University of Sheffield

Using Unsupervised Learning Methods to Analyse Magnetic Resonance Imaging (MRI) Scans for the Detection of Alzheimer's Disease

Yuyang Liu

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

The University of Sheffield
Information School

June 2023

Abstract

Background: Alzheimer's disease (AD) is the most common cause of dementia, characterised by behavioural and cognitive impairment. The manual diagnosis of AD by doctors is time-consuming and can be ineffective, so machine learning methods are increasingly being proposed to diagnose AD in many recent studies. Most research developing machine learning algorithms to diagnose AD use supervised learning to classify magnetic resonance imaging (MRI) scans. However, supervised learning requires a considerable volume of labelled data and MRI scans are difficult to label. The aim of this thesis was therefore to use unsupervised learning methods to differentiate between MRI scans from people who were cognitively normal (CN), people with mild cognitive impairment (MCI), and people with AD.

Objectives: This study applied a statistical method and unsupervised learning methods to discriminate scans from (1) people with CN and with AD; (2) people with stable mild cognitive impairment (sMCI) and with progressive mild cognitive impairment (pMCI); (3) people with CN and with pMCI, using a limited number of labelled structural MRI scans.

Methods: Two-sample t-tests were used to detect the regions of interest (ROIs) between each of the two groups (CN vs. AD; sMCI vs. pMCI; CN vs. pMCI), and then an unsupervised learning neural network was employed to extract features from the regions. Finally, a clustering algorithm was implemented to discriminate between each of the two groups based on the extracted features. The approach was tested on baseline brain structural MRI scans from 715 individuals from the Alzheimer's Disease Neuroimaging Initiative (ADNI), of which 231 were CN, 198 had AD, 152 had sMCI, and 134 were pMCI. The results were evaluated by calculating the overall accuracy, the sensitivity, specificity, and positive and negative predictive values.

Results: The abnormal regions around the lower parts of the limbic system were indicated as AD-relevant regions based on the two-sample t-test (p<0.001), and the proposed method yielded an overall accuracy of 0.842 for discriminating between CN and AD, an overall accuracy of 0.672 for discriminating between sMCI and pMCI, and an overall accuracy of 0.776 for discriminating between CN and pMCI.

Conclusion: The study combined statistical and unsupervised learning methods to identify scans of people with different stages of AD. This method can detect AD-relevant regions and could be used to accurately diagnose stages of AD; it has the advantage that it does not require large amounts of labelled MRI scans. The performances of the three discriminations were all comparable to those of previous state-of-the-art studies. The research in this thesis could be implemented in the future to help in the automatic diagnosis of AD and provide a basis for diagnosing sMCI and pMCI.

Acknowledgements

Publications

Papers

In the process of conducting the study, the following publication have been accomplished:

- Liu, Y., Mazumdar, S., & Bath, P. A. (2023). An unsupervised learning approach to diagnosing Alzheimer's disease using brain magnetic resonance imaging scans. *International Journal of Medical Informatics, 173(March)*. https://doi.org/10.1016/j.ijmedinf.2023.105027

- Liu, Y., Mazumdar, S., & Bath, P. (2022). Using machine learning techniques and brain MRI scans for detection of Alzheimer's disease. *iConference*. Retrieved from https://hdl.handle.net/2142/113739

Papers in preparation for publication:

- Liu, Y., Mazumdar, S. & Bath, P. A. (2023). An unsupervised learning approach to discriminating between stable mild cognitive impairment and progressive mild cognitive impairment using brain magnetic resonance imaging scans. For submission to *Journal of the American Medical Informatics Association (JAMIA)*

- Liu, Y., Mazumdar, S. & Bath, P. A. (2023). An unsupervised learning approach to diagnosing progressive mild cognitive impairment using brain magnetic resonance imaging scans. For submission to *Journal of the American Medical Informatics Association (JAMIA)*

Conference Presentations

- INSIGNEO: poster presentation, 08 July 2022
  - Title: Using Machine Learning Techniques and Magnetic Resonance Imaging for Detection of Alzheimer's Disease
  - Location: The Diamond, University of Sheffield, Sheffield, UK

- Dementia Futures: poster presentation, September 2022
  - Title: Using Machine Learning Techniques and Magnetic Resonance Imaging for Detection of Alzheimer's Disease
  - Location: Sheffield Institute for Translational Neuroscience (SITraN), University of Sheffield, Sheffield, UK

Contents

List of Figures

List of Tables

Declaration

I, the author, confirm that the Thesis is my own work. I am aware of the University's Guidance on the Use of Unfair Means (www.sheffield.ac.uk/ssid/unfair-means). This work has not been previously been presented for an award at this, or any other, university.

Glossary

Alzheimer's disease	The most common and prevalent cause of dementia.
Amygdala (amygdaloid body)	A structure of the brain. See Figure 1.6.
Amyloid	A type of abnormal protein that can accumulate in various tissues and organs.
Apolipoprotein E (APOE-epsilon 4)	A protein involved in the metabolism of human and it is encoded by gene APOE.
Cerebral cortex	The outer surface of the brain.
Cerebrospinal fluid (CSF)	A clear, watery, colourless fluid that flows in and around the brain and spinal cord.
Cognitively normal	A health condition.
Cortical pyramidal	A type of neuron found in the cerebral cortex.
Dementia	A clinical condition leading to cognitive and behavioural impairment.
Diffusion tensor imaging (DTI)	A type of MRI. See Section 1.4.4.2.
Diffusion-weighted imaging (DWI)	A type of MRI. See Section 1.4.4.2.
Grey matter	The darker tissue (in MRI scans) containing nerve fibres. See Figure 1.7.
Gypothalamus	A structure of the brain. See Figure 1.5.
Gyrus	A ridge on the cerebral cortex
Hippocampus	A structure of the brain. See Figure 1.6.
Limbic system	a set of structures of the brain. See Figure 1.5.
Magnetic resonance imaging (MRI)	A type of medical scans. See Section 1.4.4.
Mild cognitive impairment (MCI)	A prodromal stage of Alzheimer's disease.
Neuritic plaque	Abnormal deposits that form between neurons in the brain.
Parahippocampal gyrus	A structure of the brain. See Figure 1.6.
Progressive MCI (pMCI)	An MCI condition that develops to Alzheimer's disease within 36 months.
Slice	A plane in an MRI scan. See Figure 1.2.
Stable MCI (sMCI)	An MCI condition that does not develop to Alzheimer's disease within 36 months.
Thalamus	A structure of the brain. See Figure 1.5.
Voxel	A unit in an MRI scan. See Figure 1.2.

White matter	The pale tissue (in MRI scans) containing nerve fibres. See Figure 1.7.
White matter hyperintensity (WMH)	A kind of lesions in the brain. See Section 4.9.1.3.

Chapter 1 Introduction

1.1.   Introduction

This thesis concerns the development of machine learning methods to analyse brain magnetic resonance imaging (MRI) scans to identify people with Alzheimer's disease (AD).

This chapter provides an overview of the information and knowledge about my research, i.e., the diagnosis of Alzheimer's disease using brain MRI scans and machine learning methods, as context for the research This includes why and how this research area has emerged before (Section 1.2), the aims and objectives of this research (Section 1.3), and background information about the research (Section 1.4). Section 1.5 provides the thesis structure and summarises each chapter.

1.2.   Background of the Research

1.2.1. Impact of Dementia and Alzheimer's Disease

In recent years, dementia has received an increasing attention because according to the statistics of World Health Organisation (World Health Organisation, 2023), there are over 50 million people all over the world now suffering from dementia. Dementia is a clinical condition leading to cognitive and behavioural impairment, for example, language or memory loss, of people who are typically aged 65 years old and over (dos Santos Picanço et al., 2018), and there is currently no cure for it: treatment focuses on managing the condition for the patient and their family members and carers. Among the causes of dementia, Alzheimer's disease (AD) is the most common and prevalent cause, accounting for an approximate 60% to 80% of dementia cases (Alzheimer's Association, 2021). Although dementia has existed for hundreds of years, AD was only discovered and started to be studied in the early 20th century (Bondi, Edmonds, & Salmon, 2017). Therefore, AD is still a relatively "new" disease that still needs studying.

In addition to it being a "new" disease, the number of people who have dementia around the world is expected to continuously rise in the future, and most especially in low-income countries. Figure 1.1 (Alzheimer's Disease International, 2015) shows the estimated growth in numbers of people with dementia around the world. It can be seen that the increase in high-income countries is expected to be gradual but the low-income countries are expected to have a relatively steep rise. According to the World Alzheimer Report 2015 (Alzheimer's Disease International, 2015), the increase in people diagnosed with dementia around the world by 2030

is estimated to double from what it was in 2015. As a consequence, dementia, and AD in particular, is a risk to people all over the world, especially for older people.
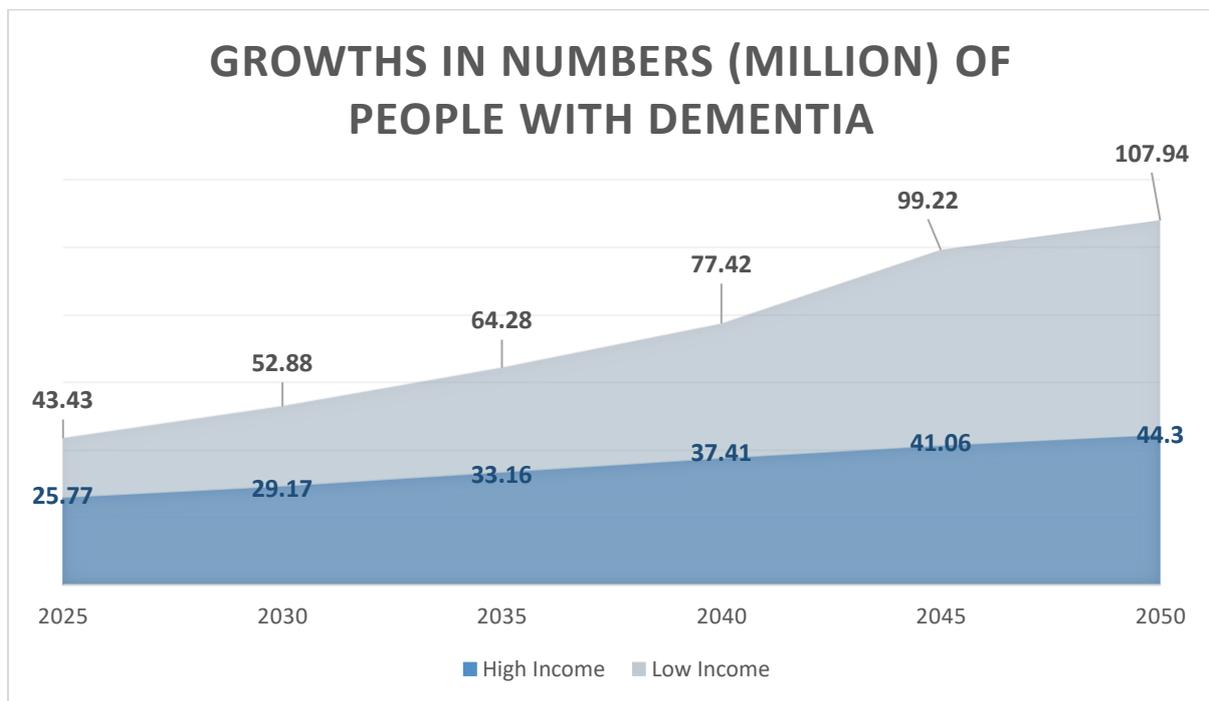
Figure 1.1 The predicted growth in the numbers (in millions) of people with dementia in high-income countries and low-income countries (Alzheimer's Disease International, 2015) (the latest version)

However, as well as the numbers being a reason for concern, our society still lacks awareness of dementia and AD, especially in low-income countries (Alzheimer's Disease International, 2019). The World Alzheimer Report 2019 reported that two thirds of people think that dementia is a consequence of ageing (Alzheimer's Disease International, 2019), which suggests that they think it is a natural phenomenon in later life. In addition, over 60% of healthcare practitioners still believe dementia is part of normal ageing (Alzheimer's Disease International, 2019). However, a quarter of people think that nothing can be done to prevent dementia (Alzheimer's Disease International, 2019). Therefore, studies about dementia and AD, and raising awareness of these conditions are crucial issues for practitioners as well as the general public.

1.2.2. Research on Alzheimer's Disease

Research on dementia and AD has been undertaken for more than a hundred years. Although AD, as a disease was first recognised by Alois Alzheimer in 1906, the realisation that the majority of cases with old-age dementia actually had Alzheimer's pathology became the

symbol of the new era in the late 1960s and 1970s (Hardy, 2006a). After that, three basic research directions were explored to understand the disease. The first was the neurochemical pathology approach to the disease; the second was to better understand the lesions, neuritic plaque, and the neurofibrillary tangle that occur; the third was identifying the genes related to the disease (Hardy, 2006a).

For the first research direction, the study undertaken by Davis et al. (1978) clarified that twisted tangles in certain brain cells are a major problem in Alzheimer's disease and a substance called glutamate might replace another substance in the brain (Hardy et al., 1987). For the second research direction, Miyakawa and Uehara (1979) reported a connection between certain brain changes and so-called neuritic plaques in AD. Glenner and Wong (1984), Goldgaber et al. (1987), and various other studies tried to identify what these brain changes were made of. As for the third research direction, after a series of studies by Miyakawa and Uehara (1979), Glenner and Wong (1984), and Goldgaber et al. (1987), Strittmatter et al. (1993) discovered a gene named apolipoprotein E epsilon 4 (APOE-epsilon 4) and a special protein named amyloid β connected to AD. Genetic analysis also showed that APOE-epsilon 4 was present in more than half of patients with Alzheimer's disease, regardless of whether the patients have a family history of dementia or not (Strittmatter et al., 1993).

However, since the 2000s, mild cognitive impairment (MCI) has been the focus of research in the field, because MCI is a prodromal stage of AD before diagnosis. MCI due to AD is characterised by memory decline or impairment, and 10% to 20% of adults who are aged 65 years old and over are diagnosed with it (Langa & Levine, 2009). In addition, MCI is generally divided into two types: stable MCI (sMCI) and progressive MCI (pMCI), they are both described in detail in Section 1.4. Although there is no drug that has shown to be effective for MCI currently, people with MCI can take steps to lower the risk of it developing to dementia. This is because, if MCI is diagnosed, some risks for negative outcomes can be reduced by counselling about issues that are related to safety of daily life and functional status (Langa & Levine, 2009). Effectively diagnosing MCI is therefore significantly important.

MCI caused by AD is typically diagnosed using neuroimaging techniques, as well as laboratory testing of body fluids, such as blood and cerebrospinal fluid (CSF) tests (Langa & Levine, 2014). For structural imaging, computed tomography (CT) and magnetic resonance imaging (MRI) are the most commonly used techniques. However, what CT reveals, such as subdural

hematomas, may be causes of reversible dementia, thus MRI is often employed because of its accurate details in changes in the cerebral cortex and hippocampus, which are mostly related to AD (Caroline dos Santos Picanço et al., 2018). To acquire functional imaging information, fluorodeoxyglucose positron emission tomography (FDG-PET) is often used for hypometabolism detection in the brain (Langa & Levine, 2009); however, obtaining information from this is considered invasive; recent efforts have thus focused more on structural imaging, such as MRI modality. These neuroimaging techniques will be described in detail in Section 1.4.2.

1.2.3. Diagnosing Alzheimer's Disease

In 1984, clinical diagnostic criteria and standards for AD were established by the National Institution of Neurologic and Communicative Disorders and Stroke and the Alzheimer's Disease and Related Disorders Association (NINCDS-ADRDA). The information that is necessary to the criteria is acquired by standard methods of examination: the medical history; neurologic, psychiatric, and clinical examinations; neuropsychological tests; and laboratory studies (McKhann et al., 1984). The following explains these criteria further.

Medical history. A medical history is essential to establish a history of the progressive deterioration of the patient. It can reveal abnormalities, including memory impairment, decline in cognitive functioning, and other activities that the patient can no longer perform well. Over time, this can also form clinicians about the progression of the disease.

Neurologic, psychiatric, and clinical examination. The examination provides data to execute inclusionary and exclusionary criteria for AD and to record symptoms such as delusion or depression. In this examination, a quantitative approach, the Mini-Mental State Examination (MMSE), is one of the most widely used tests in the world (Arevalo-Rodriguez et al., 2015).

Neuropsychological testing. Neuropsychological testing can provide additional information and confirmatory evidence for the diagnosis of AD. It may include a sequence of tests: memory evaluation; language skill examination; attention monitoring; visual perception study; problem-solving skills test; and social functioning, activities, instrumental activity assessment.

Laboratory assessments. Laboratory approaches used quantitatively in longitudinal studies can help to clarify the natural history of AD and permit the measurement of the efficacy of therapeutic interventions. It includes electrophysiologic methods, CT, PET, MRI, etc..

During this assessment, the clinical examination and neuropsychological tests often use some of the standard tests, such as the MMSE, the Recognition Span Test, the Western Aphasia Test, etc., and practitioners can relatively easily obtain results from these tests according to their diagnostic criteria. Furthermore, the medical history can be recorded by patients' families or relatives, which means that practitioners are not always necessary. However, laboratory assessment is comparatively much more difficult because analysing medical imaging scans, such as CT, PET, and MRI, requires practitioners to have considerable experience and professional knowledge; even so, analysing these medical imaging scans is still time-consuming. In addition, different practitioners may draw different conclusions due to their different levels of skills and knowledge. Therefore, more recently, artificial intelligence (AI) and computational approaches are being developed to support diagnosis using medical imaging scans and they are becoming crucial for disease detection and management.

1.2.4. Diagnosing Alzheimer's Disease Using Machine Learning and Deep Learning

As mentioned in Section 1.2.3, analysing medical imaging scans is difficult because traditional medical imaging scan analysis for AD diagnosis is undertaken manually by doctors (Bhargava, Wozniak, Mentzel, & Krumbein, 2014). This involves locating different regions of interests and analysing the condition based on their professional knowledge and experience. However, the structure of the brain is complex, and each brain scan contains thousands of voxels (as shown in Figure 1.2, the units in a 3D medical imaging scan) and massive amounts of information, so analysing it manually by experts is potentially a time-consuming task. Generally, an MRI scan is studied by a radiologist and possibly discussed with other specialists (Bhargava et al., 2014), and, as a result, the results of an MRI scan need one or two weeks to come through (National Health Service (UK), 2018). To help reduce this lengthy process, computer-aided diagnosis of AD based on medical imaging scans has become an area of immense interest (Erickson, Korfiatis, Akkus, & Kline, 2017). Machine learning has become an emerging science and computer-aided technique for AD diagnosis.
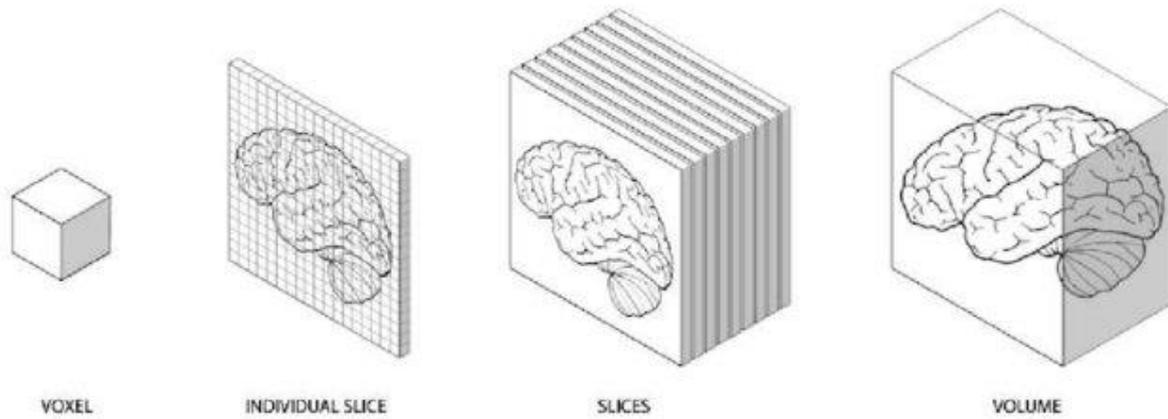
Figure 1.2 Schematic showing principle behind image acquisition in MRI (Attribution-NonComm ercial 4.0 International CC BY-NC 4.0 [available at: https://www.researchgate.net/publication/345503897_Brai n_Networks_and_Dynamics_in_Narcolepsy])

Machine learning is an interdisciplinary field, which includes computer science, artificial intelligence, and statistics (Ayodele, 2010a). It is an application of artificial intelligence (AI) that aims to enable systems to learn automatically from past experience and do the job by themselves (Mohammed, Khan, & Bashie, 2016). It is concerned with knowledge extraction and its main tasks involve recognition, prediction, data mining, etc.. In general, machine learning is categorised into four types: supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning (Mohammed, Khan, & Bashie, 2006). These are described briefly as follows:

Supervised learning: in supervised learning, data that need learning have desired outputs, and the goal of the learning is to map input data to the desired outputs, which are often presented as labels (Ayodele, 2010b);

Unsupervised learning: in unsupervised learning, training datasets have no labelled outputs, which means they do not consist of any desired outputs. The goal of unsupervised learning is usually to find the properties of the mechanism generating the data (Ayodele, 2010b);

Semi-supervised learning: in semi-supervised learning, both labelled and unlabelled data are operated on, which means that it is a hybridization of supervised and unsupervised learning methods (Sarker, 2021);

Reinforcement learning: in reinforcement learning, an agent takes actions in a particular environment in order to automatically evaluate the optimal behaviour and improve its

efficiency (Kaelbling, Littman, & Moore, 1996). It is a method based on reward and punishment with the aim of maximising the reward function.

Figure 1.3 shows a (non-exhaustive) taxonomy of the different methods in machine learning.
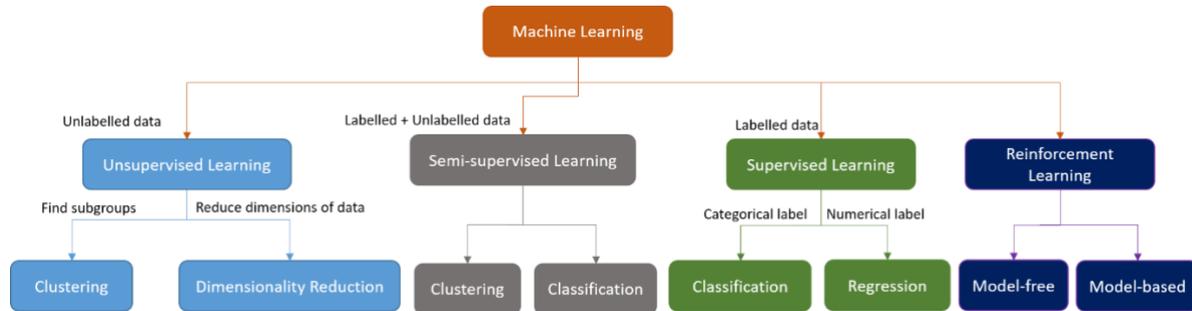
Figure 1.3 A taxonomy of the four types of machine learning (Developed by Yuyang Liu)

In machine learning, artificial neural network is one of the most widely used models (Xin Liu, Tian, Tao, & Yu, 2021), modelled on the functioning of the nervous system. It can be defined as a computational model consisting of a set of highly interconnected processing elements, called neurons (Sordo, 2002). A basic architecture of a neural network is shown in Figure 1.4. As can be seen, an artificial neural network consists of an input layer of neurons, a hidden layer of neurons, and an output layer of neurons. In recent years, deep learning has been a popular term and technique in many research fields. Deep learning refers to deep artificial neural networks (Shinde & Shah, 2018). "Deep" means it has more than one hidden layers in the neural networks (Shinde & Shah, 2018).

However, although the concept of machine learning was first developed in the 1950s, it was not widely used due to a lack of data and the performance of processors at that time. With the emergence of the Internet, there are increasing amounts of data available that can be used for research around the world. The Alzheimer's Disease Neuroimaging Initiative (ADNI) database was founded in 2004 and it provides a large amount of data about AD for researchers around the world. After that, other open and public databases for AD studies, such as the Open Access Series of Imaging Studies (OASIS), were launched. From Section 3.3, it will be seen that they have become the foundation of the application of machine learning for the diagnosis of AD. In addition, progress and breakthroughs in Graphic Processing Units (GPUs) also promote usage of machine learning (and deep learning) for its ever more powerful computing capabilities (Macedonia, 2003).
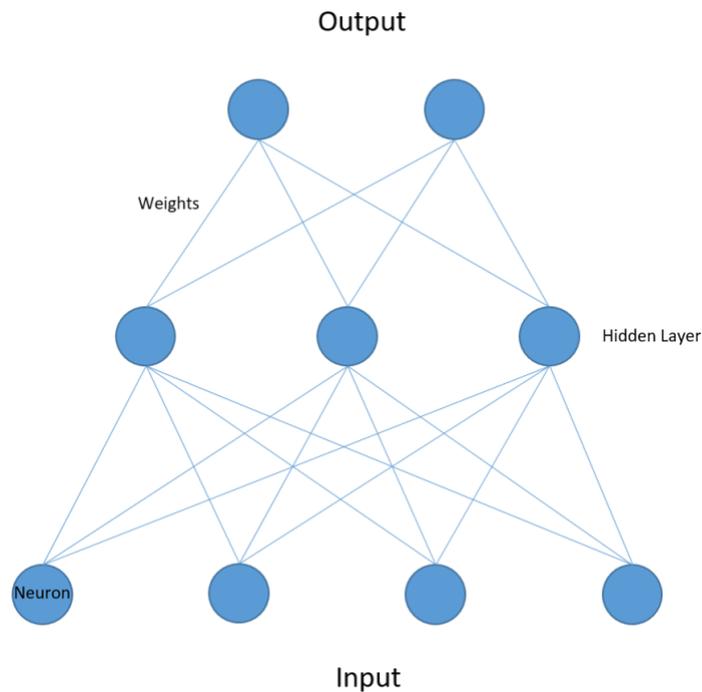
Figure 1.4 A basic architecture of a neural network (Developed by Yuyang Liu)

The MRI is one of the most commonly used medical imaging techniques for the diagnosis of AD (as discussed in Section 1.4.2). Generally, machine learning is used to classify or cluster MRI images into different diagnostic statuses, such as AD, cognitively normal (CN), and MCI, but features have to be extracted from the MRI images beforehand. From previous studies (Casanova, Hsu, Sink, Rapp, & Williamson, 2013; Eskildsen, Coupé, Fonov, Pruessner, & Collins, 2015; Weller & Budson, 2018; Lin et al., 2018; Gupta et al., 2019; Lee, Choi, Kim, & Suk, 2019; Wang et al., 2019), these extracted features can be roughly divided into two types. The first type is handcrafted features, and this includes segmented tissues (such as the hippocampus and grey matter), shape measures of tissue, pre-defined areas, etc.. Such region-based and visual features are called low-level features (Jiang, Chan, Li, & Zhang, 2005; Zhang, Li, Lu, & Cheng, 2016). The second type is to employ particular statistical methods or machine learning models to calculate and select features, for example, using the F-test or random forest model to select features, or to employ neural networks (including deep learning) to build relationships between voxels or areas in the MRI scan. The semantic features, which capture representative information, are called high-level features (Jiang et al., 2005; Zhang et al., 2016). In general, these studies have tried to find a better method to diagnose AD, and aimed to improve the accuracy of diagnosis.

