al. 2009; Landau et al. 2013], SUVRs were calculated using the ^{18}F-florbetapir target and reference regions.

During SUVR computation in sPAP, one ^{18}F-florbetapir volume, eight ^{18}F-florbetaben volumes, and three ^{11}C-PiB volumes failed to adequately register to the PET template. As a result, the registration was manually adjusted for these subjects.

5.2.2.2 SUVR Analysis

Following the computation of composite SUVRs for the ^{18}F-florbetapir test data set, classification results were obtained using an amyloid positivity threshold of SUVR > 1.12 [Hutton et al. 2015]. Similarly, classification results were obtained from the ^{18}F-florbetaben SUVRs using a threshold of composite SUVR > 1.36 [Hutton et al. 2014].

Since sPAP has not been validated for ^{11}C-PiB data, two regression equations were required to obtain an amyloid positivity threshold that is appropriate for both the tracer and the SUVR calculation method. Firstly, [Landau et al. 2013] provided a regression equation to convert the ^{11}C-PiB threshold t_{Jagust} from [Jagust et al. 2009] into a corresponding threshold for the quantification method used by [Joshi et al. 2012]:

$$t_{\text{Joshi}} = 0.67 t_{\text{Jagust}} + 0.15 \tag{5.6}$$

where the subscript denotes the study from which the threshold is acquired.

In [Hutton et al. 2015], a regression equation to convert from the Joshi method threshold into an equivalent unit for sPAP t_{sPAP} was also provided:

$$t_{\text{sPAP}} = 0.9782 t_{\text{Joshi}} + 0.04264 \tag{5.7}$$

Equations 5.6 and 5.7 can be combined to get a final equation to convert between the ^{11}C-PiB threshold t_{Jagust} and sPAP t_{sPAP}:

$$\begin{aligned} t_{\text{sPAP}} &= 0.9782(0.67 t_{\text{Jagust}} + 0.15) + 0.04264 \\ &\simeq 0.6554 t_{\text{Jagust}} + 0.1894 \end{aligned} \tag{5.8}$$

By substituting $t_{\text{Jagust}} = 1.465$ [Jagust et al. 2009] into Equation 5.8, an equivalent threshold for ^{11}C-PiB in sPAP ($t_{\text{sPAP}} = 1.15$) is obtained. Consequently, following SUVR calculation in sPAP, the accuracy of amyloid status classification in ^{11}C-PiB data was assessed using an amyloid positivity threshold of composite SUVR > 1.15.

5.2.3 Image Intensity

The SUVR and 3D HOG + SVM method were compared to the machine learning method proposed by [Vandenberghe et al. 2013]. In that work, the authors trained an SVM on voxel intensities to classify amyloid positivity in ^{18}F-flutemetamol images. Each image is a point in high-dimensional space, in which each dimension is a voxel within the brain.

To reduce the dimension of the SVM, only voxels inside the brain were used. Once all of the images were transformed into MNI space (see Section 5.1.1), a brain mask was constructed using the linear MNI152 T1-weighted MR template [Mazziotta et al. 2001] to ensure that the feature vector had the same dimension for every image. The mask was dilated by 2mm to ensure that all of the registered PET brains were wholly inside the mask. Prior to using the SVM, all of the images were normalised to have zero mean and unit variance. The SVM parameters were then optimised using the same approach as in Section 5.2.1.3. The optimum parameters were applied to

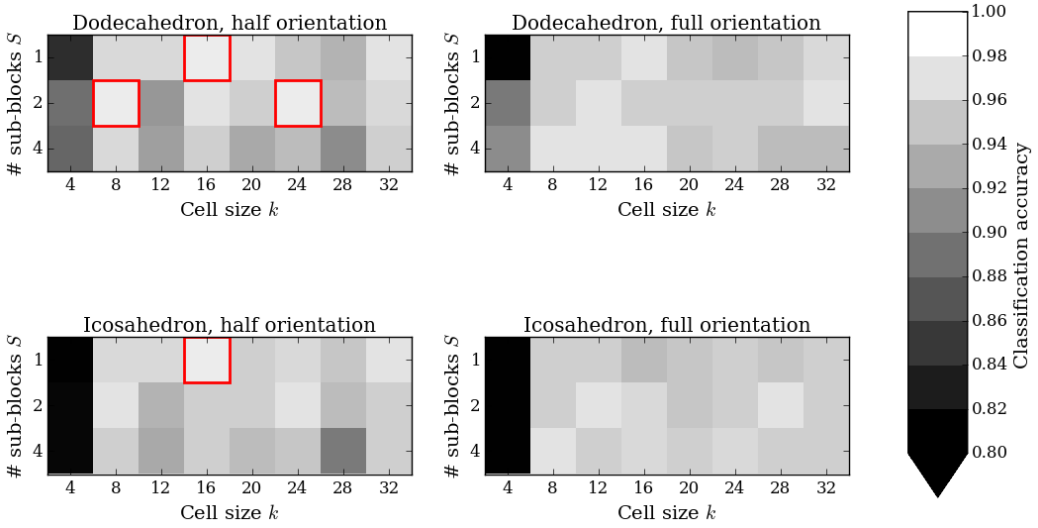


Figure 5.2: The classification accuracies achieved on the ^{18}F-florbetapir training data for each of the 96 3D HOG parameter combinations. Each sub-plot relates to one of the four histogram configurations (dodecahedron vs icosahedron, and half-orientation vs full-orientation), and shows the highest classification accuracy for each cell size k (horizontal axis) and number of sub-blocks S (vertical axis). Four parameter combinations, highlighted in red, achieved the same, highest classification accuracy (98.5%).

the ^{18}F-florbetapir test data, as well as the ^{11}C-PiB and ^{18}F-florbetaben data sets. Leave-one-out testing was used to assess the ability of the intensity-based SVMs to classify amyloid status.

5.3 Results

5.3.1 3D HOG Parameter Optimisation

As explained in Section 5.2.1.3, the 3D HOG parameters were optimised on the ^{18}F-florbetapir training data only. Figure 5.2 shows the best classification accuracies achieved on the ^{18}F-florbetapir training set for each of the 96 3D HOG parameter combinations. Each sub-plot relates to one of the four histogram configurations (dodecahedron vs icosahedron, and half-orientation vs full-orientation), and shows the classification accuracy for each cell size k (horizontal axis) and number of sub-blocks S (vertical axis). A cell size of $k = 4$ voxels universally resulted in the lowest classification accuracy. However, 68 different parameter combinations resulted in a

mean classification accuracy greater than 95%. Four parameter combinations (highlighted in red in Figure 5.2) achieved the same, highest classification accuracy of 98.5%. However, one parameter combination gave the highest combined classification accuracy, sensitivity, and specificity (98.5%, 0.973, and 1.00, respectively) on the ^{18}F-florbetapir training set: cell size $k = 16$ voxels, number of sub-blocks $S = 1$, icosahedron, half-orientation histogram. Consequently, this set of optimum 3D HOG parameters was then applied to the ^{18}F-florbetapir test set, and the ^{11}C-PiB and ^{18}F-florbetaben data.

5.3.2 Classification Results

The classification accuracy, sensitivity, specificity and area under the receiver operating characteristic curve (AUC) for the ^{18}F-florbetapir test data are shown in Figure 5.3. The black borders indicate the best results. The classification results for the ^{11}C-PiB and ^{18}F-florbetaben datasets are presented in Figures 5.4 and 5.5, respectively. For all three tracers, the 3D HOG + SVM classification method resulted in the largest classification accuracy (96.2%, 99.5%, and 96.9%, respectively) and AUC (0.962, 0.988, and 0.965, respectively). Using the DeLong method [DeLong et al. 1988] to statistically compare the AUCs of each classification method, the 3D HOG + SVM method achieved a significantly larger AUC than the intensity-based SVM method for ^{18}F-florbetapir ($p < 0.01$). Furthermore, the 3D HOG + SVM classification method also achieved a significantly larger AUC than SUVR for 11C-PiB ($p < 0.01$). Although the 3D HOG + SVM method had a larger AUC than both the other methods for ^{18}F-florbetaben, the AUCs were not significantly different. For ^{18}F-florbetaben, the intensity-based SVM had the same classification accuracy as the 3D HOG + SVM method (96.9%). The 3D HOG + SVM method had a higher specificity (0.965, 0.976, and 0.987, respectively) than the SUVR method (0.912, 0.881, and 0.867, respectively) across all of the tracers tested. Although SUVR gave the highest sensitivity for ^{18}F-florbetapir and ^{18}F-florbetaben (0.973 and 0.962, respectively), it resulted in the lowest sensitivity for ^{11}C-PiB (0.760).

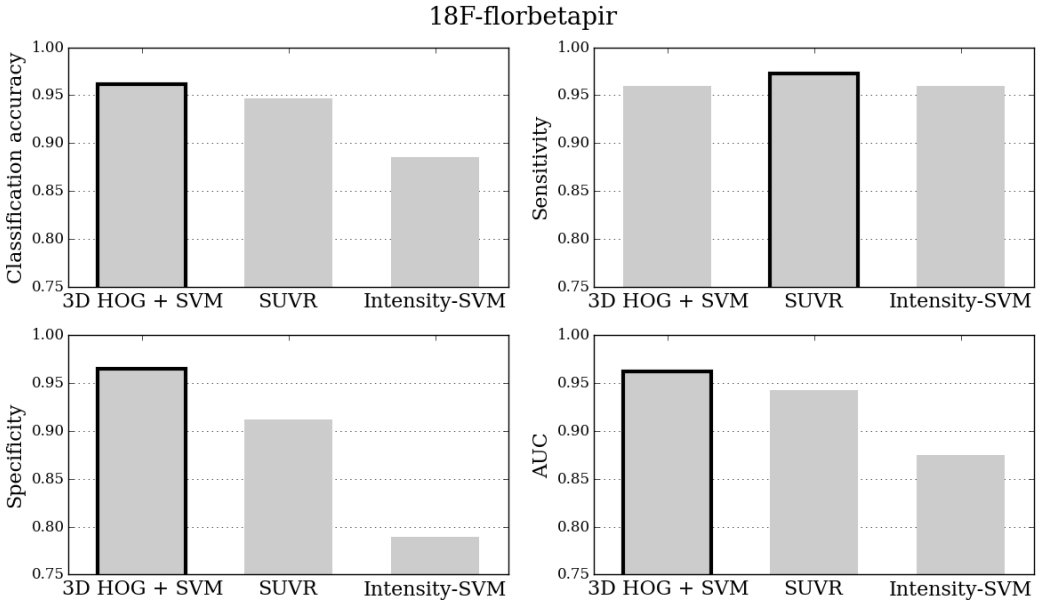


Figure 5.3: The classification accuracy, sensitivity, specificity and area under the receiver operating characteristic curve (AUC) for the ^{18}F-florbetapir test data. The best results are highlighted with a black border.

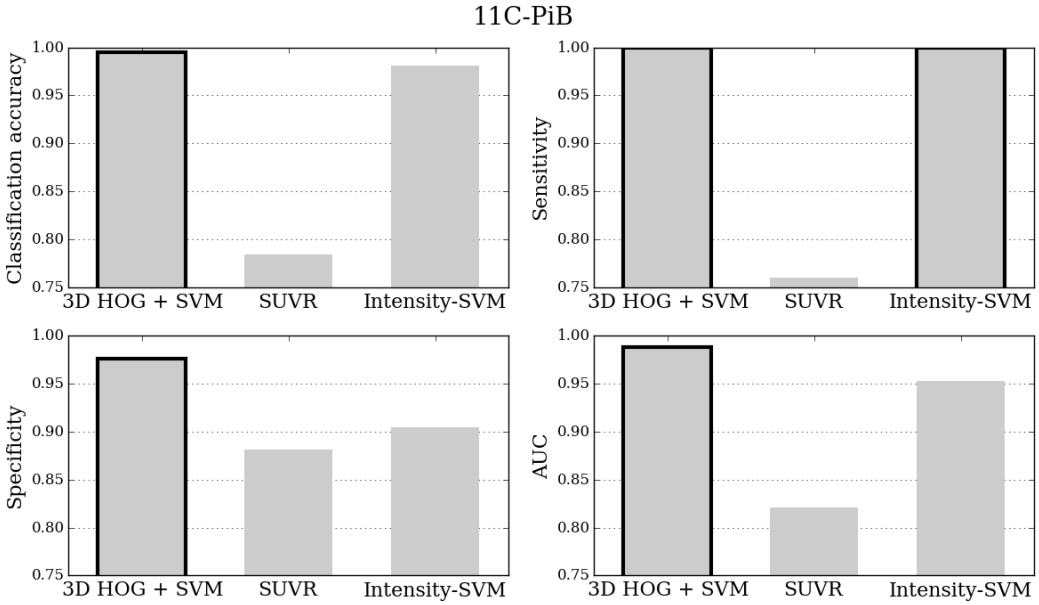


Figure 5.4: The classification accuracy, sensitivity, specificity and area under the receiver operating characteristic curve (AUC) for the ^{11}C-PiB test data. The best results are highlighted with a black border.

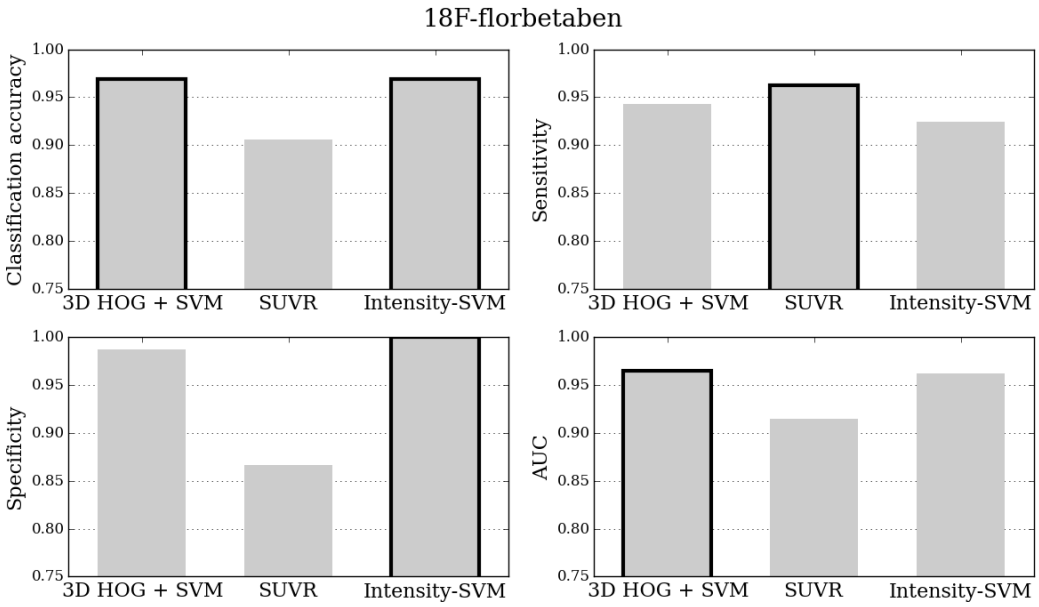


Figure 5.5: The classification accuracy, sensitivity, specificity and area under the receiver operating characteristic curve (AUC) for the ^{18}F-florbetaben data. The best results are highlighted with a black border.

5.3.3 Distance to Classification Boundary

Figures 5.6-5.8 show the distances of the test subjects from their respective classification boundary. For the SVM-based methods, the distances are the Euclidean distances to the decision hyperplane. For the SUVR method, the distances represent the subject's SUVR minus the threshold SUVR. The distances are normalised to the maximum absolute distance from the boundary. Smaller distances indicate a lower confidence in the final classification decision. In Figures 5.6-5.8, subjects in blue with positive distances were incorrectly classified by the given classification method. Similarly, subjects in red with negative distances were also misclassified.

We used a two-sided t-test, corrected for two comparisons, to examine whether the boundary distances for the 3D HOG + SVM method were significantly greater than the other two classification methods. Values of $p < 0.01$ (corrected for multiple comparisons) are considered significant.

Across all three tracers, the distances from the boundary for the 3D HOG + SVM method were found to be significantly greater than the distances of the SUVR method, for both amyloid positive and amyloid negative subjects. Furthermore, for

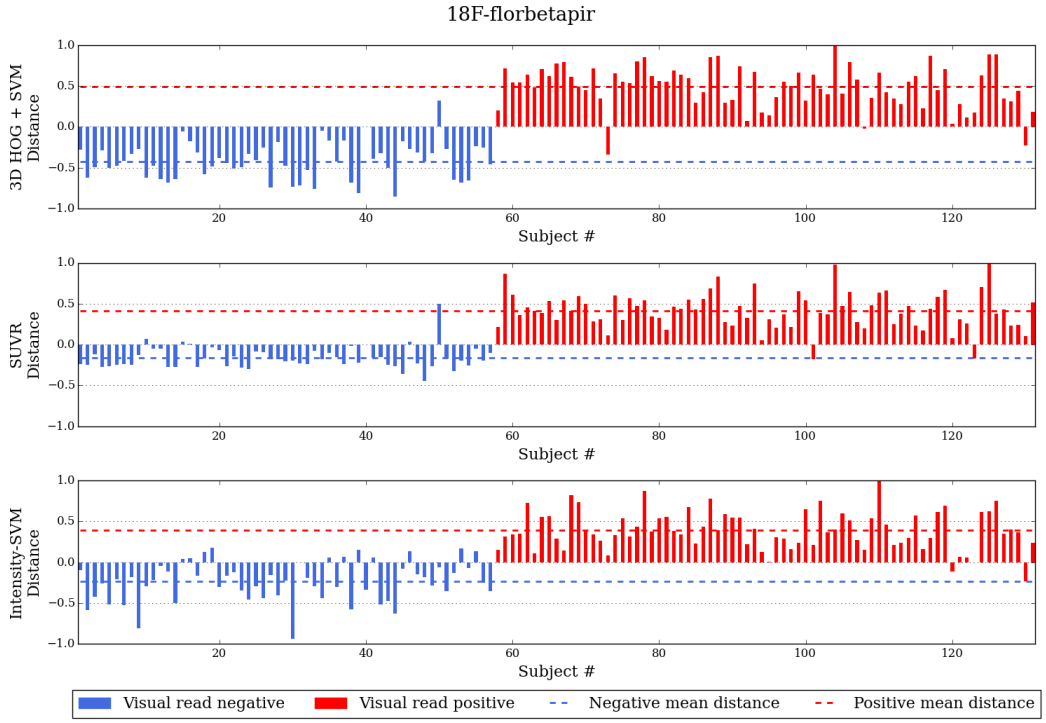


Figure 5.6: The normalised distances of ^{18}F-florbetapir test subjects from the classification boundary. Subjects in blue with positive distances were incorrectly classified by the given classification method. Similarly, subjects in red with negative distances were also incorrectly classified. The dashed lines indicate the mean distance from the boundary for the subjects visually designated as amyloid positive and amyloid negative.

both subject groups, the ^{18}F-florbetapir distances are significantly greater for the 3D HOG + SVM method than the distances of the intensity-based SVM classification method. The 3D HOG + SVM distances were also significantly greater for the amyloid positive ^{18}F-florbetaben subjects. In contrast, the distances for the amyloid positive ^{11}C-PiB subjects were significantly greater for the intensity-based SVM method compared to the 3D HOG + SVM method.

It is apparent from Figure 5.7 that one ^{11}C-PiB image was classified differently to the gold standard visual assessment by all three classification methods (image #9). Similarly, in Figure 5.8, ^{18}F-florbetaben image #104 was classified differently to the gold standard across all three methods. Axial slices from the PET and MR volumes of these outliers are shown in Figure 5.9.

Although scans with equivocal visual reads were eliminated from the three data sets prior to classification analysis, we computed the normalised distances of the

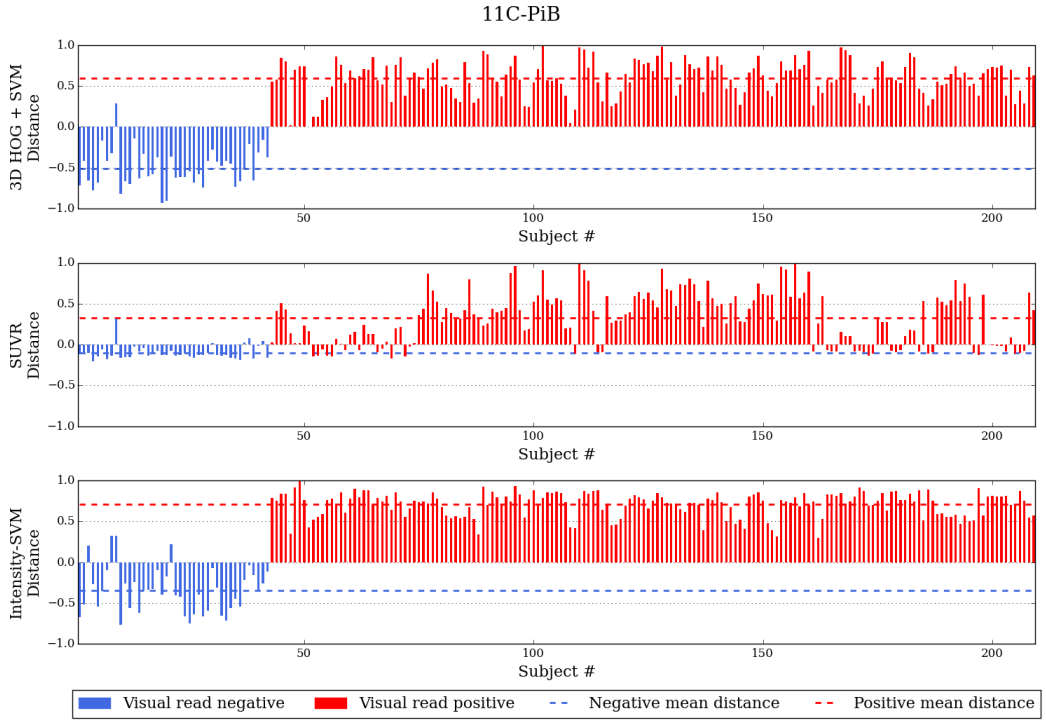


Figure 5.7: The normalised distances of ^{11}C-PiB test subjects from the classification boundary. Subjects in blue with positive distances were incorrectly classified by the given classification method. Similarly, subjects in red with negative distances were also incorrectly classified. The dashed lines indicate the mean distance from the boundary for the subjects visually designated as amyloid positive and amyloid negative.

equivocal scans from the classification boundary for each automated classification method and tracer. The 3D HOG + SVM method resulted in the largest mean absolute distance for ^{18}F-florbetapir (0.418), and the intensity-based SVM achieved the largest mean absolute distance for both ^{11}C-PiB and ^{18}F-florbetaben (0.878 and 0.700, respectively).

5.4 Discussion

5.4.1 Classification Accuracy

In this chapter an amyloid status classification method has been proposed, which is independent of the predefined regions of interest and amyloid positivity thresholds typically used to classify based on SUVR. The proposed method has been shown to generalise across multiple tracers, and could be used as an adjunct to visual

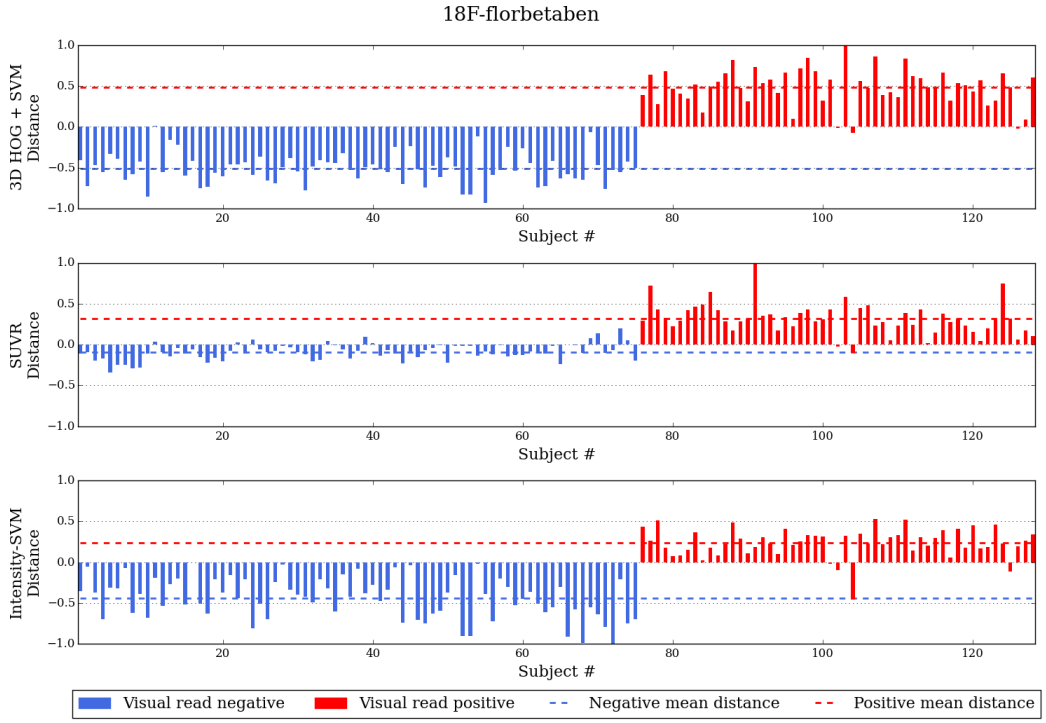
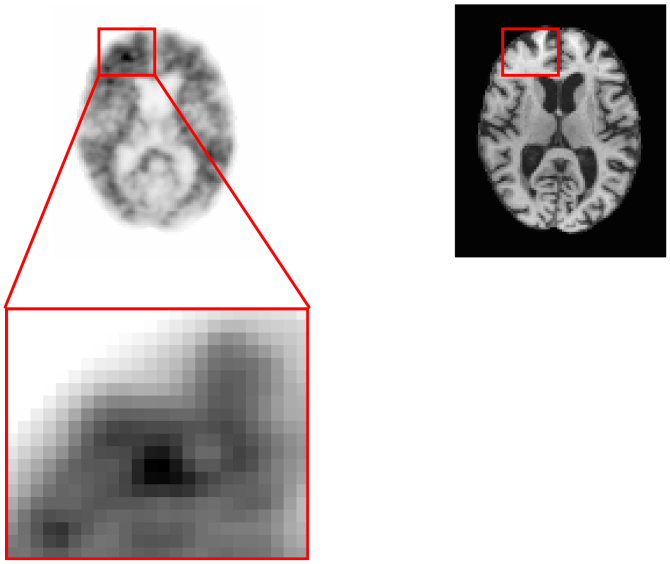


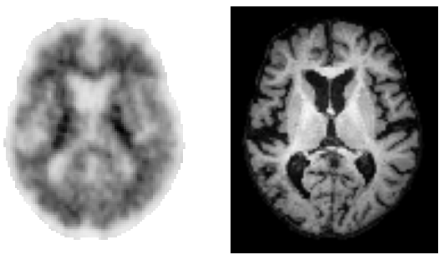
Figure 5.8: The normalised distances of ^{18}F-florbetaben test subjects from the classification boundary. Subjects in blue with positive distances were incorrectly classified by the given classification method. Similarly, subjects in red with negative distances were also incorrectly classified. The dashed lines indicate the mean distance from the boundary for the subjects visually designated as amyloid positive and amyloid negative.

interpretation of PET images, which is currently the standard method for clinical assessment. In a clinical setting, it would be straightforward to interpret the results due to the method's binary output (amyloid positive or amyloid negative). Moreover, unlike SUVR, knowledge of the specific amyloid positivity thresholds for each tracer is not required, and there is no need to check whether the target/reference regions are positioned correctly on the image.

In this work, the gold standard for amyloid status was determined using criteria based on consistent visual assessments from three image readers. Although ADNI provides a clinical diagnosis (e.g. cognitively normal, mild cognitive impairment, Alzheimer's disease) for each subject at the time of the ^{18}F-florbetapir and ^{11}C-PiB scans, these diagnoses are determined using a range of clinical tests. Consequently, the clinical diagnoses may disagree with visual interpretations of the scans [Frey 2015] (this is especially true for subjects with mild cognitive impairment), and



(a) ^{11}C-PiB, image #9



(b) ^{18}F-florbetaben, image #104

Figure 5.9: PET and MR axial slices from the two images which were classified differently to the gold standard visual assessment by all three classification methods. (a) ^{11}C-PiB image #9 was visually assessed as amyloid negative, but incorrectly classified as positive. However, on closer inspection, tracer uptake was observed in the frontal region (highlighted by the red box), suggesting that the classification should be amyloid positive. (b) ^{18}F-florbetaben image #104 was correctly assessed as positive, but automatically classified as amyloid negative.

in fact, the tracer manufacturer instructions state that a positive ^{18}F-florbetapir scan does not establish a diagnosis of AD or other cognitive disorder [Amyvid 2012]. Moreover, at present, clinical diagnosis can only be confirmed via an autopsy. Whilst ADNI strives to provide autopsy data, the majority of subjects in the database to do not have this information. The methods employed in this chapter focus on classification of amyloid status using PET images only, and therefore the clinical diagnoses were discarded. Furthermore, although the tracer manufacturers' instructions for visual assessment of amyloid images state that MR or computed tomography (CT)

images may be used to help localise tracer uptake [Amyvid 2012; Neuraceq 2014], the readers in this study were not presented with either of these scans. Both the gold standard visual assessments and the automatic amyloid status classification methods were conducted using PET images only.

Using visual assessment of the images as the gold standard, the 3D HOG + SVM method resulted in the highest classification accuracy and AUC for all three of the tracers evaluated. This could be because it uses local intensity gradients as features, rather than intensity directly, making it robust to spatially varying intensity levels. This is advantageous in PET image classification, when data that have been acquired, and reconstructed, at multiple sites and multiple scanners, may have different spatially varying intensities. Moreover, by quantising the gradients, the 3D HOG + SVM method is more robust to noise than the intensity-based SVM, which uses all of the voxels, which can include noise, as the feature vector. It may be possible to achieve a higher classification accuracy with the intensity-based SVM method by smoothing or downsampling the data to reduce the noise, though this could increase partial volume effects.

Another reason for the high classification accuracy and AUC of the 3D HOG + SVM method could be the use of cells, instead of individual voxels. Therefore, 3D HOG can cope well with minor misregistration of the brain to MNI space. When calculating SUVRs, a small misalignment of the PET brain could result in tracer uptake appearing to be outside of the region of interest. As a result, this may have some effect on SUVR.

For the ^{18}F-florbetapir and ^{18}F-florbetaben tracers the SUVR classification method exhibited the highest sensitivity. One reason for this could be the nature of the amyloid positivity thresholds. In a clinical setting, a test with a high sensitivity will rarely misdiagnose a diseased patient. Although false positives would cause unnecessary worry or treatment, a false negative patient could miss out on vital support and care. However, this notion is merely speculative, especially given that amyloid PET studies have generally not supported a clinically relevant bias towards report-

ing a scan as positive, and furthermore, there is currently no effective treatment for AD.

In contrast, the results in Figure 5.4 show a relatively lower classification accuracy (78.5%) and sensitivity (0.760) for SUVR compared to the 3D HOG + SVM and intensity-based SVM methods in ^{11}C-PiB cases. This may be due to the choice of amyloid positivity threshold. Since the sPAP quantification software that was used to calculate the SUVRs has not been validated with ^{11}C-PiB data, the threshold from [Jagust et al. 2009] (1.465) was converted to the sPAP scale (1.15). Equation 5.8 was constructed from two separate regression equations and implicitly assumed that the SUVR behaves linearly between the three quantification methods. In reality, this assumption may not be true. Moreover, not only could the original threshold of 1.465 be suboptimal, but every conversion introduces rounding errors. A slight change in the threshold can have an effect on the classification results. For example, by using a threshold of 1.16 instead of 1.15, the classification accuracy decreases from 78.5% to 78.0%. Similarly, the sensitivity decreases from 0.760 to 0.754. Although these differences appear small, on a large population a 0.5% difference in classification accuracy could mean that substantial numbers of patients are misdiagnosed. This result highlights the need for careful validation of new SUVR computation methods, and amyloid positivity thresholds, for both existing and new tracers, which is one of the goals of the Centiloid Project [Klunk et al. 2015].

5.4.2 3D HOG Parameters

The results of the 3D HOG parameter optimisation in Figure 5.2 suggest that this method is likely to give high classification accuracy, even with suboptimal parameters. This is confirmed by the fact that 68 out of 96 parameter combinations resulted in a classification accuracy greater than 95%. Cell size k had the most profound effect on classification accuracy, so very small or very large values of k should not be used. Interestingly, the optimum number of sub-blocks was $S = 1$. This is equivalent to not using sub-blocks, and only calculating gradients in the larger cells. One

possible reason for this result is that PET has a comparatively low resolution compared to the video sequences for which 3D HOG was designed. As a result, there is no need to average gradients over numerous sub-blocks. All four of the highest scoring parameter combinations used half-orientation histograms, suggesting that the sign of the gradient is uninformative in this particular application. This seems reasonable, given that the visual reading instructions for ^{18}F-florbetaben state that the images should be displayed in grey scale or inverse grey scale [Neuraceq 2014].

5.4.3 Distance from the Classification Boundary

For all cases except positive ^{11}C-PiB subjects, the 3D HOG + SVM method resulted in the greatest mean distances from the decision boundary. A large distance is desirable, since points near the classification boundary represent low-confidence classification decisions. Again, one reason for the superior performance of the 3D HOG + SVM method could be that it utilises image gradients, rather than direct voxel intensities. The resulting invariance to spatially varying intensity levels and noise robustness allows the two populations to be more easily separated than using SUVR or the intensity-based SVM method.

The small distance between amyloid negative subjects and the SUVR threshold supports its relatively lower specificity in Figures 5.3-5.5. The subjects close to the threshold are classified with a lower confidence, and are more likely to be misclassified as false positives.

The two images that were classified differently to the gold standard visual assessment (^{11}C-PiB image #9 and ^{18}F-florbetaben image #104) were visually assessed again by two of the original three readers. Although the ^{18}F-florbetaben image was again assessed to be an amyloid positive subject by the readers, the distance from the classification boundary for the SUVR and 3D HOG + SVM classification methods was small. This indicates that the different automatic classification decision has low confidence, and that this subject is a particularly difficult borderline case. After visual reassessment of the ^{11}C-PiB case, for which the gold standard

amyloid status was negative, a small region of tracer uptake was identified in the frontal lobe (highlighted by the red box in Figure 5.9(a)), suggesting that the gold standard amyloid status may have been incorrect for this case. Although thorough reader training programmes were developed in response to criticisms by the U.S. Food and Drug Administration (FDA) and European Medicines Association (EMA) regarding the consistency of F-18 amyloid image interpretation [FDA 2010; EMA 2013], our results highlight the importance of using an adjunct to visual assessment of amyloid images. In this study, the visual reads were conducted using PET volumes only. However, using MR images to help localise tracer uptake might make visual assessment more robust.

In this study, scans were given an "equivocal" visual assessment if they did not clearly fulfil the stringent definitions of amyloid positive or amyloid negative scans. Typically, scans were only designated as equivocal when readers lacked confidence in a final classification due to poor image quality. For this reason, equivocal scans are the type of scans for which automated classification could be most useful. Although there is no gold standard to which the classification results can be compared, the distances of the equivocal scans from the boundary indicate the level of confidence in the final classification of the automated classification methods. The 3D HOG + SVM method achieved the largest mean absolute distance from the boundary (0.418) for the equivocal ^{18}F-florbetapir scans, suggesting a higher level of confidence in the final classification than the SUVR and intensity-based SVM methods. For the equivocal ^{11}C-PiB and ^{18}F-florbetaben scans, the intensity-based SVM methods resulted in the largest mean absolute distance from the classification boundary (0.878 and 0.700, respectively). However, due to the small sample size (only five equivocal scans for both ^{11}C-PiB and ^{18}F-florbetaben), the distances for the intensity-based SVM method were not significantly larger ($p < 0.01$) than those achieved by the other two classification methods.

5.4.4 Methodological Considerations

Prior to computing the 3D HOG feature vectors, the PET volumes were affinely registered to MNI space to ensure that the cells generally contained the same brain regions across all subjects. A deformable registration algorithm could have been used, however, to keep the pre-processing steps of the proposed method in line with the sPAP SUVR method, affine registration was used. Furthermore, it was shown in Chapter 4 that classification of AD patients versus cognitively normal controls using SUVR is not affected by the registration method (affine versus deformable registration).

Although we used all of the ^{11}C-PiB and ^{18}F-florbetaben data that were available to us, the ADNI database contains many more ^{18}F-florbetapir scans than were used in this work. Therefore, in future, it would be useful to test our method on a larger dataset.

Many other feature descriptors have been developed in addition to histograms of oriented gradients. For example, the Scale-Invariant Feature Transform (SIFT) algorithm has been successfully used for object recognition in computer vision tasks [Lowe 1999], and has also been used in feature-based morphometry in MRI to distinguish between patients with AD and healthy controls [Toews et al. 2010]. Nevertheless, 3D HOG features were chosen for this work due to their simplicity and speed of computation. Moreover, unlike SIFT and Speeded Up Robust Features (SURF) [Bay et al. 2006], HOG operates on a dense grid of cells rather than individual points of interest. A larger set of image descriptors over a dense grid will typically offer more information than similar descriptors evaluated at a sparse set of image points.

Unlike the original intensity-based SVM method proposed in [Vandenberghe et al. 2013], which used an SVM with a linear kernel, we used an SVM with a Gaussian radial basis function kernel. Our primary reason for using a non-linear kernel was because the subjects in the original input space of the SVM might not be linearly separable. Although a linear SVM is faster to compute, and non-linear kernels can give rise to over-fitting, it has been shown that if complete model selection using

the Gaussian kernel has been conducted, there is no need to consider a linear SVM [Keerthi and Lin 2003]. In this work, we optimised the parameters of the SVM and Gaussian kernel on the ^{18}F-florbetapir training data only, using a grid search of parameters.

On a practical level, the 3D HOG + SVM method uses less memory than the intensity-based SVM method. The 3D HOG feature vector comprised 1500 elements, whereas the feature vector for the intensity-based method contained an element for each voxel inside the brain mask (290409 elements in total). As the number of elements increases, so does the time taken to train the SVM. The SUVR is also very quick to compute, but unlike the 3D HOG + SVM method, knowledge of the underlying anatomy and disease pathology is required in order to choose suitable target and reference regions.

5.4.5 Conclusion

In this chapter we have proposed a machine learning method for amyloid status classification based on histograms of oriented three-dimensional gradients. We compared our method to SUVRs obtained from clinically validated amyloid quantification software, as well as another machine learning method based solely on image intensity [Vandenberghe et al. 2013]. Across three separate amyloid tracers, the proposed 3D HOG + SVM method achieved the highest classification accuracy and area under the receiver operating characteristic curve. Moreover, the large separation between the population groups suggests that the 3D HOG + SVM method makes fewer low-confidence classification decisions. Unlike SUVR, the 3D HOG + SVM method required very little recalibration between tracers, and the results indicate that the method has the potential to produce satisfactory results even with suboptimal parameters. However, to fully assess the generalisability of the 3D HOG + SVM method, we need to analyse classification results obtained by optimising the 3D HOG parameters on other amyloid PET tracers, not just ^{18}F-florbetapir data. This is the subject of the next chapter.

Chapter 6

The Generalisability of 3D HOG for Amyloid Classification

In Chapter 5, a novel method for brain amyloid status classification was proposed. The method uses a feature vector based on histograms of oriented 3D gradients (3D HOG) as input to a support vector machine (SVM) classifier. In the experiments described in Chapter 5.2, we optimised the 3D HOG parameters using a set of ^{18}F-florbetapir PET training examples. Without recalibrating, the same 3D HOG parameters were used to classify ^{11}C-PiB and ^{18}F-florbetaben PET data. Despite this, the 3D HOG + SVM method achieved a higher classification accuracy than SUVR and a machine learning approach based on voxel intensity across all three tracers.

For an amyloid status classification method to be truly generalisable, it should be able to achieve a high classification accuracy, sensitivity and specificity, regardless of the amyloid PET tracer, and with minimal recalibration. Therefore, the aim of this chapter is to explore the generalisability of the 3D HOG + SVM method, by optimising the 3D HOG parameters on three different amyloid tracers (^{18}F-florbetapir, ^{11}C-PiB and ^{18}F-florbetaben), and assessing their classification performance when applied to unseen test data for each of those tracers. By testing three different parameter combinations on three different amyloid tracers, nine sets of results are

Table 6.1: A summary of the training and test data sets for each tracer. The ^{18}F-florbetapir training and test data are unchanged from Chapter 5.

	^{18}F-florbetapir		^{11}C-PiB		^{18}F-florbetaben	
	Train	Test	Train	Test	Train	Test
# subjects (positive/negative)	133 (75/58)	131 (74/57)	105 (84/21)	104 (83/21)	65 (27/38)	63 (26/37)
Age ± std dev.	75.8±8.6	74.6±7.2	77.6±6.9	75.6±8.0	69.9±7.5	69.9±7.4
Sex (male/female)	69/64	75/56	67/38	63/41	28/37	34/29

obtained. Analogously to Chapter 5, the results from the 3D HOG + SVM experiments were compared to SUVR and another machine learning method based on voxel intensity [Vandenberghe et al. 2013].

The remainder of this chapter is structured as follows: Section 6.1 describes the training and test PET data for each tracer. Section 6.2 briefly summarises the 3D HOG + SVM, SUVR, and intensity-based SVM classification methods. Section 6.3 presents the classification results for each method, and Section 6.4 discusses the generalisability of the 3D HOG + SVM method and concludes this work.

6.1 Materials

The images used in this work were the pre-processed PET volumes from Chapter 5. In total there were 264 ^{18}F-florbetapir volumes, 209 ^{11}C-PiB volumes, and 128 ^{18}F-florbetaben volumes.

In order to optimise the 3D HOG parameters for each tracer, the three data sets were split into training and test sets. The images comprising each set were randomly selected, but the training and test sets of a given tracer contained the same proportion of amyloid positive and amyloid negative scans as the entire data set. The ^{18}F-florbetapir training and test data were identical to those used in Chapter 5. A summary of the training and test data sets for each tracer is presented in Table 6.1

6.2 Methods

6.2.1 Parameter Optimisation

To determine the optimal 3D HOG parameters for each tracer, feature vectors were computed for 96 different parameter combinations. Similarly to Chapter 5, cell size k ranged from $k = 4$ voxels to $k = 32$ voxels in increments of 4 voxels. The number of sub-blocks S was in the set $S = \{1, 2, 4\}$, and the number of histogram bins was either 12 or 20 (i.e. the polyhedron was either a dodecahedron or icosahedron). The effect of full- and half-orientation histograms was also assessed.

For each tracer and parameter combination, an SVM classifier was trained using the feature vectors of the corresponding training data. The SVM parameters were optimised using ten-fold stratified cross-validation. As before, an SVM with a Gaussian radial basis function (RBF) was used, so both the slackness variable C and RBF free parameter γ were optimised [Boser et al. 1992; Cortes and Vapnik 1995]. For each tracer, the 3D HOG and SVM parameters that achieved the highest classification accuracy, sensitivity, and specificity on the training data were applied to the three sets of test data. Since the ^{18}F-florbetapir training data were identical to those used in Chapter 5, the optimal 3D HOG and SVM parameters for ^{18}F-florbetapir were unchanged.

6.2.2 Testing the 3D HOG + SVM Method

The procedure for testing the 3D HOG + SVM method is described in Algorithm 6.1. The testing approach in this chapter differed from Chapter 5, because we assessed three separate 3D HOG + SVM classifiers, each optimised using a different amyloid PET tracer, rather than a single 3D HOG + SVM classifier optimised for ^{18}F-florbetapir data only. In subsequent sections, the three 3D HOG + SVM classifiers are distinguished by the subscripts FBP, PIB, and FBB, which denote the amyloid PET tracer that was used to optimise the parameters (^{18}F-florbetapir, ^{11}C-PiB, and ^{18}F-florbetaben, respectively).

Algorithm 6.1 Procedure for testing the 3D HOG + SVM method

1: **for each** tracer x in {^{18}F-florbetapir,^{11} C-PiB,^{18} F-florbetaben} **do**
2: As described in Section 6.2.1, optimise the 3D HOG + SVM parameters using the training data for tracer x.
3: **for each** tracer y in {^{18}F-florbetapir,^{11} C-PiB,^{18} F-florbetaben} **do**
4: Using the optimal 3D HOG + SVM parameters for tracer x, conduct leave-one-out testing on the test data for tracer y
5: Calculate classification accuracy, sensitivity, specificity, and the area under the receiver operating characteristic curve (AUC)
6: **end for**
7: **end for**

6.2.3 Standardised Uptake Value Ratio

Identically to Chapter 5, the 3D HOG + SVM method was compared to classification results obtained using SUVR. Again, the commercially available Siemens syngo.PET Amyloid Plaque (sPAP) quantification software was used to compute the composite SUVR. The cortical regions of interest and reference region used to calculate the SUVR for each tracer are described in Chapter 5.2.2. The three amyloid positivity thresholds were:

^{18}F-florbetapir	1.12 [Hutton et al. 2015]
^{11}C-PiB	1.15 (see Chapter 5.2.2)
^{18}F-florbetaben	1.36 [Hutton et al. 2014]

Although the composite SUVR for each amyloid PET scan remains unchanged from Chapter 5, the classification accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve (AUC) were recalculated using only the test data. This is in contrast to Chapter 5, where the classification results were determined using the entire ^{11}C-PiB and ^{18}F-florbetapir data sets.

6.2.4 Image Intensity

Finally, the 3D HOG + SVM and SUVR methods were compared to the intensity-based classification method proposed by [Vandenberghe et al. 2013]. In summary, the method uses an SVM trained on the intensities of the voxels inside the brain to classify subjects as amyloid positive or amyloid negative. The method is identical to the intensity-based SVM method used in Chapter 5, however, in this work, three intensity-based SVMs were used – one for each tracer. The SVM parameters were optimised using ten-fold stratified cross-validation on the training data of a given tracer. The optimum parameters were then applied to the test data of that particular tracer, and leave-one-out testing was used to assess the ability of the intensity-based SVM to classify brain amyloid status.

6.3 Results

6.3.1 3D HOG Parameter Optimisation

Figure 6.1 shows, for a range of 3D HOG parameter settings, the highest classification accuracies achieved on the ^{11}C-PiB training data during the optimisation of the SVM parameters. Each sub-plot relates to one of the four histogram configurations (dodecahedron vs icosahedron, and half-orientation vs full-orientation), and shows the classification accuracy for each cell size k (horizontal axis) and number of sub-blocks S (vertical axis). Regardless of the number of sub-blocks and histogram configuration, a cell size of $k = 4$ resulted in the lowest classification accuracies. Nevertheless, 81 different parameter combinations resulted in a classification accuracy higher than 95%. Nineteen parameter combinations achieved the same, highest classification accuracy (99.0%). These combinations are highlighted by the red borders in Figure 6.1. Unlike Chapter 5.3.1, where one parameter combination gave the highest classification accuracy, sensitivity, and specificity, selecting a single optimal parameter combination for ^{11}C-PiB was more difficult. Fifteen of the 19 parameter combinations that achieved the highest classification accuracy also achieved the

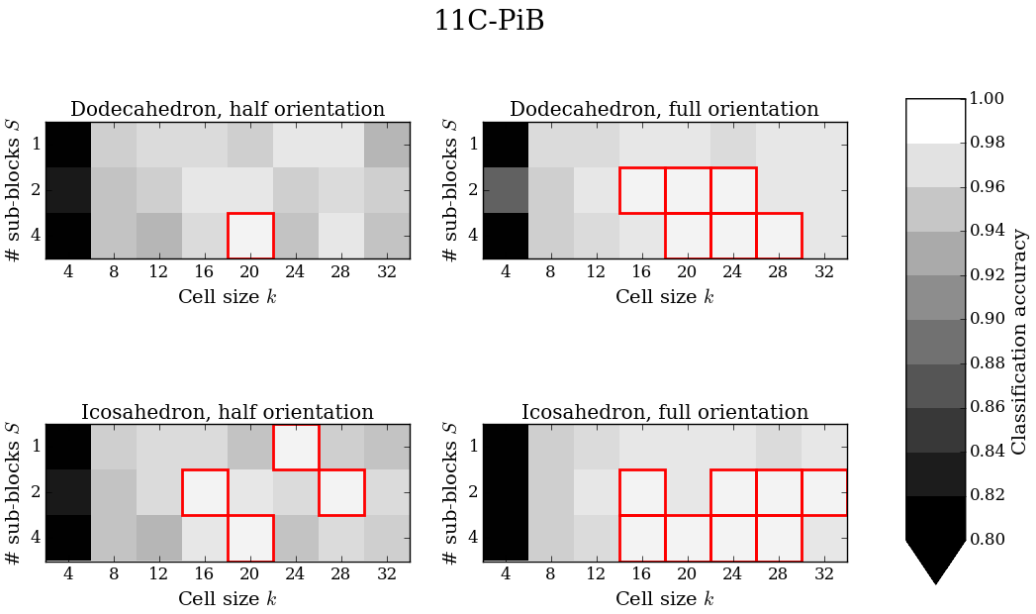


Figure 6.1: The classification accuracies achieved on the ^{11}C-PiB training data for each of the 96 3D HOG parameter combinations. Each sub-plot relates to one of the four histogram configurations (dodecahedron vs icosahedron, and half-orientation vs full-orientation), and shows the highest classification accuracy for each cell size k (horizontal axis) and number of sub-blocks S (vertical axis). Nineteen parameter combinations, highlighted in red, achieved the same, highest classification accuracy (99.0%).

same, highest sensitivity and specificity (1.00 and 0.952, respectively). Therefore, the final set of 3D HOG parameters was chosen to be the combination that yielded best classification results with the smallest feature vector. The final parameters for the ^{11}C-PiB data were as follows: $k = 28$, $S = 2$, icosahedron, half-orientation. This parameter combination results in a feature vector of 270 elements on the ^{11}C-PiB data.

Similarly to Figure 6.1, the classification accuracies achieved during the optimisation of the 3D HOG parameters for ^{18}F-florbetaben data are presented in Figure 6.2. Again, a cell size of $k = 4$ resulted in the lowest classification accuracy for all four histogram configurations. However, 81 parameter combinations achieved a classification accuracy greater than 95%. The red borders in Figure 6.1 indicate the 21 parameter combinations that resulted in the same, highest classification accuracy (100%). Therefore, similarly to the ^{11}C-PiB data, the final combination of 3D HOG parameters for ^{18}F-florbetaben data was chosen from the set of 21 on the basis of

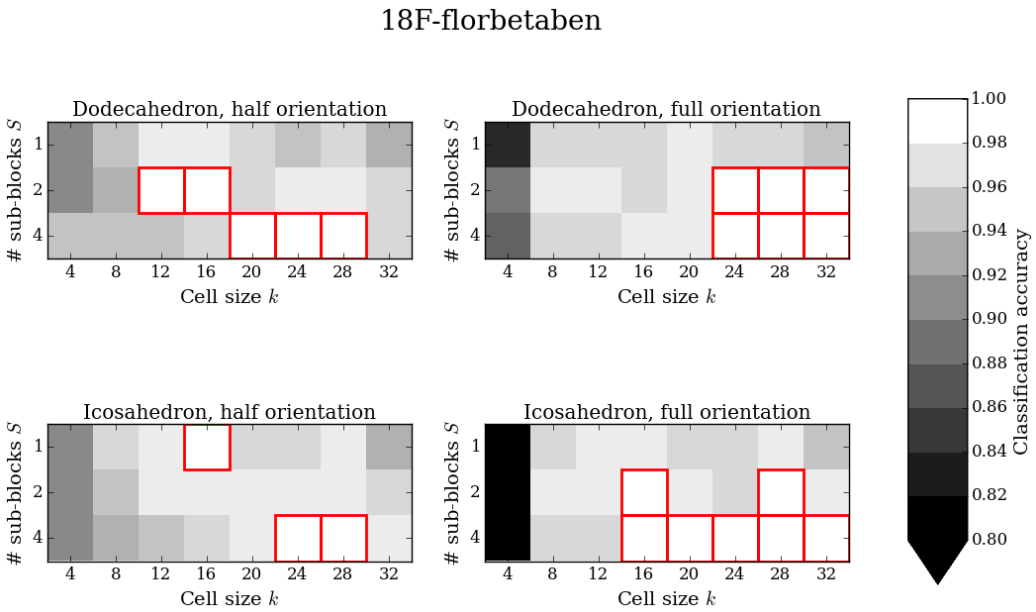


Figure 6.2: The classification accuracies achieved on the ^{18}F-florbetaben training data for each of the 96 3D HOG parameter combinations. Each sub-plot relates to one of the four histogram configurations (dodecahedron vs icosahedron, and half-orientation vs full-orientation), and shows the highest classification accuracy for each cell size k (horizontal axis) and number of sub-blocks S (vertical axis). Twenty-one parameter combinations, highlighted in red, achieved the same, highest classification accuracy (100%).

feature size. Two parameter combinations yielded a feature vector of 144 elements ($k = 32$, dodecahedron, full-orientation, with $S = 2$ or $S = 4$). As a result, the final parameter combination was chosen to be the one which is fastest to compute ($k = 32$, $S = 2$, dodecahedron, full-orientation).

The final sets of 3D HOG parameters used in the classification experiments are summarised in Table 6.2. The optimal parameters for ^{18}F-florbetapir are the same as those in Chapter 5.

6.3.2 Classification Results

Figures 6.3-6.5 show the classification accuracy, sensitivity, specificity and area under the ROC curve (AUC) of each classification method when applied to the ^{18}F-florbetapir, ^{11}C-PiB, and ^{18}F-florbetaben test data, respectively. The three coloured bars indicate the results of the 3D HOG + SVM$_{\text{FBP}}$ (red), 3D HOG + SVM$_{\text{PIB}}$ (blue), and 3D HOG + SVM$_{\text{FBB}}$ (yellow) methods, respectively. The best results

Table 6.2: The optimal 3D HOG parameters for each tracer. To test the generalisability of the 3D HOG + SVM method in the classification experiments, all three sets of parameters were used on every tracer.

Training data	Cell size k	# sub-blocks S	Polyhedron	Histogram orientation
^{18}F-florbetapir	16	1	Icosahedron	Half
^{11}C-PiB	28	2	Icosahedron	Half
^{18}F-florbetaben	32	2	Dodecahedron	Full

are highlighted with a black border. It should be noted that the results for the 3D HOG + SVM_{FBP} method differ slightly from those presented in Figures 5.4 and 5.5, since the results in Figures 6.4 and 6.5 use the ^{11}C-PiB and ^{18}F-florbetaben test data, which is a portion of the entire data set analysed in Chapter 5.

The 3D HOG + SVM_{FBP} method (red bars) achieved the highest classification accuracy across all three test sets (96.2%, 100% and 96.8%, respectively). Furthermore, for all three tracers, the three different 3D HOG + SVM methods resulted in a higher classification accuracy than the intensity-based SVM. The 3D HOG + SVM method also had a greater classification accuracy compared to SUVR in all cases except for 3D HOG + SVM_{PIB} on the ^{18}F-florbetapir data (blue bar in Figure 6.3), where it resulted in the same classification accuracy as SUVR (94.7%). For the ^{11}C-PiB data, the 3D HOG + SVM methods achieved 100% classification accuracy. Furthermore, across all three tracers, the 3D HOG + SVM methods achieved a higher specificity than SUVR, and a higher AUC than both SUVR and the intensity-based SVM method. Using the DeLong method [DeLong et al. 1988] to statistically compare the AUCs of each classification method, all three 3D HOG + SVM methods resulted in a significantly higher AUC than the SUVR and intensity-based SVM methods for the ^{11}C-PiB data ($p < 0.01$). In addition, the 3D HOG + SVM_{FBP} method achieved a significantly higher AUC than the intensity-based SVM method for the ^{18}F-florbetapir data ($p < 0.01$), and a significantly higher AUC than the SUVR method for the ^{18}F-florbetaben data ($p < 0.01$). The 3D HOG + SVM_{FBB} method also obtained a significantly higher AUC than the intensity-based SVM method for the ^{18}F-florbetapir data ($p < 0.01$). Across all three tracers, there

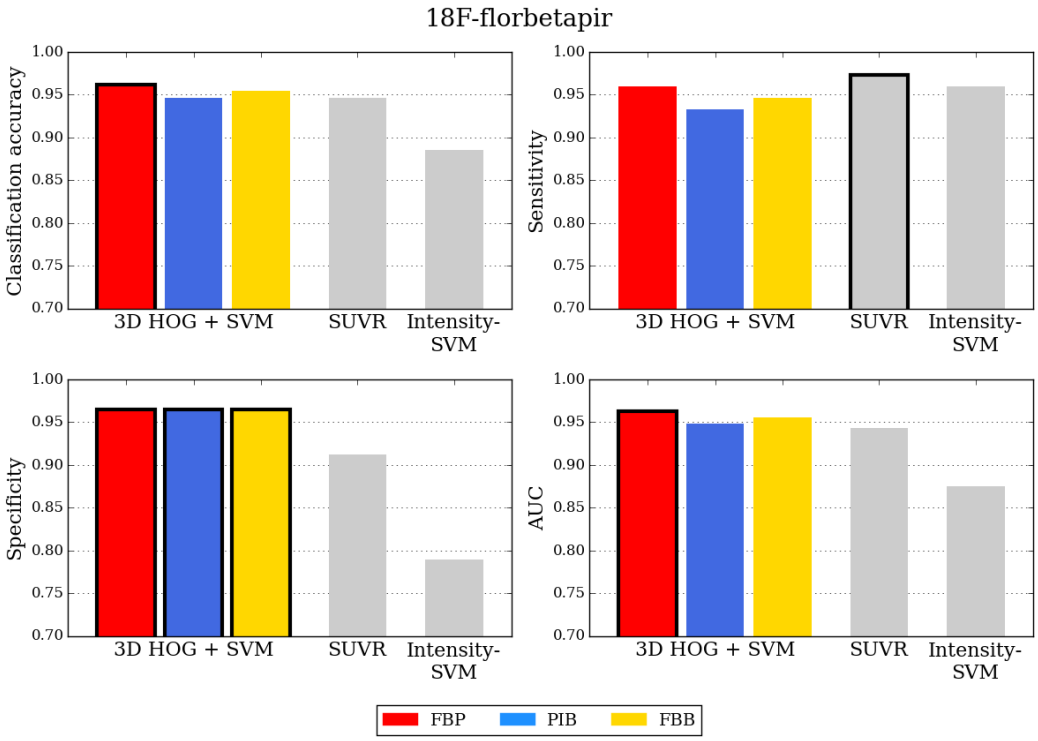


Figure 6.3: The classification accuracy, sensitivity, specificity and area under the receiver operating characteristic curve (AUC) for the ^{18}F-florbetapir test data. The coloured bars represent the three different 3D HOG + SVM methods, each optimised using data from a different amyloid PET tracer (red: 3D HOG + SVM$_{\text{FBP}}$, blue: 3D HOG + SVM$_{\text{PIB}}$, yellow: 3D HOG + SVM$_{\text{FBB}}$). The best results are highlighted with a black border.

was no statistical difference between the AUCs of the 3D HOG + SVM methods.

6.3.3 Distance to Classification Boundary

Figures 6.6-6.8 show the normalised distances of the test subjects from their respective classification boundary. The top three sub-plots of each figure denote the distances for the 3D HOG + SVM$_{\text{FBP}}$, 3D HOG + SVM$_{\text{PIB}}$, and 3D HOG + SVM$_{\text{FBB}}$ methods, respectively. The blue bars represent subjects that were visually assessed as amyloid negative, and the red bars represent subjects that were visually assessed as amyloid positive. Subjects in blue with positive distances, and subjects in red with negative distances were incorrectly classified by the given classification method. The dashed lines indicate the mean distance from the boundary for the subjects visually designated as amyloid positive and amyloid negative.

Similarly to Chapter 5, a two-tailed t-test was used to examine whether the

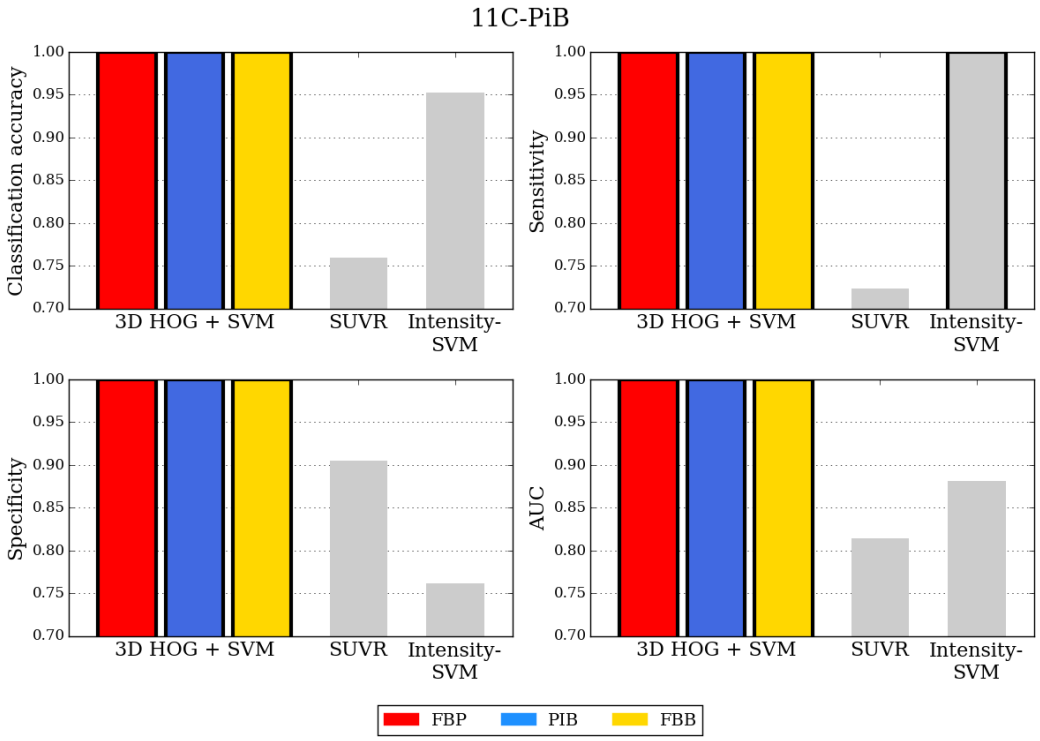


Figure 6.4: The classification accuracy, sensitivity, specificity and area under the receiver operating characteristic curve (AUC) for the ^{11}C-PiB test data. The coloured bars represent the three different 3D HOG + SVM methods, each optimised using data from a different amyloid PET tracer (red: 3D HOG + SVM$_{\text{FBP}}$, blue: 3D HOG + SVM$_{\text{PIB}}$, yellow: 3D HOG + SVM$_{\text{FBB}}$). The best results are highlighted with a black border.

boundary distances of each 3D HOG + SVM methods were significantly greater than the SUVR and intensity-based methods. Values of $p < 0.01$ (corrected for multiple comparisons) are considered significant.

The distances of the amyloid negative subjects were significantly greater for all three 3D HOG + SVM methods than SUVR, across all three tracers. The 3D HOG + SVM methods also achieved significantly greater distances than SUVR for the amyloid positive ^{11}C-PiB test data. Moreover, all three 3D HOG + SVM methods resulted in significantly larger distances than the intensity-based SVM for the ^{18}F-florbetapir and ^{11}C-PiB amyloid negative scans, and the ^{18}F-florbetaben amyloid positive scans. Table 6.3 presents a summary to indicate which of the 3D HOG + SVM methods resulted in significantly greater distances from the classification boundary than the SUVR and intensity-based SVM methods. Out of a total 36 comparisons, the 3D HOG + SVM methods resulted in significantly larger distances

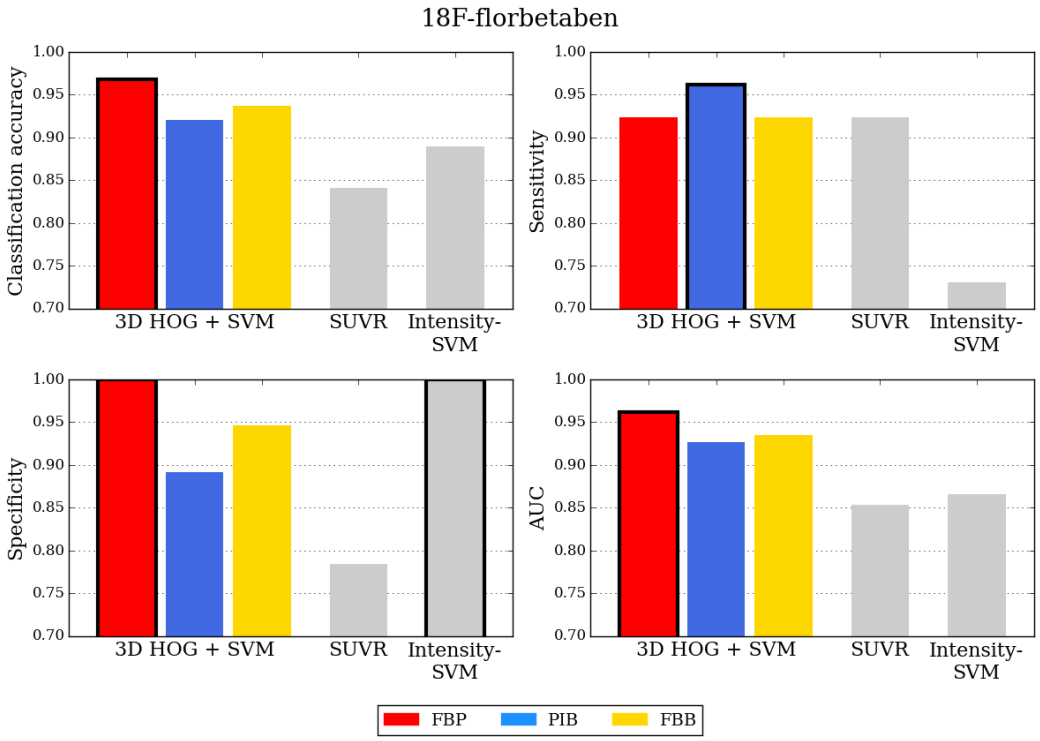


Figure 6.5: The classification accuracy, sensitivity, specificity and area under the receiver operating characteristic curve (AUC) for the ^{18}F-florbetaben test data. The coloured bars represent the three different 3D HOG + SVM methods, each optimised using data from a different amyloid PET tracer (red: 3D HOG + SVM$_{\text{FBP}}$, blue: 3D HOG + SVM$_{\text{PIB}}$, yellow: 3D HOG + SVM$_{\text{FBB}}$). The best results are highlighted with a black border.

than SUVR and the intensity-based SVM methods for 26 comparisons. Two-tailed t-tests between the 3D HOG + SVM methods showed that there was no significant difference between the distances from the classification boundary for the amyloid negative scans of all three tracers. The 3D HOG + SVM$_{\text{FBP}}$ and 3D HOG + SVM$_{\text{FBB}}$ methods achieved significantly greater distances from the classification boundary than the 3D HOG + SVM$_{\text{PIB}}$ method for amyloid positive ^{18}F-florbetapir scans. However, the 3D HOG + SVM$_{\text{PIB}}$ method resulted in significantly greater distances than the 3D HOG + SVM$_{\text{FBP}}$ and 3D HOG + SVM$_{\text{FBB}}$ methods for amyloid positive ^{11}C-PiB scans.

In addition to the distances from the classification boundary, Figure 6.6 shows that three ^{18}F-florbetapir test images were classified differently to the gold standard visual assessment by all three 3D HOG + SVM methods (images #50, #73 and #130). Moreover, one ^{18}F-florbetaben test image was classified differently to the

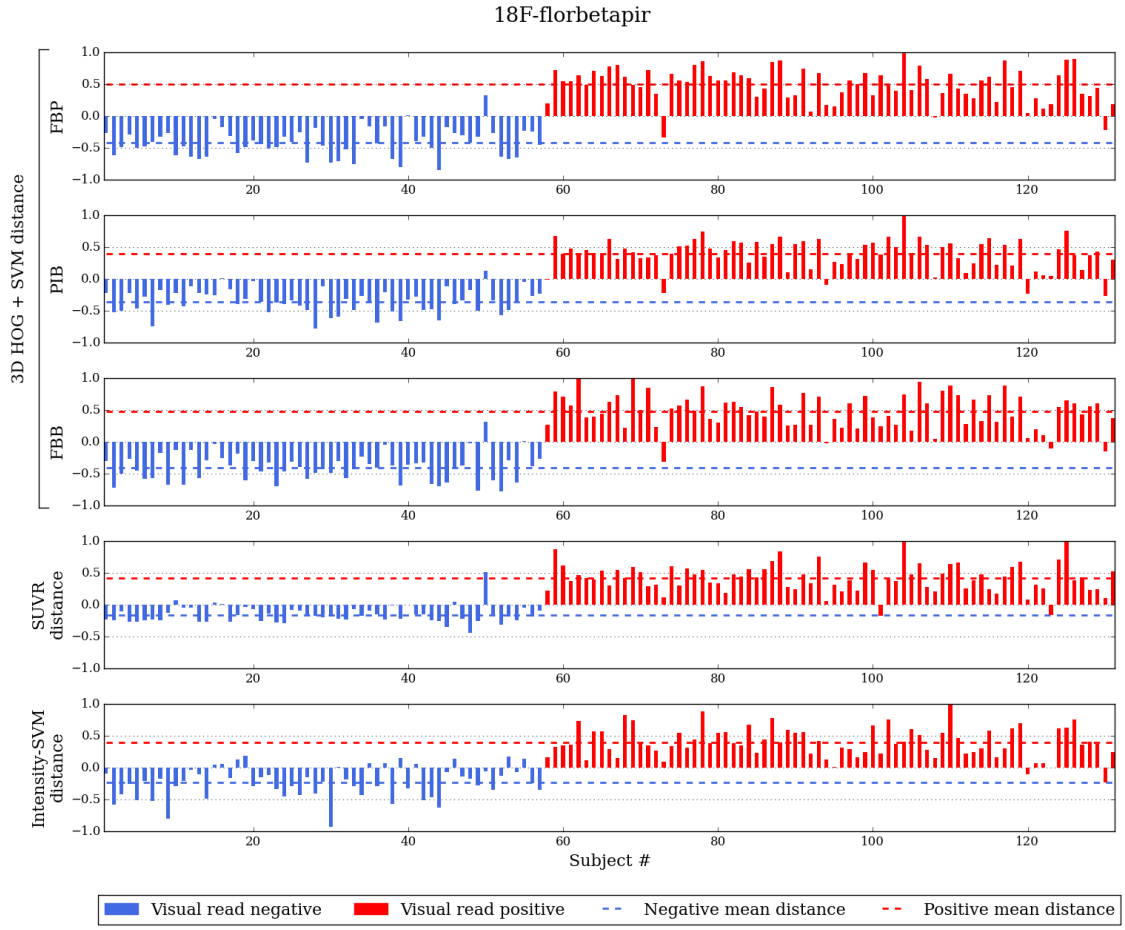


Figure 6.6: The normalised distances of ^{18}F-florbetapir test subjects from the classification boundary. Subjects in blue with positive distances were incorrectly classified by the given classification method. Similarly, subjects in red with negative distances were also incorrectly classified. The dashed lines indicate the mean distance from the boundary for the subjects visually designated as amyloid positive and amyloid negative.

visual assessment by all five classification methods (image #49). An axial slice from the PET and MR volumes of the three ^{18}F-florbetapir outliers are shown in Figure 6.9. Figure 6.10 shows the corresponding axial PET and MR slice for the ^{18}F-florbetaben image that was classified differently to the gold standard.