Although many studies (Battineni, Chintalapudi, Amenta, & Traini, 2020; Ramon Casanova et al., 2018; Zeng et al., 2018; Lama et al., 2017; Beheshti & Demirel, 2016) have achieved 80% to 95% accuracy, most or all of these studies employed supervised learning approaches. However, the number of studies using unsupervised learning methods is much less than that of studies using supervised learning, which means most research used already labelled data. Nonetheless, labelling MRI images is quite difficult because it requires the involvement of practitioners with good experience and knowledge, considerable time to interpret. Furthermore, many recent models, especially deep learning models, need a relatively large number of already labelled images to train.

In recent years, there has been a growing trend towards using pre-trained deep learning models and transfer learning to fulfil image classification. Transfer learning aims to improve performances on a new task by leveraging the knowledge from related or similar tasks in a different domain (H. E. Kim et al., 2022). The pre-trained model is either a machine learning or deep learning model that has been trained on a large dataset before being fine-tuned for a specific task. This was considered to reduce the reliance on labelled data. However, many pre-trained models in image classification tasks used the ImageNet dataset (Arafa, Moustafa, Ali, Ali-Eldin, & Saraya, 2023; Dhinagar et al., 2023; Shanmugam, Duraisamy, Simon, & Bhaskaran, 2022). ImageNet consists of a large number of images of natural objects, e.g., cats, dogs, bicycles, etc., thus it is hard to identify whether the models trained on these objects could be transferred to MRI scans and be used to diagnose AD. The structure of the brain is much more complex and the discrimination between the brain of CN, MCI, and AD is much more difficult than discrimination between the normal objects in daily life. Hence, although theoretically transfer learning can reduce the reliance on labelled date through pre-trained models, fine-tuning may still need a large volume of data in specific tasks, for example, diagnosing AD. For example, Shanmugam et al. (2022) utilised a convolutional neural network that had been pre-trained on the ImageNet dataset, but they still fine-tuned the model using 6000 (out of 7800) MRI scans; Dhinagar et al. (2023) used 2577 (out of 4098) scans to fine-tune a neural network pre-trained on the ImageNet dataset.

In addition to transfer learning, self-supervised learning can also be trained on far fewer labelled data. Self-supervised learning is a technique that learns from unlabelled data by creating a supervised learning task from the data itself (Shurrab & Duwairi, 2022). However, compared with natural images, the amount of medical images is limited, and this leads to the

fact that self-supervised learning models may struggle to learn meaningful representations. Thus, self-supervised learning models often need pre-training and transfer learning, but pretext tasks in pre-training need specific knowledge, i.e., medical knowledge in this study. Therefore, self-supervised learning could be a choice in future research when the medical image is sufficient and medical specialists are involved in the research.

Therefore, unsupervised learning methods, semi-supervised learning, or other methods (i.e., statistical methods) that need a limited number of labelled images, have become important, and require further investigation. In addition, not only is the performance related to classification or clustering crucial, but also related to the phases before them, especially feature extraction because, as described above, the high-level features contains representative information of the images. Furthermore, besides CN and AD, MCI also should be more considered when classification or clustering is operated, based on its meaningful role (to lower the risk of CN or MCI developing to dementia) as mentioned above.

Based on the background of the research on AD and machine learning, the aims and objectives of my study will be stated in the following section (Section 1.3).

1.3. Aims and Objectives

The overall aim of this research was to develop machine learning methods to discriminate between different stages of AD using unsupervised learning methods and a limited number of labelled structural MRI scans. Specifically, this research includes three studies (discrimination tasks):

Study 1: discriminating between CN and AD scans;

Study 2: discriminating between sMCI and pMCI scans;

Study 3: discriminating between CN and pMCI scans.

The objectives of the research (RO) were as follows:

RO1: To extract the most appropriate features and regions of the brain to support clustering tasks to discriminate between CN and AD (Study 1, Chapter 4) and, if possible, to discriminate between sMCI and pMCI (Study 2, Chapter 5), and between CN and pMCI (Study 3, Chapter 6);

RO2: To apply unsupervised learning methods to discriminate between CN and AD, to discriminate between sMCI and pMCI, and to discriminate between CN and pMCI;

RO3: To evaluate the performance of the methods for discriminating between CN and AD, discriminating between sMCI and pMCI, and for discriminating between CN and pMCI; to compare the methods with existing studies.

This study seeks to answer the following research questions (RQ):

RQ1: How can we discover features and regions that are meaningful to discriminating between CN and AD; discriminating between sMCI and pMCI, and between CN and pMCI?

RQ2: How does the proposed method compare with existing supervised, semi-supervised, and unsupervised learning methods?

Because this research involves professional knowledge of several fields, such as AD and MRI, to understand the research better, more background information about those areas are elaborated in the following section (Section 1.4).

1.4. Background Information

1.4.1. Brain Structure

AD is a neurodegenerative disease and it has a very close relationship with the human's brain. Therefore, to better understand the results of the study reported in the later chapters, a basic structure of the brain is shown in Figure 1.5, and Figure 1.6 shows the limbic system in the brain, which includes the hippocampus, the amygdala (amygdaloid body), the hypothalamus, etc.. Figure 1.7 shows an illustration of the white matter and the grey matter.
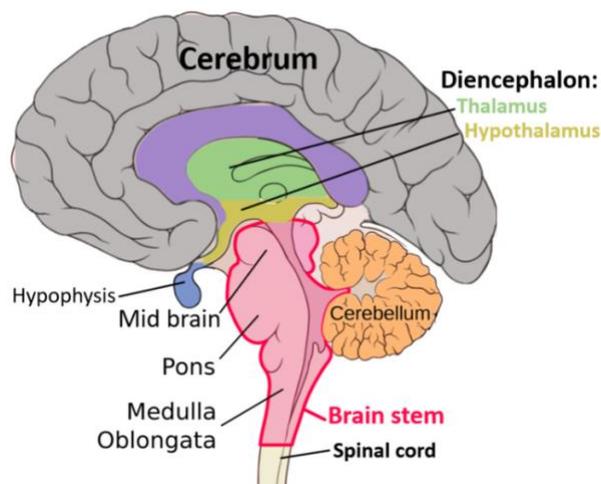
Figure 1.5 A basic structure of the brain (This image is licensed under the Creative Commons Attributio n-Share Alike 4.0 International license. [available at: https://commons.wikimedia.org/wiki/File:Basic_structure s_of_the_brain_highlighted.png])
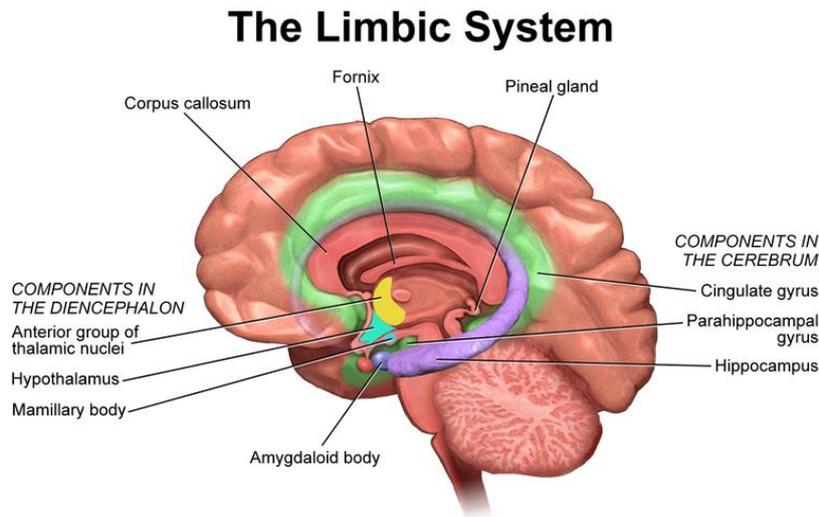
Figure 1.6 A set of structures of the brain (the limbic system) (This image is licensed under the Attribution 3.0 Unported (CC BY 3.0). [available at: https://commons.wikimedia.org/wiki/File:Blausen_0614_LimbicSystem.png])
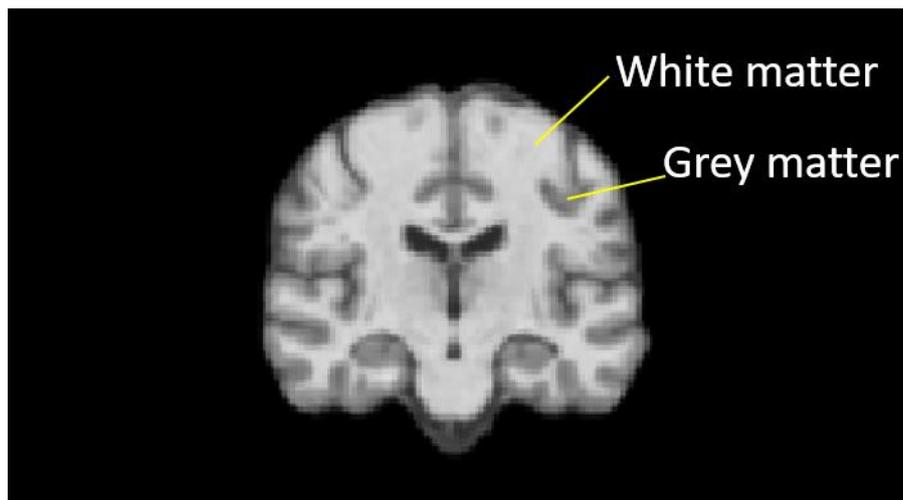
Figure 1.7 An illustration of the white matter and grey matter (Original image from ADNI adapted by Yuyang Liu)

1.4.2. Introduction to Dementia and Alzheimer's Disease

According to the definition from National Health Service (NHS) (National Health Services (UK), n.d.-a), dementia is a clinical syndrome (i.e., a group of co-occurring signs and symptoms) that involves cognitive decline and brain functioning. People with dementia will experience memory loss and difficulties in reading, speaking, or writing in the early phase of dementia, and they can experience the loss of important mental abilities, such as conversation disabilities and poor recognition, in the later phases.

Generally, there are four types of dementia (Fymat, 2018): Vascular dementia, Frontotemporal dementia, Lewy body dementia, and Alzheimer's disease. However, Alzheimer's disease (AD), a chronic degenerative condition, is the most common type, and cause about 60% to 80% of cases of dementia (Garcia-Ptacek et al., 2016). AD was considered as a rare disease or disorder until a few decades ago (Fymat, 2018); nonetheless, as was discussed in the Section 1.1, AD has now become a more prevalent public health problem all around the world.

There are two types of AD: early-onset AD and late-early onset AD. Early-onset AD, also called young-onset AD, is relatively rare and often first appears at a relatively young age, which is generally before 65 years old, and may even start in a person their 30s (Harvey, Skelton-Robinson, & Rossor, 2003). Late-onset AD, which is often called AD, is the most common type of AD, and it usually appears in people who are aged 65 and over.

Early-onset AD is generally considered to be caused by genetic factors. Mutations of at least three genes – amyloid precursor protein (APP) gene on chromosome 21, presenilin-1 (PS1) on chromosome 14, and presenilin-2 (PS2) on chromosome 1, lead to the disease (Lopera, Ardilla, Martínez, Madrigal, & Arango-viana, 1997). For late-onset AD, the more common AD, it is widely believed that the causes of it are amyloid beta plaques and neurofibrillary tangle (Hardy, 2006b), which are related to synapse loss and neurodegeneration. At the cellular level, AD is identified by a loss of cortical neurons (Korolev, 2014). In the early phase of AD, this results in damages to synaptic function and tissues in the medial temporal lobe, i.e., the hippocampus and entorhinal cortex are often the start of degeneration under the influence of AD (I. O. Korolev, 2014). Finally, the degeneration will undermine the whole neocortex and other structures of the brain. Figure 1.8 (Fymat, 2018) shows the differences between a healthy brain and a brain with severe AD.
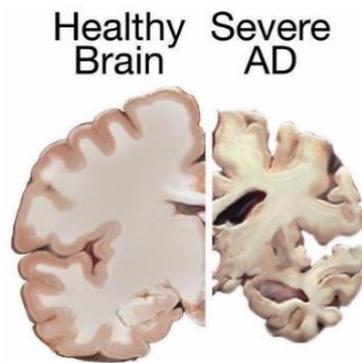
Figure 1.8 Healthy brain vs. Brain with severe AD (This work has been identified as being free of known restrictions under copyright law, including all related and neighbouring rights. Available at: https://www.flickr.com/photos/nihgov/24239522109)

In general, two types of precursors before AD can be considered: subjective cognitive impairment (SCI) and mild cognitive impairment (MCI).

1) SCI is a condition that affects a heterogeneous group of patients who show the presence of cognitive complaints with no evidence of neurodegenerative disease (Garcia-Ptacek et al., 2016). However, it still shows some non-apparent abnormalities, for example, brain shrinkage (Fymat, 2018). In addition, it may have a long duration, i.e., a decade or two (Fymat, 2018).

2) MCI is a condition in which the patient's cognitive abilities, such as memory or thinking, are slightly abnormal (Alzheimer's Society, 2017). Typically, this condition follows SCI. Approximately 5 to 20 percent of people aged over 65 have MCI (Alzheimer's Society, 2017). MCI is a crucial phase because people with it have a high probability of going on to develop dementia (or AD).

Compared with SCI, MCI has been studied more in previous research because the condition of SCI is still part of the "normal" domain in the neuropsychological testing (a cognitive complaint with normal cognitive screening test). According to *The Progression of Alzheimer's Disease and Other Dementia* by the Alzheimer's Society, 10 to 15 percent of people who are diagnosed as MCI progress to dementia every year (Alzheimer's Society, 2016). Therefore, it is important for those who have MCI to identify that it is MCI and to predict the development of AD. Such early prediction will mean that patients diagnosed with MCI can be provided with support mechanisms and can prepare in advance, and currently this is a growing area of research. In addition, as mentioned in the Section 1.2.2, MCI can be further categorised into stable MCI (sMCI) and progressive MCI (pMCI). However, the definition of thesis is slightly different in different studies. Zeng et al. (2018) defined that if a patient with MCI finally

progress to AD, the patient is diagnosed as pMCI, and otherwise the patient is sMCI; Huang et al. (2019) defined that MCI patients who develop to AD within a 36-month follow-up are pMCI patients, otherwise they are sMCI patients. In this thesis, the latter definition is used. Furthermore, because the pMCI patients finally develop to AD (even within 36 months), discriminating between pMCI and sMCI, or discriminating pMCI and CN is important and meaningful as well.

As a result, assessment and diagnosis is very important for those who have some of the early symptoms of AD and dementia. As described in Section 1.2.3, during the process of assessment and diagnosis, as well as certain tests and examinations, brain scans are crucial to monitor changes in the brain structure. Nowadays, there are various common imaging techniques for medical screening. However, not every imaging technique is suitable for diagnosing AD, thus these suitable medical imaging techniques will now be introduced in the following section (1.4.3).

1.4.3. Introduction to Medical Imaging Techniques

Before the beginning of the 20th century, the clinical equipment used for investigating health condition was still limited to exploring the external world of the living human body. What medical practitioners could use at that time was accounted for by the limited tools available, for example, the thermometer and the knife, which could not look into the organs, tissues or fluid hidden inside the living human body (Suetens, 2017). Therefore, the discovery of the X-ray by Roentgen in 1895 and the usage of it for therapy and metabolic tracer studies from 1896 were important starts. After several years, X-ray was used for medical imaging (Ganguly, Chakraborty, Balitanas, & Kim, 2010). With the development of technology, more and more medical imaging techniques emerged, and so far, the mainstream medical imaging techniques include X-ray radiography, Ultrasound, Computed Tomography, Positron Emission Tomography, and MRI. Amongst them, Computed Tomography, Positron Emission Tomography, and MRI scans are often used for diagnosing AD. They have their own respective advantages when investigating different organs or tissues, so a brief overview of them is provided below:

1) Computed Tomography

Computed Tomography (CT) images are generated by a cathode ray tube with the help of X-ray equipment with a computer and can produce a more detailed 3-D image (Atabo & Umar,

Before the beginning of the 20th century

2019). CT even needs an even higher amount of X-ray (Atabo & Umar, 2019), thus CT has the same problem as X-rays in that it makes patients exposed to ionizing radiation, which may lead to cancer in later life. In addition, based on its working principle, it does not show real time information and cannot detect intra-luminal abnormalities, and it also has poor performance without contrast. However, CT has good spatial resolution and global view of veins and, because it utilises X-rays through the body, it naturally is not painful and the diagnosis process is quick. An example CT image of a brain is shown in Figure 1.9.

Based on the advantages and risks of CT, it is usually used for examining specific parts of human body, such as brain, facial bones, teeth, etc.. (Kasban, El-bendary, & Salama, 2015). It can be also used to monitor the performance of therapy, as in cancer treatment (Kasban et al., 2015).

Figure 1.9 A CT image of a brain (This image is licensed for use under the Creative Commons Attribution-Share Alike 3.0 Unported license. [available at: https://commons.wikimedia.org/wiki/File:Brain_CT_scan.jpg])

2) Positron Emission Tomography

Positron Emission Tomography (PET) is one of the radionuclide imaging techniques. Radionuclide imaging formulates the images of the internal body using a small amount of radioactive material. PET manipulates the decay of isotopes by positron emission and tracks

the position of the emitted positron to produce the image showing information about the metabolism of a disease (Kasban et al., 2015). A PET image of a brain is shown as Figure 1.10.

PET can provide highly accurate information about the function of objects and has a good tissue contrast in the scan. It can also track the spread of cancer and the body's response to cancer treatment (Zhu, Lee, & Shim, 2011). Different to X-ray radiography, PET is a non-invasive imaging technique (Zhu et al., 2011). However, as can be expected, PET requires additional care while handling radioactive materials. In addition, PET is also expensive for the equipment and isotope production. Therefore, PET generally is used in combination with MRI to diagnose cardiovascular diseases, cancers, and some neurological disease.
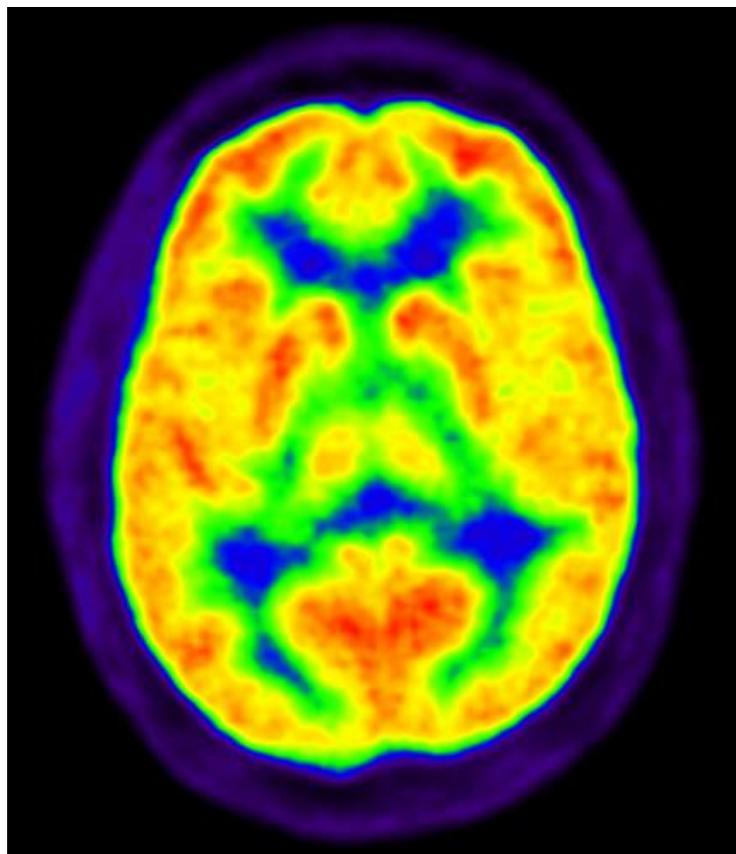
Figure 1.10 A PET image of a brain (This image is licensed under the Attribution-NonCommercial-Share Alike 2.0 Generic (CC BY-NC-SA 2.0). © Akira Kouchiyama [available at: https://www.flickr.com/photos/triumf lab/8232448893])

3) Magnetic Resonance Imaging

Magnetic Resonance Imaging (MRI) is an imaging technique for diagnosis that uses magnetic radio frequency fields to produce 3D images of tissues and chemistry in human body (Atabo & Umar, 2019). An MRI scan is shown as Figure 1.11.

Compared with X-ray radiography, CT, and PET, MRI has no ionizing radiation, and it is not painful or invasive (Atabo & Umar, 2019). Using MRI, it is easy to bind and measure flow and velocity with modern techniques. In addition, MRI has a very good performance on soft tissue contrast. However, the scanning process for MRI is long and patients have to stay in a narrow space (National Health Service (UK), 2018) as shown in Figure 1.12, which may cause some people to have feelings of claustrophobia. MRI has relatively low sensitivity and has no real time information. It also has relatively expensive cost compared to other imaging techniques.
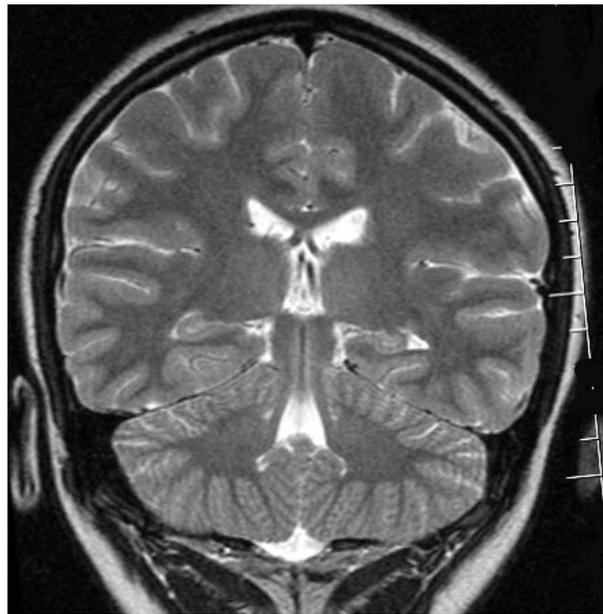
Figure 1.11 An MRI scan of a brain (This image is licensed under the Creative Commons Attribution-Share Alike 3.0 Unported license. © Nevit Dilmen [available at: https://commons.wikimedia.org/wiki/File:Brain_MRI_t2_142301.png])

Figure 1.12 An MRI machine (This image is licensed under the Creative Commons Attribution-Share Alike 4.0 International license. [available at: https://commons.wikimedia.org/wiki/File:MRI_Scanner_at_Narayana_Multispeciality_Hospital,_Jaipur.jpg])

MRI has wide applications in the medical field such as examining abnormalities in the brain and other parts of human body. It can also provide a global view of intra- and extra-cranial features.

For the diagnosis of Alzheimer's Disease, PET, CT, and MRI are all usually used. However, CT and MRI scans are used more frequently than PET (Alzheimer's Society, 2018). CT and MRI can show structural changes to the tissues in the brain but PET shows the activity of the brain. Some damaged or diseased blood vessels may result in changes in the brain, and MRI can detect such changes and indicate that a stroke has occurred at some points (Alzheimer's Society, 2018). In addition, MRI and CT are used to identify some other symptoms of dementia, e.g., tumours or build-up of fluid inside the brain (Alzheimer's Society, 2018). MRI scans are considered in many studies because it has no radiation harm for patients and is described in greater detail in Section 1.4.4.