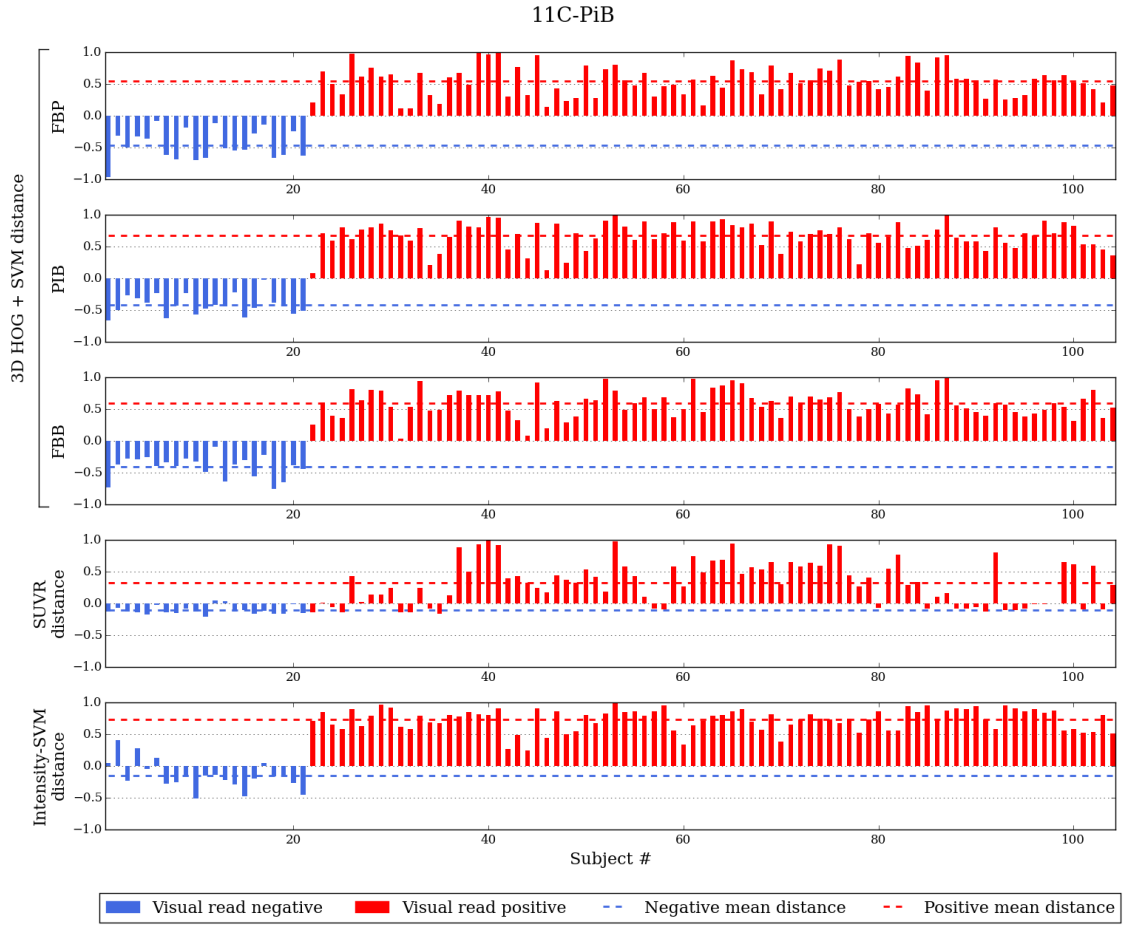


Figure 6.7: The normalised distances of ^{11}C-PiB test subjects from the classification boundary. Subjects in blue with positive distances were incorrectly classified by the given classification method. Similarly, subjects in red with negative distances were also incorrectly classified. The dashed lines indicate the mean distance from the boundary for the subjects visually designated as amyloid positive and amyloid negative.

6.4 Discussion

6.4.1 Classification Accuracy

In this chapter, the generalisability of the 3D HOG + SVM method has been evaluated, with regards to classification of brain amyloid status across three different amyloid PET tracers.

Across all three tracers, the 3D HOG + SVM$_{\text{FBP}}$ method achieved the highest classification accuracy of all the methods tested. Furthermore, the two other 3D HOG + SVM methods, 3D HOG + SVM$_{\text{PIB}}$ and 3D HOG + SVM$_{\text{FBB}}$, resulted in higher classification accuracies than the intensity-based SVM, and greater or

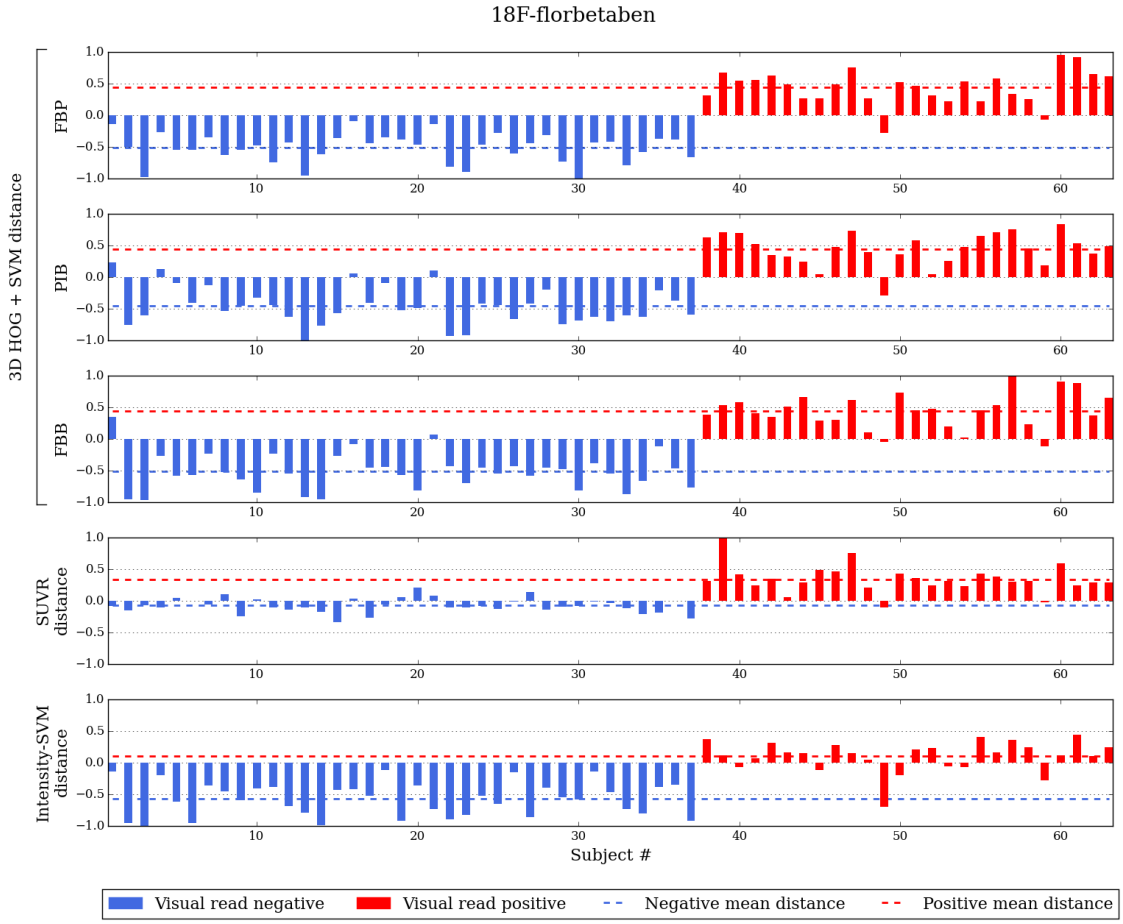


Figure 6.8: The normalised distances of ^{18}F-florbetaben test subjects from the classification boundary. Subjects in blue with positive distances were incorrectly classified by the given classification method. Similarly, subjects in red with negative distances were also incorrectly classified. The dashed lines indicate the mean distance from the boundary for the subjects visually designated as amyloid positive and amyloid negative.

equal classification accuracies to SUVR. These results demonstrate that the 3D HOG + SVM method generalises to multiple amyloid tracers, and can obtain high classification accuracies over a range of different parameter combinations.

As described in Chapter 5.4, one reason for the higher classification accuracy of the 3D HOG + SVM method compared to the other two methods could be due to its use of local intensity gradients as features, rather than voxel intensities directly. Conceptually, this is similar to visual assessment of amyloid PET scans, which utilises local loss of contrast between grey and adjacent white matter. Consequently, this makes the 3D HOG + SVM method more robust to spatially varying intensity levels.

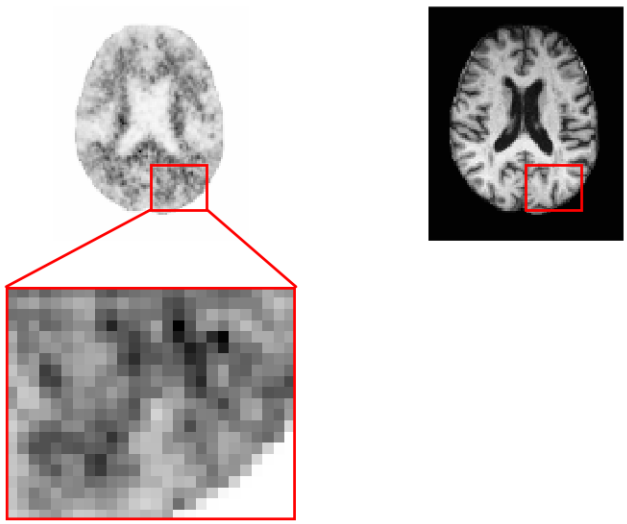
Table 6.3: A summary to indicate which of the 3D HOG + SVM methods resulted in significantly greater ($p < 0.01$, corrected) distances from the classification boundary than the SUVR and intensity-based SVM methods, for both amyloid negative (N) and amyloid positive (P) subjects.

Test data	3D HOG + SVM method	Group	Significantly greater ($p < 0.01$, corrected) distances than	
			SUVR	Intensity-SVM
^{18}F-florbetapir (Figure 6.6)	FBP	N	✓	✓
		P	✓	✓
	PIB	N	✓	✓
		P	-	-
	FBB	N	✓	✓
		P	-	✓
^{11}C-PiB (Figure 6.7)	FBP	N	✓	✓
		P	✓	✓
	PIB	N	✓	✓
		P	✓	-
	FBB	N	✓	✓
		P	✓	✓
^{18}F-florbetaben (Figure 6.8)	FBP	N	✓	-
		P	-	✓
	PIB	N	✓	-
		P	-	✓
	FBB	N	✓	-
		P	-	✓

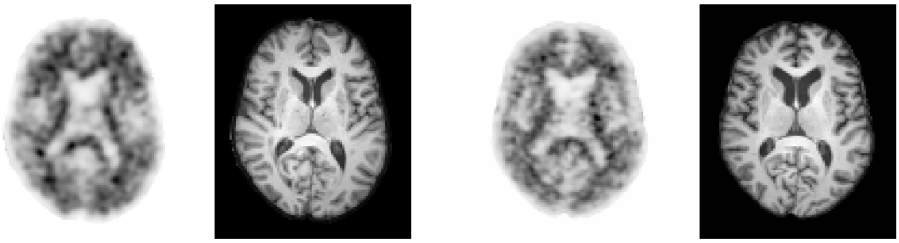
Figures 6.3-6.5 show that the three 3D HOG + SVM methods achieved a higher AUC than SUVR and the intensity-based SVM method across all three tracers. These results indicate that the 3D HOG + SVM method is more likely to agree with the visual assessment of a randomly selected subject than the SUVR or intensity-based SVM methods. Moreover, the DeLong test (Section 6.3.2) showed that there was no statistical difference between the AUCs of the three 3D HOG + SVM methods ($p < 0.01$, corrected), despite using different parameters. This result suggests that the 3D HOG + SVM method is robust to the choice of parameters and amyloid PET tracer.

6.4.2 3D HOG Parameters

The results in Figures 6.1 and 6.2 further suggest that the 3D HOG + SVM method generalises across different amyloid tracers, and is capable of delivering a high clas-



(a) Image #50



(b) Image #73 (c) Image #130

Figure 6.9: PET and MR axial slices from the three ^{18}F-florbetapir subjects which were classified differently to the gold standard visual assessment by all three 3D HOG + SVM classification methods. (a) Image #50 was visually assessed as amyloid negative, but incorrectly classified as amyloid positive. However, on closer inspection, tracer uptake was observed in the grey matter in the region highlighted by the red box, suggesting that the image could be amyloid positive. (b) and (c) Images #73 and #130 were correctly assessed as amyloid positive, but automatically classified as amyloid negative.

sification accuracy for a variety of parameter combinations. For both ^{11}C-PiB and ^{18}F-florbetaben, 81 different 3D HOG parameter combinations achieved a classification accuracy greater than 95% when optimising the SVM on the training data. Furthermore, during the optimisation on ^{18}F-florbetaben, 21 different 3D HOG parameter combinations obtained a classification accuracy of 100% on the training data.

Similarly to Chapter 5, a cell size of $k = 4$ universally resulted in the lowest classifications accuracies, suggesting that very small cell sizes should not be used. The

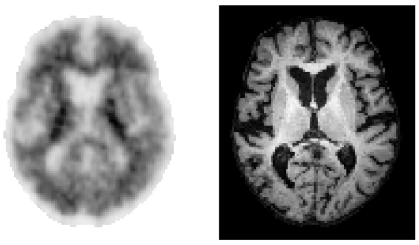


Figure 6.10: PET and MR axial slices from the ^{18}F-florbetaben subject which was classified differently to the gold standard visual assessment by all five classification methods. ^{18}F-florbetaben image #49 was visually assessed as amyloid positive, but automatically classified as amyloid negative. This is the same ^{18}F-florbetaben image that was consistently classified differently to the gold standard in Chapter 5.

majority of 3D HOG parameter combinations that gave the highest classification accuracies (the 19 red boxes in Figure 6.1 and 21 red boxes in Figure 6.2) used two or four sub-blocks. The histogram for a given cell is the sum of the quantised mean gradients of the sub-blocks in that cell. Therefore, more sub-blocks could mean that the histogram contains larger values. If the gradients in the sub-blocks are in the same general direction, this would result in a large difference between histogram elements. This large discrepancy between histogram values could be useful for classification because the histogram would be more robust to small differences in cell gradients between subjects.

In future, to fully assess the generalisability of the 3D HOG + SVM method, it would be useful train the method on a mixture of ^{18}F-florbetapir, ^{11}C-PiB, and ^{18}F-florbetaben data, rather than data from one amyloid PET tracer only. The optimum 3D HOG parameters for this case can be determined from the results of the parameter optimisations presented in Chapter 5.3.1 and Section 6.3.1. The parameters that gave the highest mean classification accuracy (98.9%) across the training data for all three tracers were: $k = 28$, $S = 2$, icosahedron, with a full-orientation histogram. Although these parameters are slightly different to the 3D HOG parameters used in this chapter, the parameter sets used in this work all achieved a mean classification accuracy greater than 97% when applied to training data comprising all three amyloid PET tracers.

6.4.3 Distance from the Classification Boundary

The 3D HOG + SVM methods achieved a significantly larger normalised distance from the classification boundary than the SUVR and intensity-based SVM methods for most of the comparisons presented in Table 6.3. A large distance from the boundary is advantageous, because points close to the boundary represent lower confidence classification decisions. As proposed in Chapter 5.4, one reason for the 3D HOG + SVM method producing larger distances than the other methods could be that it is more robust to noise because it uses quantised gradients, rather than all the voxels directly, which can include noise.

Across all three tracers, SUVR consistently resulted in the smallest mean distance from the classification boundary for amyloid negative subjects (-0.164, -0.114 and -0.080, respectively). In contrast, SUVR gave much larger mean distances for the amyloid positive subjects (0.416, 0.321 and 0.337, respectively), suggesting that the classification decisions could be made more robust by revising the amyloid positivity threshold upwards.

It is clear from Figure 6.6 and 6.8 that some subjects were consistently classified differently to the gold standard. Consequently, the four images in Figures 6.6 and 6.8 were reassessed visually. Despite being automatically classified as amyloid negative by all three 3D HOG + SVM methods, ^{18}F-florbetapir images #73 and #130 were again visually assessed to be amyloid positive. Similarly, ^{18}F-florbetaben image #49 maintained its amyloid positive visual assessment, even though it was automatically classified as amyloid negative. This is the same image that was consistently classified differently to the gold standard in Chapter 5, so it is reasonable that it would be misclassified again. In contrast, although the original gold standard assessment for ^{18}F-florbetapir image #50 was amyloid negative, following a reassessment, tracer uptake was observed in the occipital lobe grey matter (see the red box in Figure 6.10). This suggests that the image could be amyloid positive, thus agreeing with the classification decision made by the 3D HOG + SVM methods. These results highlight the need for an adjunct to visual assessment of amyloid images. If an au-

tomatic classification method disagrees with a visual read, it should prompt further examination from a clinician, leading to a more accurate assessment of the image, and ultimately, the patient.

6.4.4 Conclusion

In this chapter we have examined the generalisability of the 3D HOG + SVM method for classification of brain amyloid status. In both clinical and research settings, a generalisable amyloid status classification method would be useful in order to negate the need for tracer-specific cortical regions of interest and amyloid positivity thresholds. Our results indicate that the 3D HOG + SVM method can achieve a high classification accuracy, even on amyloid PET data for which the parameters have not been optimised. For a variety of parameter combinations, the 3D HOG + SVM method resulted in a higher classification accuracy than a machine learning method based solely on voxel intensity. Furthermore, depending on the parameters used, the 3D HOG + SVM method achieved a classification accuracy greater than or equal to SUVR. The results in this chapter also demonstrate that the 3D HOG + SVM method resulted in significantly larger distances from the classification boundary than SUVR or the intensity-based SVM method. This suggests that is it less likely to make low-confidence classification decisions.

Although our 3D HOG + SVM amyloid status classification method does not require specific tracer or anatomical knowledge, the 3D HOG image features are hand-crafted. Therefore, in the next chapter, we will explore whether it is possible to achieve the same, or better, classification results using a method for automatic feature learning.

Chapter 7

Amyloid Status Classification Using Deep Learning

In Chapters 5 and 6, we proposed and evaluated a novel machine learning method for amyloid status classification. Our method, 3D HOG + SVM, was based on the histogram of oriented gradients (HOG) feature descriptor. HOG features are known as "hand-crafted" image features because they were originally based on empirical observations and designed for a specific task (detecting pedestrians in 2D images [Dalal and Triggs 2005]). Therefore, despite the high classification accuracies obtained by our 3D HOG + SVM method in Chapters 5 and 6, HOG features may not be the optimal features for amyloid status classification in PET imaging. In contrast to using the hand-crafted HOG features, deep learning methods automatically derive image features from training data, thus eliminating the need to design suitable features for different tasks. Furthermore, it has been shown that deep learning methods can infer a hierarchy of physiologically meaningful image features [Plis et al. 2014], and achieve competitive, if not superior, segmentation and classification results to other methods [Brosch et al. 2014; Guo et al. 2014].

For these reasons, the aim of this chapter is to investigate the potential of deep learning to derive features from amyloid PET images, which can be used to discriminate between amyloid positive and amyloid negative subjects. To our knowledge,

at the time of writing this thesis, the only study that has applied deep learning to PET brain imaging for classification of AD reduced the PET images to SUVR-like features, and combined them with information from structural MR images prior to the deep learning process [Suk and Shen 2013]. Therefore, this chapter presents the first investigation into using deep learning methods to derive features directly from amyloid PET images. We also assess the effect that the depth of the deep learning method has on classification accuracy. Moreover, analogously to [Lee et al. 2009a], we explore the idea of using readily available unrelated data (i.e. non-medical images) to train the first layer of the deep learning network, in order to potentially improve the classification accuracy.

To achieve this, we constructed one-layer, two-layer, and three-layer convolutional deep belief networks (CDBNs) [Desjardins and Bengio 2008]. The CDBN is an extension to the deep belief network proposed by [Hinton et al. 2006], which was introduced in Chapter 2.3. Unlike the deep belief network, which flattens the input images into a 1D array, the CDBN is able to take into account the spatial structure of the images, and detect the same feature at multiple locations within an image. In this chapter, we adopt CDBNs for amyloid status classification because they have achieved state-of-the-art classification performance on the CIFAR-10 image data set [Krizhevsky 2010], and have been successfully used for face detection [Lee et al. 2009a]. We assessed the classification performance of each CDBN, and compared the results to four other classification methods: SUVR, a 2D and 3D implementation of the intensity-based SVM classification proposed by [Vandenberghe et al. 2013], and the 3D HOG + SVM method from Chapter 5.

The remainder of this chapter is organised as follows: Section 7.1 introduces the convolutional restricted Boltzmann machine, and its role in a CDBN. Section 7.2 gives details of the amyloid PET data used in this chapter, and Section 7.3 describes the CDBN construction and the classification experiments. In Section 7.4 we present our results, and finally, Section 7.5 discusses the applicability of CDBNs for amyloid status classification.

7.1 Methods

As explained in Chapter 2.3, each layer in a deep belief network is composed of a restricted Boltzmann machine (RBM). Analogously, each layer of a CDBN is composed of a convolutional RBM (CRBM). In the following sections we will introduce the CRBM, as originally proposed by [Desjardins and Bengio 2008], and explain how multiple CRBMs are stacked to form a CDBN [Lee et al. 2009a].

7.1.1 Convolutional Restricted Boltzmann Machines

A CRBM is the building block of a CDBN. It is a probabilistic model that defines a joint energy function between a visible layer V and a hidden layer H (see Figure 7.1). Without restricting generalisability, we will assume that the visible layer comprises a square, 2D input image of size $N_V \times N_V$. The visible units v_{ij} correspond to the pixels in that input image. The hidden layer consists of K feature detectors (known as *bases*), each one an $N_H \times N_H$ array of binary units h_{ij}. Each base H^k is connected to the visible layer by an $N_W \times N_W$ filter[1] W^k (where $N_W = N_V - N_H + 1$), represented by the coloured boxes in Figure 7.1. No connections are permitted between bases. All of the visible units share a bias a, and each of the K hidden bases has an associated bias b_k. The energy of the joint configuration $(\mathbf{v}, \mathbf{h})$ is given by [Lee et al. 2009a]:

$$E(\mathbf{v}, \mathbf{h}) = -\sum_{k=1}^{K} \sum_{i,j=1}^{N_H} \sum_{q,r=1}^{N_W} h_{ij}^k W_{qr}^k v_{i+q-1,j+r-1} - a \sum_{i,j=1}^{N_V} v_{ij} - \sum_{k=1}^{K} b_k \sum_{i,j=1}^{N_H} h_{ij}^k \qquad (7.1)$$

By changing the notation such that $\bullet$ denotes $A \bullet B = \mathrm{tr} A^T B$, and $*$ denotes discrete convolution, the energy function can be simplified:

$$E(\mathbf{v}, \mathbf{h}) = -\sum_{k=1}^{K} h^k \bullet (\tilde{W}^k * v) - a \sum_{i,j} v_{ij} - \sum_{k=1}^{K} b_k \sum_{i,j} h_{ij}^k \qquad (7.2)$$

[1] The filters are a 2D extension of the weights in the original RBM, presented in Figure 2.9 and Equation 2.17.

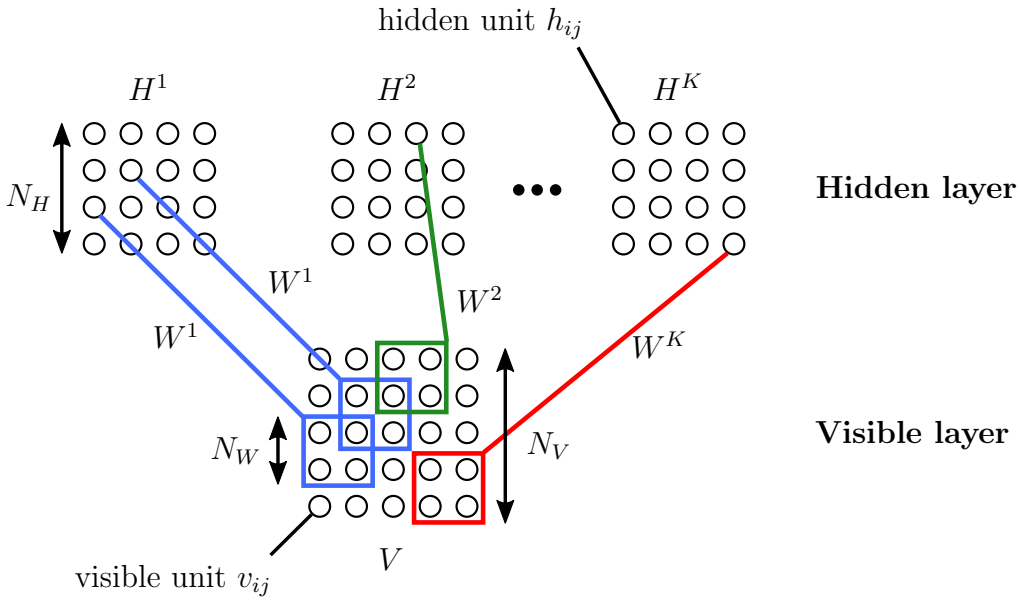


Figure 7.1: An illustration of a CRBM. For simplicity, only three bases in the hidden layer are shown. The black circles represent the individual units in each layer. Each hidden unit h_{ij} represents the presence of a particular image feature at an $N_W \times N_W$ neighbourhood of visible units. If a feature is detected within a neighbourhood, the corresponding hidden unit is activated (set to 1). The blue, green, and red lines indicate the hidden units that correspond to the neighbourhoods of the same colour. Hidden units within a base H^k represent the presence of the same feature at different locations within the visible layer V. Each filter W^k detects a different image feature.

where $\tilde{W}^k$ denotes that W^k has been reflected horizontally and vertically.

Given the partition function $Z = \sum_x \exp(-E(\mathbf{v}, \mathbf{h}))$, the probability distribution can be defined as:

$$P(\mathbf{v}, \mathbf{h}) = \frac{1}{Z} \exp(-E(\mathbf{v}, \mathbf{h})) \tag{7.3}$$

Using Equation 7.3, the probability of the visible layer $P(\mathbf{v})$ is obtained by marginalising over the hidden layer. To train the CRBM, gradient descent on the negative log-likelihood of $P(\mathbf{v})$ is performed. Since the computation of the partition function Z is mathematically intractable, so is the computation of the gradient. Therefore, identically to the RBM method presented in Chapter 2.3, the gradient is estimated using a truncated Gibbs chain. This method is also known as contrastive divergence [Hinton 2002]. The learning rate ϵ required for stable learning depends on the application and is set by the user.

Due to the specific bipartite structure of the network, where the hidden bases are conditionally independent of one another given the state of the visible layer (and vice versa), sampling a CRBM is very efficient. The probability of a single hidden unit h_{ij} being activated (i.e. set to 1) is given by:

$$P(h_{ij}^k = 1|\mathbf{v}) = \sigma(b_k + (\tilde{W}^k * v)_{ij}) \tag{7.4}$$

where $\sigma(x) = 1/(1 + \exp(-x))$.

Similarly, the probability of a visible unit v_{ij} being activated is given by:

$$P(v_{ij} = 1|\mathbf{h}) = \sigma(a + (\sum_k W^k * h^k)_{ij}) \tag{7.5}$$

It should be noted that Equations 7.4 and 7.5 have the same form as the equations in the original RBM (Equations 2.23 and 2.24, respectively).

Using the probability that a hidden unit h_{ij} will be activated, the binary value of that unit can be determined. If $P(h_{ij}^k = 1|\mathbf{v}) > 0.5$, hidden unit h_{ij} is set to 1, else hidden unit h_{ij} is set to 0. The expected values of the hidden units are known as *activations*.

7.1.2 Convolutional Deep Belief Networks

It has been demonstrated by [Bengio et al. 2014] that higher-level, more abstract features can make it easier to extract useful information about the input data when building a classifier. Higher-level image features can be obtained using CRBMs by stacking them to form a CDBN. A CDBN is constructed by using the activations from one CRBM (i.e. the values of the hidden layer) as the visible layer in another CRBM. Each CRBM is trained in isolation, and the activations of the second CRBM can be connected to a third CRBM, and so on. To reduce the computational burden of stacking CRBMs, [Lee et al. 2009a] reduced the size of the hidden layer bases H^k by a constant factor C prior to using the activations as inputs to another CRBM. This was achieved via *max-pooling*, where each unit in the max-pooled base p^k

corresponds to the maximum value in a small region in the hidden layer base H^k, i.e. over a number of hidden units. We adopt a similar approach in this chapter, and an illustration of a CDBN comprising two CRBMs with max-pooling is shown in Figure 7.2. In subsequent sections, we refer to a CDBN that consists of n CRBMs with max-pooling as an *n-layer CDBN*.

Once a CDBN has been trained, the values of the final max-pooled bases (also known as the *max-pooled activations*) can be used to train a classifier, such as a support vector machine (SVM) [Lee et al. 2009a]. In order to be used as the input to an SVM, the max-pooled bases p^k are flattened into a 1D array. This is indicated by the top dashed box in Figure 7.2. The classifier can be tested by computing the final max-pooled activations of a previously unseen image.

7.2 Materials

The 264 ^{18}F-florbetapir PET and T1-weighted MR volumes that were used in Chapter 5 were also used in this chapter. The data were pre-processed using the method described in Chapter 3.2.2. In summary, the PET and MR volumes were rigidly co-registered, skull-stripped, and then transformed to MNI space. The gold standard for the classification experiments in this chapter was the visual assessment of the amyloid PET images, as originally used in Chapter 5. The final data set consisted of 149 amyloid positive subjects and 115 amyloid negative subjects.

7.3 Experiments

7.3.1 Constructing and Training the CDBNs

In this study, we assessed the effect that the depth of a CDBN has on its ability to classify amyloid PET scans as amyloid positive or amyloid negative. To achieve this, we constructed a one-layer, two-layer, and three-layer CDBN. In practice, since each CRBM in a CDBN is trained in isolation, the one-layer CDBN is equivalent to the

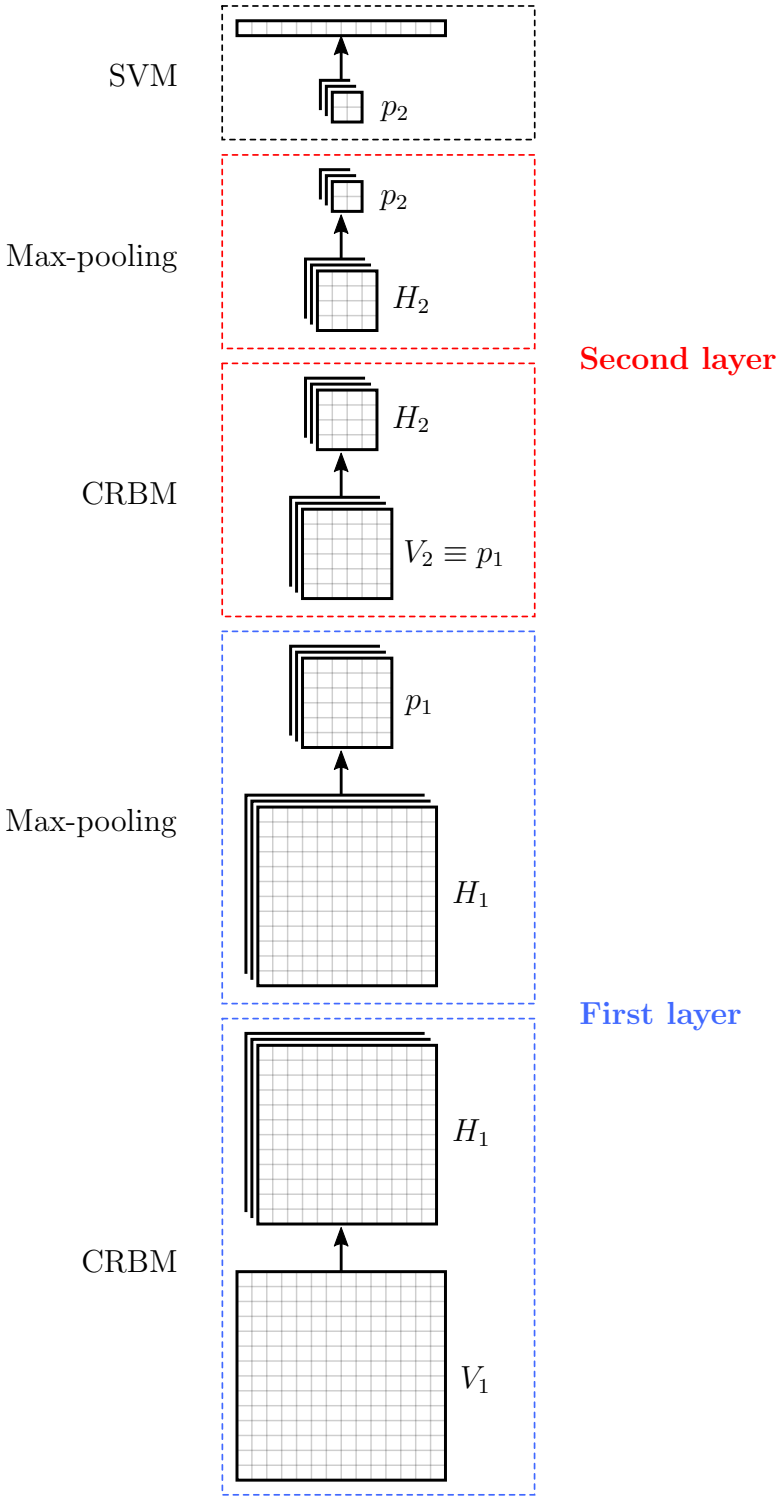


Figure 7.2: An illustration of a CDBN architecture comprising two CRBMs with max-pooling. This is known as a two-layer CDBN, and the layers are indicated by the blue and red boxes, respectively. Learning proceeds from bottom to top. Prior to using the hidden layer of the first CRBM H_1 as the input to the second CRBM V_2, max-pooling is used to reduce the size of the hidden layer bases by a constant factor C ($C = 2$ in this illustration). For classification tasks, an SVM is trained using the values of the final max-pooling layer (p_2 in this illustration). These values are referred to as *max-pooled activations*.

first layer (i.e. the first CRBM + max-pooling) of the three-layer CDBN. Similarly, the two-layer CDBN is equivalent to the first two layers of the three-layer CDBN. Therefore, in Section 7.4, the filters learnt by the one- and two- layer CDBNs are identical to the filters learnt by the first and second layers of the three-layer CDBN.

Classification experiments require a training data set and a test data set. Therefore, we split the ^{18}F-florbetapir data into two sets. The training data set comprised 40 amyloid positive and 40 amyloid negative scans (as determined via visual assessment) randomly selected from the set of 264 images. The remaining 184 scans formed the test data set.

Due to the very large number of convolution operations during the iterative CDBN training process (see Equations 7.4 and 7.5), the training is computationally expensive and very slow. Therefore, to reduce the computational burden of training with 3D amyloid PET volumes, we trained the CDBNs using a single 2D axial slice from the spatially normalised PET volume of each subject in the training set. The axial slice was selected to contain brain regions where amyloid tracer uptake should be assessed during the visual interpretation of an amyloid PET scan, such as the lateral temporal lobe and frontal lobe [Neuraceq 2014]. An example of the 2D axial slice from an amyloid negative and amyloid positive scan is shown in the top row of Figure 7.3. The corresponding 2D axial slice from the amyloid PET volumes of the test subjects were used to test the CDBNs.

Since CDBNs have not yet been explored in the context of amyloid PET analysis, the hyper-parameters used to train the CDBNs in this chapter were adopted from [Lee et al. 2009a], with the exception of the learning rate ϵ:

Layer 1	$\epsilon = 0.02$,	$K = 24$,	$N_W = 10$px,	$C = 2$,	1200 iterations
Layer 2	$\epsilon = 0.0001$,	$K = 40$,	$N_W = 10$px,	$C = 2$,	700 iterations
Layer 3	$\epsilon = 0.00001$,	$K = 24$,	$N_W = 14$px,	$C = 2$,	400 iterations

The learning rate ϵ for each layer was determined heuristically, such that unstable oscillations during the gradient descent optimisation process were avoided.

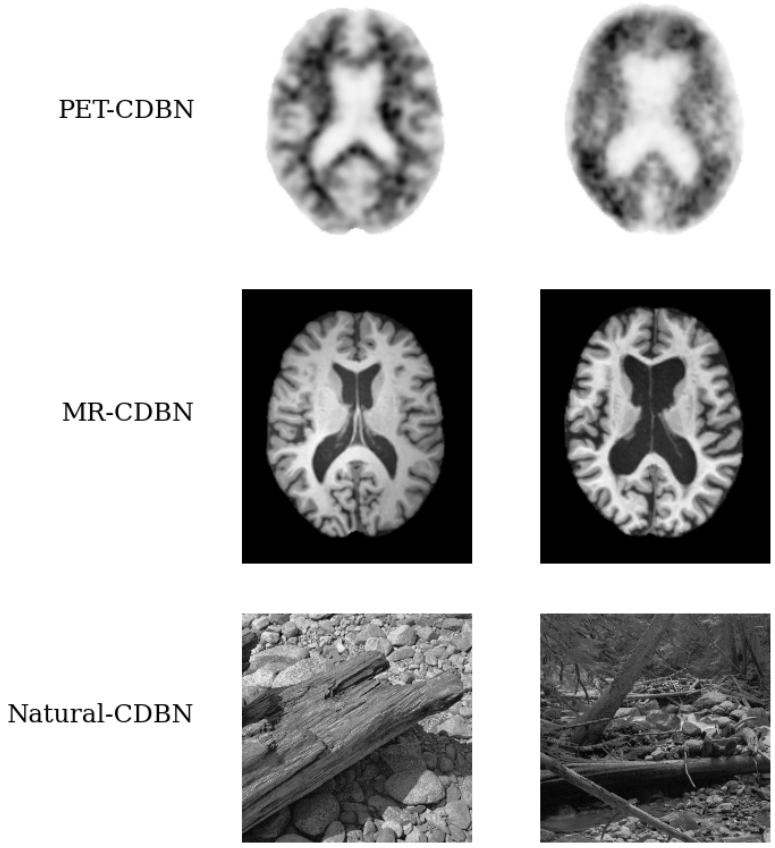


Figure 7.3: Example training data for the first layer of the PET-CDBNs, MR-CDBNs and natural-CDBNs, respectively. Top row: 2D axial slices from ^{18}F-florbetapir PET images (left: amyloid negative, right: amyloid positive). Middle row: 2D axial slices from the MR images corresponding to the same subject and anatomical location as the amyloid PET images in the top row. Bottom row: Photographs of natural scenes (as used in [Olshausen and Field 1996]). The natural images are much larger than the PET and MR images, but have been reduced in this figure. To reduce the computational cost of training with 3D volumes, the CDBNs in this chapter were trained and tested using 2D data. The higher layers of each CDBN were trained using the 2D ^{18}F-florbetapir PET training images.

7.3.2 Classification of CDBN Activations

Once the CDBNs had been trained, we trained an SVM classifier for each CDBN (one-layer, two-layer, and three-layer) using the max-pooled activations from the final layer. This is illustrated for a two-layer CDBN in Figure 7.2. The SVMs were trained with 80 sets of max-pooled activations, one for each training subject. Similarly to the SVM used in our 3D HOG + SVM method proposed in Chapter 5, we used an SVM with a Gaussian radial basis function (RBF).

The SVM hyper-parameters, i.e. the slackness variable and the free parameter of the RBF, were optimised on the PET training data using a grid search with stratified ten-fold cross-validation. The test data were classified as amyloid positive or amyloid negative using an SVM trained on the training data, with the optimal hyper-parameters. Following testing, the classification accuracy, sensitivity, and specificity were calculated. By adjusting the SVM decision boundary, receiver operating characteristic (ROC) analysis was also performed on the test data set.

7.3.3 Training CDBNs Using Non-PET Data

In addition to assessing learning depth, we also investigated the potential of using non-PET data to train the first layer of a CDBN designed to classify brain amyloid status. This technique is consistent with [Lee et al. 2009a], where photographs of natural scenery were used to train the first layer of a CDBN designed to detect human faces. A similar approach was also used in [Bar et al. 2015], where a convolutional neural network (CNN) that had been trained using non-medical data was used to identify different types of pathologies in chest x-rays.

7.3.3.1 MR-CDBN

To conduct this assessment, we constructed additional one-layer, two-layer, and three-layer CDBNs. The first layer of these CDBNs was trained using a single 2D axial slice from the MR volume of each subject in the training set. The MR slice corresponds to the same anatomical location as the 2D axial PET slice described in

Section 7.3.2. Examples of the 2D axial MR slice are shown in middle row of Figure 7.3. The higher layers of the two- and three-layer CDBNs were trained using the 2D PET data detailed in Section 7.3.2. To achieve this, the input to the second layer of the CDBN consisted of the max-pooled activations that were computed when the 2D PET images were applied to the first layer of the CDBN that had been trained using MR images. The hyper-parameters used to train each of these CDBNs were identical to those used in Section 7.3.1.

Analogously to 7.3.2, for each CDBN, an SVM was trained using the max-pooled activations from the final layer. However, the max-pooled activations used to train the SVMs were obtained using the 2D amyloid PET training data only, not the 2D MR training data. The SVMs were also tested using the max-pooled activations corresponding to the 2D PET images of the 184 test subjects.

To avoid confusion between the CDBNs described in Section 7.3.1 and the CDBNs described in this section, we will subsequently refer to them as the PET-CDBNs and MR-CDBNs, respectively, where the prefix denotes the data used to train the first layer.

7.3.3.2 Natural-CDBN

By training and testing the PET- and MR-CDBNs using 2D data, rather than 3D volumes, it is possible to train the first layer of another CDBN using readily available, non-medical 2D images, such as photographs. Therefore, in addition to assessing CDBNs where the first layer was trained using MR images, we also assessed CDBNs where the first layer was trained using non-medical data. To achieve this, we constructed a further three CDBNs (henceforth called natural-CDBNs), and trained the first layer of each CDBN using the natural scenery images[2] used in [Olshausen and Field 1996]. Example training data are shown in the bottom row of Figure 7.3. To increase the number of natural images in the training set, patches from each image were used, rather than the entire image. Identically to the MR-CDBNs, the

[2]http://redwood.berkeley.edu/bruno/sparsenet/

Table 7.1: A summary of the training data used for each layer (1,2,3) of the nine CDBNs. All nine CDBNs were tested using 2D PET slices only.

	One-layer	Two-layer	Three-layer
PET-CDBN	1: 2D PET slices	1: 2D PET slices 2: 2D PET slices	1: 2D PET slices 2: 2D PET slices 3: 2D PET slices
MR-CDBN	1: 2D MR slices	1: 2D MR slices 2: 2D PET slices	1: 2D MR slices 2: 2D PET slices 3: 2D PET slices
Natural-CDBN	1: Natural images	1: Natural images 2: 2D PET slices	1: Natural images 2: 2D PET slices 3: 2D PET slices

higher layers of the two- and three- layer natural-CDBNs were trained using the 2D PET images described in Section 7.3.2. The SVMs corresponding to each natural-CDBN were trained and tested in the same manner as the SVMs corresponding to the three MR-CDBNs.

In total, we compared nine different CDBNS: one-, two-, and three-layer PET-CDBNs, one-, two-, and three-layer MR-CDBNs, and one-, two-, and three-layer natural-CDBNs. A summary of the training data used for each layer of the nine CDBNs is presented in Table 7.1

7.3.4 Non-Deep Learning Classification Methods

In order to assess the ability of the CDBNs to classify brain amyloid status, they were compared to four other classification methods: SUVR, the intensity-based SVM method proposed by [Vandenberghe et al. 2013], a 2D equivalent of the the intensity-based SVM method, and the 3D HOG + SVM method from Chapter 5. The four methods are briefly reviewed in Sections 7.3.4.1-7.3.4.3

7.3.4.1 Standardised Uptake Value Ratio

The SUVR for each ^{18}F-florbetapir PET volume was computed using the same method as Chapter 5 (Siemens syngo.PET Amyloid Plaque software). Once again, the threshold for amyloid positivity was composite SUVR $>$ 1.12 [Hutton et al.

2015].

7.3.4.2 Intensity-Based Classification in 3D and 2D

In addition to SUVR, the CDBNs were compared to a straightforward classification technique based only on the voxel intensities of the PET brain volumes [Vandenberghe et al. 2013]. This method, known as the intensity-based SVM, is detailed in Chapter 5.2.3. In this chapter, the voxel intensities from the amyloid PET volumes of the 80 training subjects were used to train an SVM to classify subjects as amyloid positive or amyloid negative. Similarly to Section 7.3.2, the SVM hyper-parameters were optimised using stratified ten-fold cross-validation on the training data. The decision boundary was then used to classify the remaining 184 subjects excluded from training, and ROC analysis was performed on the classification results.

The classification method proposed by [Vandenberghe et al. 2013] uses 3D volumes, and not 2D images. Since the CDBNs were trained and tested using 2D images, it is reasonable to compare their classification performance to another machine learning method based on 2D axial PET slices. Therefore, we adapted the intensity-based SVM method to work with 2D images. In a similar manner to [Vandenberghe et al. 2013], only pixels inside the brain were used to train and test the intensity-based SVM. To achieve this, a brain mask was constructed from the linear MNI152 T1-weighted MR template [Mazziotta et al. 2001]. The 3D mask was dilated by 2mm, and the axial slice at the same location as the 2D axial PET slices used to train and test the CDBNs (see Section 7.3.2) was used as the final 2D brain mask. The pixels from the amyloid PET slices that were inside the 2D brain mask were reshaped into 1D arrays. Then, the 80 1D arrays corresponding to the 80 training subjects were used as training data for an SVM. Prior to using the SVM, all of the arrays were normalised to have zero mean and unit variance. As before, the SVM hyper-parameters were optimised using stratified ten-fold cross-validation, and the decision boundary was used to classify the remaining 184 subjects as amyloid positive or amyloid negative.

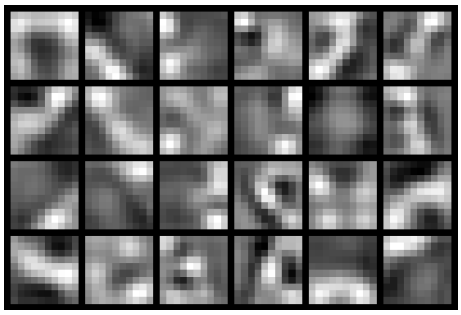
7.3.4.3 Histograms of Oriented 3D Gradients

The CDBNs were also compared to the 3D HOG + SVM classification method, as proposed in Chapter 5. An SVM was trained using the 3D HOG features derived from the 80 ^{18}F-florbetapir volumes that comprised the training set. Following optimisation of the SVM hyper-parameters (again, using stratified ten-fold cross-validation), the SVM was used to classify the 184 subjects excluded from training, based on their 3D HOG features. The 3D HOG parameters used in this work were the same as the parameters used in Chapter 5: cell size $k = 16$ voxels and number of sub-blocks $S = 1$, with an icosahedron as the platonic solid, and a half-orientation histogram.

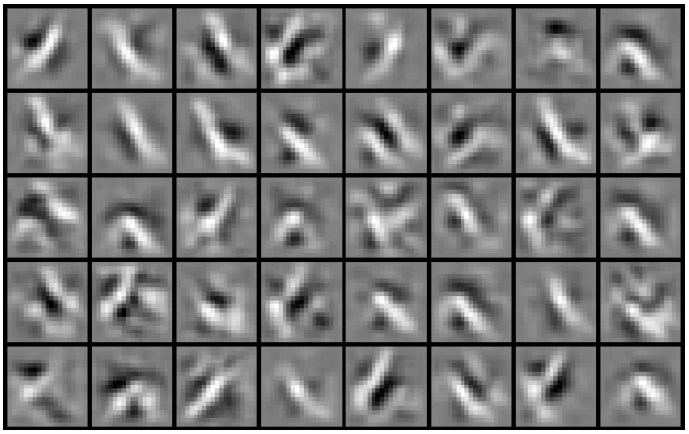
7.4 Results

7.4.1 CDBN Filters

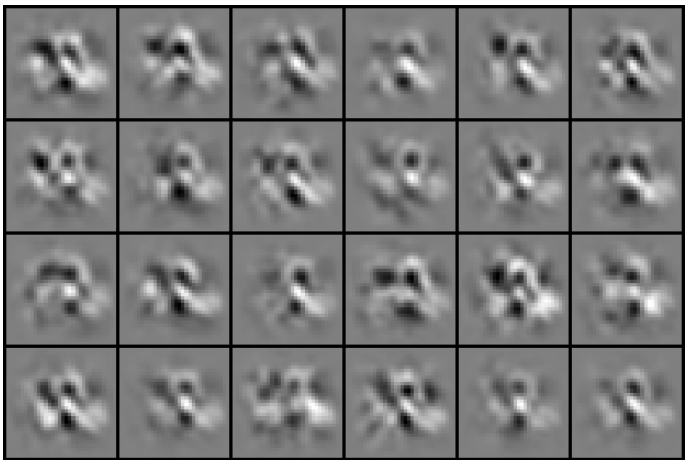
Figures 7.4-7.6 present the filters learnt by the final layer of each CDBN. The filters learnt by the one-layer PET-CDBN are less sharp than the oriented, localised edge filters learnt by the one-layer MR-CDBN and one-layer natural-CDBN. The filters learnt by all three two-layer CDBNs appear to correspond to small-scale brain features such as parts of ventricles, and the filters from the three-layer natural-CDBN resemble large-scale brain features such as whole ventricles. For example, the dark patterns in some of the filters in Figure 7.6(c) have the same general shape as the brain ventricles in an amyloid PET image. An example is presented in Figure 7.7, where the ventricle shape in a third-layer filter is highlighted and compared to the ventricles in a ^{18}F-florbetapir image. The hierarchical (local to global) feature representation demonstrated in Figures 7.4-7.6 is consistent with the deep learning literature [Lee et al. 2009a; Guo et al. 2014].



(a) One-layer PET-CDBN

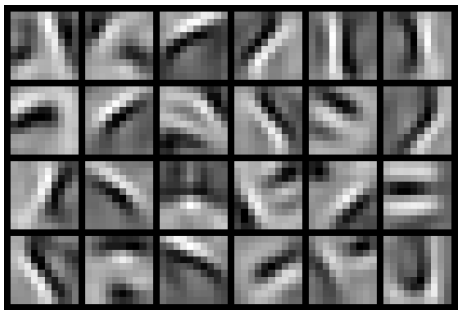


(b) Two-layer PET-CDBN

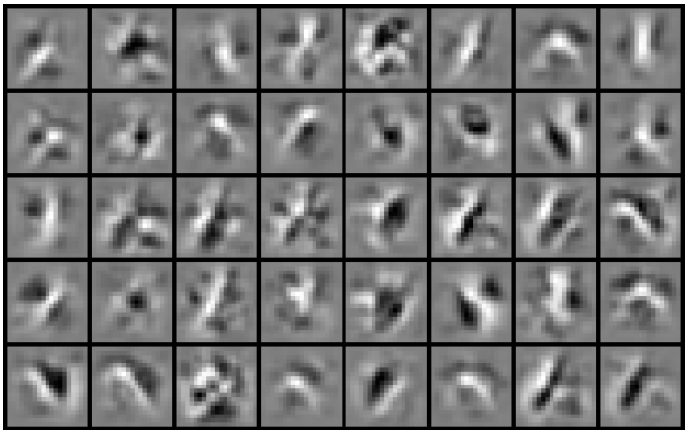


(c) Three-layer PET-CDBN

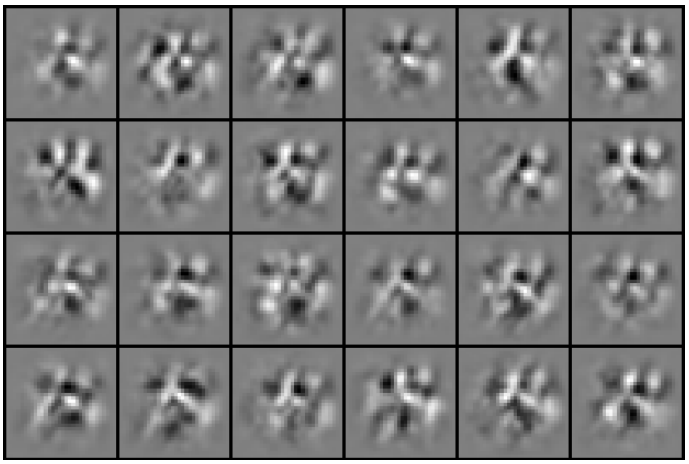
Figure 7.4: The filters learnt by the final layer of the (a) one-layer PET-CDBN, (b) two-layer PET-CDBN, and (c) three-layer PET-CDBN. Since each CRBM in a CDBN is trained in isolation, the filters learnt by the one-layer CDBN are equivalent to the filters learnt by the first layer of the two- and three-layer CDBNs. Similarly, the filters learnt by the two-layer CDBN are equivalent to the filters learnt by the second layer of the three-layer CDBN.



(a) One-layer MR-CDBN

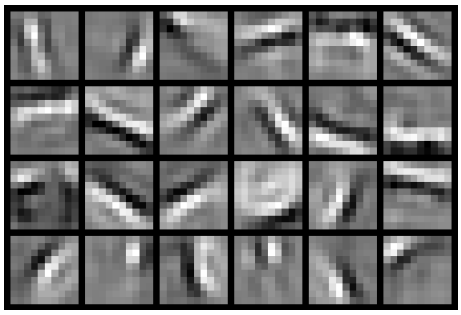


(b) Two-layer MR-CDBN

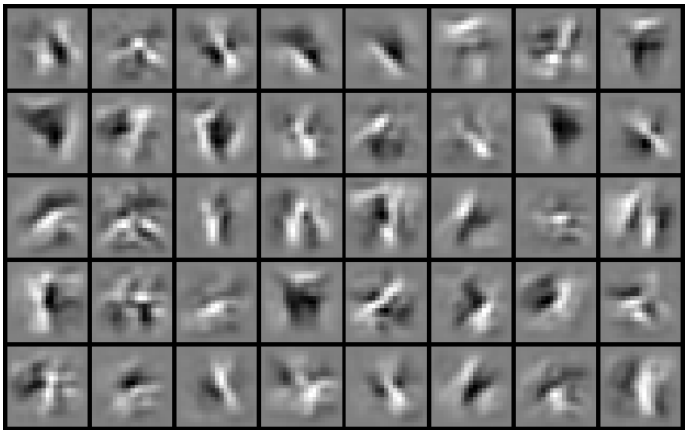


(c) Three-layer MR-CDBN

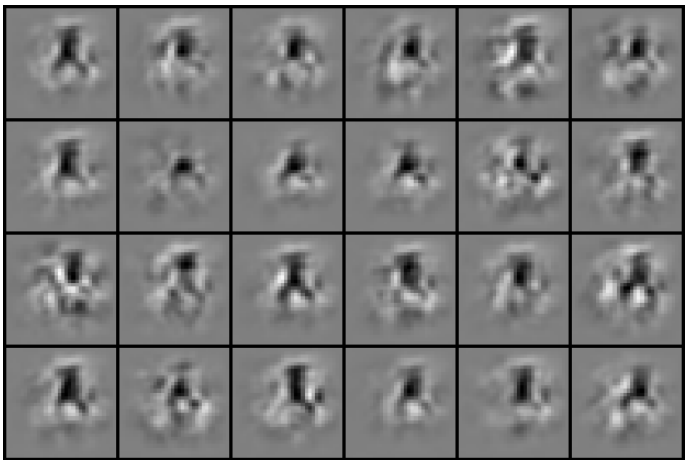
Figure 7.5: The filters learnt by the final layer of the (a) one-layer MR-CDBN, (b) two-layer MR-CDBN, and (c) three-layer MR-CDBN. Since each CRBM in a CDBN is trained in isolation, the filters learnt by the one-layer CDBN are equivalent to the filters learnt by the first layer of the two- and three-layer CDBNs. Similarly, the filters learnt by the two-layer CDBN are equivalent to the filters learnt by the second layer of the three-layer CDBN.



(a) One-layer natural-CDBN



(b) Two-layer natural-CDBN



(c) Three-layer natural-CDBN

Figure 7.6: The filters learnt by the final layer of the (a) one-layer natural-CDBN, (b) two-layer natural-CDBN, and (c) three-layer natural-CDBN. Since each CRBM in a CDBN is trained in isolation, the filters learnt by the one-layer CDBN are equivalent to the filters learnt by the first layer of the two- and three-layer CDBNs. Similarly, the filters learnt by the two-layer CDBN are equivalent to the filters learnt by the second layer of the three-layer CDBN.

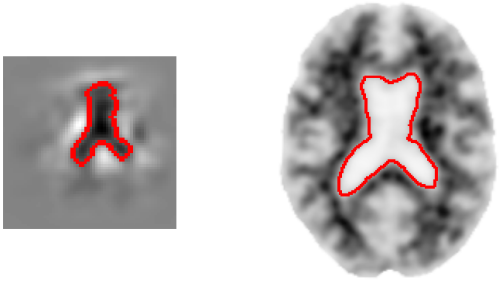


Figure 7.7: The filters learnt by the final layer of the three-layer natural-CDBN (see Figure 7.6(c)) resemble large-scale brain features. For example, the dark pattern in the filter shown on the left (enlarged for clarity) has a similar shape to the ventricles of the amyloid PET image on the right. This is highlighted by the red borders in each sub-figure.

7.4.2 Classification Results

In this chapter, we compared the ability of nine different CDBNs and four other classification methods to classify 184 ^{18}F-florbetapir PET scans as amyloid positive or amyloid negative. Figure 7.8 shows the classification accuracy, sensitivity, specificity, and area under the receiver operating curve (AUC) for each of the methods. In order to assess the effect of learning depth, Figure 7.8 presents the classification results for the one-, two-, and three-layer CDBNs for each type of training data (PET, MR, or natural images). The 3D HOG + SVM method resulted in the highest classification accuracy (94.6%), while the three-layer natural-CDBN and two-layer MR-CDBN achieved the lowest classification accuracy (78.8%). The one-layer CDBNs gave a higher classification accuracy than the corresponding two- and three-layer CDBNs, and the one-layer natural-CDBN gave the highest classification accuracy of any CDBN (91.3%). The two-layer PET-CDBN and MR-CDBN achieved a lower classification accuracy than the corresponding three-layer CDBNs (83.2% and 78.8%, compared to 83.7% and 81.0%, respectively). Conversely, the two-layer natural-CDBN resulted in a higher classification accuracy than the three-layer natural-CDBN (85.3% and 78.8%, respectively). The intensity-based SVM method using 2D axial PET slices achieved a higher classification accuracy than the identical method using the entire 3D PET volume (89.1% and 88.0%, respectively). SUVR gave the highest sensitivity (0.954), and the one-layer PET-CDBN and one-

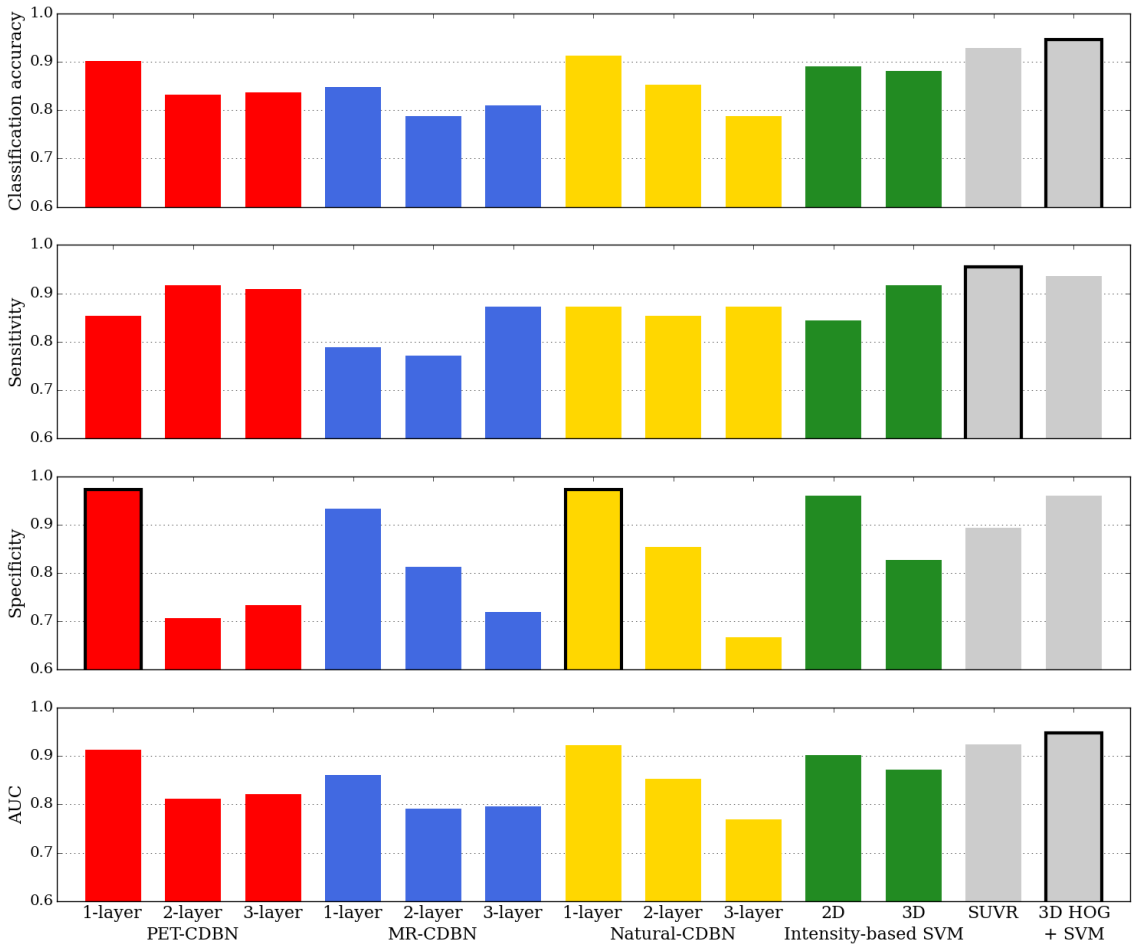


Figure 7.8: The classification accuracy, sensitivity, specificity and area under the receiver operating characteristic curve (AUC) for each classification method. The best results are highlighted with a black border.

layer natural-CDBN resulted in the highest specificity (0.973). Despite this, the 3D HOG + SVM method gave the highest AUC out of all the methods tested (0.948). In contrast, the three-layer natural-CDBN gave the lowest AUC (0.769). Using the DeLong method [DeLong et al. 1988] to statistically compare the AUCs of each classification method, the three-layer natural-CDBN achieved a significantly lower AUC than the one-layer natural-CDBN and the one-layer PET-CDBN ($p < 0.01$, corrected for multiple comparisons). The three-layer natural-CDBN also resulted in a significantly lower AUC than the 2D intensity-based SVM, SUVR, and 3D HOG + SVM methods. The two-layer PET-CDBN and the two- and three-layer MR-CDBNs also achieved a significantly lower AUC than the 3D HOG + SVM method ($p < 0.01$, corrected).

7.5 Discussion

In this chapter, we investigated the possibility of using CDBNs for amyloid status classification. We assessed the effect that the depth of the CDBN has on classification accuracy, and also evaluated how the classification results change if the first layer of the CDBN is trained on non-PET image data.

7.5.1 CDBN Classification Results

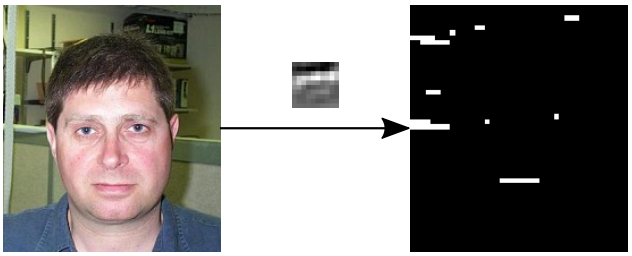
All of the CDBNs tested were able to classify ^{18}F-florbetapir PET images as amyloid positive or amyloid negative with an accuracy greater than 78%. However, the results indicate that one-layer CDBNs may be more adept at classifying amyloid PET scans than a two- or three-layer CDBN, since the one-layer CDBNs achieved a higher classification accuracy and AUC than the corresponding deeper CDBNs. Moreover, the one-layer PET-CDBN and one-layer natural-CDBN also obtained a higher classification accuracy and AUC than the 2D and 3D intensity-based SVM methods.

In contrast to the blurry edge filters learnt by the first layer of the PET-CDBNs, the first layer filters learnt by the MR-CDBNs and natural-CDBNs are sharp edge filters. This is reasonable, since the MR and natural images contain sharper edges than the amyloid PET images. Furthermore, the one-layer MR-CDBN filters in Figure 7.5(a) appear to be more curved than the one-layer natural-CDBN filters in Figure 7.6(a). This is probably due to the increased presence of strong curved edges in the MR images compared to the natural images, such as the edges of the ventricles and the brain boundary. Having filters that detect curved edges could account for the the lower classification accuracy of the one-layer MR-CDBN compared to the one-layer natural-CDBN (84.8% and 91.3%, respectively), because the filters would detect the curvature of the brain boundary, but not more discriminatory features within the brain. One reason for the relatively low sensitivity of the one-layer MR-CDBN (0.789) could be that the strong edge presented by the brain boundary was mistaken for the less distinct boundary visible between the grey and white matter in

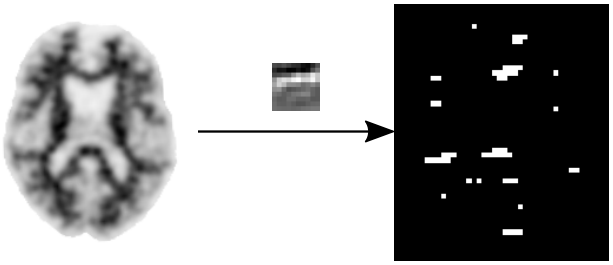
an amyloid negative scan. This may have led to fewer amyloid positive scans being correctly classified.

One reason for the relatively lower classification accuracies of the two- and three-layer CDBNs could be due to the characteristics of the amyloid PET images. Due to the lack of sharp edges in the PET data, the first layer of the PET-CDBNs learnt blurry edge filters. This could result in fewer features being detected than if sharper filters were used. Consequently, fewer hidden units would be activated (see Section 7.1.1). Since the input to the second layer of a CDBN comprises the max-pooled activations (i.e the max-pooled values of the hidden units) of the first layer, the input is sparse if few features are detected in the first layer. Therefore, the effect of the blurry edge filters would be propagated through the network. As a result, the higher layers would fail to learn useful discriminatory features. This is illustrated by the relatively poorer classification accuracy (83.7%, 81.0% and 78.8%) and AUC (0.821, 0.796, 0.769) for the all three three-layer CDBNs. Furthermore, the dominant large-scale features in ^{18}F-florbetapir PET brain images, the ventricles, are generally not discriminating features between positive and negative amyloid PET scans. Therefore, although the three-layer natural-CDBN is able to learn high-level brain features, the features miss the more subtle changes in grey matter tracer uptake which are required to correctly classify the images. This could explain why the three-layer natural-CDBN resulted in the lowest classification accuracy, specificity and AUC of all the methods tested (78.8%, 0.667 and 0.769, respectively).

The relatively poorer performance of the two- and three-layer CDBNs may also be due to the nature of the classification task in this chapter. When using a CDBN to detect faces in photographs, the first layer detects edges in the image (see Figure 2.8). Not only will edges exist within the face, but they could also be present in the background. Therefore, the higher layers are required to determine which of the detected edges could correspond to a face. There is no background information in an amyloid PET brain image, so this refinement is unnecessary, as all edges must correspond to the brain. This is demonstrated in Figure 7.9. Moreover, the



(a)



(b)

Figure 7.9: The activations of an exemplar first-layer filter when applied to (a) a photograph of a face (from the Caltech 101 data set [Fei-Fei et al. 2004]), and (b) a 2D axial slice from an amyloid PET volume. In (a), the first-layer filter detects horizontal edges within the face, such as the eyes and mouth, as well as horizontal edges in the background. There is no background in (b), so the only detected edges are within the brain. It should be noted that the filter and activations have been enlarged for clarity.

differences between amyloid positive and amyloid negative scans are more subtle than the differences between images with and without faces.

It could be argued that recognition of facial expressions is a closer computer vision task to amyloid status classification than face detection. Expression recognition requires detection of subtle changes in facial features, and moreover, the face data set used in [Ranzato et al. 2011] contains no background. However, unlike our application, the large-scale features in expression recognition (i.e. eyes, nose, mouth) are discriminatory, resulting in a greater accuracy for the higher layers of a deep learning network compared to the first layer [Ranzato et al. 2011].

7.5.2 Results of the Non-Deep Learning Methods

The 3D HOG + SVM method achieved the highest classification accuracy (94.6%) and AUC (0.948) of all the methods tested. As suggested in Chapter 5, this could be because it uses local intensity gradients as features, unlike SUVR and the intensity-

based SVM method, which use intensity directly. Therefore, the 3D HOG + SVM method is robust to spatially varying intensity levels. Since the outputs of a CDBN (i.e. the max-pooled activations of the final layer) are essentially filter responses, and not voxel intensities, the CDBNs in this chapter should also be robust to spatially varying intensity levels. However, for reasons explained in Section 7.5.1, this is not reflected in the classification results of the two- and three-layer CDBNs. Nevertheless, the CDBNs were trained and tested using 2D images, whereas the 3D HOG + SVM method used the entire PET volumes. Consequently, the CDBNs had to learn discriminating features from less information.

Although using a 2D slice discards some information in the 3D PET volume, the intensity-based SVM using 2D images achieved a higher classification accuracy than the same method using 3D volumes (89.1% and 88.0%, respectively). This could be due to the relative sizes of the feature vector and data set. The 3D brain volume produces a feature vector of 290409 elements, and the training set consists of only 80 subjects. Therefore, the classification hyperplane of the SVM is determined using very few data points in a very high-dimensional space. Although the feature vector for the 2D intensity-based SVM method comprises 5513 elements, it results in a substantially smaller dimensional space for the SVM. Consequently, the training data points are less sparsely distributed, allowing for a hyperplane that is better suited to discriminating between amyloid positive and amyloid negative subjects.

Figure 7.8 shows that SUVR resulted in the second highest classification accuracy (92.9%) and AUC (0.924). Although SUVR uses specific neuroanatomical knowledge, and the CDBNs do not, the one-layer CDBNs achieved almost equivalent results. For example, the one-layer natural-CDBN achieved a classification accuracy of 91.3% and an AUC of 0.922. This suggests that there may be more subtle differences in tracer uptake between amyloid positive and amyloid negative scans than can be detected by the SUVR target regions, and a CDBN trained using 3D volumes may be able to learn this subtle distinction.

7.5.3 Methodological Considerations

In this chapter we trained and tested the CDBNs using 2D axial slices from 3D medical image volumes. Not only did this facilitate the use of 2D natural images in the natural-CDBN training process, it also substantially reduced the computational requirements that would have been necessary to train a CDBN using 3D data. However, despite the reduction in computational burden, it took considerably longer to train the CDBNs than the intensity-based SVM or 3D HOG + SVM methods. For example, it took six hours to train the first layer of the three-layer PET-CDBN, compared to 37 seconds to obtain the features for the 2D intensity-based method and train the corresponding SVM.

Due to the long training time and the large number of variables, optimising all of the CDBN hyper-parameters is impractical. In this study we adopted the hyper-parameters from [Lee et al. 2009a], however, hyper-parameter values that have been shown to work well on 2D photographs may not be suitable for biomedical images, given both the dimension of the images and the high levels of noise. Consequently, there is no guarantee that the parameters used in this chapter are optimal for the classification task. This could also account for the relatively poorer performance of the two- and three-layer CDBNs compared to the SUVR, intensity-based SVM, and 3D HOG + SVM methods.

Many deep learning methods in the literature use extremely large sets of training data [Krizhevsky et al. 2012; Russakovsky et al. 2015]. For example, 1.2 million images were used to train the convolutional neural network (CNN) in [Krizhevsky et al. 2012]. A data set of that size is not available for our amyloid status classification task. However, only 100 training images were used by [Lee et al. 2009a] to train a CDBN to detect faces in photographs. Our amyloid PET and MR training sets consisted of 80 images, and therefore, they are a similar size to the training set used in [Lee et al. 2009a].