4) Other Medical Imaging Techniques

Electroencephalogram (EEG) and magnetoencephalogram (MEG) also provide useful indices of cognitive decline and serve as screening methods (Bamidis et al., 2014; Klados, Styliadis, Frantzidis, Paraskevopoulos, & Bamidis, 2016). EEG is a medical imaging method that captures the electrical activity produced by the brain's structures through sensors placed on the scalp (Teplan, 2002). An example of an EEG result is shown in Figure 1.13. MEG quantifies the magnetic fields originating from electric currents within the brain. MEG offers exceptionally precise temporal resolution for tracking neuronal activity. An example of an MEG scan is shown as Figure 1.14. As can be seen from Figures 1.13 and 1.14, EEG generates line charts but MEG can map line charts to brains and is presented as colour-coded images. Both of EEG and MEG are non-invasive tests.
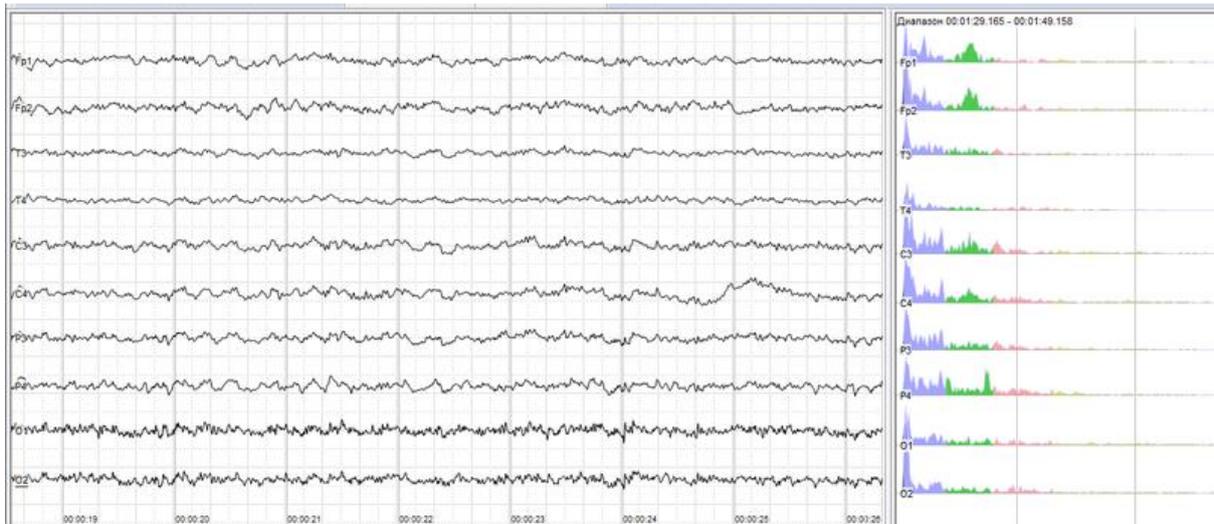
Figure 1.13 An EEG result (This image is licensed under the Creative Commons Attribution-Share Alike 4. 0 International license. [available at: https://commons.wikimedia.org/wiki/File:Human_EEG_without_alpha-rh ythm.png])
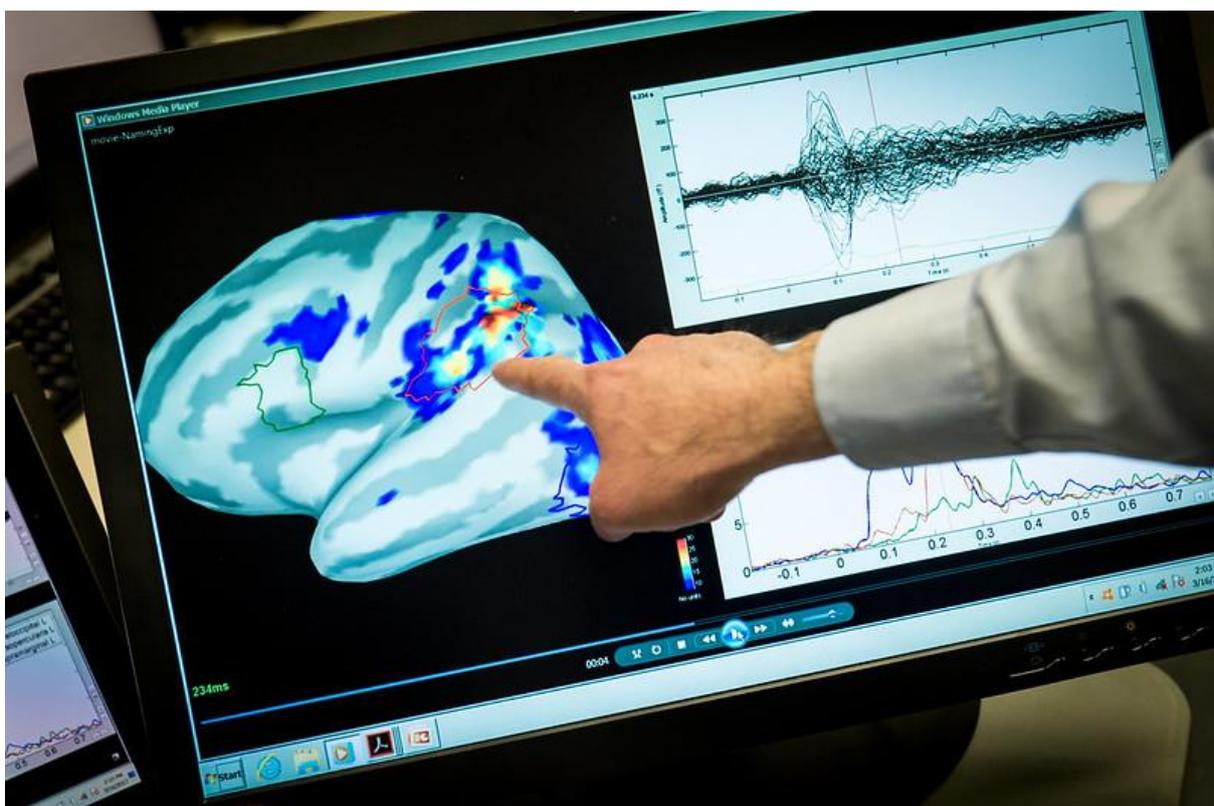
Figure 1.14 An MEG result (This image is licensed under the Attribution-NonCommercial 2.0 Generic (CC BY-NC 2.0) license. [available at: https://www.flickr.com/photos/airmanmagazine/34051669376])

Compared with EEG and MEG, MRI is more common in diagnosing AD. This is because MRI is more readily available in clinical settings and it has a long history of use in AD research and clinical practice. In addition, several biomarkers, e.g., hippocampal volume, cortical thickness,

etc., are well established in the history of use in AD diagnosis. However, EEG and MEG have their own merits and they can be used for special clinical purposes in the process of diagnosing AD. They could be potentially used more widely in future AD research.

1.4.4.   Details of Magnetic Resonance Imaging (MRI)

1.4.4.1.   *T1-weighted and T2-weighted Magnetic Resonance Imaging*

MRI, which has also been named nuclear magnetic resonance imaging (NMRI), has been widely used for imaging the brain, spinal cord, and vascular tissue (Kasban et al., 2015). Besides the fact that it does not use ionizing radiation, it is non-invasive, and is capable in identifying different soft tissues based on contrast. These reasons make MRI an important tool in the diagnosis of different diseases. Cryogenic superconducting magnets are used in MRI scanners, and magnetic fields generated by the magnets are employed to align randomly directed hydrogen proton of tissues. External Radio Frequency (RF) is then used to excite those protons out of alignment, which makes protons spin out of equilibrium. After the RF is turned off, the protons will return to the previous status, equilibrium status in only magnetic fields, and energy will be given off in this process. The MRI scanner can detect the energy changes from different tissues and display the image based on these changes (Grover et al., 2015). In summary, magnetic fields and radio waves can measure the water content in tissues based on the returned signals of energy changes after the initial RF.

After the RF is turned off, the proton returns to the previous status and this process is termed as relaxation (Grover et al., 2015). There are two types of relaxation:

- Longitudinal relaxation time (T1)
- Transverse relaxation time (T2).

Relaxation is a measure of the time taken for a proton to reach equilibrium from excitation (Hartley, Damon, Patterson, Long, & Holt, 2012). In T1-weighted images, water and cerebrospinal fluid (CSF) are dark due to long T1 values. Conversely, water and CSF are bright in T2-weighted images. T2 is the measure of time taken for energy exchange of proton spins to lose their coherence (Hartley et al., 2012). Figure 1.15 shows examples of T1-weighted and T2-weighted images. Table 1.1 shows the grey levels of different tissues in T1-weighted and T2-weighted images respectively.

As for the application, coronal T1-weighted images are commonly used for the hippocampus, which are usually outlined manually (Symms, Jäger, Schmierer, & Yousry, 2004). They also

have good performances on cross-sectional and longitudinal studies, such as identifying the changes in the volumes of the hippocampus in AD (Symms et al., 2004) over time. T2-weighted images are often used for detecting the level of iron concentration, for the detection of subarachnoid haemorrhage, and acute cerebral haemorrhages or micro-haemorrhages (Symms et al., 2004).
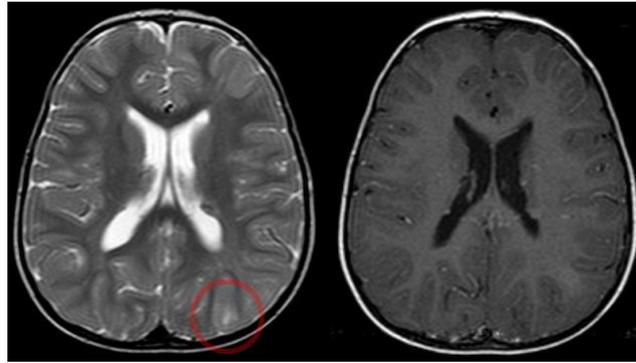
Figure 1.15 Axial planes of the brain using T2-weighted (left) and T1-weighted (right) MRI (This image is licensed under the Creative Commons Attribution-Share Alike 3.0 Unported license [available at: https://commons.wikimedia.org/wiki/File:Tuberoese_Sklerose_MRT_T2_T1_Hamartom_subcortical_und_Tube ra_subependymal.jpg])

Table 1.1 Grey level of different tissues in T1 and T2-weighted images

Tissues	T1-weighted Image	T2-weighted Image
CSF	Dark	Bright
White Matter	Light	Dark Grey
Cortex	Grey	Light Grey
Fat	Bright	Light
Inflammation	Dark	Bright
Air	Very Dark	Very Dark
Disk	Grey	Bright

In addition to relaxation time, magnetic strength is another factor that influences the quality of MRI scans. Nowadays, most clinical research employs magnet fields with strength of 1.5 Tesla (T) or 3.0T (Hartley et al., 2012). An MRI scanner with 3.0T, a higher magnetic field, can improve the signal-to-noise ratio (SNR), and provide higher spectral, and temporal resolution (Grover et al., 2015), which means it can generate higher quality images. However, although 3T MRI has been said to be more accurate than 1.5T, Wardlaw et al. (2012) found little evidence of improved diagnosis. Furthermore, considering high costs of 3T MRI, 1.5T is still more commonly used in the diagnosis of AD.

1.4.4.2.   Types of Magnetic Resonance Imaging

MRI is categorised into structural MRI (sMRI) and functional MRI (fMRI). sMRI, also referred to as "MRI", is used to examine the anatomy and pathology of the human body. fMRI is employed to indicate the metabolism changes in the regions of the brain as time goes by (Glover, 2011). However, in fact, sMRI and fMRI are not easy to be divided because structure and function are often interlinked in the brain (Symms et al., 2004).

There has been a recent trend in attempts to detect changes in tissue microstructure through MRI (D. C. Alexander, Dyrby, Nilsson, & Zhang, 2019; Cercignani & Bouyagoub, 2018; Jillings et al., 2020). Diffusion-weighted imaging (DWI) and diffusion tensor imaging (DTI) are two key microstructure imaging techniques.

DTI sometimes is simplified as diffusion MRI or dMRI. This type of MRI is used to quantify water molecule movement (Grover et al., 2015). Because more than 60% of the human body consists of water and the molecules follow Brownian motion, DWI can generate signal contrast based on differences in Brownian motion to measure molecular function and the micro-architecture of the human body (Baliyan, Das, Sharma, & Gupta, 2016). DWI is often applied to brain tumours, paediatric brain development and ageing, acute brain ischaemia, and white matter disease. White matter is one of most studied objects in the diagnosis of AD. DW MRI is sensitive to microstructure changes in white matter, so that changes in tissue orientation patterns can be detected through changes in the degree of anisotropy in MRI voxels (Baliyan et al., 2016).

In addition to DWI technique, DTI is another technique applied to MRI. DTI may be used to characterise and map the 3-D diffusion of water as function of spatial location (Alexander, Lee, Lazar, & Field, 2007). The diffusion of water within tissues changes while the tissue microstructure and organisation alter. Therefore, DTI MRI is potentially advantageous probes for characterising the effect of disease and ageing on microstructure (Alexander et al., 2007).

1.4.4.3.   Three View Planes of Magnetic Resonance Imaging

In anatomy, parts of the brain can be divided based on three views (or planes): the axial plane, sagittal plane, and coronal plane. The axial plane, or horizontal plane, is from above or below, and is a plane that divides the brain into superior (upper) and inferior (lower) parts, which is shown in Figure 1.16 (a); the sagittal plane is taken from the side and divides the brain into left and right parts, shown in Figure 1.16 (b); the coronal plane is taken from the front or back, and

separate the brain into posterior and anterior parts, shown in Figure 1.16 (c).

(a) Axial Plane   (b) Sagittal Plane   (c) Coronal Plane

Figure 1.16 Diagram of axial, sagittal, and coronal planes (Original images from ADNI adapted Yuyang Liu)

In the study described in this thesis, according to the above, 1.5T T1-weighted sMRI scans were used as the input data to diagnose AD, and its three anatomical planes and how to analyse it will be introduced in Section 3.4.

1.5. Structure of the Thesis

This thesis reports the whole process of developing machine learning and deep learning models to diagnose different stages of AD using brain MRI scans which formed the research. Overall, the thesis presents eight chapters that explain the development and experimentation of machine learning in diagnosing different stages of AD using brain MRI scans.

Chapter 2 presents a review of studies that have applied machine learning in healthcare. The chapter gives a brief description of studies that used machine learning in general healthcare fields and it then specifically describes studies that applied machine learning in diagnosing AD in more detail. The limitations and challenges in previous research for diagnosing AD using machine learning and MRI scans are also described in Chapter 2. The chapter synthesises the reviewed literature, identifies the gaps of the research area, and explain how this study attempts to fill some of those identified gaps.

Chapter 3 describes the methodology employed for this study. The study comprises three tasks (sub-studies):

Study 1 - using machine learning and brain MRI scan for discriminating between CN and AD;

Study 2 - discriminating between sMCI and pMCI;

Study 3 - discriminating between CN and pMCI.

The chapter includes the whole workflow of the three studies, which consists of data collection, data pre-processing, detection of AD-relevant regions, feature extraction, and clustering. Specifically, it presents the database that the brain MRI were collected from, the steps of MRI scan pre-processing and the pre-processing tool, the statistical method used for detecting AD-relevant regions, the deep learning model used for feature extraction, and the algorithm used for clustering. Furthermore, it also describes the measures used for estimating the performance of the model in this study.

Chapter 4 provides a detailed information of the experiments in the first task, i.e., discriminating between CN and AD. It contains details of the exact operations of each steps described in the methodology chapter (Chapter 3), including an introduction to the software and tools used in this study, how the data were filtered and collected from the database through the database search engine, how the environment was configured before using the pre-processing tools. It also includes how the two pre-processing tools were used (how the parameters were set), how the tools were used for implementing the statistical method to detect AD-relevant regions and how the regions were noted using the tool, how the deep learning model was developed and used, and how the clustering algorithm was developed and used. The results of this work are then presented in Chapter 4. The results discussed in this chapter include the comparison of the pre-processing quality between using two pre-processing tools, the detected AD-relevant regions, and clustering results. The analysis of the results and the discussion are also presented in this chapter.

Chapter 5 and Chapter 6 are broadly similar to Chapter 4, except that the tasks described are different. Chapter 5 explicitly describes the experiment of the second task, which involves discriminating between sMCI and pMCI, while Chapter 6 describes the third task, which involves discriminating between CN and pMCI.

Chapter 7 brings together the findings of the three chapters (Chapter 4, 5, and 6) and presents the discussion over the three studies (discriminating between CN vs. AD; sMCI vs. pMCI; CN vs. pMCI). It first answers the research questions posted in Section 1.3 and then holistically analyses and compares the results over the three tasks. The chapter then discusses the overall limitations of the study.

Chapter 8 presents the overall wider conclusions of the study. It first summarises the thesis followed with a summary of the methodology used in this study. It also reviews the key findings

arising from this thesis and the contribution to the new knowledge. Suggestions for future work are then presented. Finally, a conclusion rounds the thesis off, including the challenges encountered in the process of the study and the outlook from this study on the future research.

1.6. Conclusion

In this chapter, the background of the research on AD and preliminary information about AD, MRI, and machine learning (including deep learning) were introduced. According to these sections, diagnosis of AD using machine learning (including deep learning) based on brain sMRI scans has been an important study theme around the world. In the next chapter, the literature review of the related studies in this field will be described and will be discussed in detail.

Chapter 2   Literature Review

2.1.   Introduction

Chapter 1 described the impact of the dementia around the world and research on Alzheimer's disease (AD). It then concluded by stating the overall aim of this thesis, i.e., to develop machine learning models to discriminate between different stages of AD using unsupervised learning methods and a limited number of labelled magnetic resonance imaging (MRI) scans. Finally, it provided background information relevant to this research. This chapter, therefore, presents the primary literature review that considers previous research in the application of machine learning in healthcare and, more specifically, in diagnosing AD using MRI scans. The aim of this chapter is not only to review this body of research in diagnosing AD using machine learning, but also to identify the limitations and gaps that are commonly reported in the previous studies. By describing the limitations and gaps, the importance of the research in this thesis is highlighted.

The literature review in this thesis includes undertaking a comprehensive search and critical assessment of previous studies in applying machine learning to diagnosing AD using MRI scans. Articles and papers for the review were retrieved from various sources, and the search strategy is described in Section 2.2.

The literature review begins with a review research in the application of machine learning in general healthcare, which is presented in Section 2.3. It introduces the previous research in the application of machine learning in three aspects of healthcare: public health, clinical decision support systems, and others.

Section 2.4 narrows the scope of the literature, describing the application of machine learning in diagnosis using modern medical images. It presents studies in applying machine learning to medical images for diagnosis of different diseases.

Section 2.5 further narrows the scope of the literature from Section 2.4, introducing the application of machine learning in diagnosing AD using MRI scans.

2.2.   Search Strategy

The literature in this study was searched from the Internet using the following web search engines:

(1) Google Scholar: a free search engine for scholarly literature in several publishing format and disciplines;

(2) Medline: the National Library of Medicine's (NLM) premier bibliographic database that contains more than 29 million references to journal articles in life sciences with a concentration on biomedicine;

(3) StarPlus: a search engine that is provided by, and accessed through, the library at the University of Sheffield;

(4) Web of Science: a paid-access platform (accessed through the University of Sheffield subscription) that provides (typically via the internet) access to multiple databases that provide reference and citation data from academic journals, conference proceedings, and other documents in various academic disciplines.

The review has two sections. The first section is the general review of the literature about machine learning application in healthcare (Section 2.3). Therefore, the set of keywords "machine learning" and "healthcare" was used for searching the literature. In addition, the studies on it can be divided into three types: (1) public health; (2) clinical decision support systems; (3) others. Hence, for machine learning applied in public health, the keywords "machine learning" and "public health" were used in the search engines; for machine learning applied in clinical decision support systems, which is namely disease diagnosis or prediction, the keywords "machine learning" and "diagnosis" were used in the search engines; other studies searched by the keywords "machine learning" and "healthcare" but not belonging to the two types above were also included.

In the second section, the studies on machine learning based diagnosis of Alzheimer's disease using MRI scans specifically are reviewed. The concept of machine learning contains multiple types of techniques, including deep learning, supervised learning, unsupervised learning, etc., thus the studies using some methods may not be easily found if only one set of keywords is used in the search engine. As a result, different sets of keywords should be considered. In each search engine above, five sets of keywords were used. The first set of keywords includes "Alzheimer's disease", "MRI", and "machine learning"; the second set includes "Alzheimer's disease", "MRI", and "deep learning"; the third set has "Alzheimer's disease", "MRI", and "supervised learning"; the fourth set has "Alzheimer's disease", "MRI", and "unsupervised learning". However, some papers may use "clustering" instead of "unsupervised learning", thus the fifth set could be "Alzheimer's disease", "MRI", and "clustering". Table 2.1 shows the

results of the five sets of keywords in the four search engines. Specifically, Table 2.2 shows the search results of using keywords "Alzheimer's disease", "MRI", and "deep learning" using google scholar.

Table 2.1 Search results of the second section using the four sets of keywords in four search engines

Keywords	Search Engines			
	Google Scholar	Medline via Ovid	StarPlus	Web of Science
"Alzheimer's disease" "MRI" "machine learning"	15400	9546	1217	1327
"Alzheimer's disease" "MRI" "deep learning"	7320	10014	713	661
"Alzheimer's disease" "MRI" "supervised learning"	3570	10730	215	138
"Alzheimer's disease" "MRI" "unsupervised learning"	1510	3523	27	65
"Alzheimer's disease" "MRI" "clustering"	16200	4043	303	476

Table 2.2 Search results of using keywords "Alzheimer's disease", "MRI", and "deep learning" in Google Scholar

	1950 – 2022	2000 – 2022	2010 – 2022	2015 – 2022
Google Scholar	6970	6960	6890	6750

From Table 2.1, it can be seen that when the keyword "machine learning" was changed to "unsupervised learning", the number of search results had a considerable drop across all four sources. There was also a drop across all the sources when the keyword "supervised learning" was changed to "unsupervised learning". In Medline via Ovid, the result of "supervised learning" was also more than the keywords "machine learning" and "deep learning". However,

through observation of the results, some papers about unsupervised learning and deep learning were also included. This could be a reason why the search result of using "supervised learning" was more than using "machine learning" and "deep learning".

In addition, amongst those results searched using "unsupervised learning", many results were not entirely related to the key words – many of them just mentioned "unsupervised learning" in the paper instead of applying unsupervised learning methods to diagnosing AD. However, the number of search results using "clustering" was greater than using "unsupervised learning". It is because many articles involving "clustering" are about segmentation, which means that many studies used clustering techniques to segment brain MRI scans into different tissues, instead of using them to diagnose AD. As can be seen from Table 2.2, the number of the literature about "deep learning", "MRI", and "Alzheimer's disease", published between 1950 and 2022 is 6970, and 6750 out of the 6970 were published between 2015 and 2022. This suggests that deep learning started to be popular from last 5 to 10 years. Therefore, it can partially explain the fact that the number of results of using "deep learning" is also smaller than that of using "machine learning" (except Medline via Ovid) in Table 2.1. It is also because many studies that employed deep learning sometimes used the name of deep learning models directly or used the term "neural network", instead of using the term "deep learning", and this also lead to the lower number of publication using "deep learning" as a term.

In addition, as can be seen from Table 2.1, the keywords "Alzheimer's disease", "MRI", and "unsupervised learning" achieved the lowest number of studies amongst the four sets of the keywords. This indicates that fewer research in AD used unsupervised learning. Furthermore, not all of them used unsupervised learning methods to discriminate between different stages of AD. For example, Liu et al. (2013) employed an unsupervised learning algorithm, local linear embedding (LLE), for MRI data transformation instead of classification; Razavi et al. (2019) employed an unsupervised neural network structure to learning features; Baskar et al. (2019) used a fuzzy c-means method to remove suspicious data from the original dataset; Lu et al. (2018) utilised an unsupervised learning method to do pre-training and feature learning. This appears that the number of studies using unsupervised learning methods to distinguish AD is relatively low (compared to supervised learning methods).

From the results of the search, it can be summarised that the support vector machine, the k-nearest neighbours, the random forest, and the convolution neural network were the most widely-used supervised learning models for diagnosing AD using MRI scans; the semi-

supervised generative adversarial network was one of the most widely used models in semi-supervised learning; however, the number of studies using unsupervised learning methods was much smaller than using supervised and semi-supervised learning methods. The k-means model was the most widely used in unsupervised learning methods. Further detail of these previous studies and models is provided in Section 2.5 below.

2.3.   Machine Learning Application in Healthcare

In recent years, machine learning, a technique developed in 1950s, has gained much attention in various industries due to the explosion of data, and one field in particular, in which it has widespread societal impact is healthcare. With the prevalence of smart phone, smart watches, and other wearable devices that constantly collect a plethora of heath data, using machine learning to analyse them has become more and more important and a focus of research. According to Alanazi (2022), generally, the application of machine learning in different areas of healthcare can mainly be divided into two areas: public health and clinical decision support systems. However, there are other areas in addition to these two, and the literature regarding these other areas will briefly reviewed after the two main areas. The studies in these areas are described below.

2.3.1.   Public Health

In recent years, with the explosive growth of the data all over the world, machine learning has been applied in many fields, and the public health field has also benefited from it. According to Mhasawade, Zhao, & Chunara (2020), studies on the machine learning applied to the public health can be categorised into the following areas described below.

Identification of factors. It is to learn what factors contribute to a public health problem or health outcome. For example, Bhatt et al. (2013) used a boosted regression tree to estimate the occurrence of dengue with local spatial variations influenced strongly by rainfall, temperature, and the degree of urbanization. Sethi & Mittal (2020) applied three different machine learning models to correlated air pollutants and the air quality index during the lockdown period and COVID-19 fatalities in Delhi. They monitored the effect of the lockdown on the different air pollutants due to COIVD-19 pandemic and identified the ones that affect COVID-19 fatalities. Machine learning methods are likely to discover hidden factors that traditional methods cannot find, and identifying these factors could potentially provide more information for disease research and disease control.

Mining Public opinions. Machine learning has been used to learn from public opinions on a health problem. It is important for researchers to perceive popular opinions, but the popular opinions are often fragmented, especially on the Internet, which is difficult to summarise them and to obtain useful information from them. Therefore, many researchers started to use machine learning methods to mine used information: some examples of these are provided below as an illustration. Du, Xu, Song, & Tao (2017) used a hierarchical machine learning based sentiment analysis system to extract public opinions to human papillomavirus (HPV) vaccines from Twitter. Kang, Wang, Zhang, & Zhou (2017) employed opinion mining techniques to classify tweets about a new school meal policy into positive, negative, and neutral categories, and this was important for the prevention of prevention of childhood obesity. Xue et al. (2020) conducted a study to analyse COVID-19 discussions on Twitter using unsupervised learning methods and natural language processing (NLP) approaches. Gupta et al. (2021) used machine learning and NLP techniques to filter users' perspectives of the effect of weather on Twitter during COVID-19 pandemic and analysed how they evolved with respect to real world events and time.

Prediction of outcomes. Machine learning, especially supervised learning, can learn information from historical dataset and it has capabilities to generate values for unknown data and to predict the outcomes of future events. For example, Ortega Hinojosa et al. (2014) predict smoking and obesity prevalence using a lasso-based variable selection method and a two-level random effects regression with a Poisson link clustered on state and county. Dunstan et al. (2020) implemented three machine learning algorithms to predict prevalence of obesity using purchase data from 79 countries. Ahmad et al. (2021) used a Shallow Single-Layer Perceptron Neural Network (SSLPNN) and a Gaussian Process Regression (GPR) model to classify and predict the confirmed COVID-19 cases in five geographically distributed regions of Asia.

Table 2.3 summarises the examples of the studies in each category mentioned above. Based on the four categories, it can be seen that machine learning has started to be used in public health in recent years. In public health field, machine learning is especially important to health officials. It can improve their abilities to respond to future pandemics and other public crises through the predictive results or useful information behind the events learned by machine learning models.