In the computer vision literature, CDBNs are used less frequently than CNNs [Simonyan and Zisserman 2015; Russakovsky et al. 2015]. This may be because CNNs

were developed several years before CDBNs ([LeCun et al. 1998] compared to [Desjardins and Bengio 2008]), and the success of CNNs in visual recognition tasks actually drove the development of the CRBM and CDBN [Desjardins and Bengio 2008]. Despite their relatively infrequent use in the literature, CDBNs achieved state-of-the-art classification performance on the CIFAR-10 image data set [Krizhevsky 2010]. More recently, CDBNs have been used to categorise 3D objects from 2.5D depth maps [Wu et al. 2015].

7.5.4 Conclusion

In this chapter, we assessed the ability of CDBNs to classify ^{18}F-florbetapir PET scans as amyloid positive or amyloid negative. We examined the effect of learning depth on the classification performance of the CDBNs, and we also investigated the effect of training the first layer of a CDBN using non-PET data. Our results suggest that one-layer CDBNs are able to classify amyloid PET images with a higher accuracy that two- or three-layer CDBNs. Moreover, we showed that a one-layer CDBN trained using natural images can result in a higher classification accuracy than a one-layer CDBN trained using 2D axial slices from 3D medical data. Despite this, our results indicate that our 3D HOG + SVM method can achieve a higher classification accuracy and AUC than a CDBN, and takes considerably less time to train. Therefore, we conclude that non-deep learning methods may be favoured over CDBNs for classifying brain amyloid status from PET imaging.

Given the high accuracy of the 3D HOG + SVM amyloid status classification method in Chapters 5-7, it may be difficult to improve upon the classification results. Furthermore, as discussed in Chapter 2, the classification of brain amyloid status from PET imaging is just one of several biomarkers for AD. Therefore, in the next chapter we will combine the results of our 3D HOG + SVM method with other clinical biomarkers, in order to work towards a complete classification framework for AD.

Chapter 8

Towards a Complete Classification Framework for AD

The novel classification methods proposed in this thesis have all been compared to using SUVR for amyloid status classification. Research has shown that baseline SUVR is significantly lower in subjects with stable mild cognitive impairment (SMCI) than subjects with mild cognitive impairment (MCI) that later convert to AD (also known as progressive MCI, or PMCI)[Brendel et al. 2014]. Therefore, it may be possible to extend the machine learning methods proposed in Chapters 5-7 to identify subjects at risk of developing AD.

Predicting conversion from MCI to AD is a difficult task, and classification of SMCI and PMCI using MR images has plateaued around 75% accuracy, despite researchers employing new classification techniques such as deep learning [Suk and Shen 2013]. However, it may be possible to improve the results by introducing additional subject data, such as mini-mental state exam (MMSE) scores and cerebrospinal fluid (CSF) biomarkers. As discussed in Chapter 2, brain amyloid status from PET images is just one of the biomarkers that can be used to aid the diagnosis of AD. By combining the results of several tests, one could potentially gain a more comprehensive picture of the disease.

Although research into combining multiple types of data exists in the literature

(e.g. [Mattila et al. 2011; Wolz et al. 2012; Segovia et al. 2014]), the methods use different data, so it is difficult to compare between them. Therefore, as a preliminary investigation into combining multiple data types for computer-assisted AD diagnosis, we implement three existing classification frameworks, and test their classification abilities with the same set of data. Furthermore, we also propose and assess a novel method for combining multiple non-image data types, based on a technique known as multi-graph learning[1] [Wang et al. 2009]. In summary, the multi-graph learning method simultaneously optimises graphs constructed from each data source, and returns subjects from the training data that are clinically similar to the test subject. It should be emphasised that this work is purely exploratory, but could help shape future research into the topic of combining multiple types of subject data for AD diagnosis.

The remainder of this chapter is structured as follows: Section 8.1 gives a brief overview of each of the seven types of data used in this work. Section 8.2 introduces the three existing classification frameworks from the literature, and describes our novel multi-graph classification method. Section 8.3 presents the results of the classification experiments, and finally, Section 8.4 discusses the advantages and disadvantages of each method, and proposes a set of criteria that future classification frameworks must fulfil.

8.1 Materials

The data used in this chapter correspond to the 264 ^{18}F-florbetapir PET subjects originally used in Chapter 5. However, unlike Chapters 5-7, we used the clinical diagnoses from the ADNI database as the gold standard for classification. In this work, although baseline subject data were used in the classification task, PMCI subjects were defined as individuals that converted to AD within three years of their baseline MCI diagnosis. In contrast, SMCI subjects were defined as subjects that

[1]This is not to be confused with multigraphs (without the hyphen), which are graphs that permit multiple edges between two nodes.

Table 8.1: Demographic and summary of available baseline data for healthy controls (HC), subjects with stable mild cognitive impairment (SMCI), subjects with progressive mild cognitive impairment (PMCI), and patients with Alzheimer's disease (AD). The data are expressed as counts (percentages) of data available for each diagnostic group, with the exception of age, which is expressed as mean ± standard deviation.

	HC	SMCI	PMCI	AD
# subjects	50	100	49	65
Age	77.1 ± 7.2	74.1 ± 8.6	75.1 ± 6.6	75.5 ± 8.3
MMSE	49 (98%)	100 (100%)	49 (100%)	65 (100%)
ADAS-Cog	50 (100%)	100 (100%)	49 (100%)	65 (100%)
TMT	50 (100%)	100 (100%)	49 (100%)	65 (100%)
ApoE	50 (100%)	100 (100%)	49 (100%)	65 (100%)
CSF	28 (48%)	46 (46%)	35 (71%)	22 (34%)
MR	50 (100%)	100 (100%)	49 (100%)	65 (100%)
PET	50 (100%)	100 (100%)	49 (100%)	65 (100%)

did not change clinical diagnosis within three years of the baseline MCI diagnosis. The data are broadly similar to those used by [Mattila et al. 2011], with the addition of brain amyloid status, as determined from ^{18}F-florbetapir PET images using the 3D HOG + SVM method proposed in Chapter 5. Unless otherwise stated, all data are from the ADNI database. The demographics of the data set are shown in Table 8.1, and a brief summary of each type of data is presented below:

Mini-Mental State Examination (MMSE)

The MMSE was proposed by [Folstein et al. 1975] as a quick, quantitative assessment of cognitive mental status. It comprises 11 questions, and the maximum total score is 30. Typically, patients with dementia achieve a lower score than healthy controls (mean score of 9.7 and 27.6 in [Folstein et al. 1975], respectively). Analogously to [Mattila et al. 2011], we used the score from each MMSE question separately, rather than the total MMSE score.

Alzheimer's Disease Assessment Scale – Cognition (ADAS-Cog)

The original ADAS-Cog, proposed by [Rosen et al. 1984], included 11 items to quantitatively assess cognitive function. This was expanded to a 13-item scale by [Mohs et al. 1997]. The maximum score is 85, and a higher score indicates

greater cognitive impairment [Skinner et al. 2012]. Similarly to the MMSE data, the composite scores were excluded from the data set in favour of the scores from each individual ADAS-Cog item.

Trail Making Test (TMT)

The TMT is a neuropsychological test that provides information on visual search and speed of processing. Firstly, the individual must draw lines sequentially connecting 25 numbered dots randomly scattered on a sheet of paper. The second phase of the test is is identical to the first, except that the person must alternate between numbers and letters (e.g. 1, A, 2, B, 3, C, etc.) [Tombaugh 2004]. In this work, we used the time taken to complete the tasks, and the number of errors as separate variables.

Apolipoprotein E (ApoE) genotype

The polymorphic ApoE gene is closely associated with the onset of AD [Levy-Lahad and Bird 1996]. Six different genotypes exist, each comprising two alleles ($\varepsilon 2/\varepsilon 2$, $\varepsilon 3/\varepsilon 3$, $\varepsilon 4/\varepsilon 4$, $\varepsilon 3/\varepsilon 2$, $\varepsilon 4/\varepsilon 2$, $\varepsilon 4/\varepsilon 3$) [Liu et al. 2014]. Individuals with a single copy of the $\varepsilon 4$ allele have a 5-fold increased chance of developing AD and individuals with two copies of the $\varepsilon 4$ allele have a 20-fold increased risk of AD [Strittmatter 2012].

Cerebrospinal fluid (CSF) biomarkers

As discussed in Chapter 2.1.1, biochemical changes in the brain are reflected in the CSF. In this chapter, we used the levels of total tau, phosphorylated tau, and amyloid-β in the CSF as biomarkers. Elevated levels of total tau and phosphorylated tau have been reported in AD patients compared to healthy controls [Blennow et al. 2001]. In contrast, a lower level of CSF amyloid-β is found in AD patients than cognitively normal controls [Blennow et al. 2001].

MR-derived volumes

In [Mattila et al. 2011], the authors use the volumes of 13 brain regions as inputs to their AD classification framework. The paper reveals that the vol-

umetric segmentation of MR images was performed using Freesurfer [Fischl et al. 2002], however the specific brain regions are not mentioned. Therefore, in this work, we derived MR volumes for the regions used by [Schmitter et al. 2015]: left hippocampus, right hippocampus, left temporal grey matter, right temporal grey matter, total grey matter, and total CSF. These regions have been shown to be able to classify SMCI versus PMCI with an accuracy greater than 70% [Schmitter et al. 2015]. To obtain the hippocampal and temporal lobe volumes, the AAL brain atlas [Tzourio-Mazoyer et al. 2002] was registered to the 264 MR volumes (in native space) using SPM8 [Ashburner and Friston 2005]. The registered atlas was used to extract the volumes from the MR images. Grey matter and CSF tissue segmentations (also obtained using SPM8) were used to mask the tissues as required. All volumes were normalised by the intracranial volume prior to the classification tasks.

Amyloid PET status

Using the 3D HOG + SVM classification proposed in Chapter 5, the ^{18}F-florbetapir PET volume for each subject was automatically designated as amyloid positive or amyloid negative. The 3D HOG parameters used in this work were identical to the optimum parameters determined in Chapter 5: cell size $k = 16$ voxels, number of sub-blocks $S = 1$, icosahedron, half-orientation histogram. Each subject was classified using a leave-one-out testing approach.

As specified in the ADNI procedures manual[2], the MMSE is used to aid the baseline diagnosis of the subjects in the ADNI database. Healthy controls should have a total MMSE score between 24 and 30 (inclusive), and patients with probable AD should have a total score between 20 and 26 (inclusive). Consequently, the MMSE data should be able to accurately classify subjects as either healthy controls or AD patients. Despite being used to aid the clinical diagnosis of the subjects used in this chapter, we chose to include MMSE in our classification experiments. By

[2] `https://adni.loni.usc.edu/wp-content/uploads/2008/07/adni2-procedures-manual.pdf`

including a data source that should be able to accurately classify subjects according to their clinical diagnosis, we were able to assess the ability of the classification methods to distinguish between relevant and irrelevant data sources. Furthermore, the ADNI inclusion criteria state that subjects with MCI should have a total MMSE score in the same range as the healthy controls (between 24 and 30, inclusive). Therefore, the MMSE score was not used by ADNI to distinguish between healthy controls, subjects with SMCI, and subjects with PMCI.

8.2 Methods

In this work we compare four classification methods capable of combining multiple types of data: an SVM-based method proposed by [Segovia et al. 2014], a manifold learning technique proposed by [Wolz et al. 2012], the Disease State Index (DSI) proposed by [Mattila et al. 2011], and our novel method based on multi-graph learning. The four methods are described in Sections 8.2.1-8.2.4.

8.2.1 SVM-Based Classification

In [Segovia et al. 2014], the authors compared three SVM-based methods for combining PET images and neuropsychological test data for automated diagnosis of AD. The difference between the methods was the point in the pipeline at which the dimensionality-reduced PET images and neuropsychological test data were integrated. For early integration, the non-image data were combined with PET volumes in a single feature vector prior to classification with an SVM. In the intermediate integration method, the combination was performed inside the classifier by using two kernel matrices: one for the PET images, and one for the neuropsychological test data. Finally, the late integration method employed an individual classifier for each data source, and combined all the outputs to predict the final clinical diagnosis. More specifically, two predicted classes c_{NP} and c_{PET} were obtained for each subject using the neuropsychological (NP) test scores and PET image features,

respectively. The distances to the classification boundary d_{NP} and d_{PET} were also calculated, and the final class c_x (where $x \in \{\mathrm{NP}, \mathrm{PET}\}$) was selected according to $d_x = \max(d_{\mathrm{NP}}, d_{\mathrm{PET}})$.

Out of the three SVM-based classification methods presented in [Segovia et al. 2014], the late integration method achieved the highest accuracy (89.13%) for classification of MCI versus AD. Therefore, we adopted the late integration approach for our comparison of classification methods capable of combining multiple data types. To account for the increased number of data sources in this study compared to [Segovia et al. 2014], we extended the method to combine the outputs of more than two SVMs. In summary, the final predicted class c_x was selected according to $d_x = \max(d_{\mathrm{MMSE}}, d_{\mathrm{ADAS}}, d_{\mathrm{TMT}}, \ldots)$, where the subscripts denote the data sources described in Section 8.1.

Leave-one-out testing was used to assess the performance of the late integration SVM classification method. For each fold, the SVM was trained using all of the subjects except one. The remaining subject was then used as the test subject. This process was repeated until all of the subjects had been used as the test subject. The SVM slackness parameter C was optimised at each fold of the testing using stratified 10-fold cross-validation on the training data. For the multiclass classification tasks (e.g. HC vs SMCI vs PMCI vs AD), a one-vs-one scheme was employed. This method trains a classifier for every class versus every other class, resulting in $N(N-1)/2$ separate SVMs, where N is the number of classes. All of the classifiers are applied to the test data, and the class that gets the highest number of predictions is selected as the final predicted class.

8.2.2 Manifold Learning

Manifold learning is an approach to non-linear dimensionality reduction that has been used to successfully encode clinically relevant differences between MR volumes of MCI and AD patients [Wolz et al. 2010]. Given a set of N images $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\} \in \mathbb{R}^n$ that lie on or near an l-dimensional manifold, a low-dimensional

representation of the images $\mathbf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\} \in \mathbb{R}^l$ can be learnt, where $l \ll n$. In [Wolz et al. 2010], Laplacian Eigenmaps (LE) are used to achieve the non-linear dimensionality reduction $f : \mathbf{X} \rightarrow \mathbf{Y},\ \mathbf{y}_i = f(\mathbf{x}_i)$. A graph G with N nodes representing the images is defined on $\mathbf{X}$, and the weights of the edges between the nodes are defined by pairwise image similarities W_{ij}. The low-dimensional representation can be obtained by minimising the objective function:

$$\sum_{ij} \|\mathbf{y}_i - \mathbf{y}_j\|^2 W_{ij} \tag{8.1}$$

This problem can be formulated as the generalised eigenvector problem $\mathbf{L}\nu = \lambda\mathbf{D}\nu$, where the graph Laplacian $\mathbf{L} = \mathbf{D} - \mathbf{W}$ is defined by the weights matrix $\mathbf{W}$ and the diagonal degree matrix $D_{ii} = \sum_j W_{ij}$.

The LE method was extended by [Wolz et al. 2012] to learn a manifold defined by pairwise image similarities and clinical metadata available for each of the subjects. In traditional LE, neighbouring subjects are connected to force similar images to be closer in the manifold. In the method proposed by [Wolz et al. 2012], additional edges are defined that connect every subject to an additional node representing its metadata value. This forces subjects with the same metadata value to be closer in the manifold. An illustration of the method is presented in Figure 8.1. The classic LE graph is shown in black, and the additional nodes representing the two values of a single metadata variable z are shown in red and blue. For a discrete metadata variable z with M possible values (e.g. sex or genotype), M additional nodes are added to the graph. The weight $\hat{W}_{im}$ between subject i and the mth additional node is given by:

$$\hat{W}_{im} = \begin{cases} 1 & \text{if } z_i = z^m \\ 0 & \text{otherwise} \end{cases} \tag{8.2}$$

where $z_i \in \{z^1, \ldots, z^M\}$.

In the case of a continuous metadata variable $z_i \in [z^a, z^b]$, a set of M additional

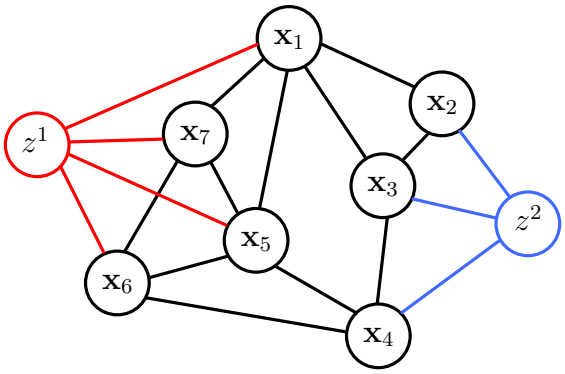


Figure 8.1: An illustration of the extended LE graph proposed by [Wolz et al. 2012]. In traditional LE, neighbouring subjects $\mathbf{x}_i$ are connected to force similar images to be close in the manifold (black nodes and edges). In the extended method, additional edges are defined that connect every subject to an additional node representing its metadata value. In this illustration with a binary metadata variable z, subjects with a metadata value z^1 are connected to the z^1 node by the red edges. Subjects with a metadata value z^2 are connected to the z^2 node by the blue edges.

nodes are defined, each representing an interval within the variable space $\bar{z}^m \in [z^{a,m}, z^{b,m}],\ m = 1, \ldots, M$. The bounds of each interval are defined using percentiles, so that each $\bar{z}^m$ has equal probability:

$$\begin{aligned} z^{a,m} &= P_z\left((m-1)\frac{100}{M}\right) \\ z^{b,m} &= P_z\left(m\frac{100}{M}\right) \end{aligned} \tag{8.3}$$

where $P_z(x)$ gives the xth percentile of interval z.

The weight $\hat{W}_{im}$ between subject i and the mth additional node is defined:

$$\hat{W}_{im} = \begin{cases} \alpha(1 + (z_i - \mu^m)^2)^{-1} & \text{if } z_i \text{ is available} \\ 0 & \text{otherwise} \end{cases} \tag{8.4}$$

where μ^m is the mean value of interval $\bar{z}^m$, and α is a constant to ensure that $\sum_m \hat{W}_{im} = 1$.

By incorporating metadata into the graph, [Wolz et al. 2012] extended the LE objective function in Equation 8.1:

$$\gamma \sum_{ij} \|\mathbf{y}_i - \mathbf{y}_j\|^2 W_{ij} + \sum_{im} \|\mathbf{y}_i - \hat{\mathbf{y}}_m\|^2 \hat{W}_{im} \tag{8.5}$$

where γ controls the influence of the classic image-based LE, and $\hat{\mathbf{y}}_m$ is the cluster centre of value z^m or interval $\bar{z}^m$.

To optimise the extended objective function by solving the generalised eigenvector problem, the weights matrix is also extended to include $\hat{\mathbf{W}}$, which defines the weights between subject i and the M additional nodes:

$$\mathbf{W}' = \begin{pmatrix} \mathbf{I} & \frac{1}{2}\hat{\mathbf{W}}^T \\ \frac{1}{2}\hat{\mathbf{W}} & \gamma\mathbf{W} \end{pmatrix} \tag{8.6}$$

Once the extended low-dimensional embedding has been obtained, classification between the different clinical subject groups can be performed. In [Wolz et al. 2012], the classification of unlabelled subjects was conducted using a linear SVM trained on the manifold representations of the subjects whose clinical diagnosis was known. Classification was performed using three different low-dimensions $l \in \{6, 7, 8\}$, and the modal classification decision was used as the final predicted class. We adopted the same approach in this work, and used leave-one-out testing to classify each subject. Similarly to the late integration SVM method, stratified 10-fold cross-validation on the training data was used to optimise the slackness parameter C of the SVM. Again, a one-vs-one approach was employed for multiclass classification tasks.

To reduce bias in our experiments, all of the classification methods being tested used exactly the same data. Therefore, we set $\gamma = 0$, such that the pairwise similarities of the MR images were excluded from the manifold learning method.

8.2.3 Disease State Index

The Disease State Index (DSI) method, proposed by [Mattila et al. 2011], derives a scalar value that denotes the AD state of a subject by measuring the similarity of subject data to previously diagnosed populations of AD patients and healthy controls. Firstly, individual patient measurement values (e.g. a single answer in the MMSE) are compared to known training data using a *fitness* function. Given

a random variable x from a distribution containing only healthy controls and AD patients, the probability density function $f(x)$ can be split into components $f_{\mathrm{HC}}(x)$ and $f_{\mathrm{AD}}(x)$, respectively:

$$\begin{aligned} f(x) &= f_{\mathrm{HC}}(x) + f_{\mathrm{AD}}(x) \\ &= f_{\mathrm{HC}}(x|\mathrm{HC})p(\mathrm{HC}) + f_{\mathrm{AD}}(x|\mathrm{AD})p(\mathrm{AD}) \end{aligned} \tag{8.7}$$

where $f_{\mathrm{HC}}(x|\mathrm{HC})$ and $f_{\mathrm{AD}}(x|\mathrm{AD})$ are the marginal distributions of healthy controls and AD patients, and $p(\mathrm{HC})$ and $p(\mathrm{AD})$ are the probabilities of being a healthy control or AD patient in the study population.

Given an actual observation a of the random variable, the left and right integrals for f_{HC} and f_{AD} can be defined (see Figure 8.2 for an illustration of the integrals):

$$\begin{aligned} L_{\mathrm{AD}}(a) &= \int_{-\infty}^{a} f_{\mathrm{AD}}(x)\,\mathrm{d}x \\ R_{\mathrm{HC}}(a) &= \int_{a}^{\infty} f_{\mathrm{HC}}(x)\,\mathrm{d}x \end{aligned} \tag{8.8}$$

The conditional probability $p(\mathrm{AD}|x)$ could be used to provide a measure of AD state for a given subject, however, as illustrated in Figure 8.2, data with wide distribution tails can cause the conditional probability to change unexpectedly. For example, if the data in Figure 8.2 represent distributions of total ADAS-Cog scores, the plot of conditional probability suggests that a subject with a maximum ADAS-Cog score has less chance of having AD than a subject with a slightly lower ADAS-Cog score. In reality, we know this is incorrect, and AD patients are more likely to have a higher ADAS-Cog Score. As a result, [Mattila et al. 2011] proposed a monotonically increasing fitness function[3]:

$$\mathrm{Fit}(a) = \frac{L_{\mathrm{AD}}(a)}{L_{\mathrm{AD}}(a) + R_{\mathrm{HC}}(a)\frac{p(\mathrm{AD})}{p(\mathrm{HC})}} \tag{8.9}$$

To determine the ability of a variable to classify subjects as healthy controls or AD patients, [Mattila et al. 2011] compute its *relevance* using values from known

[3]If the value of x is expected to be lower for AD patients than healthy controls, an analogous monotonically decreasing fitness function can be formulated.

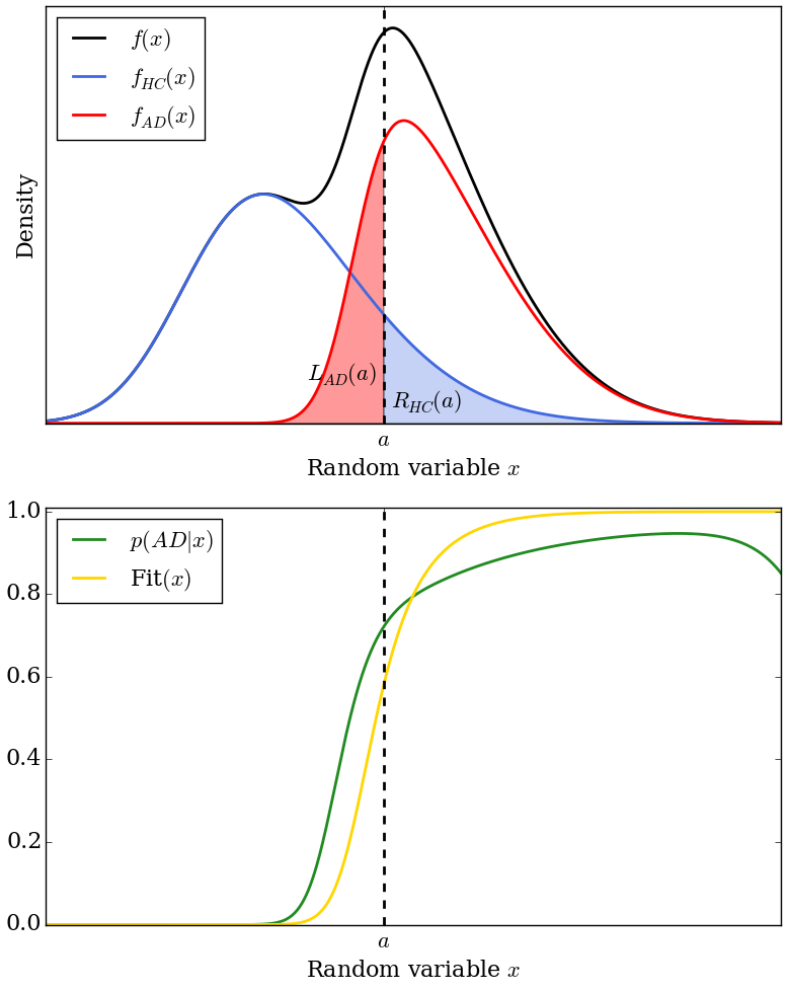


Figure 8.2: Top: An illustration of the probability density function $f(x)$ and its components $f_{\mathrm{HC}}(x)$ and $f_{\mathrm{AD}}(x)$ for the populations of healthy controls and AD patients, respectively. The left and right integrals defined in Equation 8.8 are highlighted by the red and blue shaded areas. Bottom: The conditional probability $p(\mathrm{AD}|x)$ and fitness function $\mathrm{Fit}(x)$ for the same random variable x as the top sub-figure. The large skew of $f_{\mathrm{AD}}(x)$ causes the conditional probability to suddenly decrease when x is large. In contrast, the fitness function is monotonically increasing.

healthy and AD populations only. The classification accuracy of the ith variable is estimated by applying the fitness function to the training data:

$$\mathrm{Acc}(i) = \frac{|\mathrm{HC}_T : \mathrm{Fit}(a_i) \leq 0.5| + |\mathrm{AD}_T : \mathrm{Fit}(a_i) > 0.5|}{|\mathrm{HC}_T| + |\mathrm{AD}_T|} \tag{8.10}$$

where HC_T and AD_T are the training data for the healthy controls and AD patients, respectively. The relevance of the ith variable is then defined:

$$\mathrm{Rel}(i) = \max\{0, 2(\mathrm{Acc}(i) - 0.5)\} \tag{8.11}$$

Finally, the fitness and relevance values are combined to form the DSI:

$$\mathrm{DSI}(a_1, \ldots, a_n) = \frac{\sum_i \mathrm{Rel}(i)\mathrm{Fit}(i)}{\sum_i \mathrm{Rel}(i)} \tag{8.12}$$

where a_i are the values of each variable (e.g. the individual MMSE questions), and $\mathrm{DSI}(a_1, \ldots, a_n) \in [0, 1]$. By using the DSI values of each individual data source (e.g. MMSE, ADAS-Cog, TMT) as input data, a single DSI value for a subject can be computed. Although DSI is not the probability of having AD, increasing values of DSI indicate an increasing similarity to the AD population, based on available data.

In this work, a single DSI value was computed for each SMCI and PMCI subject by training the method with all the HC and AD subjects. For each HC and AD subject, a single DSI value was obtained by training the method with all the remaining HC and AD subjects. Identically to [Mattila et al. 2011], the SMCI and PMCI subjects were never used for training. Classification was performed by comparing DSI values to interval-level diagnoses (HC = 0, SMCI = 1/3, PMCI = 2/3, AD = 1). Subjects were assigned to the class with the closest interval value to their DSI.

8.2.4 Multi-Graph Learning

Multi-graph learning was originally proposed as a method to annotate videos [Wang et al. 2009], but more recently it has been used to retrieve MR images from a medical image database to aid the diagnosis of MCI [Gao et al. 2015]. We adopt a similar approach, but apply the method to non-image data. In this work, a subject is classified as HC/SMCI/PMCI/AD based on the known clinical diagnoses of similar subjects within the training data set. The relevance of each training subject to the test subject is determined by the multi-graph framework.

Suppose we have a set of M data sources $\mathbf{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_M\}$, each containing N samples[4] $\mathbf{X}_m = \{\mathbf{x}_{m,1}, \ldots, \mathbf{x}_{m,N}\}$. For each data source m we can construct a graph G_m with N nodes, where each node represents a different data sample. The weights of the edges between the nodes in graph G_m are defined by pairwise similarities

[4]Where, for example, $\mathbf{x}_{m,i}$ could be a set of scores for a single MMSE question.

$W_{m,ij}$:

$$W_{m,ij} = \begin{cases} \exp(-\|\mathbf{x}_{m,i} - \mathbf{x}_{m,j}\|^2) & \text{if } i \neq j \\ 0 & \text{otherwise} \end{cases} \tag{8.13}$$

The classification of an undiagnosed test subject can be formulated as the minimisation of an energy function [Wang et al. 2009]:

$$Q(\mathbf{f}, \alpha) = \sum_{m=1}^{M} \alpha_m^r \left(\sum_{ij} W_{m,ij} \left| \frac{f_i}{\sqrt{D_{m,ii}}} - \frac{f_j}{\sqrt{D_{m,jj}}} \right|^2 + \mu \sum_i |f_i - y_i|^2 \right) \tag{8.14}$$

$$\arg\min_{\mathbf{f},\alpha} Q(\mathbf{f}, \alpha) \quad \text{subject to } \sum_{m=1}^{M} \alpha_m = 1 \tag{8.15}$$

where $\alpha = \{\alpha_1, \ldots, \alpha_M\}$ is a weight vector that determines the influence of each data source, and r is a relaxation parameter. The diagonal matrix $\mathbf{D}_m$ is defined by $D_{m,ii} = \sum_j W_{m,ij}$, and $\mathbf{f} = \{f_i, \ldots, f_N\}$ can be regarded as a score that indicates the relevance of each labelled subject to the test subject. The parameter μ controls the influence of the constraint of the data labels $\mathbf{y}$. For a binary classification task, y_i is set to 1 if $\mathbf{x}_i$ belongs to class A, -1 if $\mathbf{x}_i$ belongs to class B, and 0 if x_i is unlabelled (i.e. the test subject). Following minimisation of Equation 8.15, the data samples $\mathbf{x}_i$ can be classified according to the sign of f_i: class A if $f_i > 0$, and class B otherwise. In this work, we wanted to have the ability to classify subjects into one of four classes (HC, SMCI, PMCI, or AD), so an alternative construction of $\mathbf{y}$ was required. Therefore, we adopted the straightforward approach employed by [Zhao et al. 2014]:

$$y_i = \begin{cases} 1 & \text{if } i \text{ is the test subject} \\ 0 & \text{otherwise} \end{cases} \tag{8.16}$$

If α is fixed, the minimisation of $Q(\mathbf{f}, \alpha)$ has a closed form solution using the

normalised graph Laplacian $\mathbf{L}_m = \mathbf{D}_m^{-1/2}(\mathbf{D}_m - \mathbf{W}_m)\mathbf{D}_m^{-1/2}$:

$$\mathbf{f} = \left(\mathbf{I} + \frac{\sum_{m=1}^{M} \alpha_m^r \mathbf{L}_m}{\mu \sum_{m=1}^{M} \alpha_m^r} \right)^{-1} \mathbf{y} \tag{8.17}$$

Similarly, if $\mathbf{f}$ is fixed, [Wang et al. 2009] show that:

$$\alpha_m = \frac{\left(\frac{1}{\mathbf{f}^T \mathbf{L}_m \mathbf{f} + \mu |\mathbf{f} - \mathbf{y}|^2} \right)^{1/(r-1)}}{\sum_{m=1}^{M} \left(\frac{1}{\mathbf{f}^T \mathbf{L}_m \mathbf{f} + \mu |\mathbf{f} - \mathbf{y}|^2} \right)^{1/(r-1)}} \tag{8.18}$$

Equation 8.18 shows that the relaxation parameter r modulates the effect of the difference in smoothness between the graphs. For example, as $r \to 1$, the α_m of the smoothest graph is close to 1. Conversely, as $r \to \infty$, the α_m values are close to each other.

Using Equations 8.17 and 8.18, $Q(\mathbf{f}, \alpha)$ is minimised using an iterative approach, which alternates between optimising α and $\mathbf{f}$. The iterative solution is summarised in Algorithm 8.1.

Algorithm 8.1 Iterative solution for multi-graph learning.

1: Initialise $\mathbf{f} = \mathbf{y}$ and $\alpha_m = 1/M$
2: **while** not converged **do**
3: Update α using Equation 8.18
4: Based on the updated α, recalculate $\mathbf{f}$ using Equation 8.17
5: **end while**

Once Equation 8.15 has been minimised, the test subject can be classified according to the clinical diagnoses of the labelled subjects with the top k highest relevance scores f_i. For example, if the majority of the k most relevant training subjects have AD, then the test subject is classified as an AD patient.

The optimum multi-graph hyper-parameters r, μ, and k were determined by performing two classification tasks (HC vs AD and SMCI vs PMCI) on a random subset of the data, for a range of parameter values. The parameter set that gave the highest mean classification accuracy across both tasks was selected for the main

classification experiments: $r = 1.7$, $\mu = 0.5$, and $k = 9$. In order to assess the performance of the multi-graph method, leave-one-out testing was performed on the full data set.

8.2.5 Evaluation of Classification Performance

The authors of the SVM, manifold learning, and DSI methods compared their combined classification results to classification results obtained using each individual data type [Mattila et al. 2011; Wolz et al. 2012; Segovia et al. 2014]. In all cases, combining data from multiple sources resulted in a higher classification accuracy than using the data sources separately. However, a different data set was used in each paper, thus making it difficult to compare the methods directly. To address this issue, we conducted binary classification of HC vs AD subjects, and SMCI vs PMCI subjects, for each individual data type and each of the methods presented in Sections 8.2.1-8.2.4. Exactly the same data were used to test each method.

Following classification using the data sources individually, we used the methods to automatically classify each subject using all seven data types (i.e. as the methods are intended to be used). Similarly to [Schmitter et al. 2015], we conducted four binary classification tasks: HC vs AD subjects, HC vs all MCI subjects (i.e. SMCI+PMCI), all MCI subjects vs AD, and SMCI vs PMCI. In addition, we conducted two multiclass classification experiments: HC vs all MCI subjects vs AD, and HC vs SMCI vs PMCI vs AD.

The binary classification results were evaluated by calculating the classification accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve (AUC). However, for multiclass classification tasks, the sensitivity, specificity, and AUC cannot be calculated. Moreover, due to the imbalanced numbers of subjects within each diagnostic subgroup, the overall classification accuracy is misleading. For example, one could achieve 81% accuracy in the HC vs all MCI vs AD classification task by correctly classifying all of the AD and MCI subjects, and incorrectly classifying all of the healthy controls. Therefore, in this work, we present

the classification accuracy of each diagnostic subgroup separately. Furthermore, we followed [Schmitter et al. 2015] and, for all classification tasks in this chapter, we used McNemar's test to statistically assess the differences in accuracy between every classification method and every other [McNemar 1947].

8.3 Results

8.3.1 Classification Using Individual Data Sources

Figure 8.3 shows the classification accuracy, sensitivity, specificity, and AUC for the HC vs AD classification task using the data sources individually. For data types that were only available for a subset of the data (i.e. MMSE and CSF), the classification results are reported on the subset of data, not the entire data set of 264 subjects. The black borders in Figure 8.3 indicate the best results for each data type. The DSI method with ADAS-Cog data achieved the highest overall classification accuracy (96.5%). Moreover, across all four methods, the highest classification accuracy was obtained using either the MMSE or ADAS-Cog data. In contrast, the lowest classification accuracy for each method was obtained using the TMT data (SVM: 60.9%, manifold: 52.2%, DSI: 53.0%, multi-graph: 55.7%). For the CSF biomarker data, the SVM method achieved a significantly higher classification accuracy ($p < 0.01$, corrected for six comparisons) than the other three methods. Similarly, the DSI method resulted in a significantly higher classification accuracy than the manifold learning method using only ADAS-Cog data ($p < 0.01$, corrected). For all other data types, the results of McNemar's test showed that the classification accuracies of the methods were not significantly different from one another.

The classification results for the SMCI vs PMCI task using the individual data sources are presented in Figure 8.4. Again, the ADAS-Cog data achieved the highest overall classification accuracy (78.5%), albeit with the SVM classification method, rather than the DSI method. The highest classification accuracy for the multi-graph method was also obtained using the ADAS-Cog data (74.5%), whereas for

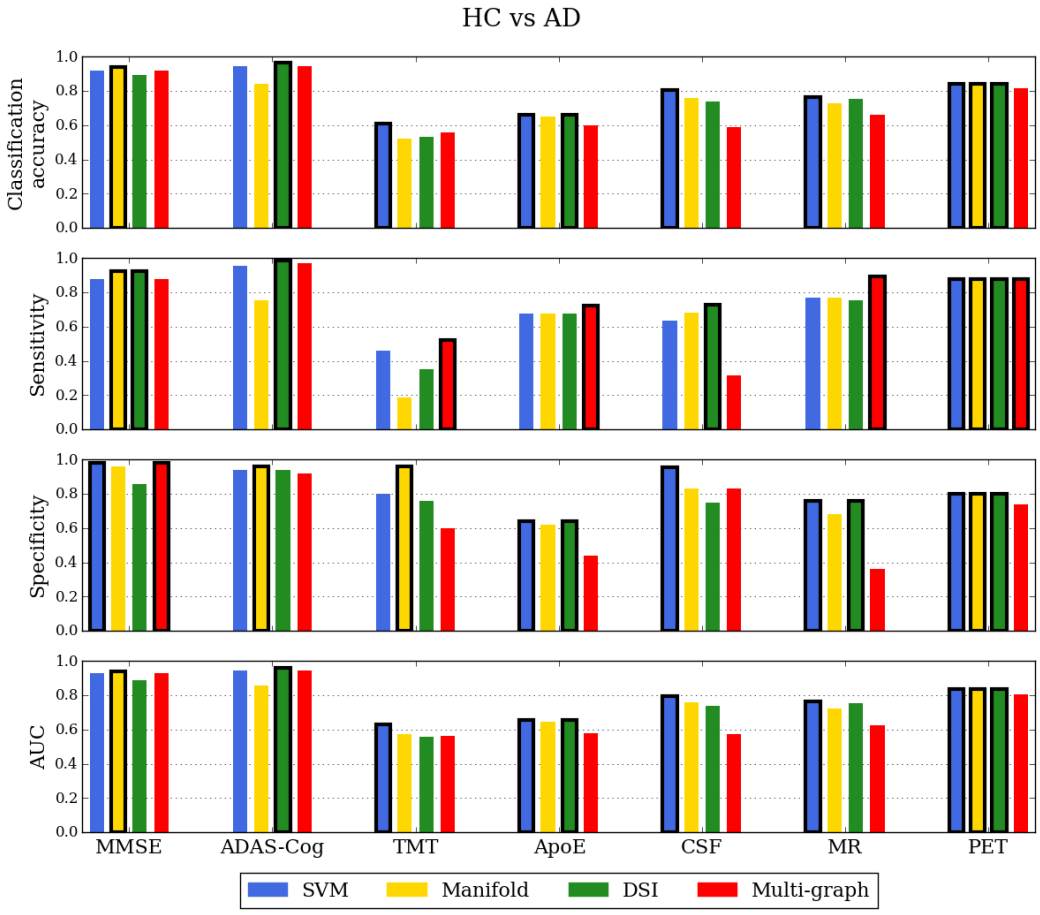


Figure 8.3: The classification accuracy, sensitivity, specificity and area under the receiver operating characteristic curve (AUC) for the HC vs AD classification task, using each data set individually. MMSE and CSF data were only available for a subset of the data, so the corresponding classification results are reported on the subset of the data, not the entire data set. The best results for each data type are highlighted with a black border.

the manifold learning and DSI methods, the highest accuracy was achieved using the CSF biomarker data (74.1% and 77.8%, respectively). In fact, the manifold and DSI methods resulted in a significantly higher classification accuracy than the multi-graph and SVM methods for the CSF biomarker data ($p < 0.01$, corrected), and significantly lower accuracy than the SVM method for the MMSE data. For the other five data types, the classification accuracies of the four methods were not significantly different from one another ($p < 0.01$, corrected). Despite this, when using the non-neuropsychological data (i.e. ApoE, CSF, MR, PET) the multi-graph method resulted in a substantially lower sensitivity (0.061, 0.371, 0.102, and 0.204, respectively) than the other three classification methods. However, the multi-graph

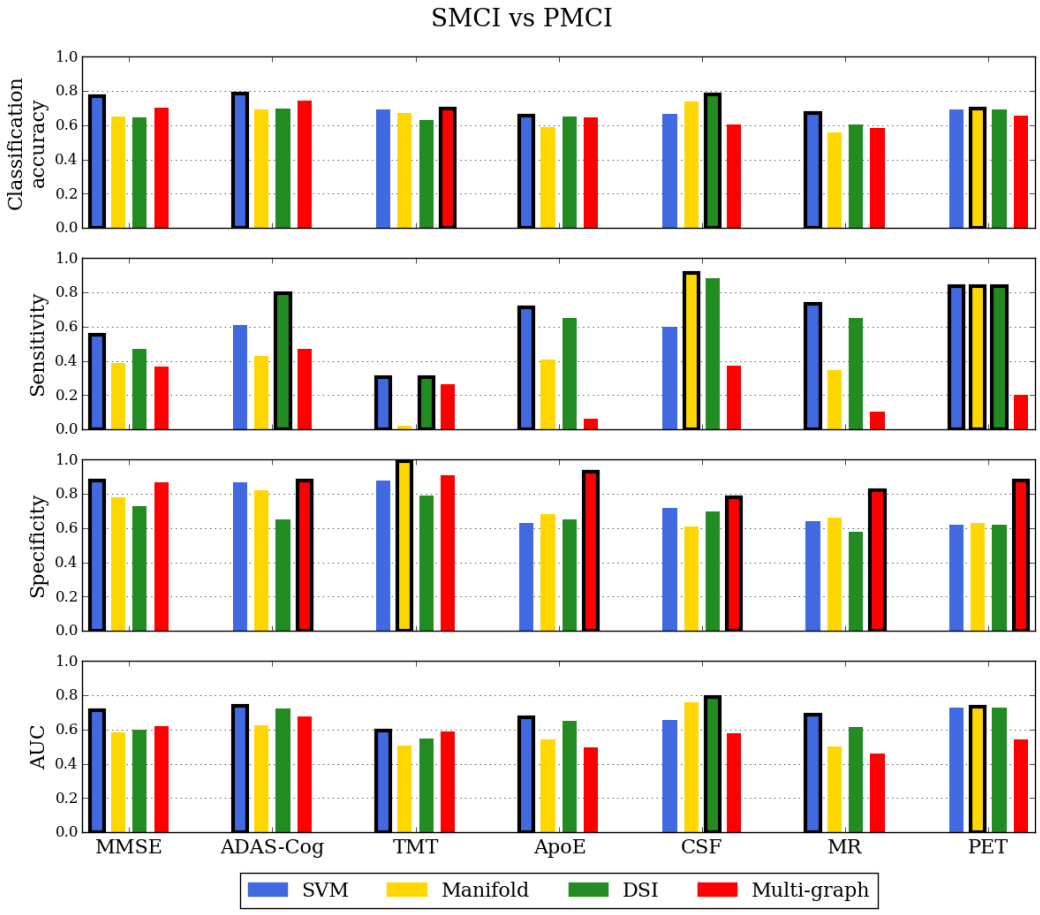


Figure 8.4: The classification accuracy, sensitivity, specificity and area under the receiver operating characteristic curve (AUC) for the SMCI vs PMCI classification task, using each data set individually. MMSE and CSF data were only available for a subset of the data, so the corresponding classification results are reported on the subset of the data, not the entire data set. The best results for each data type are highlighted with a black border.

method also resulted in a higher specificity than the other methods for the non-neuropsychological data (0.930, 0.783, 0.820, and 0.880, respectively).

8.3.2 Binary Classification Using Multiple Data Sources

The classification accuracy, sensitivity, specificity, and AUC for the HC vs AD classification task using all the data sources simultaneously are shown in Figure 8.5. All four classification methods achieved an accuracy greater than 87.0%, and the DSI and multi-graph methods achieved the same, highest classification accuracy, sensitivity, specificity, and AUC (95.7%, 0.954, 0.960, and 0.957, respectively). Despite this, McNemar's test showed that the classification accuracies of the four meth-

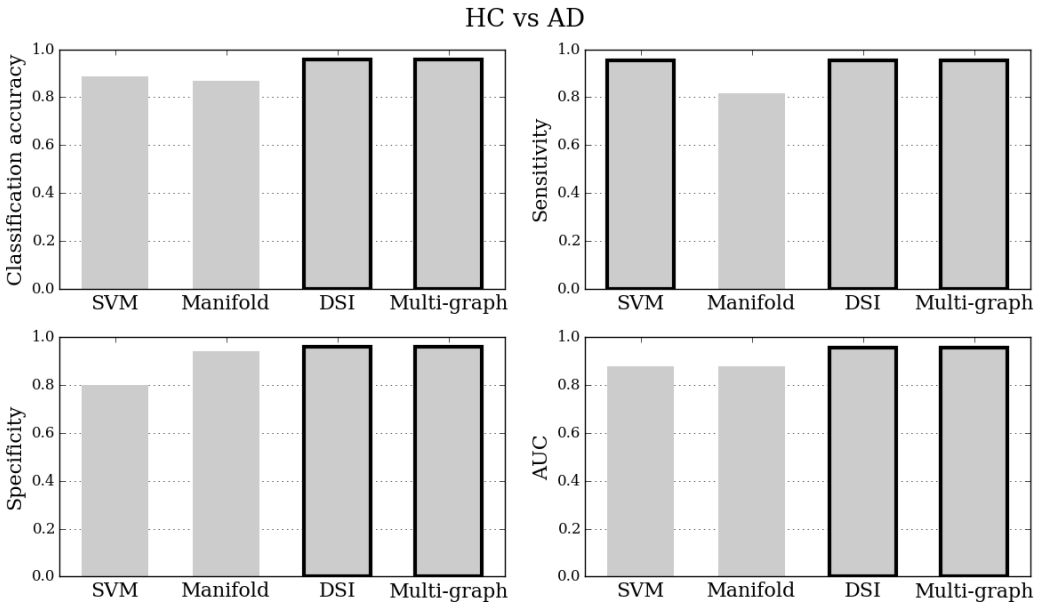


Figure 8.5: The classification accuracy, sensitivity, specificity and area under the receiver operating characteristic curve (AUC) for the HC vs AD classification task using all seven data sources. The best results are highlighted with a black border.

ods were not significantly different ($p < 0.01$, corrected). The multi-graph method resulted in a higher classification accuracy using all seven data types than using ADAS-Cog data alone (95.7% and 94.8%, respectively). In contrast, for the HC vs AD task, the other three methods achieved a higher classification accuracy using only ADAS-Cog or MMSE data than using all of the data sources.

Figure 8.6 shows the classification results for the HC vs all MCI subjects classification task using all seven data sources. The SVM method resulted in the highest classification accuracy, specificity, and AUC (82.9%, 0.800, and 0.819, respectively). Nevertheless, the classification accuracy of the SVM method was only significantly greater ($p < 0.01$, corrected) than the accuracy of the DSI method. Although the manifold learning classification method achieved perfect sensitivity, it achieved a very low specificity (0.020), indicating that it classified almost all subjects as MCI.

The classification results for the MCI vs AD classification task using all the data sources are presented in Figure 8.7. In contrast to Figure 8.6, the SVM method resulted in the lowest classification accuracy, sensitivity, specificity, and AUC (62.1%, 0.585, 0.638, and 0.611, respectively). The manifold learning, DSI, and multi-graph methods achieved a similar classification accuracy (86.9%, 85.5%, and 86.4%, respec-

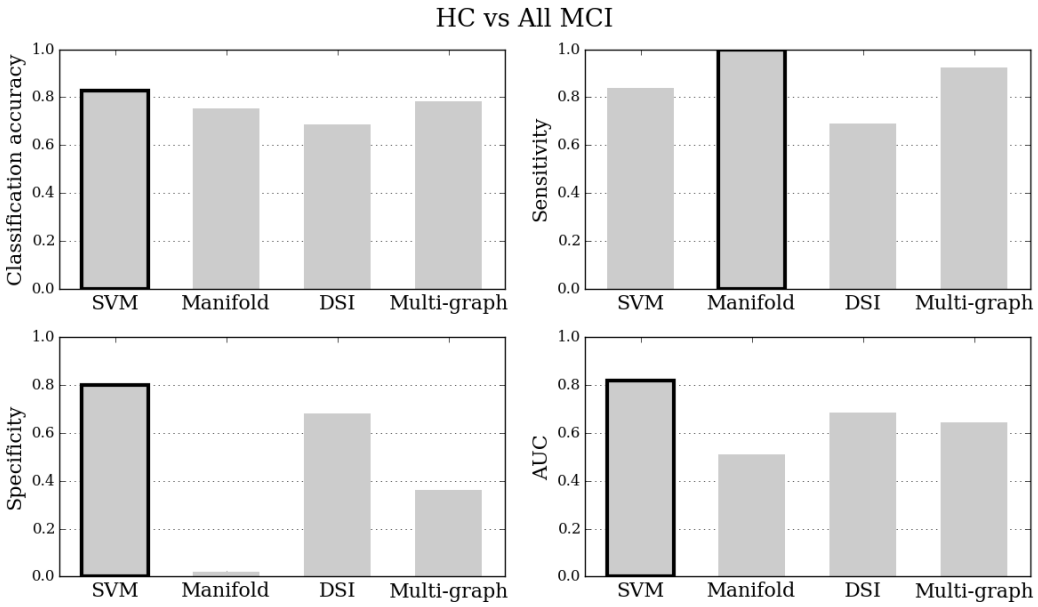


Figure 8.6: The classification accuracy, sensitivity, specificity and area under the receiver operating characteristic curve (AUC) for the HC vs all MCI subjects classification task using all seven data sources. The best results are highlighted with a black border.

tively) and AUC (0.811, 0.835, and 0.799, respectively). This is confirmed by the results of McNemar's test, which showed that the classification accuracies obtained by the manifold learning, DSI, and multi-graph methods were not significantly different from one another, but were all significantly higher that the accuracy of the SVM method ($p < 0.01$, corrected).

Finally, Figure 8.8 shows the results of the SMCI vs PMCI classification task using all seven data sources. All classification methods achieved a classification accuracy within 6% of one another. Although the classification accuracies of the four methods were not significantly different from one another ($p < 0.01$, corrected), the multi-graph method resulted in the highest classification accuracy and specificity (81.2% and 0.950, respectively). The DSI method obtained the highest sensitivity (0.735), and the SVM classification method achieved the highest AUC (0.752). Both the manifold learning method and the multi-graph method obtained a higher classification accuracy using all seven types of data (78.5% and 81.2%, respectively), than their maximum accuracy achieved using a single data source (74.1% and 74.5%, respectively).

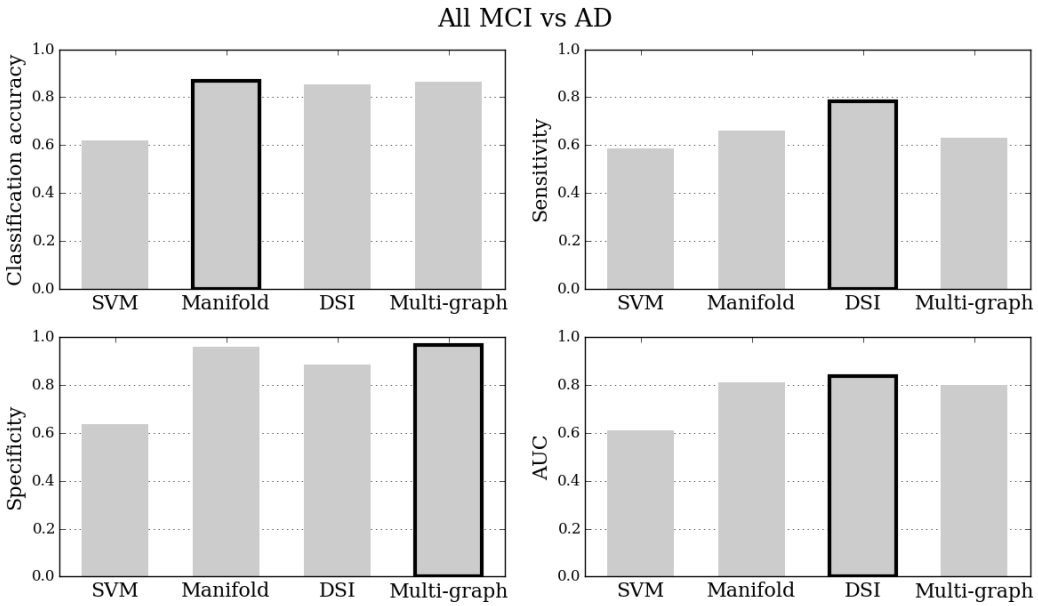


Figure 8.7: The classification accuracy, sensitivity, specificity and area under the receiver operating characteristic curve (AUC) for the all MCI subjects vs AD classification task using all seven data sources. The best results are highlighted with a black border.

8.3.3 Multiclass Classification Using Multiple Data Sources

The classification accuracy for each diagnostic subgroup following the HC vs MCI vs AD classification task using all the data sources is shown in Figure 8.9. The DSI method resulted in the highest classification accuracy for the HC and AD subjects (68.0% and 78.5%, respectively), but the lowest accuracy for MCI subjects (57.7%). Moreover, the classification accuracy of the DSI method for MCI subjects was significantly lower ($p < 0.01$, corrected), than the accuracy of the multi-graph method for the same diagnostic group (84.6%). However, the DSI method obtained a significantly higher classification accuracy than the SVM and manifold learning methods for the group of healthy controls ($p < 0.01$, corrected).

Figure 8.10 presents the classification accuracy for each diagnostic subgroup following the HC vs SMCI vs PMCI vs AD classification task using all seven data sources. The highest classification accuracy for each diagnosis subgroup was achieved by a different classification method. The manifold learning resulted in the highest classification accuracy for healthy controls (64.0%), and the multi-graph method achieved a significantly higher accuracy than all other methods for the SMCI sub-

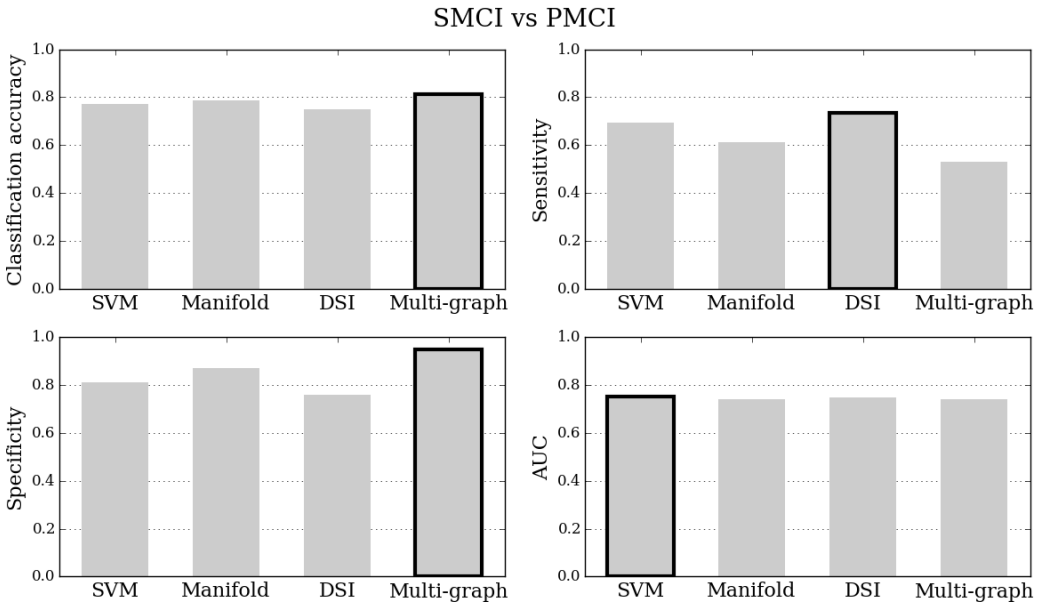


Figure 8.8: The classification accuracy, sensitivity, specificity and area under the receiver operating characteristic curve (AUC) for the SMCI vs PMCI classification task using all seven data sources. The best results are highlighted with a black border.

group ($p < 0.01$, corrected). The DSI method obtained the highest accuracy for PMCI subjects (55.1%), which was significantly higher than the accuracy achieved by the SVM method ($p < 0.01$, corrected). Furthermore, for the PMCI subgroup, the SVM method achieved a lower classification accuracy (16.3%) than the expected accuracy of a random classifier. Despite this result, the SVM method obtained the highest classification accuracy for the AD patients (78.5%), which was significantly greater than the accuracy of the DSI method for the same diagnosis subgroup ($p < 0.01$, corrected). Nevertheless, similarly to the HC vs MCI vs AD classification task using all the data sources, the DSI method achieved the most consistent classification accuracies across the different diagnosis subgroups (50.0%, 49.0%, 55.1%, and 52.3%, respectively).

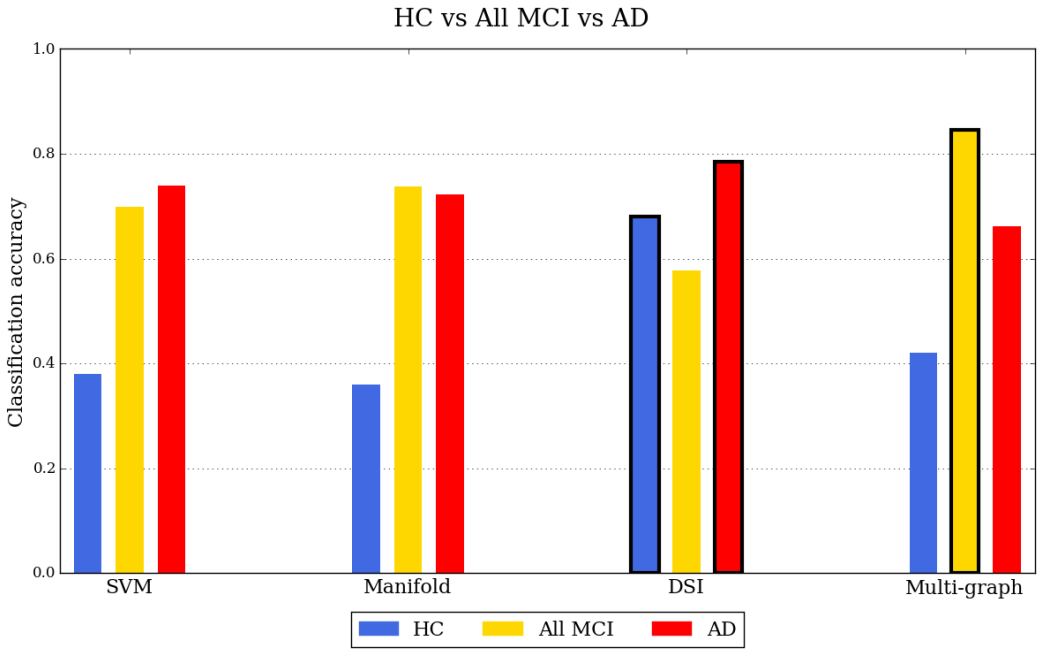


Figure 8.9: The classification accuracy for each diagnostic subgroup following the HC vs all MCI subjects vs AD classification task using all seven data sources. The highest classification accuracy for each diagnosis subgroup is highlighted with a black border.

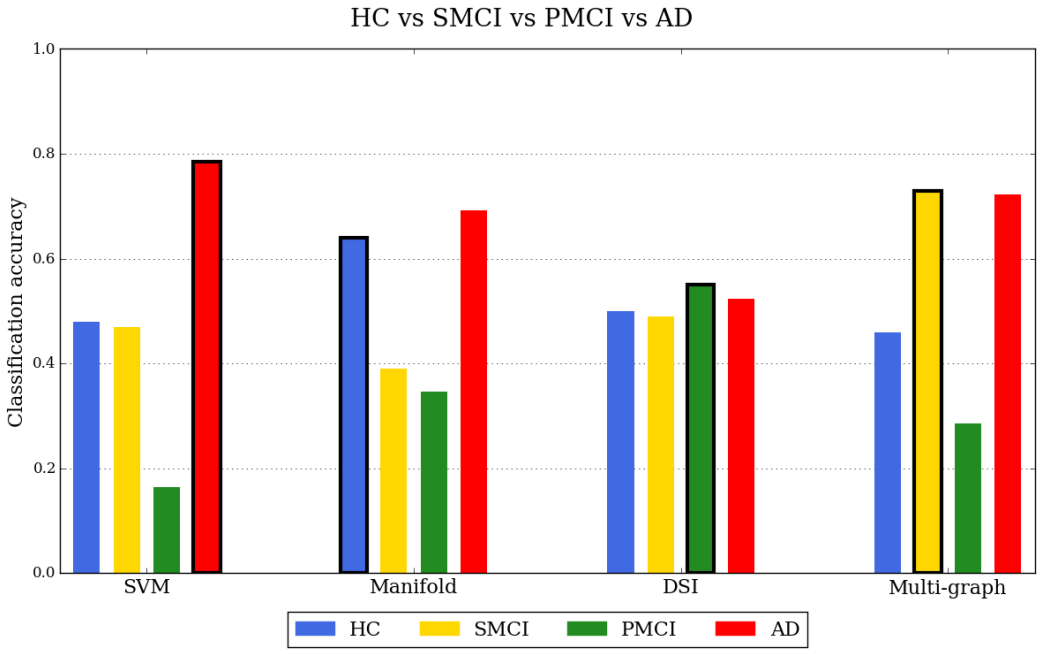


Figure 8.10: The classification accuracy for each diagnostic subgroup following the HC vs SMCI vs PMCI vs AD classification task using all seven data sources. The highest classification accuracy for each diagnosis subgroup is highlighted with a black border.

8.4 Discussion

8.4.1 Classification Performance

The classification results shown in Figure 8.3 demonstrate that MMSE and ADAS-Cog are the most adept data sources for classifying AD patients and healthy controls. This result agrees with the findings of [Mattila et al. 2011], which showed that the best combinations of data sources for discriminating between HC and AD subjects all contained MMSE and ADAS-Cog data. Moreover, as discussed in Section 8.1, the total MMSE scores form part of the baseline clinical diagnosis criteria used in the ADNI study for AD subjects and healthy controls . Consequently, it is unsurprising that MMSE, and a similarly comprehensive neuropsychological test (i.e. ADAS-Cog), resulted in the highest classification accuracy in the HC vs AD classification task. As explained in Chapter 2.1, research has shown that incorrect diagnosis of AD is common [Beach et al. 2012]. This could account for the discrepancies between the automated classification results obtained using MMSE data and the gold standard diagnoses provided by ADNI. In order to assess the ability of the MMSE to classify subjects according to their clinical subgroup independently of the ADNI diagnoses, an alternative gold standard, such as post-mortem examination, would be required. However, at the time of this thesis, autopsy data are unavailable for most subjects in the ADNI database.

Figure 8.5 showed that all four classification methods are capable of classifying healthy controls versus AD patients with high accuracy ($> 87.0\%$), despite the inclusion of relatively uninformative data (such as TMT data), as indicated by the results in Figure 8.3. However, for the HC vs AD classification task, only the multi-graph method achieved a higher classification accuracy using all seven data sources than a single data type. Similarly, for the SMCI vs PMCI classification task, only the manifold learning and multi-graph methods achieved a higher accuracy using all of the data sources compared to a single data source. Not only does this demonstrate that a higher classification accuracy is possible with the available data, but it also

suggests that the classification methods are not able to fully assess the relevance of each individual data source.

Similarly to the results reported by [Mattila et al. 2011], none of our data sources had zero relevance in the DSI framework (see Equation 8.11). Since all seven data sources can be used clinically to aid the diagnosis of AD, this result is not surprising. However, data sources with a low relevance can have a detrimental effect on the final DSI value, and subsequently, the final predicted clinical diagnosis. This is demonstrated in Figures 8.3 and 8.5, which show that the DSI method achieved a higher classification accuracy using only the ADAS-Cog data (96.5%) compared to using all seven data sources (95.7%). Although the difference in classification accuracy is small, the difference could be exacerbated if there were more data sources with a low relevance. For example, suppose we have a data source that can correctly classify all subjects. According to Equation 8.11, this perfect data source would have a relevance Rel = 1. Furthermore, suppose we also have four additional data sources, each with a relevance Rel = 0.1. If we have a test subject with AD, it will have a fitness Fit = 1 for our perfect data source (see Equation 8.9). For our remaining data sources, assume that their fitness values are all Fit = 0, suggesting that the subject is actually a healthy control (which is possible, given their low relevance). Using Equation 8.9, the resulting DSI value is 0.741. Given that the first clinical test has a relevance of Rel = 1 and a fitness Fit = 1, one would expect a final DSI value DSI = 1. Even if the data sources with low relevance have fitness Fit > 0.5, the final DSI value is less than one unless all the fitness values are Fit = 1. This is a trivial example, and small variations in DSI may not affect the final predicted clinical diagnosis, but it highlights that highly relevant data sources are weakened by less informative data sources.

The results presented in Figure 8.6 show that, for the HC vs MCI classification task, the manifold learning method achieved perfect sensitivity and a very low specificity (0.020) compared to the other methods. This suggests that the manifold learning method favoured the MCI subgroup over the healthy controls. A similar,

though less pronounced, trend was observed for the MCI vs AD classification task in Figure 8.7. The manifold learning method achieved a higher classification accuracy for the MCI subjects than the AD patients (i.e. a higher specificity than sensitivity). One reason for this could be the imbalanced data set; there are substantially more MCI subjects than healthy controls or AD patients (149 subjects compared to 50 or 65 subjects, respectively). In the manifold learning framework, subjects are classified using an SVM following dimensionality reduction of the data via a manifold. Although SVMs are fairly insensitive to the relative numbers of subjects in each class [Drucker et al. 1999], it has been shown that when the class sizes are very different, SVMs may favour the larger class, leading to poor accuracy in the smaller class [Lin and Chen 2013]. Nevertheless, Figures 8.6 and 8.7 indicate that the SVM method proposed by [Segovia et al. 2014] did not suffer from the same problem as the manifold learning method.

The SVM and manifold learning methods achieved significantly lower classification accuracies for the HC vs MCI vs AD classification task than the DSI method (see Figure 8.9). Moreover, the accuracies of the SVM and manifold learning methods were only marginally better than a random classifier for a three-class classification problem (0.380 and 0.360, respectively). In the HC vs SMCI vs PMCI vs AD task, the SVM method resulted in a lower classification accuracy than the expected accuracy of a random classifier for the PMCI subgroup (0.163). These relatively poorer results could be because the SVM and manifold frameworks are unable to adjust the influence of each data source. Therefore, data sources that are less relevant to the classification task have as much influence on the final classification decision as highly informative data sources. One straightforward way to address this issue in the SVM method could be to use a majority vote approach when combining the predicted classes of each data source, rather than selecting the most confident predicted class. Alternatively, data could be excluded prior to the training of the SVM or manifold learning methods. For example, data sources with no statistically significant difference between healthy controls and AD patients could be omitted from

the classification framework.

Figures 8.9 and 8.10 suggest that the DSI method is the most consistent classifier in the multiclass classification tasks. For example, across the four diagnostic subgroups in the HC vs SMCI vs PMCI vs AD classification task, the classification accuracy of the DSI method varied by only 6.1%. In contrast, for the same classification task, the difference between the highest and lowest classification accuracy for the SVM method was 62.1%. The reason for the small range of classification accuracies for the DSI method could be due to the training data. Unlike the other three methods, which use all of the available data during the training phase, the DSI method only uses data for AD patients and healthy controls. Due to the broad definition of MCI, the data for MCI subjects are more variable. Therefore by including MCI data in the training set, it could be more difficult to train a classifier to separate the clinical subgroups. Furthermore, by only training with subjects at the extremes of the clinical spectrum, it is easier to model the progression of the disease. For example, the volume of the hippocampus is known to decrease with AD progression [Frisoni et al. 2010], but the rate of volumetric change from HC to MCI could be very different to the rate of change from MCI to AD. The DSI method assumes that disease progression is linear, and fundamentally interpolates between healthy controls and AD patients in order to diagnose subjects with MCI. Although this approach has achieved favourable results in the experiments presented in this chapter, the biomarker cascade model proposed by [Jack et al. 2010] implies that the development of dementia due to AD is non-linear. Therefore, it may be possible to improve the classification accuracy of the DSI method by changing the interval-levels used in the classification assessment. Rather than setting values at regular intervals (i.e. HC = 0, SMCI = 1/3, PMCI = 2/3, and AD = 1, as used in this work), a higher classification accuracy may be achieved by reducing the interval between HC and SMCI, and PMCI and AD. However, the biomarker cascade model also suggests that the level of non-linearity is different for different data sources. Consequently, it may be difficult to translate the non-linear nature of the data into

suitable intervals for classification of the final DSI values.

When predicting the clinical diagnoses in the four-class classification task, a different classification method obtained the highest classification accuracy for each diagnosis subgroup. Therefore, an interesting extension to this work could be to combine the outputs of the four methods, in order to assess the classification performance of a "super-framework". However, this would lead to another issue, namely, how to combine the results of each method in the best possible manner.

8.4.2 Methodological Considerations

Although the original SVM method proposed by [Segovia et al. 2014] cannot intrinsically account for missing data, we modified the method to handle the incomplete MMSE and CSF data. In short, missing data points were omitted from the comparison of distances from the classification boundary. For example, for a subject with data from all seven sources, the final predicted class was the predicted class obtained by the data source in which the subject was the furthest from the classification boundary. If the test subject did not have CSF biomarker data, only the remaining six data sources were used to determine the final predicted clinical diagnosis.

In the paper by [Wolz et al. 2012], the manifold learning method combines non-imaging information with MR imaging. The extended Laplacian Eigenmaps objective function (Equation 8.5) comprises two terms. The first term maps subjects with similar MR images close together in the embedding space, and the second term maps subjects close to the cluster centres of groups of subjects with similar meta-data. The influence of the MR imaging term is controlled by a weighting parameter γ. In this work, we set $\gamma = 0$, such that the pairwise similarities between the MR images were ignored. This will probably have had an effect on the classification performance of the manifold learning framework, since it is contrary to the original design of the method. However, in order to ensure an unbiased comparison between the classification methods tested in this chapter, all four methods used exactly the

same data.

Unlike the SVM and DSI methods, the manifold learning and multi-graph methods have to be completely retrained if new data sources or test data are introduced. This is because a new graph (or set of graphs, for the multi-graph method) has to be constructed. For the relatively small dataset used in this work, this was not a problem, however, if more data sources and training data were incorporated into the frameworks, reconstructing the graph and solving the model could become very time-consuming. Nevertheless, one advantage of the multi-graph method is that it returns subjects that are similar to the test subject. This could have benefits beyond classification, such as collecting similar subjects for clinical studies.

Compared to the time taken to compute the final predicted clinical diagnoses using the SVM, manifold learning, or multi-graph methods, the DSI method was very slow. This is due to the recursive nature of the DSI calculations, and the design of our experiments, which used leave-one-out testing. In order to obtain the final DSI value for a single test subject, the DSI values of all the training data must also be computed. In a clinical setting, where the training data are fixed, the DSI values of the training data can be calculated in advance prior to deploying the framework. However, in settings where the training data are not fixed (such as a leave-one-out test), the DSI values of the training data have to be continually recalculated.

8.4.3 Conclusion

In this investigatory chapter we have compared the performance of four classification frameworks capable of utilising multiple types of data: the late integration SVM method proposed by [Segovia et al. 2014], a manifold learning technique proposed by [Wolz et al. 2012], the Disease State Index proposed by [Mattila et al. 2011], and our novel method based on multi-graph learning. Our results suggest that the ability of the four methods to classify subjects according to clinical diagnosis is highly variable, and depends on the clinical subgroups within the data. We achieved a high classification accuracy with all four methods when classifying healthy controls

versus AD patients. However, for multiclass classification tasks, the classification accuracy of the individual diagnosis subgroups was typically much lower.