Table 2.3 Examples of studies on machine learning application in public health

Types of study	Studies	Application of Machine Learning

Identification of factors	Bhatt et al. (2013)	Estimate the dengue influenced by multiple factors.
	Du, Xu, Song, & Tao (2017)	Extract public opinions to HPV vaccine from Twitter.
	Kang, Wang, Zhang, & Zhou (2017)	Classify opinions from Twitter about a school new meal policy.
Mining Public Opinions	Xue et al. (2020)	Analyse COVID-19 discussions on Twitter.
	Gupta et al. (2021)	Filter tweets about effect of weather during COVID-19 pandemic and analyse them.
Prediction of Outcomes	Ortega Hinojosa et al. (2014)	Predict smoking and obesity prevalence.
	Dunstan et al. (2020)	Predict prevalence of obesity.
	Ahmad et al. (2021)	Classify and predict COVID-19 cases in five regions of Asia.
	Sethi & Mittal (2020)	Identify the effect of different air pollutants that affect COVID-19 fatalities.

2.3.2. Clinical Decision Support System

Detecting and diagnosing diseases using machine learning are increasingly studied in recent years. Machine learning methods can achieve higher performances and assist clinicians to support clinical decisions. Vaka, Soni, & K. (2020) detected breast cancer using a Deep Neural Network with Support Value (DNNS). They achieved the accuracy of more than 90% in four segmentation images. Xie et al. (2021) employed six machine learning models to discover diagnostic biomarkers of early lung cancer. They achieved a sensitivity of 98% and a specificity of 100%. Machine learning methods can potentially increase the performance of the diagnosis or early diagnosis and this is important for those patients with fatal diseases.

Additionally, machine learning can diagnose multiple cases once and this can save considerable time, especially for pandemic, e.g., COVID-19. Elaziz et al. (2020) used fractional multichannel exponent moment to extract features from the chest X-ray images and classified the extracted features into COVID-19 and non-COVID-19 classes using K-nearest neighbours (KNN) classifier. They achieved accuracies of 96.09% and 98.09% respectively in the two datasets. Zoabi, Deri-Rozov, & Shomron (2021) trained a decision tree on records of sex, age, symptoms of COVID-19, and known contact with an individual confirmed to have COVID-19, and then predicted the COVID-19.

Furthermore, machine learning methods can be also applied to diagnose other diseases. Zhang et al. (2020) employed a support vector machine to prognose hepatocellular carcinoma using gene expression profiles. It yielded an accuracy of 100%, a sensitivity of 100%, and a

specificity of 100%. Kim, Cho, & Oh (2017) collected the data of retina nerve fiber layer thickness and visual field, and then employed a random forest, a support vector machine, and a KNN classifiers to diagnose glaucoma. The random forest achieved the best performance: accuracy of 0.98, sensitivity of 0.983, specificity of 0.975. Sontakke et al. (2017) used a support vector machine (SVM) and a back propagation neural network to diagnose liver disease. The proposed method achieved an accuracy of 71% using SVM and an accuracy of 73.2% using the neural network. Wu et al. (2019) employed multiple machine learning methods to predict fatty liver disease, and the random forest outperformed the other methods, which yielded an accuracy of 87.48%. As can be seen, these studies achieved satisfactory performances by using machine learning techniques. Figure 2.4 summarises the studies above in this section.

Machine learning methods have potentials to increase the performance of the diagnosis, as well as saving time and human resource, and the most importantly, saving lives, especially for those people with fatal diseases, e.g., cancer, etc.

Table 2.4 Examples of studies on machine learning application in clinical decision support systems

Studies	Application of Machine Learning
Vaka, Soni, & K. (2020)	Detect breast cancer.
Xie et al. (2021)	Discover diagnostic biomarkers of early lung cancer.
Elaziz et al. (2020)	Extract features from chest X-ray images and diagnose COVID-19.
Kim, Cho, & Oh (2017)	Diagnose glaucoma.
Zoabi, Deri-Rozov, & Shomron (2021)	Predict COVID-19.
Zhang et al. (2020)	Prognose hepatocellular carcinoma.
(Sontakke et al., 2017)	Diagnose liver diseases.
(Wu et al., 2019)	Predict fatty liver disease.

2.3.3. Other Applications of Machine Learning in Healthcare

In addition to the areas above, studies have been undertaken for other purposes. The first can be called diagnosis aid, e.g., segmentation of medical images. Segmentation can aid lesion or tumour localisation or measurement and assessment of treatment response, which is important for the clinical diagnosis (though it is not directly helpful in the diagnosis). Xu et al. (2019) applied a convolutional neural network to segment breast ultrasound images into four major tissues: skin, fibroglandular tissue, mass, and fatty tissue. Almajmaie et al. (2022) proposed a new deep learning model, combining SegNet and U-net techniques, to segment brain tissues. Kharghanian et al. (2012) employed an SVM classifier to segment retinal blood vessels based on Gabor wavelets and line operators. The best accuracy they achieved was 96.39%. Bonte et

al. (2018) presented a novel approach using a random forest model combining voxel-wise texture and abnormality features and then using a clustering algorithm to segment brain tumours from the brain MRI scans. This is pertinent to the research reported in this thesis because discovering abnormalities related to AD in the brain and segment them from the MRI scans are important for diagnosing AD using MRI scans. With the help of machine learning, medical images can be segmented automatically instead of manually by practitioners. This saves much time and human resource for this stage or even the whole study involving segmentation.

Another area can be summarised as pathological research aid. Studies in this area used machine learning to help study diseases. Alexander et al. (2020) proposed using four clustering algorithms to identify subtypes of AD: k-means, kernel k-means, affinity propagation, and latent class analysis, based on patients' electronic health records. This is highly relevant to the research reported here because stable mild cognitive impairment (sMCI) and progressive mild cognitive impairment (pMCI) can be regarded as subtypes of AD and their study can provide potentially ideas to diagnose sMCI and pMCI cases. Ganggayah et al. (2019) applied a decision tree, a random forest, neural networks, an extreme boost, a logistic regression, and a support vector machine, to predict factors for survival of breast cancer. As can be seen, machine learning becomes a new tool for pathological research and it can potentially provide different views to the research, which may lead to new discoveries. Figure 2.5 summarises the studies discussed in this section.

Table 2.5 Examples of studies on machine learning applications in other research purposes

Studies	Application of Machine Learning
Xu et al. (2019)	Segment breast ultrasound images into four tissues.
Almajmaie et al. (2022)	Segment the brain tissues.
Alexander et al. (2020)	Identify subtypes of AD.
Ganggayah et al. (2019)	Identify factors for survival of breast cancer.
Kharghanian & Ahmadyfard (2012)	Retinal blood vessel segmentation.
Bonte et al. (2018)	Segment brain tumours in MRI scans.

As can be seen from Section 2.3.1 to 2.3.3, machine learning has become an important role in healthcare research. It can help health officials make decisions for public and help researchers study the health problems in depth. In addition, it aids clinical practitioners to diagnose the diseases accurately as well. These all benefit public to survive and to prevent from diseases.

To apply machine learning methods in the healthcare field, data are necessary, and the data have different types and formats. Both non-imaging and imaging data can be used to predict or diagnose neurocognitive diseases, e.g., dementia and AD. In Section 2.4, research literature on the application of machine learning in diagnosis of AD using non-imaging data will be described.

2.4. Application of Machine Learning in Diagnosis of Alzheimer's Disease Using Non-imaging Data

Previous studies have also shown that non-imaging data can help and provide good diagnostic value to problems of identifying AD. Huan Wang et al. (2022) applied different machine learning and deep learning models to non-imaging data (nursing data and demographic data) and selected non-imaging features to diagnose AD. The best overall accuracy they achieved was 86.24%. L. Liu et al.(2020) used speech data and different machine learning models to diagnose AD. The best results were yielded by LogisticRegressionCV model: an accuracy of 83.3% for one dataset and an accuracy of 84.4% for the other dataset was obtained.

In addition to traditional clinical non-imaging data, the data collected from serious games (in-game metrics) have potential to help diagnosing AD. Although no study using in-game metrics to diagnose AD was found in the literature, studies used in-game metrics to diagnose other neurocognitive diseases and dementia were identified and are described as follows.

Petsani et al. (2022) predicted well-being digital biomarkers from Parkinson patients during interactions with serious games. They utilised decision tree classifiers and Gini index to classify the patients into different classes. Their method achieved an accuracy of 84.6% and showed that in-game metrics can effectively classify participants into groups of different cognitive and physical states.

Konstantinidis et al. (2021) collected in-game metrics (built-in game performance measures) and correlated these metrics with classical neuropsychological tests. Multilayer perceptron classifiers were used and they achieved an accuracy of 73.53% on classification of CN and mild dementia and an accuracy of 70.69% on classification of CN and MCI.

In Section 2.5, research literature on the application of machine learning in diagnosis of AD using MRI scans will be described.

2.5. Application of Machine Learning in Diagnosis of Alzheimer's Disease Using MRI Scans

Section 2.3 has presented the studies applying machine learning approaches to public health. One of the applications in public health is the clinical decision support system, and medical images are often used in the clinical decision support system to diagnose diseases. Therefore, previous studies applying machine learning techniques to medical images were reviewed in Section 2.4. MRI scans are commonly used for diagnosing AD combined with machine learning approaches. Hence, in this section, previous studies that have used machine learning models and MRI scans to diagnose AD are discussed in detail.

2.5.1. Diagnosis (Classification/Clustering)

This section presents different studies that have used machine learning techniques and MRI scans to diagnose AD according to different types of machine learning: supervised learning, semi-supervised learning, and unsupervised learning. For each type of the machine learning, the studies are described based on different commonly used models for each type.

2.5.1.1. Supervised Learning Methods used to Diagnose Alzheimer's Disease

Thus far, four supervised learning models have been frequently used in previous studies to diagnose AD using MRI scans: support vector machine, k-nearest neighbour, random forest, and convolutional neural network. The four models are also presented with the reviewing of the studies.

Support Vector Machine (SVM)

SVM is a method developed by Cortes et al. (1995), which tries to find an optimal hyperplane in N-dimensional space with the highest margin between classes (Figure 2.1). SVM has been employed in a wide range of real-world problems, e.g., handwritten digital recognition, image classification, and object detection. It is increasingly used because of its pure mathematical foundation and this method seems to perform well, at least to a certain extent, in various real-world problems (Chandra & Bedi, 2021). Figure 2.1 shows an illustration of the linear SVM algorithm, and if the performance of the linear model is not sufficiently high, non-linear SVMs or SVMs with kernels are often considered. Using kernels, SVM can enlarge the feature space using mapping procedure to transmit data to a space where the data can be linearly separable (Boateng, Otoo, & Abaye, 2020), which means that it uses a linear classifier to solve a non-linear problem.
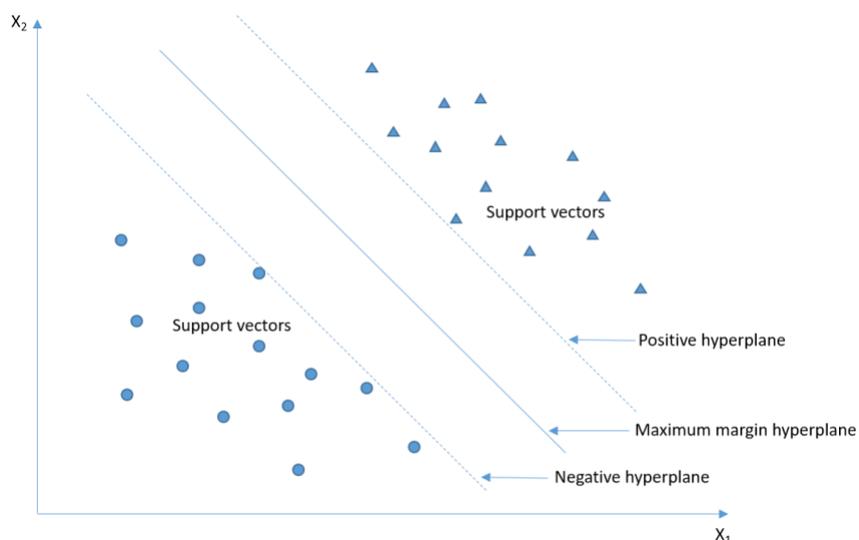
Figure 2.1 A simple illustration of the SVM algorithm in two-dimensions (Developed by Yuyang Liu)

SVM (and SVM with kernels) have been used for diagnosing AD using MRI scans in many previous studies. Lee et al. (2019) applied a linear SVM classifier to features extracted from a neural network structure. They achieved an overall accuracy of 92.75% for the classification of AD and CN, an accuracy of 88.52% for the classification of sMCI and pMCI, and an accuracy of 71.18% for triple classification (i.e., AD, MCI, and CN).

In Lin et al.'s (2022) study, multimodal MRI features of 7 brain networks and 90 ROIs were fed into an SVM classifiers to diagnose subjective cognitive impairment. They achieved accuracies of 79.94% and 83.13%, respectively, in Chinese and Alzheimer's Disease Neuroimaging Initiative (ADNI) cohorts, using the features of MRI alone. Bigham et al. (2022) applied various types of SVMs to features in superficial white matter from diffusion tensor imaging (DTI) scans to discriminate between three stages of AD (CN vs. AD, AD vs. MCI, and CN vs. MCI). The linear SVM became the best classifier for AD and CN, with an accuracy of 95.8%; the quadratic SVM achieved the best accuracy for classification of CN and MCI, which is 83.3%; the best classifier for MCI and AD was Gaussian SVM, whose accuracy was 83.3%. Vichianin et al. (2021) applied an SVM algorithm to different combinations of tissues of the brain (features) and the highest accuracy they achieved was 90.74% with a ROC of 0.925.

Although these studies achieved relatively good performance in diagnosing AD using linear (traditional) SVM, they were required to ensure that the features were linearly separated. For

example, Lee et al. (2019) employed a linear SVM directly because they worked much on feature extraction and feature reduction, which made the data suitable for SVM. Hence, the traditional SVM highly requires good feature extraction methods before it is applied. However, the data can be separated only using a hyperplane in linear SVM so, in many complex cases, traditional linear SVM cannot be used directly. Therefore, kernels, which map the input into another space by a linear or non-linear function, have been used in many previous studies.

The radial basis function (RBF) is one of the most commonly used non-linear kernels for the SVM. Khedher et al. (2015) first used partial least squares (PLS) and principle component analysis (PCA) respectively to extract features from grey matter and white matter of the MRI scan, and they then applied linear SVM and SVM with RBF to the extracted features to discriminate between CN and AD, between CN and MCI, and between MCI and AD. PLS combined with SVM with RBF achieved the best performance on classifying CN and AD, whose accuracy was 88.49%, when using both grey matter and white matter. However, when classifying CN and MCI, using PLS combined with linear SVM had the best performance, this achieved an accuracy of 81.89% using both grey matter and white matter. When PLS was combined with linear SVM, this also yielded the best accuracy of 81.89% for discriminating between MCI and AD. Kumari et al. (2022) applied an SVM with RBF to T1-weighted MRI scans (they did not provide details of the features extracted from the scans), and they achieved an accuracy of 96.92% for classification of CN and AD, an accuracy of 86.36% for classification of CN and MCI, and an accuracy of 77.27% for classification of MCI and AD. Lazli et al. (2018) applied a linear SVM, a polynomial SVM, and an SVM with RBF to MRI scans with different percentage of additive noise. The accuracy yielded by the proposed method for the noisiest MRI scans was 75% using RBF-based SVM. Ortiz et al. (2013) also used a kernel based SVM, with RBF, to discriminate between AD and CN. They achieved an accuracy of over 90% and a sensitivity of 95%.

Guo et al. (2017) combined the brain region features and subgraph features together and applied a multi-kernel SVM to the features for classification. The accuracy of the proposed method was 91.6% for classification of CN and AD. Beheshti & Demirel (2016) employed both linear and non-linear SVMs to discriminate between AD and CN using features of the grey matter extracted using a t-test based feature-ranking approach. Their proposed method achieved a mean accuracy of 86.76% using both linear and non-linear SVMs. After improvement helped by data fusion of the features (clusters) in their proposed method, the linear SVM achieved the

best accuracy of 96.32% and non-linear SVM achieved an accuracy of 95.59%. Bron et al. (2021) applied a linear kernel SVM to modulated grey matter maps to classify AD and CN data. The method yielded an AUC of 0.940. They also used the SVM to predict the conversion in MCI and achieved an AUC of 0.756, which was significantly higher than CNN in the study.

Although the kernel-based SVM can solve some problems of the basic SVM, choosing an appropriate kernel function becomes a non-trivial task (although most studies chose RBF as the kernel). In other words, when the data dimension is high, it cannot be known where the data should be mapped. Therefore, this will lead to a problem that, when a kernel-based SVM does not perform well compared with other models, it may not be easy to identify whether the SVM itself is not suitable for the data or the right kernel has not been selected. For example, Y. Wang et al. (2019) employed a random forest, logistic regression, and an SVM with a linear kernel in their study to predict progression from MCI to AD. From different angles, the author did not report the SVM as the model having the best performance. However, it cannot be confirmed that the SVM is not an appropriate classifier in their study because an inappropriate kernel might have been chosen, which greatly influenced on the performance of the SVM. In addition to choosing appropriate kernels, kernel-based SVMs are prone to overfitting or sensitive due to noises or outliers. For example, in Lazli et al.'s (2018) study, the SVM did not have good performance on data with noise.

Therefore, some studies employed some optimised versions of SVM algorithms. Zeng et al. (2018) proposed a new switching delayed particle swarm optimisation (SDPSO) algorithm to optimise the SVM algorithm, and they applied the optimised SVM to features extracted from regions of interest (ROIs) by PCA. Their method yielded an accuracy of 81.25% for the classification of CN and AD, 69.23% for the classification of sMCI and pMCI, and 85.71% for the classification of CN and pMCI. Alam et al. (2017) employed a twin SVM, which was developed by Jayadeva et al. (2007), to diagnose AD. The accuracy of the proposed method yielded up to 92.65% over one of the datasets and it yielded up to 96.68% over the other dataset. Kumari et al. (2022) used a hyperparameter tuning-twin SVM (HPT-TSVM) for three binary classifications (AD vs. CN; MCI vs. CN; MCI vs. AD). Their proposed method achieved an accuracy of 96.92% for the classification of AD and CN; an accuracy of 86.36% for the classification of MCI and CN; an accuracy of 77.27% for the classification of MCI and AD.

K-Nearest Neighbour

K-nearest neighbour (KNN) is an algorithm used to classify objects based on the closeness to training examples in the feature space (Boateng et al., 2020). An object (the star in Figure 2.2) is classified by a major vote of its neighbours, which means that it is assigned to the class that is most common amongst its k-nearest neighbour (Figure 2.2), where the value of k is a positive integer. The parameter in KNN is the distance between the object and its neighbours. The measures of the distance that are widely used are: Euclidean distance, Cosine similarity, Hamming distance, Manhattan distance (city block distance), the Minkowski distance, etc.
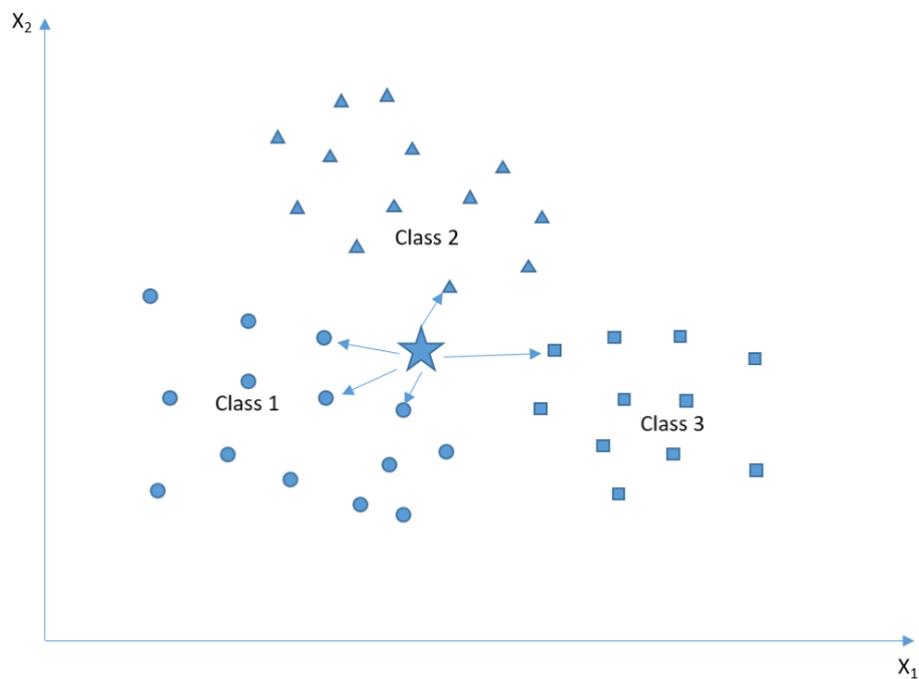
Figure 2.2 A simple pictorial overview of k-nearest neighbour (KNN) algorithm (Developed by Yuyang Liu)

KNN is also commonly used in the diagnosis of AD because of the simple idea of its algorithm. Aruchamy et al. (2020) applied a traditional KNN algorithm to the extracted and reduced features. The best accuracy the method yielded was 93.18% for the classification of CN and AD. The optimal *k* values in four datasets varied in their study, which ranged from 4 to 6, and the Euclidean distance and Manhattan distance outperformed the other distance measures. Battineni et al. (2020) used four classifiers to discriminate between CN and AD, and the KNN performed best for the selective features, which yielded an accuracy of 91.32%. In Battineni et al.'s (2020) study, the *k* value was set as 1 and the distance was calculated by the Euclidean function. Even when applied to selective features, KNN was ranked second amongst the four

classifiers, which yielded an accuracy of 95.92%. Kamal et al. (2022) employed five classifiers to diagnose AD using denoised MRI scans. The accuracy that KNN yielded was more than 93%, which ranked third amongst the four classifiers; however, its sensitivity and recall ranked first, both of which were more than 99%. Deepa et al. (2022) used KNN algorithms, where k values were set as 1, with various distance measures, to diagnose AD, and the best accuracy was obtained by the KNNs with Euclidean distance and Minkowski distance, which was 84%.

Traditional KNN assigns equal weight to all features, and this can lead to a problem that it may reject points that are close in some dimensions but far from others. Therefore, the data may be incorrectly classified due to those non-AD relevant features because the features leading to AD may only occupy a small proportion but non-AD relevant features may be a large proportion of all of the features.

As a consequence, some studies employed weight-based KNN algorithms or fuzzy version of the KNN algorithms. Arabi et al. (2022) used four classifiers to diagnose AD based on the features extracted from MRI scans. The weighted-KNN yielded a testing accuracy of 100% only using seven features, which was the best amongst the four classifiers. Balamurugan et al. (2017) proposed a novel dimension reduction-based KNN and applied it to features extracted from MRI scans. The average accuracy that the model achieved was 99%. Emmanuel et al. (2022) enhanced the traditional KNN algorithm by proposing an enhanced fuzzy KNN (EFKNN) and they applied this enhanced algorithm to features extracted from structural MRI scans. The algorithm achieved an accuracy of more than 90% for classification of AD, CN, and MCI in all runs.

The concept of the KNN algorithm is simple and it is easy to implement. In addition, the KNN algorithm is much faster than some other supervised learning methods, e.g., SVM, linear regression, etc., because it does not have a training phase. It only uses training data when adding a new sample and calculate its k nearest neighbours. However, it does not work well with high dimensions and with noise. The image data often consist of high dimensions, thus features are necessary to be selected and reduced before they are fed into the KNN algorithm. The noise in the raw medical images also needs to be removed. Therefore, from the reviewed literature above, Aruchamy et al. (2020) used extracted and reduced features; Battineni et al. (2020) used selective features; Arabi et al. (2022) only used seven features in their study; Balamurugan et al. (2017) employed a dimension reduction based KNN; Kamal et al. (2022) used denoised MRI scans. Although KNN algorithm is simple, the extra work on the denoising

and reducing features may make the whole process more complicated. Further, the inappropriate feature reduction may lead to information loss and this may add risks to the overall performance of the diagnosis.

Random Forest

Random forest is one of ensemble learning methods of classification, regression, and other tasks, this is, the method that generates multiple classifiers and aggregate their results (Boateng et al., 2020). A random forest consists of a number of decision trees, and each tree grows using some form of randomisation (as shown in Figure 2.3). The leaf nodes of each tree are labelled by estimates of the posterior distribution over the image classes, and each internal node contains a test that best splits the space of data to be classified (Boateng et al., 2020). The process of the classification of an image is sending the image down every tree in the random forest and aggregating the reached leaves through majority voting.
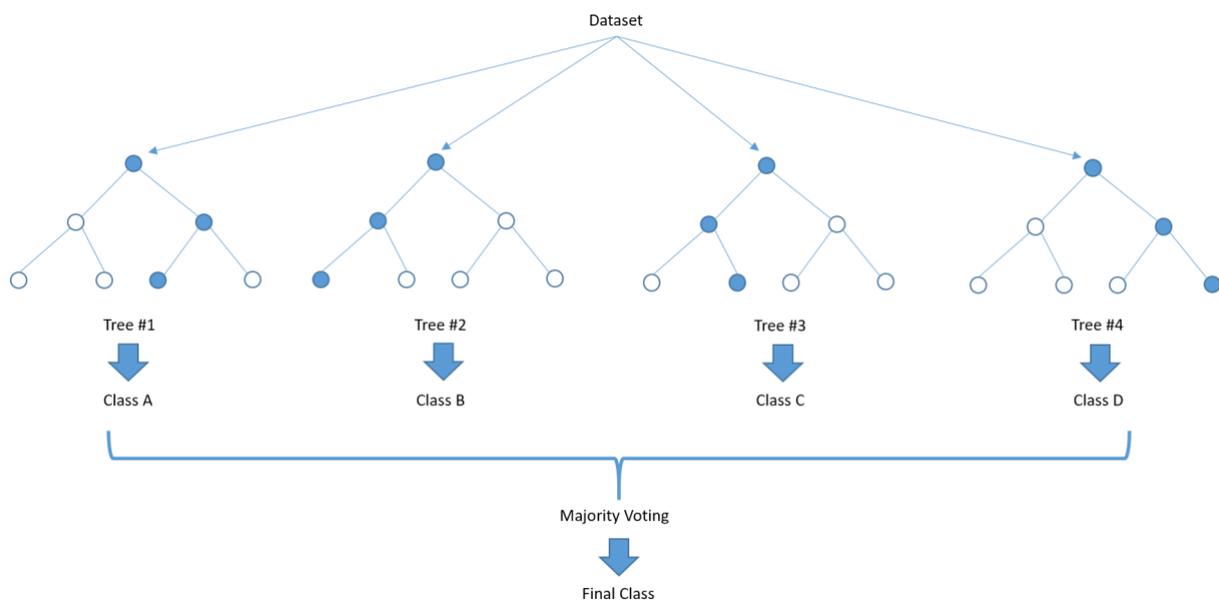
Figure 2.3 A pictorial overview of the random forest algorithm (Developed by Yuyang Liu)

The random forest has also been used for the diagnosis of AD based on brain MRI scans in previous studies. Saim et al. (2022) applied a random forest to the features extracted from MRI scans by transfer learning architectures for the classification of AD, sMCI, and CN in two datasets. The random forest yielded an accuracy of 73.4% in one dataset and yielded an accuracy of 77% in the other dataset. Das et al. (2021) employed multiple models to extract features from MRI scans and fed the features to distinct machine learning models to classify

AD. Amongst the machine learning models, the random forest achieved the best performance, which yielded an accuracy of 95.15% with 100 decision trees.