The aim of this work was to compare existing methods for automated diagnosis of AD using a standardised set of data from a multitude of clinical sources. Subsequently, we have set benchmark classification results to which any future developments should be compared. We have also identified key problems regarding the design and implementation of each of the four methods, and therefore, we propose that any future work towards a complete classification framework for AD should, at minimum, fulfil the following criteria:

- The framework must be able to individually assess the relevance of each data source, and disregard irrelevant data.
- Imbalanced class sizes in the training data should have little effect on the final classification performance.
- The framework must be able to cope with missing or incomplete data. For example, a subject may have MMSE scores, but no CSF biomarker data.
- It must be straightforward to incorporate additional data sources into the framework, as they become available (e.g. further neuropsychological test scores).
- When used clinically, it must be possible to classify previously unseen subjects without having to retrain the entire framework.
- The output of the framework (i.e. the predicted clinical diagnosis of a test subject, or a numerical value indicating disease severity) must be quickly and easily interpretable.
- In all classification tasks, the framework should achieve a similar, high classification accuracy for each clinical subgroup.

The work in this chapter has explored the possibility of using a single framework for multiclass classification of the state of AD. The problems we identified and the

criteria that we proposed could help shape future research into combining multiple biomarkers for computer-aided AD diagnosis. In the next, and final, chapter we will summarise the key contributions of the research presented in this thesis. Moreover, we will outline further topics for future research, in addition to working towards a complete classification framework for AD.

Chapter 9

Conclusion

This thesis has presented novel methods for the analysis of amyloid PET images. The methods include frameworks for the joint registration of PET-MR image pairs to a template space, and approaches to amyloid status classification. In this final chapter, we summarise the key contributions of our research, and discuss some of its clinical limitations. We also indicate how the methods developed in this work could be extended, and encourage research advancing our work towards a complete classification framework for AD.

9.1 Key Contributions

The first two methods chapters in this thesis considered the quantitative analysis of brain amyloid PET images. We explained that amyloid PET tracer uptake is typically quantified using the standardised uptake value ratio (SUVR), which is the ratio of mean tracer uptake in a region of interest, to uptake in a reference region (where the uptake is non-specific). Prior to SUVR calculation, the PET volume is registered to a template space in which the regions of interest are defined. Our first contribution was to investigate the possibility of combining PET and MR information during this registration process, thus potentially capturing changes present in both modalities, and leading to more precise SUVRs (see Chapter 3). To achieve this, we proposed an extension to the demons registration framework, in which two,

single-modality registration updates are combined to simultaneously register PET and MR images to a template space. We also introduced a weighting parameter to control the influence of each modality during the registration process. We compared our novel method to affine registration, non-linear PET-based registration, and non-linear MR-based registration. Our results showed that our novel combined PET-MR registration method could perform at least as well as single-modality methods in terms of registration accuracy. Moreover, our results also suggested that SUVRs acquired using our combined PET-MR registration method could be better suited to distinguishing between populations of AD patients and healthy controls than SUVRs obtained using the single-modality registration methods. We also compared the registration accuracy and SUVRs acquired using a synthetic PET template and a PET template constructed from real ^{18}F-florbetapir data. We demonstrated that the real PET template can achieve significantly higher registration accuracy than the synthetic PET template when PET is the dominant modality in our combined PET-MR registration method. However, the SUVRs obtained using our PET-MR registration method with the synthetic PET template may be marginally better suited to discriminating between groups of healthy controls and AD subjects than those acquired using the real PET template.

The main contribution in Chapter 4 was our assessment of the robustness of SUVR calculation to the image registration method. We also extended our method proposed in Chapter 3 by posing the registration framework as the minimisation of a single energy function, and introduced a locally adaptive modality weighting. Although the MR-only and joint PET-MR registration methods could outperform the PET-only registration method in terms of registration accuracy, our results showed that there was a significant separation between the SUVRs of populations of AD patients and healthy controls, regardless of the registration method used to align the PET volumes to the template.

Following our conclusion that the accuracy of amyloid status classification using SUVR is not affected by the registration method, Chapters 5-7 of this thesis focused

on developing novel methods for the automated classification of brain amyloid status, which could overcome the shortcomings of SUVR, namely, the need to predefine cortical regions of interest and an amyloid positivity threshold. In Chapter 5, we proposed a method for classification of brain amyloid status based on histograms of oriented 3D gradients. Our method, 3D HOG + SVM, achieved a higher classification accuracy than SUVR and another machine learning method based on voxel intensity directly [Vandenberghe et al. 2013]. Moreover, we trained our proposed classification method using ^{18}F-florbetapir PET images, and showed that it could be applied to other amyloid PET tracers with little recalibration.

In Chapter 6 we assessed the generalisability of our proposed 3D HOG + SVM amyloid status classification method. We retrained our method three times, each time using training data from a different amyloid PET tracer. Following each successful training phase, our classification method was tested using data from all three tracers. In all cases, our proposed 3D HOG + SVM method resulted in a higher classification accuracy than another machine learning method from the literature [Vandenberghe et al. 2013], and achieved a classification accuracy greater than or equal to SUVR. We also showed that our method is capable of making fewer low-confidence classification decisions.

A further contribution to the classification of amyloid status using PET imaging was presented in Chapter 7. We investigated the potential of convolutional deep belief networks (CDBNs) as an alternative classification method to SUVR and our proposed 3D HOG + SVM method from Chapter 5. Our results indicated that one-layer CDBNs can classify brain amyloid status more accurately than two- and three-layer CDBNs. Our results also suggested that a one-layer CDBN trained using non-medical images could more accurately classify brain amyloid status than a similar CDBN trained using 2D amyloid PET data. However, all of the CDBNs that we tested were unable to achieve a higher classification accuracy than the amyloid status classification method proposed in Chapter 5. For this reason, in addition to the long training time and the impracticality of optimising the CDBN hyper-parameters, we

concluded that, at present, our 3D HOG + SVM method may be favoured over CDBNs for classifying brain amyloid status from PET imaging. Nevertheless, it should be noted that the CDBNs in Chapter 7 were only trained using 2D data, whereas the 3D HOG + SVM method used 3D PET volumes. Therefore, it may be possible to improve the classification accuracy of the CDBNs by by leveraging more computational power to train them using 3D data.

Finally, in Chapter 8, we conducted investigatory research towards a classification framework for AD that is capable of exploiting data from multiple clinical tests. Using a standardised data set, we compared three existing frameworks from the literature, and a novel classification method based on multi-graph learning. All four methods were able to classify healthy controls versus AD patients with a very high degree of accuracy, however for multiclass classification tasks, the classification accuracy of the individual diagnostic subgroups was much lower. In addition to presenting our preliminary classification results, we identified drawbacks in the design and implementation of each classification framework. Consequently, we proposed a set of criteria that any future work towards a complete classification framework for AD should fulfil.

9.2 Limitations of this Work

Two problems associated with nearly all PET image analysis methods that have not been considered in this thesis are partial volume effects (PVEs) and noise. The PVE occurs when a region of an image contains part of one anatomical structure or tissue type and part of another, mixed together, such that the structures or tissues are indistinguishable. This leads to an apparent loss in activity in some regions, and an increase in activity in adjacent regions. Correcting for the PVE is very likely to change the SUVRs in Chapters 3 and 4. However, since PVE correction requires anatomical segmentations and accurate registration to anatomical images, any errors in these steps will lead to errors in the PVE correction. Moreover, it is not completely clear from the literature that PVE corrected SUVRs can more

accurately classify amyloid status than non-corrected SUVRs. For example, [Rullmann et al. 2016] demonstrated that PVE correction increases the ability of SUVRs to discriminate between AD patients and healthy controls in ^{18}F-florbetaben scans, but [Villemagne et al. 2012] concluded that the PVE has not proven to be a major obstacle in quantitative analysis of amyloid PET imaging. Noise in the amyloid PET images may have had an influence on the registration results, but because the SUVR calculation is based on mean tracer uptake in regions of interest, SUVRs should be less affected by noise. Moreover, since the same amyloid PET data were used when comparing the different registration and classification methods, any impact of the errors from noise and the PVE can be considered to be consistent across the different methods. Whilst these two problems may not be significant enough to affect decisions about patient diagnosis and management, they could affect research efforts for early disease detection.

Except for the ^{18}F-florbetaben PET data used in Chapters 5 and 6, the data used in this thesis were exclusively from the ADNI database. Whilst ADNI is a global research effort to aid investigation into the progression of AD, the participants are only recruited from North America. The ADNI inclusion criteria[1] also state that participants must be fluent in English and/or Spanish, thus excluding some North American sub-populations from the study. Consequently, the results and conclusions presented in this work are only applicable to a small subset of the global population at risk of developing AD. Following the success of the North American ADNI study, further initiatives have been established in other countries and regions. The methods for conducting imaging scans and collating data are standardised by the umbrella organisation, World Wide Alzheimer's Disease Neuroimaging Initiative[2] (WW-ADNI). In future, it will be essential to apply our methods, and other research methods, to other populations in order to assess their global effectiveness.

In Chapters 4 and 8, we used the clinical diagnoses from ADNI as the gold

[1] `http://adni.loni.usc.edu/wp-content/uploads/2008/07/adni2-proceduresmanual.pdf`

[2] `http://www.alz.org/research/funding/partnerships/WW-ADNI_overview.asp`

standard for our classification experiments. However, as explained in Chapter 2, definitive diagnosis of AD can currently only be achieved via an autopsy. Whilst ADNI aims to provide post-mortem confirmation of the disease state, autopsy information is unavailable for the majority of subjects in the database. Therefore, although autopsy confirmation would be the ideal gold standard for classification, the lack of available data would have limited our experiments.

The largest single data set in used in this thesis consisted of 294 ^{18}F-florbetapir PET images from 294 subjects (though this was reduced to 264 subjects following visual assessment of the images). Ideally, due to the high dimensionality of the data, our proposed classification methods should be trained using more training samples [Batmanghelich et al. 2012]. However, the time taken to pre-process and, in the case of Chapters 5-7, visually assess medical image data, often limits the size of the data sets used in medical image analysis. We used a subset of the total number of ADNI participants with ^{18}F-florbetapir PET images, and since ADNI is a longitudinal study, some subjects now have multiple ^{18}F-florbetapir PET scans available in the database. Nevertheless, even if we consider images from the same subject separately, the total number of ^{18}F-florbetapir PET images in the ADNI database is significantly smaller than the number of images in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) data set used in computer vision research (2122 compared to 1461406 [Russakovsky et al. 2015], respectively[3]). Attempts have been made in the literature to artificially increase the number of training samples for machine learning in medical imaging. For example, in order to generate more samples, [Roth et al. 2015] randomly applied non-rigid deformations to their training data. Although the deformations were chosen such that the resulting warped images resembled plausible physical variations of the original images, caution is required when using this data augmentation approach. There needs to be a consensus on the validity of the "biologically plausible deformations," and the degree of deformation is likely to vary depending on the image modality and machine learning task.

[3]Correct on Feb 15, 2016.

The clinical value of amyloid imaging is dependent on the quality of the images and accuracy of interpretation. Although we demonstrated in Chapters 5 and 6 that our proposed 3D HOG + SVM classification method can achieve a very high agreement with visual assessment of amyloid PET images, there are currently other factors affecting the clinical relevance of our work. In September 2013 the U.S. Centers for Medicare and Medicaid Services (CMS) determined that there is insufficient evidence to conclude that amyloid PET imaging is reasonable and necessary for the diagnosis or treatment of dementia[4]. Consequently, the cost of amyloid PET imaging is not covered under the Social Security Act. Moreover, the CMS declared that they would only cover the cost of one amyloid PET image per patient in clinical trials, thus restricting longitudinal AD studies. However, almost two years after their decision to not reimburse amyloid PET images, the CMS approved a four-year, $100 million study called Imaging Dementia – Evidence of Amyloid Scanning (IDEAS) [Frantz and Farley 2015]. The aim of the IDEAS study is to examine how amyloid PET imaging can aid physicians in diagnosing and treating AD and other dementias, outside of tertiary academic care and clinical trials. As a result, robust and accurate automated tools for identifying, quantifying, and classifying amyloid will be required for assessing whether or not amyloid PET scans have an effect on short-term patient management, and hence justify the use/cost of amyloid PET imaging.

9.3 Future Research

In Chapter 3 we compared registration results following registrations of PET-MR image pairs to a synthetic PET template and a template constructed from real ^{18}F-florbetapir PET volumes. The synthetic PET template was created from a weighted combination of the MR template tissue segmentations, where the weightings were selected to produce a synthetic PET template that had similar grey-white matter

[4] https://www.cms.gov/medicare-coverage-database/details/nca-decision-memo.aspx?NCAId=265

contrast to a range of amyloid positive and negative PET images. Nevertheless, the final choice of parameters was heuristic, and a different set of tissue weightings could improve the synthetic PET template registration results. Consequently, one area of future work could be to comprehensively analyse the effect of the tissue weighting parameters used in the synthetic PET template construction on the registration accuracy and SUVR.

Although we exploited multiple clinical biomarkers in our preliminary investigation into a complete classification framework for AD (Chapter 8), we only considered amyloid PET and T1-weighted MR imaging in this thesis. In future, it would interesting to extend our novel amyloid status classification method, proposed in Chapter 5, to other modalities or clinical problems. For example, classifying the extent of neurofibrillary tangles imaged using the ^{11}C-PBB3 PET tracer [Maruyama et al. 2013]. Moreover, other MR-based techniques, such as arterial spin labelling (ASL), have shown promise as diagnostic biomarkers for AD [Musiek and Chen 2012]. With more data becoming available, our proposed amyloid status classification method could be applied to the classification of AD using ASL images.

As discussed in Chapter 5, it may be possible to classify brain amyloid status more accurately using a CDBN trained with 3D data compared to using a CDBN trained with a single 2D axial slice from 3D PET volumes. However, it is very computationally expensive to train a CDBN using 3D data, making it impractical to evaluate its classification performance. Therefore, future research into deep learning methods for amyloid status classification, or other medical image analysis tasks, could focus on developing novel methods for reducing the computational cost of training. Although some methods exist in the computer vision literature for speeding up the training of deep learning methods [Rigamonti et al. 2013; Jaderberg et al. 2014], they are applied to deep learning methods trained using 2D data, and there is no guarantee that they would scale to 3D data. By substantially reducing the training time, not only would it be possible to optimise the hyper-parameters of the CDBN, but it would also be possible to investigate other deep learning methods,

such as convolutional neural networks (CNNs), that would otherwise be impractical to train using 3D data. Moreover, a reduction in computational cost would allow the classification performance of state-of-the-art deep architectures, such as the 19-layer CNN proposed by [Simonyan and Zisserman 2015], to be assessed in a medical image analysis task using 3D data. The outcomes of this research could have applications beyond medical image analysis, since general methods for efficiently and quickly training deep learning methods would allow them to be applied to other high-dimensional data, such as video sequences.

In the absence of an AD screening programme, it is reasonable to assume that most individuals only undergo clinical testing for dementia when they start to exhibit memory problems. However, it has been shown that AD biomarkers become abnormal in a temporally ordered manner as the disease progresses [Jack et al. 2010]. For example, brain amyloid deposition occurs early in the disease, before the appearance of clinical symptoms [Villemagne et al. 2011b], and MR imaging is the last biomarker to become abnormal, retaining a close relationship with cognitive performance late into the disease [Jack et al. 2010]. Therefore, in future, longitudinal subject data could be used to develop frameworks for the classification and assessment of AD. Not only could longitudinal data result in an increased classification accuracy of disease state, but by analysing the rates of change in each biomarker, it could be possible to provide a more accurate prognosis for each subject. For example, with sufficient longitudinal training data, it may be possible to predict the probability that a subject with mild cognitive impairment will develop AD within five years of the baseline assessment.

We believe that without a more objective gold standard than visual assessment, it may be difficult to achieve clinically significant improvements in the classification accuracy of amyloid status. Therefore, the majority of future research in this field should be directed by the initial results presented in Chapter 8. We have proposed design requirements for future AD classification frameworks that use data from multiple clinical tests, and have set benchmark classification results for them to

surpass.

9.4 Closing Remarks

The aim of this thesis was to work towards the development of an amyloid PET image analysis tool for aiding the accurate and consistent diagnosis of AD. We started, in Chapters 3 and 4, by proposing a novel method for joint PET-MR registration, and assessing the robustness of SUVR, the current clinical method for brain amyloid quantification. We showed that SUVR is robust to the registration method used to align the PET images to the template space, and therefore, registration has little effect on the ability of SUVR to classify subjects as healthy or diseased. In Chapters 5-7, we investigated several other methods for amyloid status classification, and proposed a machine learning method that can achieve high classification accuracy regardless of the amyloid PET tracer.

It may be difficult to achieve more accurate amyloid status classification results than those presented in this thesis. However, knowledge of amyloid status alone is not sufficient to diagnose AD. For accurate diagnosis of AD, careful patient history and examination are required in addition to brain amyloid status. Therefore, we have conducted preliminary research into combining our amyloid status classification results with data from several other clinical tests, in order provide a more complete overview of an individual's disease state. The results presented in Chapter 8 should serve as a benchmark for subsequent research, and any future AD classification frameworks should fulfil the criteria proposed in our conclusion to that work.

List of Publications

Cattell, L., J.A. Schnabel, J. Declerck, and C. Hutton (2014). "Combined PET-MR Brain Registration to Discriminate between Alzheimer's Disease and Healthy Controls." In: *Biomedical Image Registration - 6th International Workshop, WBIR*, pp. 134–143.

— (2015). "Investigation of Single- Versus Joint-Modality PET-MR Registration for 18F-Florbetapir Quantification: Application to Alzheimers Disease". In: *Computational Methods for Molecular Imaging*. Vol. 22, pp. 197–205.

Cattell, L., G. Platsch, R. Pfeiffer, J. Declerck, J.A. Schnabel, and C. Hutton (2016). "Classification of Amyloid Status Using Machine Learning with Histograms of Oriented 3D Gradients". In: *NeuroImage: Clinical*. In press.

Bibliography

ADI (2009). *World Alzheimer Report*. Tech. rep. Alzheimer's Disease International.

ADNI (2008). *ADNI-2: Defining Alzheimer's Disease*. Tech. rep. Alzheimer's Disease Neuroimaging Initiative.

Altman, N.S. (1992). "An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression". In: *The American Statistician* 46.3, pp. 175–185.

Alzheimer, A. (1907). "Über eine eigenartige Erkrankung der Hirnrinde". In: *Allgemeine Zeitschrift fur Psychiatrie und Psychisch-gerichtliche Medizin* 64, pp. 146–148.

Alzheimer's Association (2014). "Alzheimer's Association Report: 2014 Alzheimer's disease facts and figures". In: *Alzheimer's and Dementia* 10.2, e47–e92.

Amyvid (2012). *Amyvid (florbetapir F 18 injection) highlights of prescribing information*. `http://www.accessdata.fda.gov/drugsatfda_docs/label/2012/202008s000lbl.pdf`. Accessed: Sep 17, 2015.

Anoop, A, Pradeep K Singh, Reeba S Jacob, and Samir K Maji (2010). "CSF Biomarkers for Alzheimer's Disease Diagnosis." In: *International journal of Alzheimer's disease* 2010, pp. 1–12.

Arai, T., K. Iked, H. Akiyama, Y. Shikamoto, K. Tsuchiya, S. Yagishita, T. Beach, J. Rogers, C. Schwab, and P.L. McGeer (2001). "Distinct isoforms of tau aggregated in neurons and glial cells in brains of patients with Pick's disease, corticobasal degeneration and progressive supranuclear palsy". In: *Acta Neuropathologica* 101.2, pp. 167–173.

Ashburner, J. (2007). "A fast diffeomorphic image registration algorithm". In: *NeuroImage* 38.1, pp. 95–113.

Ashburner, J. and K.J. Friston (2005). "Unified segmentation". In: *NeuroImage* 26, pp. 839–851.

Avants, B.B., C.L. Epstein, M. Grossman, and J.C. Gee (2008). "Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain". In: *Medical Image Analysis* 12.1, pp. 26–41.

Bar, Y., I. Diamant, L. Wolf, S. Lieberman, E. Konen, and H. Greenspan (2015). "Chest pathology detection using deep learning with non-medical training". In: *IEEE 12th International Symposium on Biomedical Imaging (ISBI) 2015*, pp. 294–297.

Barthel, H., H.J. Gertz, S. Dresel, O. Peters, P. Bartenstein, K. Buerger, F. Hiemeyer, S.M. Wittemer-Rump, J. Seibyl, C. Reininger, and O. Sabri (2011). "Cerebral amyloid-β PET with florbetaben (18F) in patients with Alzheimer's disease and healthy controls: A multicentre phase 2 diagnostic study". In: *Lancet Neurology* 10.5, pp. 424–435.

Batmanghelich, N.K., B. Taskar, and C. Davatzikos (2012). "Generative-discriminative basis learning for medical imaging". In: *IEEE Transactions on Medical Imaging* 31.1, pp. 51–69.

Bay, H., T. Tuytelaars, and L. Van Gool (2006). "SURF: Speeded Up Robust Features". In: *Ninth European Conference on Computer Vision*. Vol. 3951, pp. 404–417.

Beach, T.G, S.E Monsell, L.E. Phillips, and W. Kukull (2012). "Accuracy of the Clinical Diagnosis of Alzheimer Disease at National Institute on Aging Alzheimer's Disease Centers, 20052010". In: *Journal of Neuropathology and Experimental Neurology* 71.4, pp. 266–273.

Belongie, S. and J. Malik (2000). "Matching with shape contexts". In: *Workshop on Content-Based Access of Image and Video Libraries*, pp. 20–26.

Bengio, Y., A. Courville, and P. Vincent (2014). "Representation Learning: A Review and New Perspectives". In: *arXiv preprint arXiv:1206.5538*, pp. 1–30.

Bierer, L.M., P.R. Hof, D.P. Purohit, L. Carlin, J. Schmeidler, K.L. Davis, and D.P. Perl (1995). "Neocortical Neurofibrillary Tangles Correlate with Dementia Severity in Alzheimer's Disease". In: *Archives of Neurology* 52.1, pp. 81–88.

Birks, J (2006). "Cholinesterase inhibitors for Alzheimers disease". In: *Cochrane Database of Systematic Reviews* 1.

Blennow, K., A. Wallin, H. Agren, C. Spenger, J. Siegfried, and E. Vanmechelen (1995). "CSF total tau, Aβ42 and phosphorylated tau protein as biomarkers for Alzheimer's disease." In: *Molecular and Chemical Neuropathology* 3.3, pp. 231–245.

Blennow, K., E. Vanmechelen, and H. Hampel (2001). "CSF total tau, Aβ42 and phosphorylated tau protein as biomarkers for Alzheimer's disease." In: *Molecular neurobiology* 24.1-3, pp. 87–97.

Bookstein, F.L. (1989). "Principal Warps: Thin-Plate Splines and the Decomposition of Deformations". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 11.6, pp. 567–585.

Boser, B.E., I.M. Guyon, and V.N. Vapnik (1992). "A training algorithm for optimal margin classifiers". In: *Fifth Annual Workshop on Computational Learning Theory*, pp. 144–152.

Bossa, M., M. Hernandez, and S. Olmos (2007). "Contributions to 3D diffeomorphic atlas estimation: application to brain images". In: pp. 667–674.

Braak, H. and E. Braak (1989). "Neuropathological stageing of Alzheimer-related changes". In: *Acta Neuropathologica* 82, pp. 239–259.

Breiman, L. (2001). "Random Forests". In: *Machine Learning* 45.1, pp. 5–32.

Brendel, M., M. Högenauer, A. Delker, P. Bartenstein, and A. Rominger (2014). "Longitudinal amyloid PET in mild cognitive impaired patients". In: *Journal of Nuclear Medicine* 55.supplement 1, p. 193.

Bro-Nielsen, M. and C. Gramkow (1996). "Fast fluid registration of medical images". In: *Proceedings of the International Conference on Visualization in Biomedical Computing (VBC)*. Vol. 1131, pp. 267–276.

Brookmeyer, R., E. Johnson, K. Ziegler-Graham, and H.M. Arrighi (2007). "Forecasting the global burden of Alzheimer's disease". In: *Alzheimer's and Dementia* 3.3, pp. 186–191.

Brosch, T., Y. Yoo, D.K. Li, A. Traboulsee, and R. Tam (2014). "Modeling the variability in brain morphology and lesion distribution in multiple sclerosis by deep learning". In: *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention* 17.Pt 2, pp. 462–469.

Cachier, P., E. Bardinet, D. Dormont, X. Pennec, and N. Ayache (2003). "Iconic feature based nonrigid registration: the PASHA algorithm". In: *Computer Vision and Image Understanding* 89, pp. 272–298.

Cattell, L., J.A. Schnabel, J. Declerck, and C. Hutton (2014). "Combined PET-MR Brain Registration to Discriminate between Alzheimer's Disease and Healthy Controls." In: *Biomedical Image Registration - 6th International Workshop, WBIR*, pp. 134–143.

— (2015). "Investigation of Single- Versus Joint-Modality PET-MR Registration for 18F-Florbetapir Quantification: Application to Alzheimers Disease". In: *Computational Methods for Molecular Imaging*. Vol. 22, pp. 197–205.

Chang, C.-C. and C.-J. Lin (2011). "LIBSVM: A library for support vector machines". In: *ACM Transactions on Intelligent Systems and Technology* 2.3. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`, 27:1–27:27.

Chefd'hotel, C., G. Hermosillo, and O. Faugeras (2002). "Flows of diffeomorphisms for multimodal image registration". In: *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 753–756.

Christensen, G.E. (1994). "Deformable shape models for anatomy". PhD thesis. Washington University, Server Institute of Technology.

Clarfield, A.M. (2003). "The Decreasing Prevalence of Reversible Dementias: An Updated Meta-analysis". In: *Archives of Internal Medicine* 163.18, pp. 2219–2229.

Cohen, A.D., W. Mowrey, L.A. Weissfeld, H.J. Aizenstein, E. McDade, J.M. Mountz, R.D. Nebes, J.A. Saxton, B. Snitz, S. DeKosky, J. Williamson, O.L. Lopez, J.C. Price, C.A. Mathis, and W.E. Klunk (2013). "Classification of amyloid-positivity in controls: Comparison of visual read and quantitative approaches". In: *NeuroImage* 71, pp. 207–215.

Cortes, C. and V. Vapnik (1995). "Support-Vector Networks". In: *Machine Learning* 20.3, pp. 273–297.

Cselényi, Z. and L. Farde (2015). "Quantification of blood flow-dependent component in estimates of beta-amyloid load obtained using quasi-steady-state standardized uptake value ratio". In: *Journal of Cerebral Blood Flow & Metabolism* 35, pp. 1485–1493.

Cummings, J.L., T. Morstorf, and K. Zhong (2014). "Alzheimer's disease drug-development pipeline: few candidates, frequent failures." In: *Alzheimer's research & therapy* 6.4, p. 37.

Dalal, N. and D. Triggs (2005). "Histograms of oriented gradients for human detection". In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Vol. 1, pp. 886–893.

Davis, K.L., L.J. Thal, E.R. Gamzu, C.S. Davis, R.F. Woolson, S.I. Gracon, D.A. Drachman, L.S. Schneider, Whitehouse P.J., Hoover T.M., Morris J.C., Kawas C.H., Knopman D.S., Kumar V., Doody R.S., Earl N.L., Taylor S.V., and Hinman-Smith E.A. (1992). "A double-blind, placebo-controlled multicenter study of tacrine for Alzheimer's disease". In: *New England Journal of Medicine* 327.18, pp. 1253–1259.

DeLong, E.R., D.M. DeLong, and D.L. Clarke-Pearson (1988). "Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach." In: *Biometrics* 44.3, pp. 837–845.

Deng, L. and D. Yu (2013). "Deep Learning: Methods and Applications". In: *Foundations and Trends in Signal Processing*. Vol. 7. 3–4. Now Publishers.

Desjardins, G. and Y. Bengio (2008). *Empirical Evaluation of Convolutional RBMs for Vision*. Tech. rep. 1327. Département d'Informatique et de Recherche Opérationnelle, Université de Montréal, pp. 1–13.

Drucker, H., D. Wu, and V. Vapnik (1999). "Support vector machines for spam categorization". In: *IEEE Transactions on Neural Networks* 10.5, pp. 1048–1054.

Edison, P., S.F. Carter, J.O. Rinne, G. Gelosa, K. Herholz, A. Nordberg, D.J. Brooks, and R. Hinz (2013). "Comparison of MRI based and PET template based approaches in the quantitative analysis of amyloid imaging with PIB-PET". In: *NeuroImage* 70, pp. 423–433.

EMA (2013). *Public Assessment Report: Neuraceq Florbetaben 18 F.* http://www.ema.europa.eu/docs/en_GB/document_library/EPAR_-_Public_assessment_report/human/002553/WC500162593.pdf. Accessed: Sep 17, 2015.

Farinha, R.J.P.C., U. Ruotsalainen, J. Hirvonen, L. Tuominen, J. Hietala, J.M. Fonseca, and J. Tohka (2009). "Segmentation of striatal brain structures from high resolution PET images". In: *International journal of biomedical imaging* 2009, pp. 1–12.

FDA (2010). *FDA Advisory Committee Briefing Document: New Drug Application 202-008 Amyvid (Florbetapir F 18 Injection).* http://www.fda.gov/downloads/Advisory%20Committees/CommitteesMeetingMaterials/Drugs/Peripheraland%20CentralNervousSystemDrugsAdvisoryCommittee/UCM240265.pdf. Accessed: Sep 15, 2015.

— (2012). "FDA Approves 18F-Florbetapir PET Agent". In: *Journal of Nuclear Medicine* 53.6, 15N.

— (2013). *FDA Press Release: Vizamyl approval.* http://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/ucm372261.htm. Accessed: Sep 15, 2015.

Fei-Fei, L., R. Fergus, and P. Perona (2004). "Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories." In: *IEEE Conference on Computer Vision and Pattern Recognition 2004, Workshop on Generative-Model Based Vision.*

Feldman, H., S. Gauthier, J. Hecker, B. Vellas, P. Subbiah, and E. Whalen (2001). "A 24-week, randomized, double-blind study of donepezil in moderate to severe Alzheimer's disease". In: *Neurology* 57.4, pp. 613–620.

Felzenszwalb, P.F., R.B. Girshick, D. McAllester, and D. Ramanan (2010). "Object detection with discriminatively trained part-based models". In: *Transactions on Pattern Analysis and Machine Intelligence* 32.9, pp. 1627–1644.

Fergus, R., P. Perona, and A. Zisserman (2005). "Object Class Recognition by Unsupervised Scale-Invariant Learning". In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition.* Vol. 2, pp. 264–271.

Fischl, B., D.H. Salat, E. Busa, M. Albert, M. Dieterich, C. Haselgrove, A. van der Kouwe, R. Killiany, D. Kennedy, S. Klaveness, A. Montillo, N. Makris, B. Rosen, and A.M. Dale (2002). "Whole Brain Segmentation: Automated Labeling of Neuroanatomical Structures in the Human Brain". In: *Neuron* 33.3, pp. 341–355.

Fleisher, A.S., K. Chen, X. Liu, A. Roontiva, P. Thiyyagura, N. Ayutyanont, A.D. Joshi, C.M. Clark, M.A. Mintun, M.J. Pontecorvo, P.M. Doraiswamy, K.A. Johnson, D.M. Skovronsky, and E.M. Reiman (2011). "Using Positron Emission Tomography and Florbetapir F 18 to Image Cortical Amyloid in Patients With Mild Cognitive Impairment or Dementia Due to Alzheimer Disease". In: *Archives of Neurology* 68.11, pp. 1404–1411.

Fodero-Tavoletti, M.T., C.C. Rowe, C.A. McLean, L. Leone, Q.X. Li, C.L. Masters, R. Cappai, and V.L. Villemagne (2009). "Characterization of PiB binding to white matter in Alzheimer disease and other dementias." In: *Journal of Nuclear Medicine* 50.2, pp. 198–204.

Folstein, M.F., S.E. Folstein, and P.R. McHugh (1975). ""Mini-mental state". A practical method for grading the cognitive state of patients for the clinician". In: *Journal of Psychiatric Research* 12.3, pp. 189–198.

Forsberg, D., Y. Rathi, S. Bouix, D. Wasserman, H. Knutsson, and C.F. Westin (2011). "Improving Registration Using Multi-channel Diffeomorphic Demons Combined with Certainty Maps". In: *Multimodal Brain Image Analysis*. Vol. 7012, pp. 19–26.

Frantz, N. and S. Farley (2015). *Major New Research Study to Demonstrate Value of PET Scans in Alzheimer's Disease Diagnosis*. `http://www.alz.org/documents_custom/IDEAS_study_news_release_041615.pdf`. Accessed: Feb 16, 2016.

Frey, K.A. (2015). "Amyloid Imaging in Dementia: Contribution or Confusion?" In: *Journal of Nuclear Medicine* 56.3, pp. 331–333.

Frisoni, G.B., N.C. Fox, C.R. Jack, P. Scheltens, and P.M. Thompson (2010). "The clinical use of structural MRI in Alzheimer disease". In: *Nature Reviews Neurology* 6.2, pp. 67–77.

Fung, G., S. Sandilya, and R.B. Rao (2005). "Rule Extraction from Linear Support Vector Machines". In: *Proceedings of the 11th International Conference on Knowledge Discovery in Data Mining*, pp. 32–50.

Gao, Y., E. Adeli-m, and M. Kim (2015). "Medical Image Retrieval Using Multigraph Learning for MCI Diagnostic Assistance". In: *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015* 9350, pp. 86–93.

Goedert, M. and M.G. Spillantini (2000). "Tau mutations in frontotemporal dementia FTDP-17 and their relevance for Alzheimers disease". In: *Biochimica et Biophysica Acta (BBA) – Molecular Basis of Disease* 1502.1, pp. 110–121.

Grabner, G., A.L. Janke, M.M. Budge, D. Smith, J. Pruessner, and D.L. Collins (2006). "Symmetric Atlasing and Model Based Segmentation: An Application to the Hippocampus in Older Adults". In: *Medical Image Computing and Computer-Assisted Intervention MICCAI 2006*. Vol. 4191, pp. 58–66.

Guimond, A., J. Meunier, and J.-P. Thirion (2000). "Average brain models: A convergence study". In: *Computer Vision and Image Understanding* 77.2, pp. 192–210.

Guimond, A., A. Roche, N. Ayache, and J. Meunier (2001). "Three-Dimensional Multimodal Brain Warping Using the Demons Algorithm and Adaptive Intensity Corrections". In: *Transactions on Medical Imaging* 20.1, pp. 58–69.

Guo, Y., G. Wu, L.A. Commander, S. Szary, V. Jewells, W. Lin, and D. Shen (2014). "Segmenting hippocampus from infant brains by sparse patch matching with deep-learned features". In: *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention* 17.Pt 2, pp. 308–315.

Hang, Y., P. Chiao, J. Sevigny, L. Williams, X. Miao, J. OGorman, and J. Ferrero (2015). "Pharmacokinetic and Pharmacodynamic (PK/PD) Assessment and Covariate Analysis for Aducanumab (BIIB037) in Patients with Prodromal or Mild Alzheimers Disease: Interim Results of a Randomized, Double-Blind, Placebo-Controlled, Phase 1b Study". In: *Proceedings of the Alzheimer's Association International Conference*.

Heckemann, R.A., J.V. Hajnal, P. Aljabar, D. Rueckert, and A. Hammers (2006). "Automatic anatomical brain MRI segmentation combining label propagation and decision fusion". In: *NeuroImage* 33, pp. 115–126.

Hinton, G.E. (2002). "Training products of experts by minimizing contrastive divergence". In: *Neural Computation* 14.8, pp. 1771–1800.

— (2012). "A Practical Guide to Training Restricted Boltzmann Machines". In: *Neural Networks: Tricks of the Trade*. Vol. 7700. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 599–619.