Dinu et al. (2021) applied two machine learning models, a random forest and a tree bagger (another ensemble of decision trees for classification and regression), to combined extracted features to diagnose AD. The results indicated that the random forest classifier outperformed the tree bagger, which yielded an accuracy of 98.42% (the tree bagger achieved an accuracy of 98.14%). Ali et al. (2021) proposed an optimised random forest, modified random forest, individualise between CN people and people with risks of having AD. Compared with other machine learning algorithms (SVM, KNN, adaptive boosting, etc.), the modified random forest achieved the best performance, and it yielded an accuracy of 96.43%. Kim et al. (2021) developed a random forest-based model to classify CN, MCI, and AD using a combination of multiple types of data. The random forest model showed high diagnostic performance for differentiating CN from AD and from MCI using structural MRI scans, patient characteristics, and tests of cognitive function. The accuracy of classification of CN and AD was 93.5% (with an AUC of 0.99); the accuracy of discriminating between CN and MCI was 80.8% (with an AUC of 0.88).

Tang et al. (2021) applied multiple machine learning algorithms to brain MRI characteristic indices to classify and predict the course of AD. The random forest outperformed the other two machine learning models in all classifications: for classification of CN and AD, the random forest achieved an accuracy of 96.14%; for classification of early MCI and AD, the random forest yielded an accuracy of 90.15%; for discriminating between CN and late MCI, the accuracy was 87.56% using the random forest; the accuracy of classifying late MCI and AD was 84.54%, which was 10% greater than that using the other two models.

Bloch et al. (2019) applied volumetric features of a combination of baseline MRI scans and follow-up MRI scans to predict the future conversion to AD using a random forest algorithm with 500 decision trees. The accuracy for the random forest trained with cognitive tests and MRI volumetric features of the follow-up was 83.55%. Khan et al. (2019) used two random forest models, an imputation random forest and a non-imputation random forest, of missing data values to diagnose AD based on longitudinal MRI data. The imputation random forest achieved the best performance, in which the accuracy was 86.84%.

Ardekani et al. (2017) utilised a random forest to discriminate between sMCI and pMCI based on a novel measure of hippocampal volumetric integrity from structural MRI scans. The overall accuracy of the classification was 82.3% and the accuracy in women was 89.1%. Huang et al. (2016) presented a nonlinear supervised sparse regression-based random forest to predict a variety of longitudinal AD clinical scores using MRI scans. They first estimated the missing scores with the proposed soft-split sparse regression-based random forest and then utilised the proposed random forest to predict multiple AD clinical scores. Their proposed method showed the best performance in both of the measures (the mean absolute error and weighted Pearson's correlation coefficient) for all 4 types of clinical scores, which means that the two measures yielded the least value for all the clinical sores compared with those predicted by other models.

Lebedev et al. (2014) applied a random forest model to a combination of cortical thickness and volumetric measures extracted from baseline brain MRI scans to detect and predict AD. The proposed method achieved sensitivity of 88.6% and specificity of 92.0% in test dataset for the classification of AD and CN; the accuracy of the prediction of MCI-to-AD was from 61.5% to 78.6% according to different conversion times. Moradi et al. (2015) employed a random forest to discriminate sMCI and pMCI, and the proposed method achieved an accuracy of 76.61% with an AUC of 0.7661.

Random forests have performed well in diagnosing AD in previous studies because they can handle thousands of input variables and identify the most significant variables, which are important to the diagnosis. However, as a random forest builds numerous tree structures and combines their outputs, it requires a large amount of computational power and resources (much memory storage and runtime) when a large number of decision trees are needed or the input has high dimensions. MRI scans are 3-D images, which have even higher dimensions than 2-D images, thus the random forest will take a long time to train them. However, few studies (including the studies above) reported the resources usage in the results when they used random forests. In addition, although random forests had better performance when diagnosing AD in many studies and they are not very sensitive to the noise, and they may lead to overfitting when there is relatively a large amount of noise. Thus, decreasing the amount of noisy data should be considered; however, few studies reported that in their papers.

Convolutional Neural Network

Artificial neural networks (ANN) were introduced in the Section 1.2.4 and this is currently one of the most widely used models. The convolutional neural network (CNN) is an outstanding

branch of neural network applications, which is widely used in images (Bernal et al., 2019). Generally, the CNN has three layers: a convolutional layer, a pooling layer (or sub-sampling layer namely), and a fully connected layer. Convolutional and pooling layers are used to extract representative features, and the fully connected layer then converts the 2-dimensional structure features from previous layer into a vector (Bernal et al., 2019). A basic structure of a CNN shows in Figure 2.4.

| Convolution Layer | Sub-sampling Layer | Convolution Layer | Sub-sampling Layer | Fully Connected Layer |
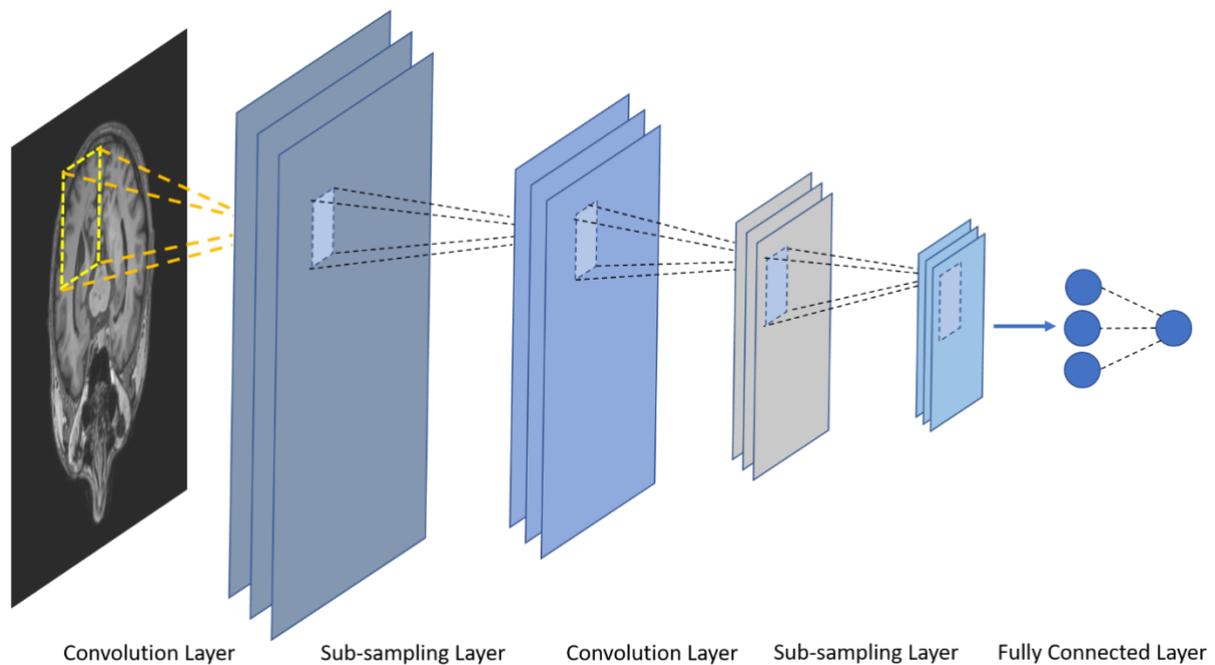
Figure 2.4 A basic structure of CNN (Developed by Yuyang Liu)

Nowadays, deep neural networks (deep learning) are increasingly used because of the enhanced performance of computers. Therefore, different CNN architectures with deeper layers have been developed. The most popular CNN architectures are AlexNet, Visual Geometry Group (VGG) Network, Residual Network (ResNet), LeNet, etc. (Bezdan & Bačanin Džakula, 2019). However, a drawback of the CNN is that there is no specific CNN architecture that is only suitable for a particular situation, thus researchers have employed different CNN architectures in their studies in diagnosing AD using MRI scans.

Some previous studies employed ResNet to diagnose AD. For example, Y. Zhang et al. (2022) used a ResNet with attention mechanism to diagnose three stages of AD: i.e., CN, MCI, and AD. They achieved an accuracy of 90% for discriminating between AD and CN, and achieved an accuracy of 82.5% for discriminating between AD and MCI. Massalimova et al. (2021) applied a ResNet-18, a 72-layer architecture, to both structural MRI scans and Diffusion tensor

imaging (DTI) scans to distinguish between CN, MCI, and AD. The method yielded an accuracy of 96% using both MRI scans and DTI scans. Valliani et al. (2017) pre-trained an 18-layer ResNet with data augmentation to improve the diagnosis of AD using MRI scans, and they achieved an accuracy of 81.3% for binary classification (classification of AD and CN) and an accuracy of 56.8% for 3-way classification (classification of AD, CN, and MCI).

In addition to ResNet, VGG networks were also used in previous studies and yielded good performances. Mehmood et al. (2021) used a VGG-19 network, which consists of 16 convolutional layers, 5 pooling layers, and 3 fully connected layers, to classify CN and AD, and classify early MCI and late MCI. The highest accuracy for classification of AD and CN they achieved was 98.73%, and the accuracy of classification of late MCI and early MCI was 83.87%.

As for the AlexNet, Spasov et al. (2019) applied a AlexNet to a combination of structural MRI, demographic, neuropsychological, and APOe4 genetic data and discriminate between sMCI and pMCI, and CN and AD. The method was able to distinguish MCI patients who developed to AD within 3 years from those patients with stable MCI over the same time-period with an AUC of 0.925 and an accuracy of 86% when using 10-fold cross-validation.

Some studies employed multiple CNN architectures to compare the performances of them. Muhammed et al. (2021) employed multiple models: an SVM, an AlexNet, a VGG-16, a VGG-19, and a GoogleNet, to classify CN, MCI, and AD. In their study, the four deep learning models (AlexNet, VGG-16, VGG-19, and GoogleNet) outperformed SVM, and the VGG-19 achieved the best average accuracy (the mean value of the accuracies for predicting MCI, AD, and CN), which was 90.02%. Korolev et al. (2017) used two CNN architectures, VoxCNN and ResNet, to diagnose different stages of AD. ResNet outperformed VoxCNN for the classification of CN and AD and for the classification of early MCI and CN, which achieved accuracies of 80% and 56%, respectively; VoxCNN had better performance on classification of AD and early MCI, AD and late MCI, late MCI and NC, and late MCI and early MCI, with accuracies of 66%, 61%, 67%, and 57%, respectively. Farooq et al. (2017) used three CNN architectures: GoogleNet, ResNet-18, and ResNet-152, to undertake 4-way classification for AD (classification of AD, late MCI, MCI, and CN). GoogleNet yielded an accuracy of 98.88%; ResNet-18 yielded an accuracy of 98.01%; ResNet-152 yielded an accuracy of 98.14%. Sarraf et al. (2016) employed a well-known architecture of CNN, LetNet-5, and successfully classified structural MRI scans of AD from CN where the accuracy of test data on trained data

was 98.4%. Aderghal et al. (2017) proposed a 2-D+ε approach, where only a limited amount of consecutive slices of the MRI scans were used for training, and they combined the approach with a CNN architecture. The proposed method yielded accuracies of 82.8% for the classification of CN and AD, 66% for the classification of MCI and CN, 62.5% for the classification of AD and MCI.

The studies described above all applied CNN architectures to a number of slices of each MRI scan. However, MRI scans are 3D-images and it will result in information loss in 3-D space if only each 2-D slice was computed by CNN architectures. Therefore, some studies have proposed 3-D CNN architectures to diagnose AD using MRI scans before the research reported in this thesis started.

Wang et al. (2019) proposed an ensemble of 3-D densely connected CNN (3D-DenseNet) to diagnose AD and MCI based on MRI scans. For multi-classification (classification of CN, MCI, and AD), the 3D-DenseNet achieved an accuracy of more than 90% with different hyper-parameters and architectures. Basaia et al. (2019) applied a 3-D CNN to MRI scans with data augmentation and transfer learning. They achieved a high accuracy, which was 99%, for discriminating between CN and AD using one test dataset; the sensitivity and specificity were 98.9% and 99.5%, respectively. They achieved an accuracy of 98%, a sensitivity of 98.1%, and a specificity of 98.3%, for the classification of CN and AD using two test datasets. The proposed method also yielded an accuracy of up to 75% for the classification of sMCI and pMCI using either one or two test datasets. The sensitivity and specificity were both close to 75%, respectively. Asl et al. (2018) proposed to use a deep supervised algorithm to improve a 3-D adaptive CNN and the proposed method outperformed several other approaches, which yielded an accuracy of 94.8% for the multi-classification (i.e., classification of CN, MCI, and AD). It yielded an accuracy of 99.31%, a sensitivity of 100%, a specificity of 98.63%, a positive predictive value (PPV) of 98.63%, and a negative predictive value (NPV) of 100%, for discriminating between CN and AD. Backstrom et al. (2018) applied a 3-D CNN architecture to MRI scans and achieved a good performance with a test accuracy of 98.74% AD detection rate. However, they did not report other measures, and this cannot evaluate the method comprehensively.

After the research reported in this thesis started (from 2019), there were other studies employing 3-D and they also achieved good performances. Zhang et al. (2021) employed a 3-D CNN (i.e., a 3-D residual attention network) to diagnose AD. They achieved an accuracy of

91.3%, a sensitivity of 91.0%, and a specificity of 91.9% for the classification of CN and AD; they achieved an accuracy of 82.1%, a sensitivity of 81.2%, and a specificity of 80.9% for the classification of sMCI and pMCI. Katabathula (2021) developed a lightweight 3-D CNN, DenseCNN, to diagnose AD using features from hippocampus segment. The proposed method yielded an accuracy of 92.52%, a sensitivity of 88.2%, and a specificity of 94.9% for the classification of CN and AD.

While the CNN architectures have consistently yielded high performances on discriminating between different stages of AD, it is still uncertain whether the high accuracy results from appropriate problem modeling rather than the exploitation of correlative factors presenting in the data (Nigri, Ziviani, Cappabianco, Antunes, & Veloso, 2020). Therefore, some studies have proposed explainable CNN architectures. Nigri et al. (2020) proposed an explainable method combined with two CNN architectures, AlexNet and VGG-16, to diagnose AD based on MRI scans. The 2-D AlexNet achieved the best performance, which yielded an accuracy of 92.3%. Yu et al. (2022) applied attention mechanisms to multi-scale features to learn discriminative representations and proposed a novel explainable tool to explain the predictors. They combined these with CNN and generate their own neural network, which was called MAXNet. Finally, their proposed method achieved accuracies of 92.8%, 95.3%, and 95.4% for classification of AD and CN in three datasets, respectively. Kim et al. (2022) employed a deepbrain AD (DBAD) CNN based model to diagnose AD. The model yielded an accuracy of 87.1%, a sensitivity of 93.3%, and a specificity of 85.5%. Lin et al. (2018) used a CNN and an extreme learning machine for classification of CN and AD and prediction of AD conversion. The method yielded an accuracy of 88.79% for the classification of CN and AD; it achieved an accuracy of 79.9%, a sensitivity of 84%, and a specificity of 74.8% for the prediction of AD conversion.

CNN architectures indeed have shown their strong abilities and good performances in image classification. However, they contain back-propagation, which is not an efficient algorithm because it is highly dependent on the massive amounts of data. As mentioned in Chapter 1, the medical imaging scans are difficult to obtain and label, thus some studies divided a single scan into multiple slices and used different slices of a single scan as both training dataset and test dataset, which can lead to data leakage. The problem of data leakage is discussed in the Section 4.10.1.

Table 2.7 summarises the studies (using supervised learning methods to diagnose AD) discussed above in this section. In summary, the accuracies of classification on CN and AD in these studies ranged from 77% to 99.31%, and the classification of MCI (MCI vs. AD, MCI vs. CN, etc.) yielded lower accuracies in the studies; however, most of studies only reported accuracies in their studies.

Table 2.6 Examples of studies using supervised learning methods used to diagnose AD

Studies	Models/Algorithms	Feature Extraction	Performances
SVM			
Lee et al. (2019)	Basic SVM	ROI, a deep learning structure	Accuracy: 92.75% (AD vs. CN) Accuracy: 88.52% (sMCI vs. pMCI) Accuracy: 71.18% (AD vs. MCI vs. CN)
Bigham et al. (2022)		Superficial white matters	Accuracy: 95.81% (AD vs. CN) Accuracy: 83.3% (CN vs. MCI) Accuracy: 83.3% (AD vs. MCI)
Lin et al. (2022)		90 ROIs and 9 brain networks	Accuracy: 79.94% (for ADNI) Accuracy: 83.13% (for Milan cohort)
Vichianin et al. (2021)		Brain cortical and subcortical brain volumes	Accuracy: 90.74%
Beheshti & Demirel (2016)	Linear and non-linear SVM	ROIs	Accuracy: 96.32% (linear) Accuracy: 95.59% (non-linear)
Khedher et al. (2015)	Linear and RBF SVM	PCA, PLS	Accuracy: 88.49% (AD vs. CN) Accuracy: 81.89% (CN vs. MCI) Accuracy: 81.89% (AD vs. MCI)
Ortiz et al. (2013)	Kernel-based SVM	Probability histograms	Accuracy: 90% Sensitivity: 84% Specificity: 96%
Guo et al. (2017)		A hyper network	Accuracy: 91.6%
Kumari et al. (2022)		Not reported	Accuracy: 96.92% (AD vs. CN) Accuracy: 86.36% (MCI vs. CN)

			Accuracy: 77.27% (MCI vs. AD)
Lazli et al. (2018)		Tissue qualification (possibilistic c-means)	Accuracy: 75% (for noisiest MRI scans)
Bron et al. (2021)		Modulated grey matter maps	AUC: 0.940
Zeng et al. (2018)	SDPSO algorithm optimised SVM	ROIs	Accuracy: 81.25% (AD vs. CN) Accuracy: 69.23% (sMCI vs. pMCI) Accuracy: 85.71% (pMCI vs. CN)
Alam et al. (2017)	Twin SVM	Dual-tree CWT	Accuracy: 92.65% (one dataset) Accuracy: 96.68% (the other dataset)
KNN			
Aruchamy et al. (2020)	Basic KNN	Bidimensional Empherical Model Decomposition	Accuracy: 93.18% (the best)
Battineni et al. (2020)		Wrapping techniques	Accuracy: 95.92%
Kamal et al. (2022)		Haar wavelet transformation	Accuracy: 93%+
Deepa et al. (2022)		Auto colour correlogram Filter techniques	Accuracy: 84%
Arabi et al. (2022)	Weighted KNN	Volumetric and surface-based structures	Accuracy: 100% Sensitivity: 100% Specificity: 100%
Balamurugan et al. (2017)	Feature reduce based KNN	Not reported	Accuracy: 99%
Emmanuel et al. (2022)	Fuzzy KNN	Hippocampus	Accuracy: 93%
Random forest			
Saim et al. (2022)		VGG16, PCA	Accuracy: 73.4% (for one dataset) Accuracy: 77% (for the other dataset)
Das et al. (2021)		Live neurons, GLCM, random forest mapping	Accuracy: 95.15%
Dinu et al. (2021)	Basic random forest	Histogram of Oriented Gradients (HOG)	Accuracy: 98.42$
Kim et al. (2021)		ROIs	Accuracy: 93.5% (AD vs. CN) Accuracy: 80.8% (CN vs. MCI)
Tang et al. (2021)		ROIs	Accuracy: 96.14% (AD vs. CN)

			Accuracy: 90.15% (early MCI vs. AD) Accuracy: 87.56% (CN vs. late MCI) Accuracy: 84.54% (late MCI vs. AD)
Bloch et al. (2019)		34 cortical areas per hemisphere, 34 subcortical areas, and eTIV	Accuracy: 83.55%
Khan et al. (2019)		Not reported	Accuracy: 86.84
Ardekani et al. (2017)		Hippocampal volumetric integrity (HVI)	Accuracy: 82.3% (sMCI vs. pMCI)
Huang et al. (2016)		90 cerebral regions (ROIs)	-
Lebedev et al. (2014)		Volumetric measurements	Sensitivity: 88.6% Specificity: 92%
Moradi et al. (2015)		Semi-supervised LDS	Accuracy: 76.61%
Ali et al. (2021)	Optimised random forest		Accuracy: 96.43%
CNN Architectures			
Y. Zhang et al. (2022)	ResNet	Grey matter coronal slices, TresNet (improved ResNet)	Accuracy: 90% (AD vs. CN) Accuracy: 82.5% (AD vs. MCI)
Massalimova et al. (2021)		ResNet	Accuracy: 96%
Valliani et al. (2017)		ResNet	Accuracy: 81.3% (AD vs. CN) Accuracy: 56.8% (AD vs. MCI vs. CN)
Mehmood et al. (2021)	VGG	Grey matter, CNN	Accuracy: 98.73% (AD vs. CN) Accuracy: 83.87% (late vs. early MCI)
Muhammed et al. (2021)	Alex, VGG (best performance), GoogleNet	CNN architectures	Accuracy: 90.2%
Spasov et al. (2019)	Alex	Jacobian determinant maps	sMCI vs. pMCI Accuracy: 86% Sensitivity: 87.5% Specificity: 85%
Korolev et al. (2017)	VoxCNN, ResNet (best performance)	CNN	Accuracy: 80% (AD vs. CN) Accuracy: 56% (MCI vs. CN) Accuracy: 66% (AD vs. early MCI) Accuracy: 61% (AD vs. late MCI) Accuracy: 67% (late MCI vs. CN)

			Accuracy: 57% (early vs. late MCI)
Farooq et al. (2017)	GoogleNet, ResNet	Grey matter, CNN	Accuracy: 98.88% (GoogleNet) Accuracy: 98.01% (ResNet-18) Accuracy: 98.14% (ResNet-154)
Aderghal et al. (2017)	a 2-D+ε approach CNN	Hippocampus, CNN	Accuracy: 82.8% (AD vs. CN) Accuracy: 66% (MCI vs. CN) Accuracy: 62.5% (AD vs. MCI)
Sarraf et al. (2016)	LetNet	Grey matter, white matter, CSF, CNN	Accuracy: 98.4%
Zhang et al. (2021)	3D residual attention network	3D residual attention network	Accuracy: 91.3% (CN vs. AD) Sensitivity: 91.0% (CN vs. AD) Specificity: 91.9% (CN vs. AD) Accuracy: 82.1% (sMCI vs. pMCI) Sensitivity: 81.2% (sMCI vs. pMCI) Specificity: 80.9% (sMCI vs. pMCI)
Katabathula (2021)	DenseCNN	A ROI (hippocampus)	Accuracy: 92.52%
Wang et al. (2019)	3D-DenseCNN	3D-DenseCNN	Accuracy: 90%+ (CN vs. MCI vs. AD)
Basaia et al. (2019)	3-D CNN	CNN	Accuracy: 98% (CN vs. AD) Accuracy: 75% (sMCI vs. pMCI)
Asl et al. (2018)	3-D adaptive CNN	Autoencoder	Accuracy: 94.8% (CN vs. MCI vs. AD) Accuracy: 99.31% (CN vs. AD)
Backstrom et al. (2018)	3-D CNN	CNN	Accuracy: 98.74%
Nigri et al. (2020)	AlexNet (best performance), VGG	AlexNet, VGG	Accuracy: 92.3%
Yu et al. (2022)	MAXNet	MAXNet	Accuracy: 92.8% (the first dataset) Accuracy: 95.3% (the second dataset)

			Accuracy: 95.4% (the third dataset)
Kim et al. (2022)	DBAD-CNN based model	CNN	Accuracy: 87.1% Sensitivity: 93.3% Specificity: 85.5%
Lin et al. (2018)	CNN with extreme learning machine	AlexNet	Accuracy: 88.79%

2.5.1.2. *Semi-supervised Learning Methods used to Diagnose Alzheimer's Disease*

In Section 1.2.4, the concept of semi-supervised learning was introduced, which is when the learning is used with partially labelled data. For diagnosis using semi-supervised learning methods, semi-supervised classification is often used, where the training data consist of both labelled data and unlabelled data (i.e., using labelled data for ground predictions and using unlabelled data to learn the larger data distribution). Its goal is to train a classifier from both labelled data and unlabelled data, such that it is better than the supervised classifier trained on the labelled data alone.

Generative adversarial networks

Generative adversarial network (GAN) is a class of machine learning frameworks developed by Goodfellow et al. (2014). It is a special type of neural network model where two network structures are trained simultaneously, with one focus on image generation and the other one centred on discrimination (Yi, Walia, & Babyn, 2019). It automatically discovers and learns the patterns and regularities from the input data in such a way that the model can be used to generate new examples that could have been plausibly acquired from the original dataset. Figure 2.5 shows a process of the generative adversarial network. Random images are input into a generator. The generator learns to generate plausible data and the generated instances become negative training examples for the discriminator. The discriminator distinguishes the generator's fake images from real images, and it penalises the generator for producing implausible results.
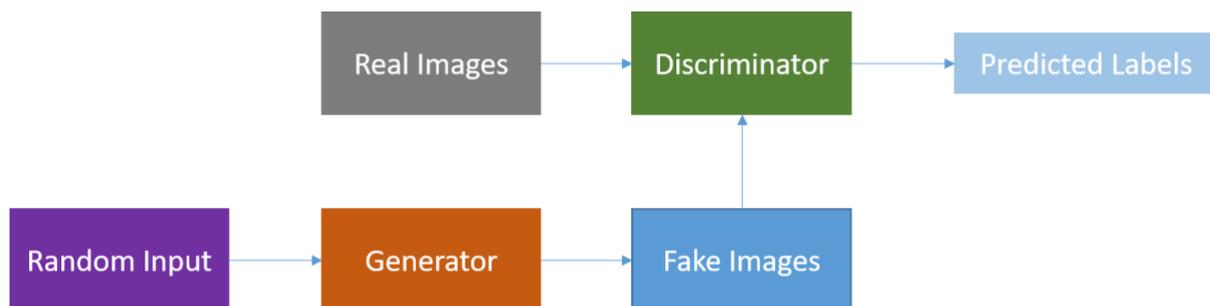
Figure 2.5 A pictorial overview of the generative adversarial networks (Developed by Yuyang Liu)

Although GAN is an unsupervised learning method, some studies have extended it to a semi-supervised version that involves the simultaneous training of a supervised discriminator, unsupervised discriminator, and a generator model, and they used these semi-supervised GANs to diagnose AD in their studies. Saravanakumar et al. (2022) employed a semi-supervised generative adversarial network (GAN) classifier to predict AD using brain MRI scans. Saravanakumar et al. (2022) achieved the maximum accuracy of 97% for diagnosing AD. Saravanakumar & Saravanan (2022) used semi-supervised GAN in their another study to early detect AD and achieved a maximum accuracy of 97%. Baskaran (2021) also designed a semi-supervised GAN to automatically diagnose the presence of AD from brain MRI scans. They fed hippocampal volumes and features extracted from a CNN architecture to the semi-supervised GAN, and they achieved a maximum of accuracy of 96%.

GAN, essentially, consists of two networks: the generator and discriminator. This adds uncertainty and risk to the performance, because the correct functioning of a GAN requires the harmonic cooperation between the generator and the discriminator (Laino et al., 2022). However, when the generator generates an image that is extremely different from the original one due to the random generation, it may lead that the discriminator yielded the optimal solution immediately and it cannot further update the parameter, which is often called the gradient vanishing problem. Therefore, this may result in the failure of the training process. Therefore, more advanced versions could be considered, e.g., the Wassertein GAN. In addition, the biased networks due to under or over-representation of certain findings can cause misdiagnosis of medical conditions, because the composition of the source and target domains can bias the image transformation to result an unwanted feature hallucination (Cohen, Luck, & Honari, 2018; Laino et al., 2022).

In summary, the GAN highly depends on both the generator and the discriminator, and its performance is also influenced by the quality of generated images. These may lead to unstable training process; however, it is important to perform them stably for diagnosing diseases.

Other semi-supervised methods

In addition to the GAN, there are some other semi-supervised methods used by previous studies. Khajehnejad et al. (2017) proposed to use a label propagation in a manifold-based semi-supervised learning framework to diagnose very early stages of AD based on brain MRI scans. They applied a principal component analysis (PCA)-based dimension reduction algorithm to the AD-relevant feature extracted from the MRI scans, and presented a hybrid

learning framework which embedded the feature vectors in a subspace. Next, the label propagation method was applied in the generated manifold space to predict the labels of remaining images by using a small volume of labelled data. Label propagation is a semi-supervised learning method that assigns labels to previous unlabelled data points. The proposed method yielded an accuracy of 93.86% and a lower error rate of 3% for classification compared to the best existing methods listed in their manuscript. However, label propagation is not stable because the newly labelled data depend a lot on the pre-labelled data and this process might be biased or different in different runs.

Moradi et al. (2014) employed both a low density separation method with the help of a domain adaption method to discriminate between CN, MCI, and AD. The accuracy of the classification of their proposed method was 76.7%. Generally, the semi-supervised low density separation often employs self-learning strategy to classify the data. It is trained by labelled data, and then unlabelled data are fed into the trained low density separation model and the unlabelled data are assigned labels (pseudo-labels), and finally a subset of the newly labelled data are used to train the model. Therefore, the performance of the model is much influenced by the choice of the subset of the data with pseudo-labels. In addition, the size of the dataset in Moradi et al.'s (2014) was only 30 (including all CN, MCI, and AD data), and such a limited volume of data is not large enough to indicate whether their proposed method was effective or not.

Zhu et al. (2014) proposed a co-training based semi-supervised learning method to classify CN and AD using both MRI and PET scans. Co-training is one of the semi-supervised learning techniques which trains two classifiers based on two different views of data (Katz, 2017). Their proposed method yielded an accuracy of 92.91% and an AUC of 0.964 using both PET and MRI scans. The co-training is beneficial only if the two views provided by the two classifiers are different, ideally independent given the category (Qiao, Shen, Zhang, Wang, & Yuille, 2018). However, this cannot be guaranteed. In addition, in order to achieve the different views by the two classifiers, the selection of the two classifiers is also not easy.

Ye et al. (2011) applied a semi-supervised classifier, the linear Laplacian SVM (LapSVM), to the morphological features extracted from MRI scans to discriminate between MCI patients who converted to AD within 15 follow-up months and who had not yet converted to AD yet (classifying sMCI and pMCI). The proposed method achieved an accuracy of 56.1%; however, the sensitivity of the method was 94.1% and the specificity was 40.8%, which was extremely unbalanced. In addition, the number of sMCI scans was 169 and that of pMCI was 68, and this

means that the data were unbalanced. Therefore, the effectiveness of the proposed method (LapSVM) cannot be comfirmed. D. Zhang et al. (2011) used MCI data to aid classification of AD and CN: they treated MCI as unlabelled data and then used a semi-supervised learning technique, multi-modal Laplacian Regularised Least Squares (mLapRLS), to solve the classification problem. Using their method, the best AUC that was obtained was 98.5% and the best accuracy, specificity, and sensitivity they obtained were all more than 94%. However, the study only used 51 AD scans and 52 CN scans for the classification, and such volume of the data are not enough to approve the effectiveness of the method.

Table 2.8 summarises the studies (using semi-supervised learning methods to diagnose AD) discussed above in this section. In summary, similar to Table 2.7, most studies only reported accuracy in their studies. In addition, the number of studies using semi-supervised learning methods is not many as using supervised learning methods.

Table 2.7 Examples of studies using semi-supervised learning methods used to diagnose AD

Studies	Models/Algorithms	Feature Extraction	Performances
GAN			
Saravanakumar et al. (2022)	Semi-supervised GAN	CNN	Accuracy: 97%
Saravanakumar et al. (2022)		CNN	Accuracy: 97%
Baskaran (2021)		CNN	Accuracy: 96%
Others			
Khajehnejad et al. (2017)	Label propagation	ROIs, PCA	Accuracy: 93.86%
Moradi et al. (2014)	Semi-supervised low density separation	Grey matter, white matter, CSF	Accuracy: 76.7%
Zhu et al. (2014)	Co-training based semi-supervised learning	93 ROIs	Accuracy: 92.91%
Ye et al. (2011)	LapSVM	RAVEN maps, ISOMAP algorithm	Accuracy: 56.1% (sMCI vs. pMCI) Sensitivity: 94.1% (sMCI vs. pMCI) Specificity: 40.8% (sMCI vs. pMCI)
D. Zhang et al. (2011)	mLapRLS	93 ROIs	Accuracy: 94%+ Sensitivity: 94%+ Specificity: 94%+

In Section 1.2.4, the concept of unsupervised learning was also presented, that is, it is the learning without using labelled data. Some studies used unsupervised learning methods to diagnose AD based on MRI scans as well.

K-means

The k-means clustering algorithm is a classic unsupervised learning method developed by Macqueen (1967). This algorithm partitions an N-dimension population into k groups in which each sample belongs to the cluster with the nearest cluster centroid. Figure 2.6 roughly shows a process of the clustering using the k-means algorithm: Figure 2.6 (1) is an initial stage; in Figure 2.6 (2), each point is assigned to its closest centroid (shown as cross signs), and a new centroid (position) is chosen for each cluster based on the average of all points in that cluster. As a result, the blue and the red cluster centroids (cross signs) move to the position shown in Figure 2.6 (3). After two rounds, as shown in Figure 2.6 (4), the clusters have reached a steady state, and would not change further through an infinite number of iterations.

The choice of k value is usually influenced by prior knowledge regarding the nature of the data. However, in diagnosis of AD, k value is often fixed because most studies aim to cluster two or more of AD, CN, and MCI (or subtypes of MCI).
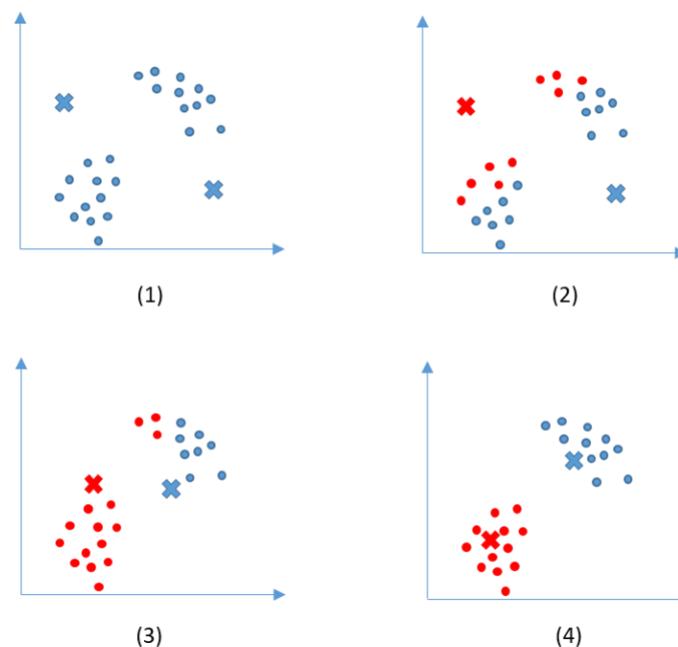
Figure 2.6 A simple pictorial overview of k-means algorithm (Developed by Yuyang Liu)

Several studies have employed k-means to diagnose AD using MRI scans. Raza et al. (2021) compared the performance of two unsupervised learning algorithms, the k-means and the k-medoids, on clustering voxel-based morphometry features extracted from MRI scans. In their study, the k-means algorithm achieved the best performance, where the accuracy was 76.3% and the runtime was 43.9 seconds. Bi et al. (2019) applied a k-means algorithm to features extracted from three slices respectively corresponding to sagittal, coronal, and axial planes of an MRI image. The proposed method achieved an accuracy of 89.15% for the classification of AD and CN, an accuracy of 97.1% for the classification of AD and MCI, and an accuracy of 91.25% for three-class classification of AD, CN, and MCI. Escudero et al. (2011) applied a k-means clustering algorithm to data features to divide the subjects into pathological and non-pathological groups. The types of data they used involved mini-mental state examination (MMSE), the Alzheimer's disease assessment scale-cognitive subscale (ADAS-cog), APOE-4, cerebrospinal fluid, and MRI scans. Finally, 79.1% of AD patients were correctly clustered into the pathological group; 48.3% of MCI patients were clustered into the pathological group; 15.1% of CN subjects were mistakenly clustered into the pathological group, i.e., the false positive rate was 15.1%. However, all of these studies only reported accuracies but did not report other measures. Using accuracies alone cannot evaluate performances of studies comprehensively.

K-means is relatively simple to implement and guarantees convergence. As mentioned above, although the *k* value needs to be selected manually, it can easily be chosen when diagnosing AD. Thus, it became a potential method to diagnose AD using MRI scans. However, on the one hand, its performance depends on the initial position of the centroid. Hence, optimised versions of k-means (e.g., k-means++, etc.) to solve the centroid initialisation problem need to be considered. On the other hand, the performance is also influenced by the density of the data and is sensitive to the outliers, so appropriate feature extraction is necessary.

Generative adversarial networks
GAN has been introduced in the Section 2.5.1.2 and some studies have employed unsupervised versions of GAN to diagnose AD. Cabreza et al. (2022) employed an unsupervised deep learning approach to alleviate the need for large amounts of labelled data. They proposed an unsupervised model based on the deep convolutional generative adversarial network that performed anomaly detection on brain MRI scans to diagnose AD. The model yielded an AUC of 0.7951 and an accuracy of 74.44%. Han et al. (2019) proposed a two-step method using

GAN-based multiple adjacent brain MRI slice reconstruction to detect AD at various stages. The proposed method yielded an AUC of 0.780 for the detection of AD at a very early stage and it achieved an AUC of 0.917 for the detection of AD at a very late stage.

GAN used in the unsupervised learning has similar problems as that mentioned in the Section 2.5.1.2, i.e., the performance of the GAN is influenced by the cooperation of the discriminator and the generator, and this increases the uncertainty and risks to the performance.

Other unsupervised learning methods

Long et al. (2010) applied the quick shift clustering method to the embedded MRI scans to partition them into different groups. Using the white matter alone, 94.67% of patients were clustered into the appropriate group; after adding the grey matter, 97.33% of subjects were correctly clustered. The quick shift is a clustering method based on the density of the data (Heinrich Jiang, 2017). It is a simple and fast clustering algorithm. However, the number of clustering cannot be controlled so that this may cause over-clustering.

Table 2.9 summarises the studies (using unsupervised learning methods to diagnose AD) discussed above in this section. In summary, the k-means method and the unsupervised GAN were the most commonly used techniques in unsupervised learning methods. However, all these studies only provided one measure (accuracy, sensitivity, or AUC) to evaluation the performances in their studies.

Table 2.8 Examples of studies using unsupervised learning methods used to diagnose AD

Studies	Models/Algorithms	Feature Extraction	Performances
K-means			
Farouk & Rady (2020)	Basic k-means	ROIs	Accuracy: 76.3%
Bi et al. (2019)		PCANet	Accuracy: 89.15% (AD vs. CN) Accuracy: 97.1% (AD vs. MCI) Accuracy: 91.25% (CN vs. MCI vs. AD)
Escudero et al. (2011)		Hippocampal volumes	Sensitivity: 79.1%
GAN			
Cabreza et al. (2022)	Unsupervised GAN	GAN	Accuracy: 74.4%
Han et al. (2019)		GAN	AUC: 0.780 (detection of AD at early stage) AUC: 0.917 (detection of AD at late stage)
Others			

Long et al. (2010)	Quick shift clustering method	Symmetric log-domain diffeomorphic demons algorithm, spectral embedding algorithm	Accuracy: 97.33%

2.5.2. Feature Extraction

Feature extraction before classification and clustering also play an important role in the diagnosis of AD using machine learning techniques and brain MRI scans. Different studies might have different goals in this step: some extracted meaningful regions directly from the origin scans; some mapped the low-level features (voxel-level features) to higher-level features of the scan; some did both above. Higher-level features refer to a more complex and a more abstract representation of the input raw data that is learned by neural networks, and are truly helpful for the discrimination between different classes.

Table 2.9 Feature extraction methods used in examples of studies to diagnose AD based on MRI scans

Studies	Feature extraction methods
H. Lin et al. (2022)	90 ROIs and 9 brain networks
Raghavaiah et al. (2022)	Autoencoders
Yanteng Zhang et al. (2022)	Grey matter coronal slices, tresnet (improved ResNet)
Kamal et al. (2022)	Haar wavelet transformation
Emmanuel & Jabez (2022)	Hippocampus
AlSaeed & Omar (2022)	ResNet
Sharma, Goel, Tanveer, & Murugan (2022)	ResNet
Bigham et al. (2022)	Superficial white matters
Arabi et al. (2022)	Volumetric and surface-based features
Bloch & Friedrich (2021b)	34 cortical areas per hemisphere, 34 subcortical areas, and eTIV
Katabathula et al. (2021)	A ROI (hippocampus)
Jin et al. (2021)	Autoencoders
Vichianin et al. (2021)	Brain cortical and subcortical brain volume
Saratxaga et al. (2021)	BraiNet2D
Muhammed Raees & Thomas (2021)	CNN architectures (or variations)
Mehmood et al. (2021)	Grey matter, CNN
Salunkhe et al. (2021)	Hippocampus, GLCM
Das et al. (2021)	Live neurons, GLCM, random forest mapping
Massalimova & Varol (2021)	ResNet
Kim et al. (2021)	ROIs
Tang et al. (2021)	ROIs
Sathiyamoorthi et al. (2021)	ROIs and 2D-GLCM
Zeng, Li, & Peng (2021)	Deep belief network

Nigri et al. (2020)	AlexNet, VGG
Aruchamy et al. (2020)	Bidimensional Empherical Model Decomposition
(Ma et al., 2020)	Morphometric features, Jacobian determinant values
Battineni et al. (2020)	Wrapping technique
H. Wang et al. (2019)	3D-DenseNets
Jabason et al. (2019)	Autoencoders
Basaia et al. (2019)	CNN architectures (or variations)
Spasov et al. (2019)	Jacobian determinant maps
Petrone et al. (2019)	Jacobian determinant maps
Y. Wang et al. (2019)	Morphometric measures
Bi et al. (2019)	PCANet
Hongming Li, Habes, Wolk, & Fan (2019)	ROI (hippocampus), CNN
Y. Gupta et al. (2019)	ROIs
Lee et al. (2019)	ROIs, a deep learning structure
Baskar et al. (2019)	ROIs, texture and shape features
Razavi et al. (2019)	Unsupervised feature learning structures
L. Lin, Zhang, & Wu (2018)	AlexNet
Asl et al. (2018)	Autoencoders
Backstrom et al. (2018)	CNN architectures (or variations)
W. Lin et al. (2018)	CNN, morphological information
Zeng et al. (2018)	ROIs
Ramon Casanova et al. (2018)	ROIs
Guo et al. (2017)	A hyper network
Alam et al. (2017)	Dual-tree CWT
Long, Chen, Jiang, & Zhang (2017)	Embedding algorithm
Farooq et al. (2017)	Grey matter, CNN
Aderghal et al. (2017)	Hippocampus, CNN
Lama, Gwak, Park, & Lee (2017)	Morphometric measures
Valliani & Soni (2017)	ResNet
Aderghal et al. (2017)	ROI (hippocampus), CNN
Moradi, Hallikainen, Hänninen, & Tohka (2017)	ROIs
S. Korolev et al. (2017)	VoxCNN, ResNet
Sarraf & Tofighi (2016)	Grey matter, white matter, CSF, CNN
I. O. Korolev, Symonds, & Bozoki (2016)	Morphometric measures
Beheshti & Demirel (2016)	ROIs
Yudong Zhang et al. (2015)	Eigenbrain
Papakostas, Savio, Graña, & Kaburlasos (2015)	Jacobian determinant maps
Khedher et al. (2015)	PCA, PLS
Moradi et al. (2015)	Regularised logistic regression
S. Liu et al. (2014)	Autoencoders
Xin Liu et al. (2013)	Local linear embedding
Aguilar et al. (2013)	Not report
R Casanova et al. (2013)	ROIs

| Ye et al. (2011) | grey map, RAVENS maps |
| D. Zhang, Wang, Zhou, Yuan, & Shen (2011) | ROIs |

Feature extraction aims to select features (lower-level or higher-level features) that are helpful for the discrimination. The region of interest (ROI) extraction is one of the most commonly-used methods because of its specificity. Table 2.10 lists some examples of studies and the feature extraction methods they used in the studies. As the table shows, feature extraction has been used in many previous studies, and features extracted from the ROI around the hippocampus, entorhinal cortex, amygdala, and other tissues in the brain, perform effectively in recognising AD or MCI cases (Cuingnet et al., 2011; I. O. Korolev, Symonds, & Bozoki, 2016; Almdahl et al., 2017; Almdahl et al., 2017).

The ROI features from these areas can potentially improve the accuracy of classification and reduce the number of false positive diagnoses. H. Lin et al. (2022), Bigham, Zamanpour, & Zare (2022), Bloch & Friedrich (2021), Sathiyamoorthi et al. (2021) Katabathula et al. (2021), Tang et al. (2021), Kim et al. (2021), Li et al. (2019), Baskar et al. (2019), Gupta et al. (2019), Lee et al. (2019), Zeng et al. (2018), Ramon Casanova et al. (2018), Moradi et al. (2017), Aderghal et al. (2017), Beheshti & Demirel (2016), R Casanova et al. (2013), and Zhang et al. (2011) all extracted ROIs from the pre-processed MRI scans. Emmanuel et al. (2022), Katabathula et al. (2021), Li et al. (2019), Baskar et al. (2019), Wang et al. (2019), Aderghal et al. (2017), Casanova et al. (2013), and Zhang et al. (2011) clearly indicated that they included the hippocampus as a ROI in their studies. However, although some studies (Gupta et al., 2019; Lee et al., 2019; Zeng et al., 2018; Ramon Casanova et al., 2018; Moradi et al., 2017; Farooq et al., 2017) did not report using the hippocampus, reported using the grey matter, which includes hippocampus. The features in Bloch & Friedrich (2021) were from 34 cortical areas per hemisphere of the Desikan-Killiany-Tourville (DKT) atlas, 34 subcortical areas, and the estimated Total Intracranial Volume (eTIV). Bigham et al. (2022) extracted superficial white matters from DTI scans and employed a tractography approach and a connectivity and network analysis to obtain the features for classification.

H. Lin et al. (2022) used features extracted from 90 ROIs and 9 brain network based on a brain connectivity toolbox. Arabi et al. (2022) employed 71 volumetric and 68×3 surface-based features, i.e., cortical thickness, sulcal depth, and gyrification index. Vichianin et al. (2021) extracted brain cortical and subcortical brain volume. Wang et al. (2019) extracted morphometric features including hippocampal subdivisions, and thickness, surface area, and

volume of cortical and subcortical regions. Korolev et al. (2016) used morphometric measures computed from 452 ROIs extracted from individual MRI scans as potential predictors of MCI-to-AD progression.

Although ROIs selected in these studies facilitated the discrimination, they require that researchers have professional knowledge of brain anatomy. In addition, extracting ROIs directly means that researchers rely on the pathology of AD, which was studied by previous AD pathologists. This means that the ROI selection is highly influenced by the development of the study of AD.

In contrast to using ROIs, some studies have employed the whole brain, employing neural networks to extract features. Feature extraction is sometimes not an independent step when a neural network is employed, which means it accomplishes both feature extraction and classification directly. Therefore, in this section, only feature extraction functions of the neural network are described, although the same neural networks may have been discussed in previous sections (for the classification purpose).

CNN architectures

In addition to the classification, CNN architectures were also commonly used for feature extraction in previous studies as they can map voxel values to a higher level. AlSaeed et al. (2022) applied a Residual Neural Network (ResNet) to the whole brain MRI scans to extract features from them. As described in 2.5.1.1, ResNet is a variation of convolutional neural network (CNN) developed by He et al. (2016) in 2015. It tried to solve the degradation (of training accuracy) problem when the depth of basic CNN increases (He et al., 2016). Sharma et al. (2022) also applied ResNet to the whole brain scans to extract features in their study but they only utilised the sagittal plane of the scans. Valliani et al. (2017) also used a deep ResNet architecture consisting of 18 layers to extract features directly from the whole MRI scans. Massalimova & Varol (2021) used multi-modalities, T1-weighted MRI and diffusion tensor imaging (DTI), as the input data in their study, and employed ResNet to extract features from the scans.

Lin et al. (2018) employed another CNN architecture variation, AlexNet, to learn features from training data. Basaia et al. (2019) applied a CNN model to whole brain MRI scans and Backstrom et al. (2018) employed a 3D CNN and the whole brain MRI scans to extract features. Nigri et al. (2020) extracted features from brain MRI scans through applying another two

variations of CNN, AlexNet and Visual Geometry Group (VGG). Muhammed et al. (2021) compared the performance of diagnosis of AD using four CNN architecture variations, i.e., AlexNet, VGG16, VGG19, and GoogleNet. Therefore, the features were extracted by the four CNN architectures.

In addition to the well-known CNN architectures, some studies developed their own CNN architectures to extract features of the data. Korolev et al. (2017) compared the performance of two CNN architectures, a newly designed CNN architecture named VoxCNN, which had four volumetric convolutional blocks for extracting features, and the ResNet. Hence, they used the two architectures to learn the features from MRI scans, respectively. Saratxaga et al. (2021) proposed a small custom network named BrainNet2D, which was designed based on the CNN, and used it to extract features MRI scans. In addition to supervised neural networks above, some studies employed unsupervised CNN to extract features without labels for the data. Bi et al. (2019) used a PCANet, an unsupervised CNN model, to extract features from a sagittal slice, coronal slice, and an axial slice of the MRI scan. PCANet is a combination of CNN and principle component analysis (PCA), and this will be described in detail in the Section 3.6.

CNN architectures are commonly-used methods to map lower-level features to higher-level features in the computer vision field. Lower-level features in images are straight properties, i.e., colours, edges, contours, etc.. Higher-level features are more abstract semantic representations learned from the neural networks. However, higher-level features computed by CNNs are often difficult to perceive and interpret for humans because neural networks are effectively black boxes. Additionally, none of the studies described above reported on the effectiveness of the CNN architectures by comparing the performance measures of discrimination achieved using higher-level features and using low-level features.

CNN architectures have some limitations. Similar to the application as a classifier, it also needs an enormous amounts of data to learn features (Patil & Rane, 2021). However, to ensure the performance of the feature extraction, multiple convolutional layers and multiple pooling layers were often constructed in the previous studies, e.g., VGG16, VGG19, ResNet50, etc.. The adding of the two layers make the unproductive backpropagation even more time-consuming during the training phase. This requires computers with optimal hardware and high performance. In addition, the CNN cannot encode the position of objects in the image because it predicts by checking whether certain components are present in the image or not. However,

the position of objects is important in the field of diagnosing because the similar object in different positions might lead to or be due to different diseases.