Hinton, G.E., S. Osindero, and Y. Teh (2006). "A fast learning algorithm for deep belief nets". In: *Neural Computation* 18.7, pp. 1527–1554.

Holmes, C.J, R. Hoge, L. Collins, R. Woods, A.W. Toga, and A.C. Evans (1998). "Enhancement of MR images using registration for signal averaging". In: *Journal of Computer Assisted Tomography* 22.2, pp. 324–333.

Horn, B.K.P. and B.G. Schunck (1981). "Determining optical flow". In: *Artificial Intelligence* 71, pp. 185–203.

Hulstaert, F., K. Blennow, A. Ivanoui, H.C. Schoonderwaldt, M. Riemenschneider, P.P. De Deyn, C. Bancher, P. Cras, J. Wiltfang, P.D. Mehta, K. Iqbal, H. Pottel, E. Vanmechelen, and H. Vanderstichele (1999). "Improved discrimination of AD patients using beta-amyloid(1-42) and tau levels in CSF". In: *Neurology* 52.8, pp. 1555–1562.

Hutton, C., L. Sibille, S. Bullich, A. Catafau, N. Koglin, R. Pfeiffer, and J. Declerck (2014). "Comparison of two methods, with and without MRI, for quantification

of florbetaben (18F) PET". In: *Annual Congress of the European Association of Nuclear Medicine*.

Hutton, C., J. Declerck, M.A. Mintun, M.J. Pontecorvo, M.D. Devous Sr, and A.D. Joshi (2015). "Quantification of 18F-florbetapir PET: comparison of two analysis methods". In: *European Journal of Nuclear Medicine and Molecular Imaging* 42.5, pp. 725–732.

Jack, C.R., D.S. Knopman, W.J. Jagust, L.M. Shaw, P.S. Aisen, M.W. Weiner, R.C. Petersen, and J.Q. Trojanowski (2010). "Hypothetical model of dynamic biomarkers of the Alzheimer's disease pathological cascade". In: *Lancet Neurology* 9.1, pp. 119–128.

Jack, C.R. Jr, V.J. Lowe, M.L. Senjem, S.D. Weigand, B.J. Kemp, M.M. Shiung, D.S. Knopman, B.F. Boeve, W.E. Klunk, C.A. Mathis, and R.C. Petersen (2008). "11C PiB and structural MRI provide complementary information in imaging of Alzheimer's disease and amnestic mild cognitive impairment." In: *Brain* 131(Pt3), pp. 665–680.

Jaderberg, M., A. Vedaldi, and A. Zisserman (2014). "Speeding up Convolutional Neural Networks with Low Rank Expansions". In: *Proceedings of the British Machine Vision Conference*, pp. 1–7.

Jagust, W.J., S.M. Landau, L.M. Shaw, J.Q. Trojanowski, R.A. Koeppe, E.M. Reiman, N.L. Foster, R.C. Petersen, M.W. Weiner, J.C. Price, and C.A. Mathis (2009). "Relationships between biomarkers in aging and dementia". In: *Neurology* 73.15, pp. 1193–1199.

Jenkinson, M. and S.M. Smith (2001). "A global optimisation method for robust affine registration of brain images". In: *Medical Image Analysis* 5.2, pp. 143–156.

Jenkinson, M., P.R. Bannister, J.M. Brady, and S.M. Smith (2002). "Improved optimisation for the robust an accurate linear registration an motion correction of brain images". In: *NeuroImage* 17.2, pp. 825–841.

Joshi, A.D., M.J. Pontecorvo, C.M. Clark, A.P. Carpenter, D.L. Jennings, C.H. Sadowsky, L.P. Adler, K.D. Kovnat, J.P. Seibyl, A. Arora, K. Saha, J.D. Burns, M.J. Lowrey, M.A. Mintun, and D.M. Skovronsky (2012). "Performance characteristics of amyloid PET with florbetapir F 18 in patients with alzheimer's disease and cognitively normal subjects". In: *Journal Nuclear Medicine* 53.3, pp. 378–384.

Keerthi, S. and C. Lin (2003). "Asymptotic behaviors of support vector machines with Gaussian kernel." In: *Neural Computation* 15.7, pp. 1667–89.

Kepe, V., M.C. Moghbel, B. Langstrom, H. Zaidi, H.V. Vinters, S.C. Huang, N. Satyamurthy, D. Doudet, E. Mishani, R.M. Cohen, P.F. Hoilund-Carlsen, A. Alavi, and J.R. Barrio (2013). "Amyloid-β Positron Emission Tomography Imag-

ing Probes: A Critical Review". In: *Journal of Alzheimer's Disease* 36.4, pp. 613–631.

Kiebel, S.J., J. Ashburner, J.-B. Poline, and K.J. Friston (1997). "MRI and PET Coregistration – A Cross Validation of Statistical Parametric Mapping and Automated Image Registration". In: *NeuroImage* 5.4, pp. 271–279.

Kläser, A., M. Marszalek, and C. Schmid (2008). "A Spatio-Temporal Descriptor Based on 3D-Gradients". In: *Proceedings of the British Machine Vision Conference*, pp. 1–10.

Klein, A., J. Andersson, B.A. Ardekani, J. Ashburner, B. Avants, M.C. Chiang, G.E. Christensen, D.L. Collins, J. Gee, P. Hellier, J.H. Song, M. Jenkinson, C. Lepage, D. Rueckert, P. Thompson, T. Vercauteren, R.P. Woods, J.J. Mann, and R.V. Parsey (2009). "Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration." In: *NeuroImage* 46.3, pp. 786–802.

Klunk, W. E., H. Engler, A. Nordberg, Y. Wang, G. Blomqvist, D. P. Holt, M. Bergstrom, I. Savitcheva, G. F. Huang, S. Estrada, B. Ausen, M. L. Debnath, J. Barletta, J. C. Price, J. Sandell, B. J. Lopresti, A. Wall, P. Koivisto, G. Antoni, C. A. Mathis, and B. Langstrom (2004). "Imaging brain amyloid in Alzheimer's disease with Pittsburgh Compound-B". In: *Annals of Neurology* 55.3, pp. 306–319.

Klunk, W.E., R.A. Koeppe, J.C. Price, T.L. Benzinger, M.D. Sr Devous, W.J. Jagust, K.A. Johnson, C.A. Mathis, D. Minhas, M.J. Pontecorvo, C.C. Rowe, D.M. Skovronsky, and M.A. Mintun (2015). "The Centiloid Project: standardizing quantitative amyloid plaque estimation by PET". In: *Alzheimers & Dementia* 11.1, pp. 1–15.

Knopman, D.S., S.T. DeKosky, J.L. Cummings, H. Chui, J. Corey-Bloom, N. Relkin, G.W. Small, B. Miller, and J.C. Stevens (2001). "Practice parameter: diagnosis of dementia (an evidence-based review)– Report of the Quality Standards Subcommittee of the American Academy of Neurology". In: *Neurology* 56.9, pp. 1143–1153.

Komaroff, A. (2015). *How does Alzheimers wreak so much havoc in the brain?* `http://www.askdoctork.com/how-does-alzheimers-wreak-so-much-havoc-in-the-brain-201506258011`. Accessed: Sep 23, 2015.

Krizhevsky, A. (2010). "Convolutional deep belief networks on CIFAR-10". `http://www.cs.toronto.edu/~kriz/conv-cifar10-aug2010.pdf`.

Krizhevsky, A., I. Sutskever, and G.E. Hinton (2012). "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems*. Vol. 25, pp. 1097–1105.

Landau, S.M., C. Breault, A.D. Joshi, M. Pontecorvo, C.A. Mathis, W.J. Jagust, and M.A. Mintun (2013). "Amyloid-β imaging with Pittsburgh compound B and

florbetapir: comparing radiotracers and quantification methods." In: *Journal of Nuclear Medicine* 54.1, pp. 70–77.

LeCun, Y., L. Bottou, Y. Bengio, and P. Haffner (1998). "Gradient-Based Learning Applied to Document Recognition". In: *Proceedings of the IEEE* 86.11, pp. 2278–2324.

Lee, H., R. Grosse, R. Ranganath, and A.Y. Ng (2009a). "Convolutional Deep Belief Networks for Scalable Unsupervised Learning of Hierarchical Representations". In: *Proceedings of the 26th international conference on Machine learning*, pp. 609–616.

Lee, H., J. Lee, N. Kim, I.K. Lyoo, and Y.G. Shin (2009b). "Robust and fast shell registration in PET and MR/CT brain images". In: *Computers in Biology and Medicine* 39.11, pp. 961–977.

Levy-Lahad, E. and T.D. Bird (1996). "Genetic factors in Alzheimer's disease: a review of recent advances". In: *Annals of neurology* 40.127, pp. 829–840.

Lin, W.J. and J.J. Chen (2013). "Class-imbalanced classifiers for high-dimensional data". In: *Briefings in Bioinformatics* 14.1, pp. 13–26.

Lipton, S.A. (2004). "Paradigm shift in NMDA receptor antagonist drug development: molecular mechanism of uncompetitive inhibition by memantine in the treatment of Alzheimer's disease and other neurologic disorders". In: *Journal of Alzheimer's disease* 6.6 Suppl. S61–S74.

Liu, C., J. Yuen, A. Torralba, J. Sivic, and W.T. Freeman (2008). "SIFT Flow: Dense Correspondence across Different Scenes". In: *Proceedings of the European Conference on Computer Vision*. Vol. 3, pp. 28–42.

Liu, M., C. Bian, J. Zhang, and F. Wen (2014). "Apolipoprotein E gene polymorphism and Alzheimer's disease in Chinese population: a meta-analysis". In: *Nature: Scientific Reports* 4.4383, pp. 1–7.

Lopresti, B.J., W.E. Klunk, C.A. Mathis, J.A. Hoge, S.K. Ziolko, X. Lu, C.C. Meltzer, K. Schimmel, N.D. Tsopelas, S.T. DeKosky, and J.C. Price (2005). "Simplified Quantification of Pittsburgh Compound B Amyloid Imaging PET Studies: A Comparative Analysis". In: *Journal of Nuclear Medicine* 46.12, pp. 1959–1972.

Lorenzi, M., N. Ayache, G. Frisoni, and X. Pennec (2013). "LCC-Demons: A robust and accurate symmetric diffeomorphic registration algorithm". In: *NeuroImage* 81, pp. 470–483.

Lowe, D.G. (1999). "Object recognition from local scale-invariant features". In: *Proceedings of the International Conference on Computer Vision*. Vol. 2, pp. 1150–1157.

Lundqvist, R., J. Lilja, B.A. Thomas, J. Lotjonen, V. L. Villemagne, C.C. Rowe, and L. Thurfjell (2013). "Implementation and Validation of an Adaptive Template Registration Method for 18F-Flutemetamol Imaging Data". In: *Journal of Nuclear Medicine* 54.8, pp. 1472–1478.

Maruyama, M., H. Shimada, T. Suhara, H. Shinotoh, B. Ji, J. Maeda, M.R. Zhang, J.Q. Trojanowski, V.M.Y. Lee, M. Ono, K. Masamoto, H. Takano, N. Sahara, N. Iwata, N. Okamura, S. Furumoto, Y. Kudo, Q. Chang, T.C. Saido, A. Takashima, J. Lewis, M.K. Jang, I. Aoki, H. Ito, and M. Higuchi (2013). "Imaging of tau pathology in a tauopathy mouse model and in Alzheimer patients compared to normal controls". In: *Neuron* 79.6, pp. 1094–1108.

Mattila, J., J. Koikkalainen, A. Virkki, A. Simonsen, M. van Gils, G. Waldemar, H. Soininen, and J. Lötjönen (2011). "A disease state fingerprint for evaluation of Alzheimer's disease." In: *Journal of Alzheimer's Disease* 27.1, pp. 163–76.

Mattson, M.P. (2004). "Pathways Towards and Away from Alzheimers Disease". In: *Nature* 430.7000, pp. 631–639.

Mazziotta, J., A. Toga, A. Evans, P. Fox, J. Lancaster, K. Zilles, R. Woods, T. Paus, G. Simpson, B. Pike, C. Holmes, L. Collins, P. Thompson, D. MacDonald, M. Iacoboni, T. Schormann, K. Amunts, N. Palomero-Gallagher, S. Geyer, L. Parsons, K. Narr, N. Kabani, G.L. Goualher, D. Boomsma, T. Cannon, R. Kawashima, and B. Mazoyer (2001). "A probabilistic atlas and reference system for the human brain: International Consortium for Brain Mapping (ICBM)". In: *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 356.1412, pp. 1293–1322.

McKhann, G., D. Drachman, M. Folstein, R. Katzman, D. Price, and E.M. Stadlan (1984). "Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease". In: *Neurology* 34.7, pp. 939–944.

McKhann, G.M., D.S. Knopman, H. Chertkow, B.T. Hyman, C.R. Jack Jr, C.H. Kawas, W.E. Klunk, W.J. Koroshetz, J.J. Manly, R. Mayeux, R.C. Mohs, J.C. Morris, M.N. Rossor, P. Scheltens, M.C. Carrillo, B. Thies, S. Weintraub, and C.H. Phelps (2011). "The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute of Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease". In: *Alzheimer's and Dementia* 7.3, pp. 263–269.

McNemar, Q. (1947). "Note on the sampling error of the difference between correlated proportions or percentages". In: *Psychometrika* 12.2, pp. 153–157.

Mehta, M., A. Adem, and M. Sabbagh (2012). "New acetylcholinesterase inhibitors for Alzheimer's disease." In: *International Journal of Alzheimer's Disease* 2012, pp. 1–8.

Mohamed, A., G.E. Dahl, and G. Hinton (2012). "Acoustic Modeling Using Deep Belief Networks". In: *Transactions on Audio, Speech, and Language Processing* 20.1, pp. 14–22.

Mohs, R.C., D. Knopman, R.C. Petersen, S.H. Ferris, C. Ernesto, M. Grundman, M. Sano, L. Bieliauskas, D. Geldmacher, C. Clark, and L.J. Thal (1997). "Development of cognitive instruments for use in clinical trials of antidementia drugs: additions to the Alzheimer's Disease Assessment Scale that broaden its scope. The Alzheimer's Disease Cooperative Study." In: *Alzheimer Disease and Associated Disorders* 11.Suppl. 2, S13–21.

Mott, R.T. and C.M. Hulette (2011). "Neuropathology of Alzheimers Disease". In: *Neuroimaging Clinics of North America* 15, pp. 755–765.

Musiek, E.S. and Y. Chen (2012). "Direct Comparison of FDG-PET and ASL-MRI in Alzheimer's Disease". In: *Alzheimer's & Dementia* 8.1, pp. 51–59.

Nayate, A.P., J.G. Dubroff, J.E Schmitt, I. Nasrallah, R. Kishore, D. Mankoff, and D.A. Pryma (2015). "Use of Standardized Uptake Value Ratios Decreases Interreader Variability of 18F Florbetapir PET Brain Scan Interpretation". In: *American Journal of Neuroradioly*, pp. 1–8.

Neuraceq (2014). *NEURACEQ (florbetaben F 18 injection) highlights of prescribing information*. `http://www.accessdata.fda.gov/drugsatfda_docs/nda/2014/204677Orig1s000Lbl.pdf`. Accessed: Sep 17, 2015.

Olshausen, B.A. and D.A. Field (1996). "Emergence of simple-cell receptive field properties by learning a sparse code for natural images". In: *Nature* 381, pp. 607–609.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

Pennec, X., P. Cachier, and N. Ayache (1999). "Understanding the "Demon's Algorithm": 3D Non-Rigid registration by Gradient Descent". In: *Proceedings of the International Conference on Medical Image Computing and Computed-Assisted Intervention (MICCAI)*, pp. 597–605.

Perl, D.P. (2010). "Neuropathology of Alzheimer's Disease". In: *Mt Sinai Journal of Medicine* 77.1, pp. 32–42.

Peskind, E.R., R. Riekse, J.F. Quinn, J. Kaye, C.M. Clark, M.R. Farlow, C. DeCarli, C. Chabal, D. Vavrek, M.A. Raskind, and D. Galasko (2005). "Safety and Acceptability of the Research Lumbar Puncture". In: *Alzheimer Disease & Associated Disorders* 19.4, pp. 220–225.

Peyrat, J-.M., H. Delingette, M. Sermesant, C. Xu, and N. Ayache (2010). "Registration of 4D cardiac CT sequences under trajectory constraints with multichannel diffeomorphic demons". In: *Transactions on Medical Imaging* 29.7, pp. 1351–1368.

Peyrat, J-.M., A. Joshi, M. Mintun, and J. Declerck (2012). "An automatic method for the quantification of uptake with Florbetapir imaging". In: *Journal of Nuclear Medicine Meeting Abstracts* 53.1 Meeting Abstracts, p. 210.

Pike, V.W. (2009). "PET Radiotracers: crossing the blood-brain barrier and surviving metabolism". In: *Trends in Pharmacological Sciences* 30.8, pp. 431–440.

Piramal Enterprises Ltd (2014). *FDA approves Piramal Imaging's Neuraceq (florbetaben F18 injection) for PET imaging of beta-amyloid neuritic plaques in the brain.* `http://www.piramal.com/imaging/pdf/FDA-Approval-Press-Release.pdf`. Accessed: Sep 15, 2015.

Plis, S.M., D.R. Hjelm, R. Salakhutdinov, E.A. Allen, H.J. Bockholt, J.D. Long, H.J. Johnson, J.S. Paulsen, J.A. Turner, and V.D. Calhoun (2014). "Deep learning for neuroimaging: a validation study". In: *Frontiers in Neuroscience* 8, pp. 1–11.

Pontecorvo, M., M. Devous, A. Arora, M. Devine, M. Lu, A. Joshi, C. Breault, D. Skovronsky, M. Mintun, and S. Heun (2014). "Can incorporation of a quantitative estimate of cortical to cerebellar SUVr as an adjunct to visual interpretation improve the accuracy and reliability of florbetapir PET scan interpretation?" In: *Journal of Nuclear Medicine Meeting Abstracts* 55.1, p. 245.

Prince, M., M. Knapp, M. Guerchet, P. McCrone, M. Prina, A. Comas-Herrera, R. Wittenberg, B. Adelaja, B. Hu, D. King, A. Rehill, and D. Salimkumar (2014). *Dementia UK: Update.* Tech. rep. Alzheimer's Society.

Qizilbash, N., A. Whitehead, J. Higgins, G. Wilcock, L. Schneider, and M. Farlow (1998). "Cholinesterase inhibition for alzheimer disease: A meta-analysis of the tacrine trials". In: *Journal of the American Medical Association* 280.20, pp. 1777–1782.

Raniga, P., P. Bourgeat, J. Fripp, O. Acosta, V.L. Villemagne, C. Rowe, C.L. Masters, G. Jones, G. O'Keefe, O. Salvado, and S. Ourselin (2008). "Automated ^{11}C-PiB Standardized Uptake Value Ratio". In: *Academic Radiology* 15.11, pp. 1376–1389.

Ranzato, M'A., J. Susskind, V. Mnih, and G. Hinton (2011). "On deep generative models with applications to recognition". In: *24th Conference on Computer Vision and Pattern Recognition*, pp. 2857–2864.

Reisberg, B., R. Doody, A. Stöffler, F. Schmitt, S. Ferris, and H.J. MÖbius (2003). "Memantine in Moderate-to-Severe Alzheimer's Disease". In: *New England Journal of Medicine* 348.14, pp. 1333–1341.

Rigamonti, R., A. Sironi, V. Lepetit, and P. Fua (2013). "Learning separable filters". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2754–2761.

Roberts, R.O., Y.E. Geda, D.S. Knopman, R.H. Cha, V.S. Pankratz, B.F. Boeve, R.J. Ivnik, E.G. Tangalos, R.C. Petersen, and W.A. Rocca (2008). "The Mayo Clinic Study of Aging: Design and Sampling, Participation, Baseline Measures and Sample Characteristics". In: *Neuroepidemiology* 30.1, pp. 58–69.

Rogers, S.L., M.R. Farlow, R.S. Doody, R. Mohs, and L.T. Friedhoff (1998). "A 24-week, double-blind, placebo-controlled trial of donepezil in patients with Alzheimer's disease". In: *Neurology* 50.1, pp. 136–145.

Rosen, W.G., R.C. Mohs, and K.L. Davis (1984). "A new rating scale for Alzheimer's disease". In: *American Journal of Psychiatry* 141.11, pp. 1356–1364,

Rösler, M., R. Anand, A. Cicin-Sain, S. Gauthier, Y. Agid, P. Dal-Bianco, H.B. Stähelin, R. Hartman, and M. Gharabawi (1999). "Efficacy and safety of rivastigmine in patients with Alzheimer's disease: international randomised controlled trial." In: *BMJ (Clinical research edition)* 318, pp. 633–640.

Roth, H.R., L. Le, A. Farag, S. Hoo-Chang, L. Jiamin, E.B. Turkbey, and R.M. Summers (2015). "DeepOrgan: Multi-Level Deep Convolutional Networks for Automated Pancreas Segmentation". In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Vol. 9351, pp. 349–357.

Rowe, C.C. and V.L. Villemagne (2013). "Brain Amyloid Imaging". In: *Journal of Nuclear Medicine Technology* 41, pp. 11–18.

Rowe, C.C., U. Ackerman, W. Browne, R. Mulligan, K.L. Pike, G. O'Keefe, H. Tochon-Danguy, G. Chan, S.U. Berlangieri, G. Jones, K.L. Dickinson-Rowe, H.P. Kung, W. Zhang, M.P. Kung, D. Skovronsky, T. Dyrks, G. Holl, S. Krause, M. Friebe, L. Lehman, S. Lindemann, L.M. Dinkelborg, C.L. Masters, and V.L. Villemagne (2008). "Imaging of amyloid beta in Alzheimer's disease with 18F-BAY94-9172, a novel PET tracer: proof of mechanism". In: *Lancet Neurology* 7.2, pp. 129–135.

Rueckert, D., S.C. Hayes, D.L.G. Hill, O.M. Leach, and D.J. Hawkes (1999). "Non-rigid Registration Using Free-Form Deformations: Application to Breast MR Images". In: *IEEE Transactions on Medical Imaging* 18.8, pp. 712–721.

Rullmann, M., J. Dukart, K.T. Hoffmann, J. Luthardt, S. Tiepolt, M. Patt, H.J. Gertz, M.L. Schroeter, J. Seibyl, W.J. Schulz-Schaeffer, O. Sabri, and H. Barthel (2016). "Partial-Volume Effect Correction Improves Quantitative Analysis of 18F-Florbetaben β-Amyloid PET Scans". In: *Journal of Nuclear Medicine* 57.2, pp. 198–203.

Russakovsky, O., J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A.C. Berg, and L. Fei-Fei (2015). "ImageNet Large

Scale Visual Recognition Challenge". In: *International Journal of Computer Vision (IJCV)*, pp. 1–42.

Saint-Aubert, L., P. Nemmi, P. Peran, E.J. Barbeau, P. Payoux, F. Chollet, and J. Pariente (2014). "Comparison Between PET Template-Based Method and MRI-Based Method for Cortical Quantification of Florbetapir (AV-45) Uptake In Vivo." In: *European Journal of Nuclear Medicine and Molecular Imaging* 41.5, pp. 836–843.

Salakhutdinov, R., A. Mnih, and G.E. Hinton (2007). "Restricted Boltzmann machines for collaborative filtering". In: *Proceedings of the 24th International Conference on Machine learning*, pp. 791–798.

Schmitter, D., A. Roche, B. Maréchal, Delphine Ribes, A. Abdulkadir, M. Bach-Cuadra, A. Daducci, C. Granziera, S. Klöppel, P. Maeder, R. Meuli, and G. Krueger (2015). "An evaluation of volume-based morphometry for prediction of mild cognitive impairment and Alzheimer's disease". In: *NeuroImage: Clinical* 7, pp. 7–17.

Schneider, L.S., K.S. Dagerman, J.T. Higgins, and R. McShane (2011). "Lack of evidence for the efficacy of memantine in mild alzheimer disease". In: *Archives of Neurology* 68.8, pp. 991–998.

Segovia, F., C. Bastin, E. Salmon, J.M. Gorriz, J. Ramirez, and C. Phillips (2014). "Combining PET Images and Neuropsychological Test Data for Automatic Diagnosis of Alzheimers Disease". In: *PloS one* 9.2, pp. 1–8.

Sevigny, J., P. Chiao, L. Williams, T. Chen, L. Ling, J. OGorman, C. Hock, R.M. Nitsch, and A. Sandrock (2015). "Pharmacokinetic and Pharmacodynamic (PK/PD) Assessment and Covariate Analysis for Aducanumab (BIIB037) in Patients with Prodromal or Mild Alzheimers Disease: Interim Results of a Randomized, Double-Blind, Placebo-Controlled, Phase 1b Study". In: *Proceedings of the Alzheimer's Association International Conference*.

Simonyan, K. and A. Zisserman (2015). "Very Deep Convolutional Networks for Large-Scale Image Recognition". In: *International Conference on Learning Representations*.

Sivic, J. and A. Zisserman (2003). "Video Google: A Text Retrieval Approach to Object Matching in Videos". In: *Proceedings of the International Conference on Computer Vision*. Vol. 2, pp. 1470–1477.

Sjögren, M, P Davidsson, M Tullberg, L Minthon, A Wallin, C Wikkelso, A K Granérus, H Vanderstichele, E Vanmechelen, and K Blennow (2000). "CSF levels of tau, beta-amyloid(1-42) and GAP-43 in frontotemporal dementia, other types of dementia and normal aging". In: *Journal of Neural Transmission* 107.5, pp. 563–579.

— (2001). "Both total and phosphorylated tau are increased in Alzheimer's disease". In: *Journal Neurology Neurosurgery and Psychiatry* 70.5, pp. 624–630.

Skinner, J., J.O. Carvalho, G.G. Potter, A. Thames, E. Zelinski, P.K. Crane, and L.E. Gibbons (2012). "The Alzheimer's Disease Assessment Scale-Cognitive-Plus (ADAS-Cog-Plus): An expansion of the ADAS-Cog to improve responsiveness in MCI". In: *Brain Imaging and Behavior* 6.4, pp. 489–501.

Sotiras, A., C. Davatzikos, and N. Paragios (2013). "Deformable Medical Image Registration: A Survey". In: *IEEE Transactions on Medical Imaging* 32.7, pp. 1153–1190.

Stankoff, B., L. Freeman, M.S. Aigrot, A. Chardain, F. Dolle, A. Williams, D. Galanaud, L. Armand, S. Lehericy, C. Lubetzki, B. Zalc, and M. Bottlaender (2011). "Imaging central nervous system myelin by positron emission tomography in multiple sclerosis using [methyl-11C]-2-(4'-methylaminophenyl)- 6-hydroxybenzothiazole." In: *Annals of Neurology* 69.4, pp. 673–680.

Steele, L.S. and R.H. Glazier (1999). "Is donepezil effective for treating Alzheimer's disease?" In: *Canadian Family Physician* 45, pp. 917–919.

Stella, F., M. Radanovic, P. R. Canineu, V. J. R. de Paula, and O. V. Forlenza (2015). "Anti-dementia medications: current prescriptions in clinical practice and new agents in progress". In: *Therapeutic Advances in Drug Safety* 6.4, pp. 151–165.

Strittmatter, W.J. (2012). "Old Drug, New Hope for Alzheimers Disease". In: *Science* 335, pp. 1447–1448.

Studholme, C., D. Hill, and D. Hawkes (1996). "Automated 3-D registration of MR and CT images of the head". In: *Medical Image Analyis* 1, pp. 163–175.

Suk, H.I. and D. Shen (2013). "Deep Learning-Based Feature Representation for AD/MCI Classification". In: *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention* 16.Pt 2, pp. 583–590.

Suotunen, T., J. Hirvonen, P. Immonen-Räihä, S. Aalto, I. Lisinen, E. Arponen, M. Teräs, K. Koski, R. Sulkava, M. Seppänen, and J.O. Rinne (2010). "Visual assessment of [11C]PIB PET in patients with cognitive impairment". In: *European Journal of Nuclear Medicine and Molecular Imaging* 37.6, pp. 1141–1147.

Szegedy, C., W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich (2012). "Going deeper with convolutions". In: *arXiv preprint arXiv:1409.4842*, pp. 1–12.

Tariot, P.N., P.R. Solomon, J.C. Morris, P. Kershaw, S. Lilienfeld, C. Ding, and the Galantamine USA-Study Group (2000). "A 5-month, randomized, placebo-controlled trial of galantamine in AD". In: *Neurology* 54.12, pp. 2269–2276.

Taylor, G.W., G.E. Hinton, and S.T. Roweis (2007). "Modeling human motion using binary latent variables". In: *Advances in Neural Information Processing Systems* 19, pp. 1345–1353.

Terry, R.D. (1986). "Interrelations among the lesions of normal and abnormal aging of the brain". In: *Progress in Brain Research* 70, pp. 41–48.

Teune, L.K., A.L. Bartels, and K.L. Leenders (2013). *FDG-PET Imaging in Neurodegenerative Brain Diseases, Functional Brain Mapping and the Endeavor to Understand the Working Brain*. InTech.

Thie, J.A. (2004). "Understanding the Standardized Uptake Value, Its Methods, and Implications for Usage". In: *Journal of Nuclear Medicine* 45.9, pp. 1431–1434.

Thirion, J.-P. (1998). "Image matching as a diffusion process: an analogy with Maxwell's demons". In: *Medical Image Analysis* 2.3, pp. 243–260.

Toews, M., W.M. Wells III, D.L. Collins, and T. Arbel (2010). "Feature-based Morphometry: Discovering Group-related Anatomical Patterns". In: *NeuroImage* 49.3, pp. 2318–2327.

Tombaugh, T.N. (2004). "Trail Making Test A and B: Normative data stratified by age and education". In: *Archives of Clinical Neuropsychology* 19.2, pp. 203–214.

Toyn, Jeremy (2015). "What lessons can be learned from failed Alzheimers disease trials?" In: *Expert Review of Clinical Pharmacology* 8.3, pp. 267–269.

Tzourio-Mazoyer, N., B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, and M. Joliot (2002). "Automated Anatomical Labelling of Activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single subject brain". In: *NeuroImage* 15.1, pp. 273–289.

Vandenberghe, R., N. Nelissen, E. Salmon, A. Ivanoiu, S. Hasselbalch, A. Andersen, A. Korner, L. Minthon, D.J. Brooks, K. Van Laere, and P. Dupont (2013). "Binary classification of (18)F-flutemetamol PET using machine learning: Comparison with visual reads and structural MRI". In: *NeuroImage* 64, pp. 517–525.

Čížek, J., K. Herholz, S. Vollmar, R. Schrader, J. Klein, and W.-D. Heiss (2004). "Fast and robust registration of PET and MR images of human brains". In: *NeuroImage* 22.1, pp. 434–442.

Vercauteren, T., X. Pennec, A. Perchant, and N. Ayache (2007). "Non-parametric Diffeomorphic Image Registration with the Demons Algorithm". In: *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 319–326.

— (2008). "Symmetric log-domain diffeomorphic registration: a demons-based approach." In: *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 754–761.

— (2009). "Diffeomorphic demons: Efficient non-parametric image registration". In: *NeuroImage* 45.1, S61–S72.

Villemagne, V.L., K. Ong, R.S. Mulligan, G. Holl, S. Pejoska, G. Jones, G. O'Keefe, U. Ackerman, H. Tochon-Danguy, J.G. Chan, C.B. Reininger, L. Fels, B. Putz, B. Rohde, C.L. Masters, and C.C. Rowe (2011a). "Amyloid imaging with (18)F-florbetaben in Alzheimer disease and other dementias". In: *Journal of Nuclear Medicine* 52.8, pp. 1210–1217.

Villemagne, V.L., K.E. Pike, G. Chételat, K.A. Ellis, R.S. Mulligan, P. Bourgeat, U. Ackermann, G. Jones, C. Szoeke, O. Salvado, R. Martins, G. O'Keefe, C.A. Mathis, W.E. Klunk, D. Ames, C.L. Masters, and C.C. Rowe (2011b). "Longitudinal assessment of Aβ and cognition in aging and Alzheimer disease". In: *Annals of Neurology* 69.1, pp. 181–192.

Villemagne, V.L., W.E. Klunk, C.A. Mathis, C.C. Rowe, D.J. Brooks, B.T. Hyman, M.D. Ikonomovic, K. Ishii, C.R. Jack, W.J. Jagust, K.A. Johnson, R.A. Koeppe, V.J. Lowe, C.L. Masters, T.J. Montine, J.C. Morris, A. Nordberg, R.C. Petersen, E.M. Reiman, D.J. Selkoe, R.A. Sperling, K. Van Laere, M.W. Weiner, and A. Drzezga (2012). "A-β imaging: Feasible, pertinent, and vital to progress in Alzheimer's disease". In: *European Journal of Nuclear Medicine and Molecular Imaging* 39.2, pp. 209–219.

Vincent, P., H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol (2010). "Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion". In: *Journal of Machine Learning Research* 11.3, pp. 3371–3408.

Viola, P. and M. Jones (2001). "Rapid object detection using a boosted cascade of simple features". In: *Computer Vision and Pattern Recognition*.

Wang, M., X.S. Hua, R. Hong, J. Tang, G.J. Qi, and Y. Song (2009). "Unified video annotation via multigraph learning". In: *IEEE Transactions on Circuits and Systems for Video Technology* 19.5, pp. 733–746.

Wang, S. and R. Summers (2012). "Machine learning and radiology". In: *Medical Image Analysis* 16.5, pp. 933–951.

Wechsler, D. (1987). *Manual for the Wechsler Memory Scale-Revised*. Tech. rep. The Psychological Corporation.

Winblad, B., L. Kilander, S. Eriksson, L. Minthon, S. Batsman, A.L. Wetterholm, C. Jansson-Blixt, and A. Haglund (2006). "Donepezil in patients with severe Alzheimer's disease: double-blind, parallel-group, placebo-controlled study". In: *Lancet* 367.9516, pp. 1057–1065.

Wolk, D.A., Z. Zhang, S. Boudhar, C. Clark, M. Pontecorvo, and S. Arnold (2012). "Amyloid imaging in Alzheimer's disease: comparison of Florbetapir and Pittsburgh Compound-B PET". In: *Journal of Neurology, Neurosurgery & Psychiatry* 83.9, pp. 923–926.

Wolz, R., P. Aljabar, J.V. Hajnal, and D. Rueckert (2010). "Manifold Learning for Biomarker Discovery in MR Imaging". In: *Machine Learning in Medical Imaging*. Vol. 6357, pp. 116–123.

Wolz, R., P. Aljabar, J.V. Hajnal, J. Lötjönen, and D. Rueckert (2012). "Nonlinear dimensionality reduction combining MR imaging with non-imaging information". In: *Medical Image Analysis* 16.4, pp. 819–830.

Wong, K.-P., D. Feng, S.R. Meikle, and M.J. Fulham (2002). "Segmentation of dynamic PET images using cluster analysis". In: *Transactions on Nuclear Science* 49.1, pp. 200–207.

Wu, Z., S. Song, A. Khosla, Y. Fisher, L. Zhang, X. Tang, and J. Xiao (2015). "3D ShapeNets: A Deep Representation for Volumetric Shapes". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1912–1920.

Yushkevich, P.A., B.B. Avants, S.R. Das, J. Pluta, M. Altinay, and C. Craige (2010). "Bias in Estimation of Hippocampal Atrophy using Deformation-Based Morphometry Arises from Asymmetric Global Normalization: And Illustration in ADNI 3 Tesla MRI Data". In: *NeuroImage* 50.2, pp. 434–445.

Zasadny, K.R and R.L. Wahl (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". In: *Radiology* 189, pp. 847–850.

Zhao, Sicheng, Hongxun Yao, You Yang, and Yanhao Zhang (2014). "Affective Image Retrieval via Multi-Graph Learning". In: *Proceedings of the ACM International Conference on Multimedia - MM '14*, pp. 1025–1028.

Zijdenbos, A.P., B.M. Dawant, R.A. Margolin, and A.C. Palmer (1994). "Morphometric analysis of white matter lesions in MR images: method and validation". In: *Transactions on Medical Imaging* 13.4, pp. 716–724.