Autoencoder networks

Autoencoders, first introduced by Rumelhart et al. (1987), are a special type of a neural network, mainly developed to encode the input into a compressed and meaningful representation, and then decoded it such that the reconstructed input is similar as possible to the original one (Bank, Koenigstein, & Giryes, 2020). Figure 2.7 shows the overall process of the autoencoder: an encoder maps the input to a compressed representation and a decoder reconstructs the input from the compressed representation. An optimal autoencoder performs as close to perfect reconstruction as possible.
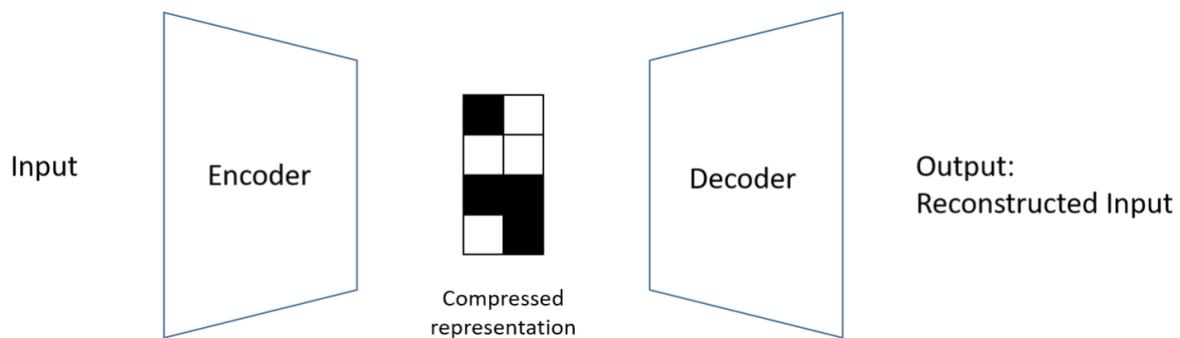
Figure 2.7 A pictorial overview of the autoencoder (Developed by Yuyang Liu)

The autoencoder is, in fact, a feature extractor for the classification of images. Raghavaiah et al. (2022) proposed a deep enhanced stacked autoencoder (DESAE) classifier to diagnose AD. The proposed method achieved an accuracy of 98.68% for multi-classification of AD, CN, and MCI. Jin et al. (2021) applied an unsupervised deep network based on adversarial autoencoders to brain MRI scans to portray atrophy of AD patients, thereby distinguishing AD patients from CN patients. They combined variational autoencoders and GAN to learn the potential distribution and trained a generator. Their proposed method yielded an accuracy of 94%, and the structural similarity, which was used to judge robustness of the proposed model, was 0.86 in both training and the external validation. Asl et al. (2018) employed a 3-D convolutional autoencoder to extract a few co-aligned scalar feature maps for a set of input 3-D images with scalar or vectorial voxel-wise signals by combining data encoding and decoding.

Similar to the GAN, as an unsupervised learning model, autoencoder can also be extended to a semi-supervised version and used to extract features. Jabason et al. (2019) proposed a

classification framework which employed an unsupervised autoencoder network to select the subset from the structural MRI scans and clinical data by exploring the linear and non-linear relationships among them, followed by a supervised multinomial logistic layer to automatically distinguish the patients with AD, MCI, and CN. Jabason et al. (2019) achieved a mean overall accuracy of 98.78% for the classification of CN, MCI, and AD using both MRI scans and clinical data. S. Liu et al. (2014) designed a deep learning architecture, which consisted of stacked autoencoders and a softmax output layer. The architecture can be extended to use unlabelled training samples. For the binary classification, the proposed method achieved an accuracy of 87.76% for the classification of AD and CN and achieved an accuracy of 76.92% for classification of MCI and CN; as for the 4-class classification, the method yielded an accuracy of 47.42% for distinguishing between AD, CN, sMCI, and pMCI.

Autoencoders can extract feature and reduce the number of features, but their algorithms are too specialised, which namely means that the generated features are often unique to the data that they are trained on. This may cause the autoencoders to be sensitive to noise and it requires representative data, especially in diagnosing AD. In addition, a typical autoencoder has multilayers of progressively fewer neurons for encoding the input, which is called a bottleneck layer. However, the algorithm may have important dimension missing problems if the bottleneck layer is too narrow. Therefore, different sizes of the bottleneck layer need to be tested; however, the aforementioned studies involving autoencoders did not test this.

Other unsupervised and semi-supervised feature extraction methods
Razavi et al. (2019) utilised a sparse filtering method, i.e., a two-layer network, to learn the expressive features of the MRI scans. They achieved an accuracy of 87.6% for classification of CN and AD using the softmax regression classifier. The only required parameter in the sparse filtering method is the number of features, thus this indicates that a proper number of features need to be determined. However, it may not be feasible.

Other feature extraction methods
Some studies involves both ROIs and neural networks. Yanteng Zhang et al. (2022) applied an improved ResNet to the grey matter coronal slices of the MRI scans. They also employed an attention mechanism to indicate the pathological regions, which can partially solve the problem that the CNN is difficult to explain. Mehmood et al. (2021) used CNN to extract features from the grey matter. Li et al. (2019) extracted features from hippocampus using CNN; Lee et al. (2019) partitioned the brain into multiple regions and divided each region into a number of sub-

spaces and piped those sub-spaces from different regions into a deep neural network model which tried to discover the relationship between those sub-spaces and could map them to high level features. Aderghal et al. (2017) also extracted hippocampus patches and fed them into a CNN architecture to learn the features. Sarraf et al. (2016) segment the MRI scans into grey matter, white matter, and cerebrospinal fluid (CSF), and fed them into a CNN to learn features. This indeed has good performance on feature extraction because it combines the advantages of both methods (ROIs and CNN). However, it has all disadvantages of the two: it requires professional knowledge of the medical and brain, and still requires considerable volumes of data to train.

In addition to directly using ROIs and applying neural networks, there are some other approaches to learn features (or convert original representation of features to other formats), such as the wavelet transformation, co-occurrence matrix, mapping methods, etc..

Kamal et al. (2022) applied a Haar wavelet transformation to the MRI scans pre-processed by histogram equalisation. Wavelet transformation is a mathematical tool for analysing data where features vary over different scales, and Haar wavelet transformation is one of the simplest wavelet transformation (Kamal et al., 2022). Alam et al. (2017) applied another type of wavelet transformation, dual-tree complex wavelet transformation (Dual-Tree CWT), to the input MRI images. The Dual-Tree CWT engaged two real diffusion weighted tensor images and provided the real and imaginary components of the wavelet transform, respectively (Alam et al., 2017). Using the wavelet transformation, the features are not presented as voxel values but signals, e.g., frequency resolution. Such feature representation may have information that voxel values do not have and thus may enhance the performance of the discrimination.

In Sathiyamoorthi et al.'s (2021) study, features were computed using second order 2-Dimension Grey Level Co-occurrence Matrix (2D-GLCM). The co-occurrence matrix is a matrix defined over an image to be the distribution of co-occurring pixel values (greyscale values or colours) at a given offset (Nanni, Brahnam, Ghidoni, Menegatti, & Barrier, 2013). Salunkhe et al. (2021) also applied a co-occurrence matrix to the segmented hippocampus to extract the features. The sources of the features in Das et al. (2021) involved a GLCM as well. Das et al. (2021) employed three approaches: a live neuron, a GLCM, and a random forest mapping, to learn the features respectively, and combined the extracted features together.

Ma et al. (2020), Petrone et al. (2019), Spasov et al. (2019), and Papakostas et al. (2015) all involved Jacobian determinant maps in their studies. The Jacobian determinant is the determinant of the Jacobian matrix, which is the matrix of all its first order partial derivatives. It is computed for the deformation field between a reference and a target image (Leow et al., 2006). Therefore, it can capture changes structural changes in the brain between two points in time and so that improve accuracy in the diagnostic classification of AD or MCI versus CN (Plant et al., 2010; Wolz et al., 2011). Ye et al. (2011) employed another map, RAVENS maps, to interpreted the features of the MRI scans. Long et al. (2017) and Liu et al. (2013) both used embedding algorithms transform multivariate MRI data to a lower-dimension space and represent the MRI images as vectors.

Long et al. (2017) applied a metric multi-dimensional scaling (MDS) algorithm and X. Liu et al. (2013) used a locally linear embedding (LLE) algorithm for embedding in their studies, respectively. H. Lin et al. (2022) used 90 ROIs and 7 brain networks calculated from structural MRI scans and 7 brain networks and 4 parameters computed from diffusion weighted tensor images as features. The brain network (connectivity) is a pattern of anatomical links, of statistical dependencies, or of casual interactions between distinct units within a nervous system. The template of the brain network used in H. Lin et al. (2022) was designed by Yeo et al. (2015).

Different from using neural networks to extract features, i.e., mapping lower-level features to higher-level features, these studies transformed voxel values to other formats. These formats can represent the original images in a different way from voxels and they can potentially provide more information than what the original images provide, e.g., voxel values, shapes, etc..

2.6.   Synthesis and Limitations of Previous Studies

Feature extraction and classification (or clustering) both have a great influence on the performance of the final diagnosis of AD. As can be seen from Section 2.5.2, the ways to extract features can be divided into two types: using ROIs and using the whole brain (or slices). As can be seen from Table 2.10, most studies that have used ROIs extracted voxel values or information on morphological features from the ROIs without any assumptions being stated or explanation. Directly using features extracted from ROIs requires that the users of the models have extensive clinical experience or medical knowledge so that they can accurately locate the regions relevant to AD: this is clearly not possible for those who are not experienced medical

practitioners. In addition, directly using voxel values of the ROI also leads to another problem: even though the ROIs should present the characteristics of the AD, they may be not easily divisible in the space, which indicates that the voxel values of the ROI are lower-level features. Therefore, as described in the section 2.5.2, some studies (Aderghal et al., 2017; Lee et al., 2019; Hongming Li et al., 2019; Mehmood et al., 2021; Sarraf & Tofighi, 2016; Yanteng Zhang et al., 2022) combined both ROIs and neural networks to locate the AD-relevant regions and transfer them to a higher level, because higher-level features are more abstract semantic representations learned from the neural networks and they are expected to provide more invariance to intra-class variability (Chan et al., 2015).

In addition to using ROIs, many studies started to employ neural networks to detect AD-relevant regions within the whole brain areas. One merit of the neural networks is that the whole brain can be fed into them directly and the meaningful features (the most disease relevant features) can be detected automatically based on the weight of neurons. As can be seen in the section 2.5.2, most studies that applied neural networks employed the CNN architecture or its variations. CNN-based neural networks truly have shown a great enhancement and performance on various visual tasks previously, e.g., image classification, image segmentation (Bazgir et al., 2020). Training basic CNN architectures (including CNN architecture variations), which are supervised and need backpropagation, requires large amounts of labelled training data (Xialei Liu, Weijer, & Bagdanov, 2019). Researchers may have to use nearly half of the data as the training dataset, e.g., Li et al. (2019) employed 803 out of 2146 MRI scans as the training dataset.

Some studies (Xin Liu et al., 2013; Ma et al., 2020; Papakostas et al., 2015; Petrone et al., 2019; Spasov et al., 2019) have tried alternative approaches to transfer the MRI images to other spaces or other patterns of the data. However, the overall performance of the diagnosis of using such methods was still not as good as using neural networks.

To overcome the disadvantage of supervised neural networks, unsupervised neural networks have been considered to diagnose AD using MRI scans. However, due to the lack of backpropagation, unsupervised neural networks may have weaker performance than supervised neural networks if they are applied to whole brain MRI scans. Therefore, feeding ROIs into unsupervised deep learning architectures could be a solution to this ineffectiveness. To pre-select ROIs before feeding the data into unsupervised deep learning architectures, in addition to selecting based on professional experience and knowledge, selecting ROIs using statistical

methods could be another approach. However, only a limited number of studies have combined statistical models and unsupervised neural networks before. In addition, the number of unsupervised neural networks is not as large as that of supervised neural networks. This might then be the reason why most studies have not used unsupervised learning methods to diagnose AD.

Some studies have used morphometric features of different tissues or ROIs in the brain, instead of voxel values, as the extracted features before feeding them to the classifiers. However, studies using morphometric features alone, e.g., Ma et al. (2020), I. O. Korolev et al. (2016), and Lama et al. (2017), did not achieve good performance as those using voxel-based features. Although Y. Wang et al. (2019) used morphometric features in their study and achieved an accuracy of 90% and above for all classifications, they combined connectome (a comprehensive map of neural connections in the brain) as the features. Hence, methods involving multiple types of features have the potential to improve the methods that use only one type of features. Thus, multiple types of features from a single data modality could be considered further in the future studies.

After feature extraction, classification (supervised learning) or clustering (unsupervised learning) is implemented, and the choice of the algorithms is crucial to the performance as well. As presented in the Section 2.5.1.1, amongst the supervised learning methods, SVM and CNN architectures are the most commonly-used models applied in the diagnosis of AD using MRI scans. The performance of SVM is highly influenced by the distribution of data in the space. Although this can be solved by using kernels, selecting kernels also depends on whether the data are linearly separable or not, which requires setting parameters of the kernel. Thus, to some extent, the distribution of data (or features) in the space determines the choice of kernel. However, the SVM has no function for feature extraction, thus it depends much on the feature extraction before the classification.

Unlike SVM, CNN and its other architectures can both extract features and classify data in the process, which has been shown to have powerful performance in the field of image recognition. However, their advantageous performance is based on having a considerable volume of data to train on. As mentioned in Chapter 1, both acquiring medical imaging scans and labelling them are quite difficult because both of them are time-consuming and need medical practitioners, which has a labour cost. Therefore, some studies have started to use semi-supervised learning and unsupervised learning.

Amongst application of semi-supervised learning methods, multiple models have been used by different studies and semi-supervised GAN is one of commonly used semi-supervised methods in Table 2.8. Nevertheless, GAN is not stable because it relies a lot on both the generator and the discriminator, but stability is important in diagnosing diseases. As for some other semi-supervised learning methods, e.g., label propagation algorithms and co-training, they utilise pseudo-labels to maximise the use of the unlabelled data. However, the performance of such methods is much determined by the consistency between the pseudo-labels and the real labels, and this becomes one of unstable factors as well. Therefore, it is necessary to think how the pseudo-label generators are given and this often requires comprehensive knowledge of the research field.

As for the unsupervised learning methods, both the k-means algorithm and the unsupervised GAN have been commonly used. Compared with GAN, k-means algorithm has fewer parameters, that is, it only has one parameter, k, the number of clusters to be created. In diagnosing AD, k is often fixed because the discrimination of stages of AD is generally clarified in the studies. However, unlike GAN, k-means can only split the data into groups but cannot extract features. Therefore, it is necessary to apply an appropriate feature extraction method before applying the k-means algorithm, especially an unsupervised feature extraction method.

As can be seen from Section 2.5.1, despite the lack of labelled MRI scans, the number of studies using supervised learning methods is still much larger than the number of studies using unsupervised learning methods. Therefore, unsupervised learning methods could be more considered in future research. However, due to the lack of capabilities to learn from labels of data and output measures in the unsupervised learning methods, it will not always contain the expected grouping (clustering) results. Hence, the potentially meaningful features, which are related to AD, should be identified before the clustering algorithms are applied. In this way, feature extraction is more important in the process if the clustering algorithms are used for discriminating AD. In addition, none of the studies (using unsupervised learning methods) shown in Table 2.9 discriminated between sMCI and pMCI scans.

As can be seen from the previous studies, both feature extraction and classification/clustering are important. In brief, feature extraction is used to discover pathological (AD-relevant) regions and classification/clustering is to discriminate data based on the extracted features.

Based on the review of literature above, there are several gaps and limitations can be identified as follows:

(1) Single modality, e.g., structural MRI, functional MRI, etc., was employed by most studies. Namely, most studies used structural MRI scans or functional MRI scans only, instead of a combination of them or other modalities;

(2) Analysis of ROIs and feeding the whole brain MRI scans (or whole slices) into neural networks are commonly used to extract features; however, the number of studies using the combination of ROIs and neural networks is still relatively small;

(3) The number of studies using unsupervised learning or semi-supervised learning methods is much less than that of studies using supervised learning methods. Most studies therefore still need a considerable volume (nearly half or more) of labelled MRI scans to train on;

(4) The studies that have used unsupervised learning methods focussed on discriminating between AD and CN; none of them discriminate between sMCI and pMCI;

(5) Despite using supervised learning methods, previous research on discriminating between sMCI and pMCI report relatively lower accuracy compared with discriminating between AD and CN;

(6) When discriminating between sMCI and pMCI, few studies have used CN and AD data;

(7) Few studies performed multi-classification (AD, CN, and MCI). Although some researchers have tried multi-classification, the results of the classification were not good (accuracies were less than 60%) as for binary classification.

2.7.   Conclusion

The research literature was searched mainly using four search engines: Google Scholar, Medline via Ovid, StarPlus, and Web of Science. As can be seen from the results of the search, deep learning started to be widely used in diagnosing AD in the last five to ten years; the number of studies using unsupervised learning methods is much less compared with supervised learning methods.

In Section 2.3, the machine learning techniques for healthcare that were described were mainly applied in public health, clinical decision support systems, and others fields such as medical image segmentation and auxiliary medical study. Section 2.4 narrowed the scope of the literature: studies applying machine learning techniques to diagnose diseases, i.e., cancer and COVID-19, using medical imaging scans were reviewed. Section 2.5 further narrowed the scope from Section 2.4 and it reviewed studies that applied machine learning techniques to

diagnose AD using MRI scans. Most studies employed supervised learning methods to discriminate AD, and SVM, KNN, random forests, and CNN architectures are the most commonly used models of supervised learning methods developed for the diagnosis of AD using MRI scans; k-means and GAN were the most commonly used in unsupervised learning methods; semi-supervised GAN was also popular in the semi-supervised methods. As can be seen from Table 2.7 to Table 2.9, the number of studies using supervised learning methods is much greater than those using the other two methods. Section 2.5.2 introduced the methods used to extract features from MRI scans. The methods of feature extraction can be summarised as four types: extracting ROIs, feeding the whole scan into neural networks, feeding ROIs into neural networks, and converting voxel value data to other formats of data, e.g., determinant maps or morphometric features. Section 2.6 synthesised the methods used for diagnosing AD based on brain MRI scans and listed limitations and gaps of the research in this field.

In next chapter, the methodology of the whole study will be presented. The methodology will optimise some of the limitations and gaps mentioned above, and it will describe the steps other than feature extraction and discrimination, i.e., data collection, data pre-processing, etc.

Chapter 3   Methodology

3.1.   Introduction

In the previous chapter, the literature in the areas related to this research was reviewed and their research methods were introduced and summarised. From the literature review, it can be seen that supervised learning methods, including supervised deep learning models, often require a considerable amount of labelled data to train and learn. Therefore, employing supervised learning methods results high labour costs because much data needs labelling manually but some types of data are difficulty to label, especially medical imaging scans. Although crowdsourcing, a way to collection information from a group of people via the Internet, can offer a mechanism to label large volumes of data, e.g., ecological observations, satellite imagery, etc., labelling medical imaging scans need professional clinical knowledge, which is not feasible with trained clinicians. In addition, although pre-trained transfer learning and self-supervised learning can also reduce the reliance on labelled data, as described in Section 1.2.4, the models have to be pre-trained on natural images and still need fine-tuning using a large number of medical images. The pretext task in the self-supervised learning needs much specific medical knowledge, which was not feasible in this study. Therefore, methods that can learn from unlabelled data, which are unsupervised learning methods, could potentially be explored as an alternative. In addition, compared with supervised learning methods, the number of studies that have applied unsupervised learning methods to medical imaging scans to diagnose Alzheimer's disease (AD) is limited. This study, therefore, focussed on the application of unsupervised methods to discriminate between the different stages of AD.

This chapter gives an outline of research methods that were followed in the whole study. Figure 3.1 illustrates the workflow of this study: data collection, data pre-processing, detection of changes between two groups, feature extraction, and clustering. The later sections in this chapter will comprehensively present the steps in Figure 3.1 in order. Section 3.2 presents the research ethics issues in this study. Section 3.3 presents the data collection approach for this study, and it indicates how and from where the data were acquired. Section 3.4 presents the approach used to pre-process the data in this study, including the pre-processing tools and the steps of the pre-processing. Section 3.5 introduces the ways to detect abnormality (AD-relevant regions) detection between each two groups of the data (cognitively normal vs. AD; stable mild cognitive impairment vs. progressive mild cognitive impairment; cognitively normal vs. progressive mild cognitive impairment). It provides a method to detect differentiated regions

between each two groups of the data. Section 3.6 introduces the approach of feature extraction, which indicates the models or algorithms that were used to map features from a low level to a higher level using the pre-processed data. Section 3.7 presents the algorithm used for clustering data. In the final section, Section 3.8 presents the measures for evaluating the performance of the methods in this thesis.
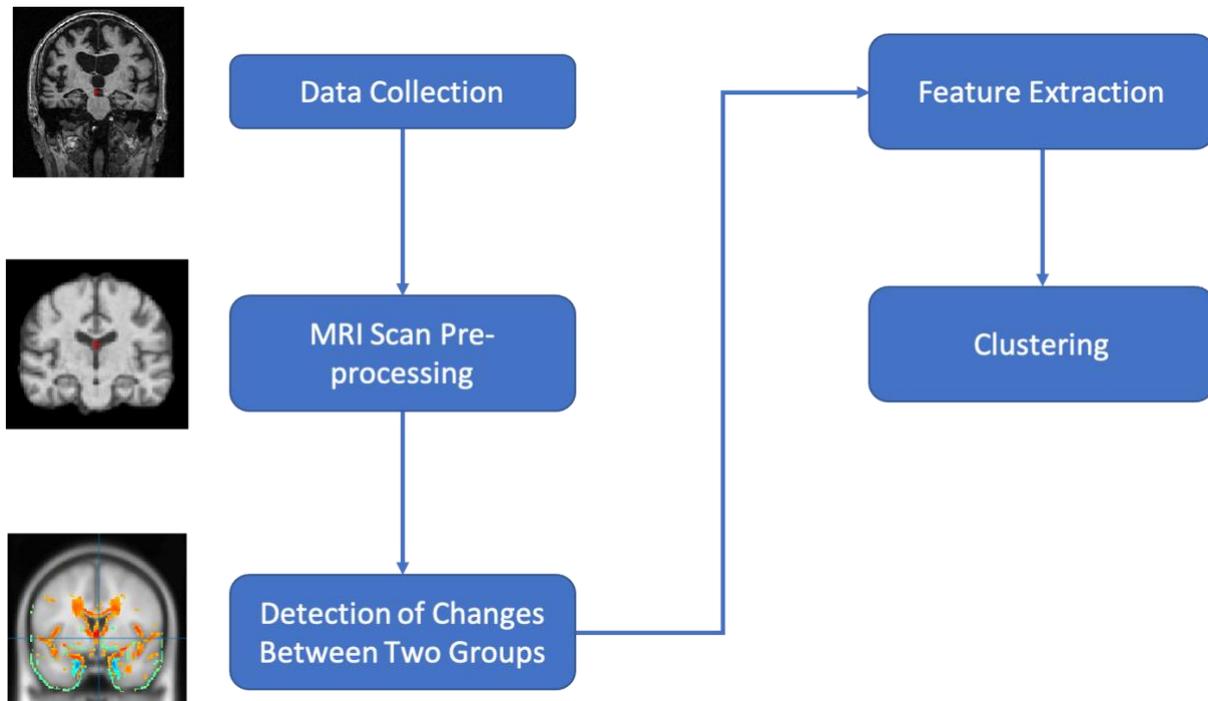
Figure 3.1 Workflow of the study - these stages are explained in further detail in the later chapter sections (MRI images from ADNI adapted by Yuyang Liu)

3.2. Research Ethics

The data used in this research were from the Alzheimer's Disease Neuroimaging Initiative (ADNI). The ADNI database contains completely anonymous data and therefore it is impossible to identify the participants from the information. In addition, ethics approval was obtained for the original study through the Institution Review Board process in the US (*Alzheimer's Disease Neuroimaging Protocol (ADNI)*, 2008). For these reasons, further ethics approval was not required from the University of Sheffield, as confirmed in the letter in Appendix A.

3.3. Data Collection

Public databases for studies of AD have been made available since the 2000s. The most commonly used databases include the Alzheimer's Disease Neuroimaging Initiative (ANDI), the Open Access Series of Imaging Studies (OASIS), AddNeuroMed, Minimal Interval

Resonance Imaging in Alzheimer's Disease (MIRAD), the Australian Imaging, Biomarkers, and Lifestyle Flagship Study of Ageing (AIBL), etc.. Researchers can acquire AD-related data from these public databases conveniently. In addition, data in AD-related studies have sometimes been obtained from private databases, e.g., National Health Insurance Service – Ilsan Hospital (NHIS-IH). Therefore, a range of datasets are available that can be potentially used to study AD.

As the earliest launched public database in the field of study on AD, ADNI has the most comprehensive data compared with other major AD databases, including the largest amounts of data, multiple data types, and several modalities. In addition, as can be seen in Table 3.1, a large number of the studies chose to use ADNI (or use ADNI as one of databases in their studies). Therefore, the data used in this study were obtained from ADNI.

Table 3.1 A summary of databases used by reviewed studies

References	Database
(Yanteng Zhang et al., 2022)	ADNI
(AlSaeed & Omar, 2022)	ADNI & MIRIAD
(Sharma et al., 2022)	ADNI
(Katabathula et al., 2021)	ADNI
(Muhammed Raees & Thomas, 2021)	ADNI
(Pang et al., 2021)	ADNI
(Massalimova & Varol, 2021)	OASIS
(Saratxaga et al., 2021)	OASIS
(Sathiyamoorthi et al., 2021)	ADNI
(Petrone et al., 2019)	ADNI
(Lee et al., 2019)	ADNI
(Hongfei Wang et al., 2019)	ADNI
(Hongming Li et al., 2019)	ADNI&AIBL
(Spasov et al., 2019)	ADNI
(Y. Wang et al., 2019)	NHIS-IH&ADNI
(Basaia et al., 2019)	ADNI
(Razavi et al., 2019)	ADNI
(Ramon Casanova et al., 2018)	ADNI
(Zeng et al., 2018)	ADNI
(W. Lin et al., 2018)	ADNI

(Backstrom et al., 2018)	ADNI
(Asl et al., 2018)	ADNI
(Lama et al., 2017)	ADNI
(S. Korolev et al., 2017)	ADNI
(Moradi et al., 2017)	ADNI
(Aderghal et al., 2017)	ADNI
(Beheshti & Demirel, 2016)	ADNI
(Khedher et al., 2015)	ADNI
(Yudong Zhang & Wang, 2015)	OASIS
(R Casanova et al., 2013)	ADNI
(Aguilar et al., 2013)	AddNeuroMed

3.3.1. Alzheimer's Disease Neuroimaging Initiative

As can be seen from Table 3.1, Alzheimer's Disease Neuroimaging Initiative (ADNI) is the most widely used source of data for neuroimaging research. It is a longitudinal study funded by the National Institution of Health (NIH) in the United States, which aims at studying elderly people with normal condition, mild cognitive impairment, or Alzheimer's disease (Jack et al., 2008). The imaging data in ADNI study has MRI and PET scans. In addition to MRI and PET scans, the study also assesses urine, serum, cerebrospinal fluid (CSF), to monitor other biomarkers as well as clinical information (Jack et al., 2008).

ADNI started in 2004 and is a private-public partnership created by 20 companies and two foundations. Up to this point, ADNI study has been undertaken in four phases: ADNI-1, ADNI-GO, ADNI-2, and ADNI-3. ADNI-1 started in 2004 and last 5 years; ADNI-GO is an extension of ADNI-1, which last for 2 years from 2009; ADNI-2 started in 2011 and ADNI-3 started in 2016.

3.3.2. Data Preparation

In this study, data from ADNI-1 were used due to its relatively larger amounts of data, from 800 participants, compared with the other three study phases. The modality of the data in this study was 1.5T T1-weighted structural magnetic resonance imaging (MRI), and the baseline dataset consisted of 231 cognitively normal (CN) scans, 198 AD scans, and 286 mild cognitive impairment (MCI) scans. However, in Study 2, sMCI and pMCI scans were discriminated, thus the MCI scans needed to be divided into the two subgroups (sMCI and pMCI) beforehand. The

MCI scans were divided based on their AD progression over 36 months following the baseline using the diagnostic summary document provided by the ADNI. The "baseline" in the ADNI database means the first visit and the first MRI scans for the subjects (patients). Section 5.3 in Chapter 5 will provide the detailed operations of how the MCI data were divided. Figure 3.2 shows a summary of the "data collection" step within the workflow of the study.
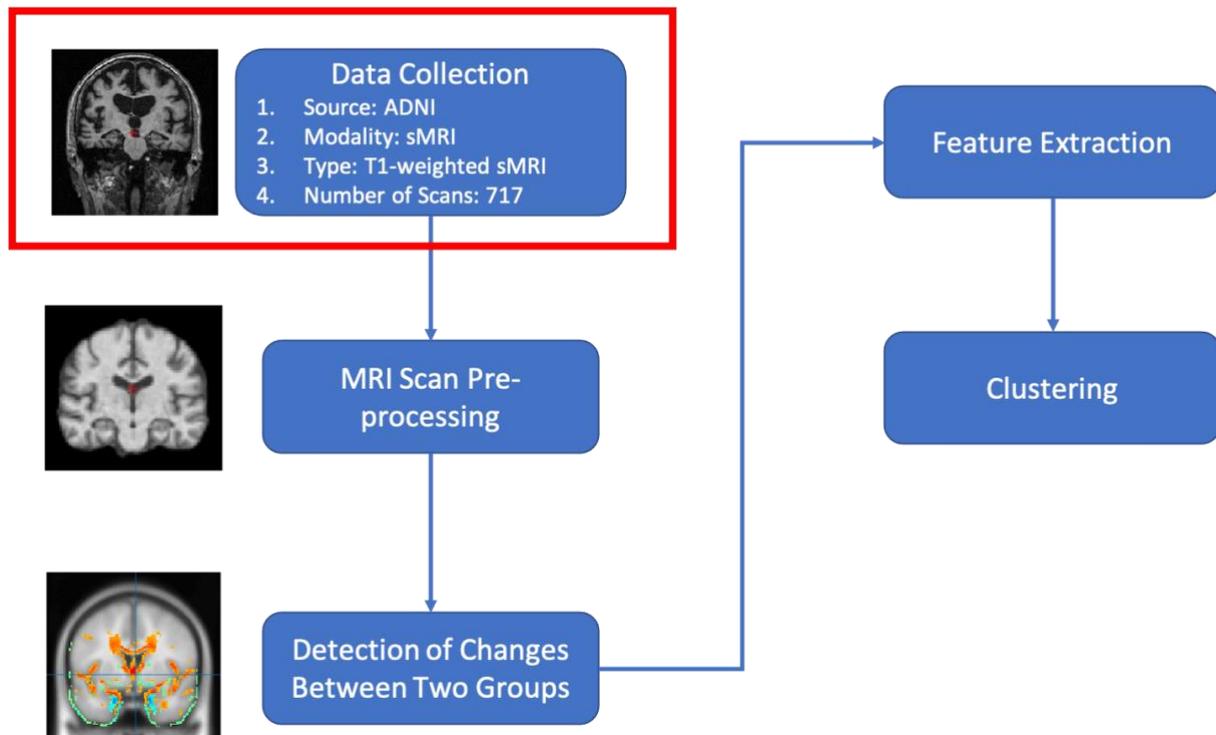
Figure 3.2 A summary of "Data Collection" step in the workflow (MRI images from ADNI adapted by Yuyang Liu)

To obtain data from the ADNI, the search engine of the ADNI (shown in Figure 3.3) was used, and the corresponding choices were selected. However, as can be seen in Figure 3.4, there are several baseline scans for each subject, because the ADNI provides different raw MRI scans and different pre-processed versions of an MRI scan. In this study, raw scans were used and the scans whose "Description" column is "MP-RAGE" sequence were chosen from the ADNI-1. The full name of MP-RAGE is Magnetization Prepared-Rapid Gradient Echo, and is a popular sequence for brain imaging in clinical settings (J. Wang, He, Zheng, & Lu, 2014). It can generate images with high contrast of tissues and provides high spatial resolution with the whole brain coverage in a short scan time which is generally under 10 minutes (Jack et al., 2008).

Figure 3.3 ADNI image search engine (Screenshot by Yuyang Liu)

Figure 3.4 Search results (Screenshot by Yuyang Liu)

3.4. Data Pre-processing

In this study, the MRI scans were pre-processed using the following steps:

(1) Anterior commissure-posterior commissure alignment (AC-PC alignment);

(2) Motion correction;

(3) Intensity normalisation;

(4) Template registration;

(5) Skull stripping.

The steps (2) to (5), motion correction to skull stripping, were completed using the pre-processing software while step (1), AC-PC alignment, had to be conducted manually using Statistical Parametric Mapping 12 (SPM12) beforehand. The commissures are fibre tracts connecting two hemispheres and they are important landmarks when brains are registered into a template (e.g., Talairach space) (Ardekani & Bachman, 2009; Volkau et al., 2012). As shown in Figure 3.5, AC-PC alignment is to set the brain so that the anterior commissure (in red) and

the posterior commissure (in yellow) are on a horizontal line. In step (2), the motion correction aims to remove the motion of the patient, e.g., shake or rotation, during the process of MRI scan acquisition. Intensity normalisation (step (3)) aims to normalise the intensities of voxels to a common scale across people because the sources of the MRI scans were arbitrary. Template registration (step (4)) was carried out to register different MRI images collected from different sources (i.e., from different participants or different MRI machines) into a common template or 3D space. The final step, skull stripping (step (5)), was to remove the skull of the brain. To implement the aforementioned pre-processing steps (from motion correction to skull stripping), two tools, FreeSurfer software suite (version 6.0.0) and Statistical Parametric Mapping (SPM), were utilised independently. The performances of pre-processing using the two sets of software were compared and it is shown in Section 4.9.1.
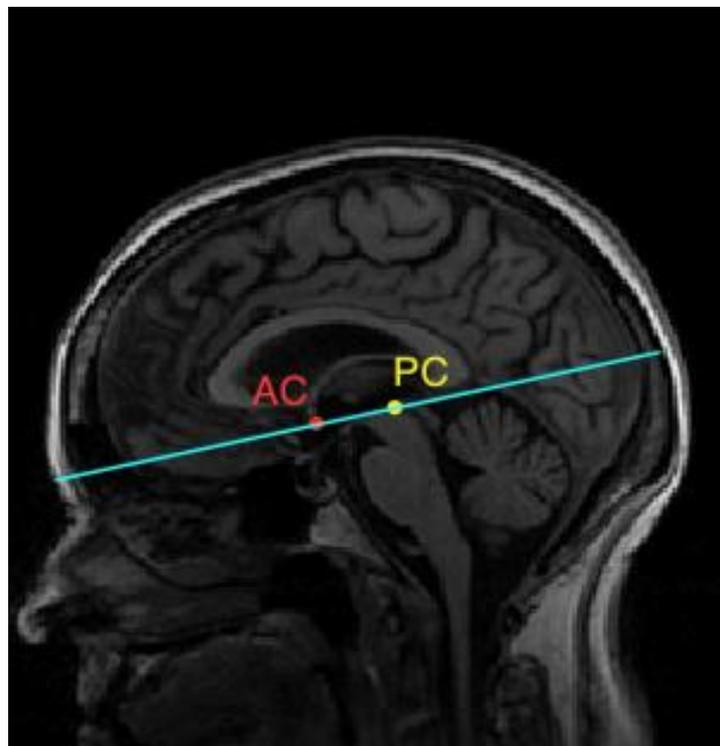
Figure 3.5 Anterior commissure-posterior commissure alignment (Original image from ADNI adapted by Yuyang Liu)

3.4.1. FreeSurfer

FreeSurfer (Laboratory for Computational Neuroimaging at the Athinoula A. Martinos Center for Biomedical Imaging, n.d.-a), developed by the Laboratory for Computational Neuroimaging at the Athinoula A. Martinos Centre for Biomedical Imaging, is an open-source software utilised to the analyse and visualise structural and functional neuroimaging data

derived from cross-sectional or longitudinal studies. FreeSurfer is the structural MRI analysis software of choice for the Human Connectome Project. It registers brains into a common space – Talairach Template in the first step and normalises the scans. In final step, FreeSurfer was used to strip the skulls (i.e., the bone material in the images) from all of the MRI scans.

FreeSurfer (version 6.0.0) was downloaded from the website https://surfer.nmr.mgh.harvard.edu/fswiki/rel6downloads. Because it has no graphical user interface (GUI), every operation was performed through the command line. Although FreeSurfer provides a sequence of image processing workflows, in this part of the study, only the following the five features were used:

(1) Motion Correction and Conform;

(2) Non-uniform Intensity Normalisation;

(3) Talairach Transform Computation;

(4) Intensity Normalisation 1;

(5) Skull Strip;

In the FreeSurfer, the command "recon-all -all" was used to run the whole pre-processing workflow (the whole pre-processing workflow includes 31 steps). However, to implement the operations (1) – (5) above, the command "recon-all -autorecon1" was used. An example of the command line is as follows:

```
recon-all -s FolderName -i Path -autorecon1,
```
where the "FolderName" is the name of the destination folder where the results will save and the "Path" is the directory path of the image file that needs pre-processing. However, FreeSurfer needs environment configuration before starting to use. The environment configuration on the MacOS system was undertaken as follows:

```
export FREESURFER_HOME=/Applications/freesurfer/7.1.1
source $FREESURFER_HOME/SetUpFreeSurfer.sh.
```
A Matlab script was written to execute command lines for multiple image files because FreeSurfer does not have GUI and multiple MRI scans need to be pre-processed, and the script is made available in Script A.1 in Appendix A.

3.4.2. Statistical Parametric Mapping

SPM (Wellcome Centre for Human Neuroimaging, 2020) is another open source and free software designed for analysing brain imaging data sequences. It refers to software created by the Wellcome Department of Imaging Neural Science at University College London. However,

SPM was originally designed for functional MRI, Positron emission tomography (PET), single-photon emission computerized tomography (SPECT), etc., which are time-series images from the same subject. Thus, to pre-process the structural MRI scans in this research, a toolbox of SPM – Computational Anatomy Tool (CAT), which is an extension to SPM to provide computational anatomy, was used. The graphical user interfaces (GUIs) of SPM and CAT are shown in Figures 3.6 and 3.7, respectively (however, FreeSurfer has no GUI).
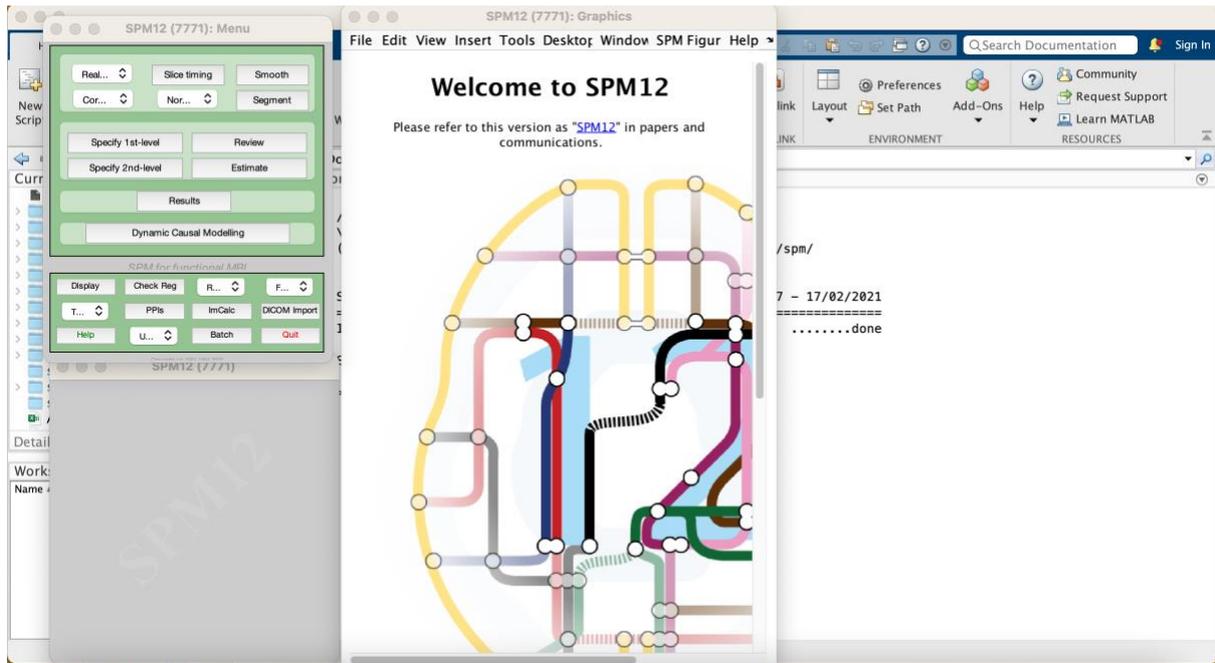
Figure 3.6 GUI of SPM *(Screenshot by Yuyang Liu)*

SPM (version 12, SPM12) was downloaded from the website https://www.fil.ion.ucl.ac.uk/spm/software/download/. Matlab has to be installed beforehand because SPM12 is based on it. However, the pre-processing function was implemented by the CAT (version 12, CAT12) toolbox in the SPM12, which was downloaded from http://www.neuro.uni-jena.de/cat/index.html#DOWNLOAD. For pre-processing, CAT12 provides several features, e.g., segment, segment longitudinal, and some other features for extracting values after pre-processing; however, in this phase, only the segment feature was employed (the segment feature in CAT12 includes all pre-processing steps needed in this study). Figure 3.8 shows an example of the settings of the segment and all the settings were the default values.
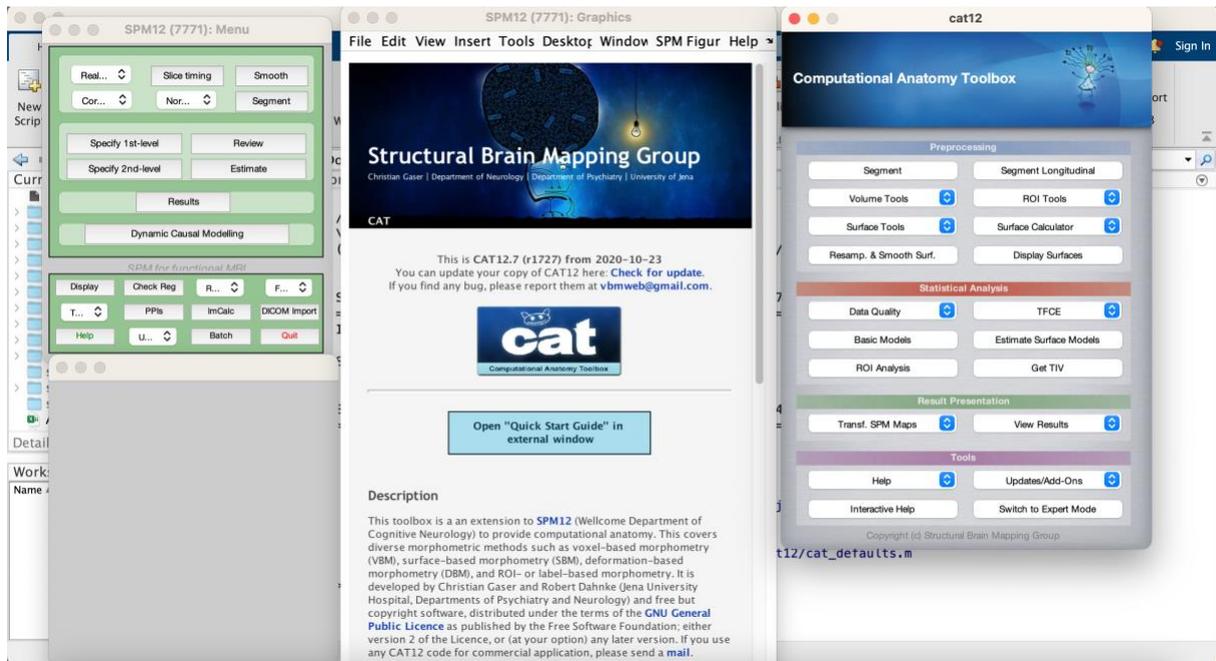
Figure 3.7 GUI of CAT (Screenshot by Yuyang Liu)
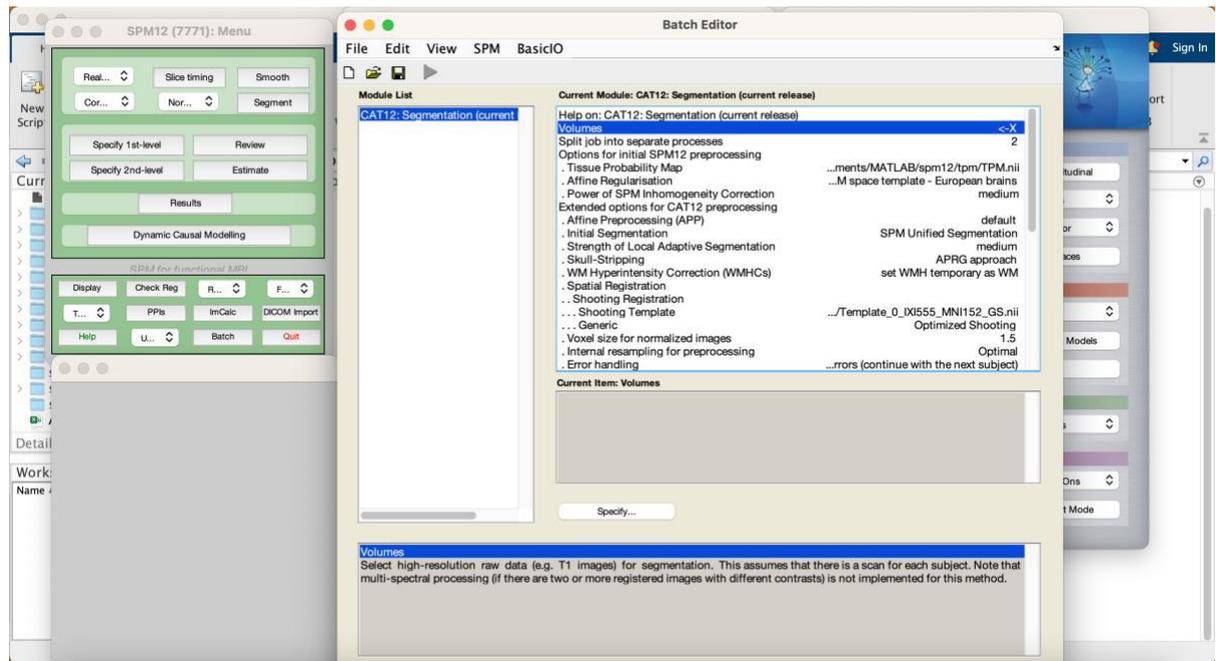
Figure 3.8 An example of pre-processing settings in CAT12 (Screenshot by Yuyang Liu)

The pre-processing in each of the three studies will be described in Sections 4.4, 5.4, and 6.4. Following the pre-processing steps using each of the two tools, a comparative analysis was conducted that assess the performances. This is further discussed in Section 4.9.1. Figure 3.9 shows a summary of "data pre-processing" step within the workflow of the study.
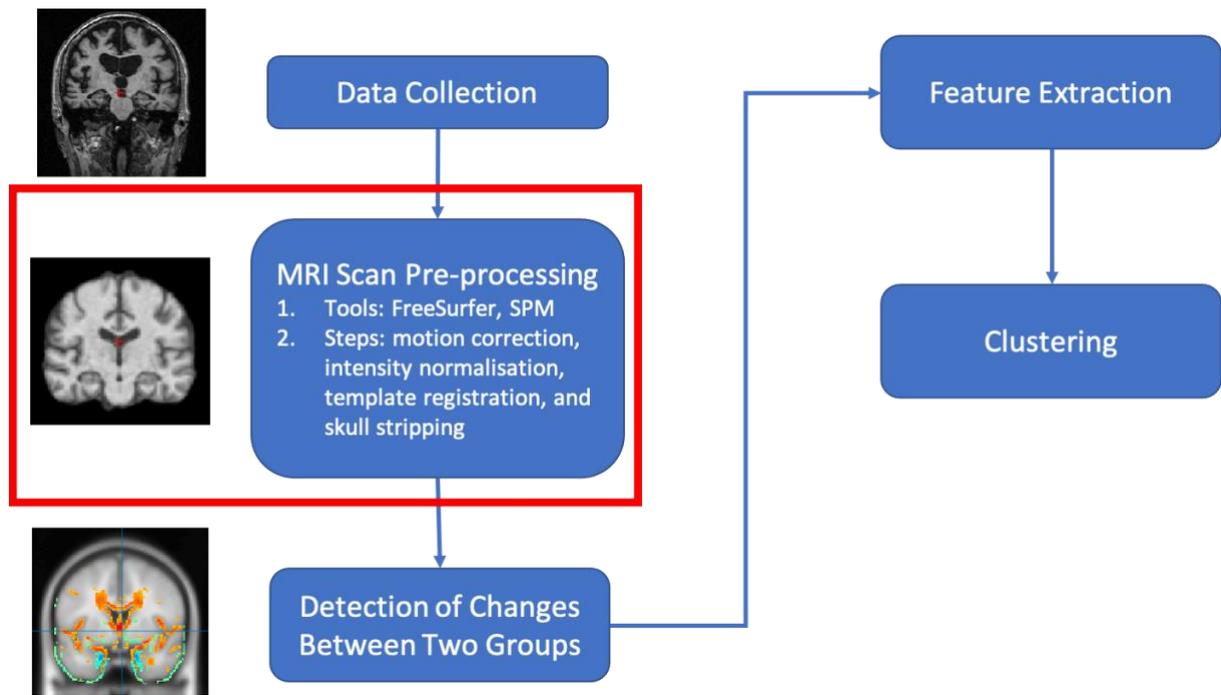
Figure 3.9 A summary of "MRI Scan Pre-processing" in the workflow (MRI images from ADNI adapted by Yuyang Liu)

3.5.   Detection of Abnormalities Between Two Groups

The aim of this phase is to discover regions that are visibly different in each set of two groups (AD vs. CN, sMCI vs. pMCI, and CN vs. pMCI) at the voxel level. Such regions are often called regions of interest (ROIs) and they are important because from a clinical perspective, it is of paramount significance to understand which parts of the brain are meaningful to diagnose AD and differentiate them from brain from CN people (Feng, Zhang, & Chen, 2022).

3.5.1.   Alzheimer's Disease vs. Cognitively Normal Brains

To discover the ROIs between AD and CN scans, a two-sample t-test was applied. Figure 3.10 shows a summary of the detection of abnormalities between two groups in the workflow.
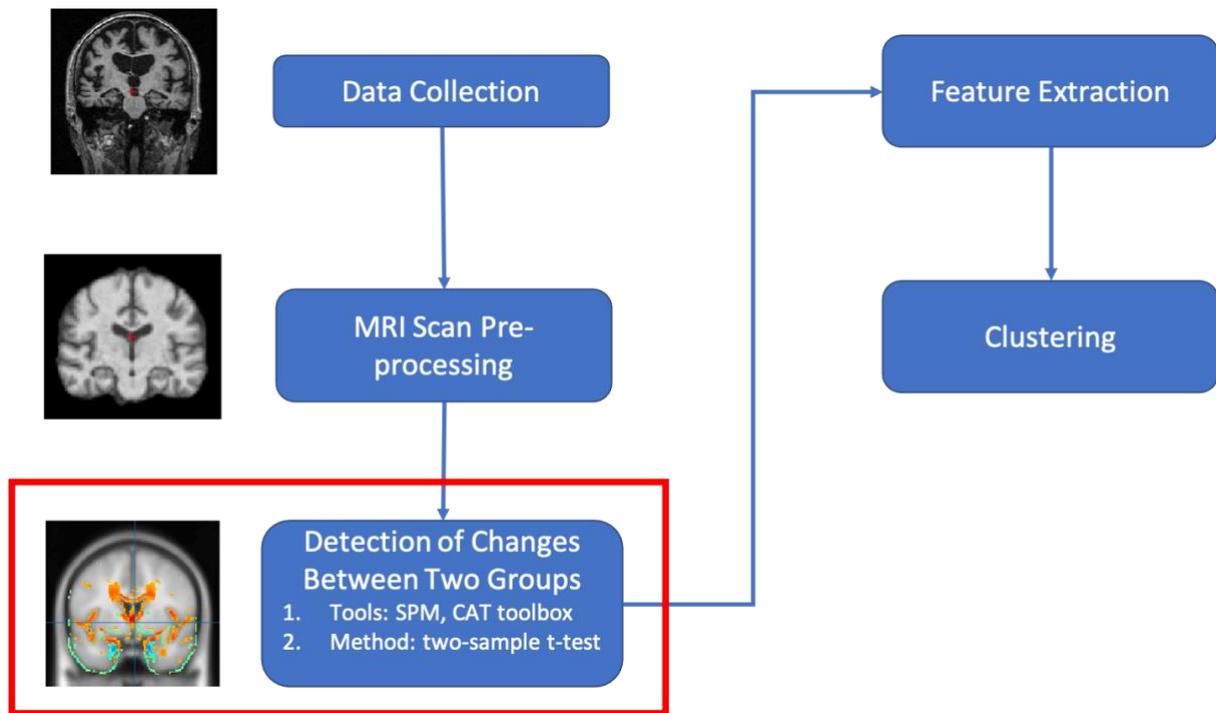
Figure 3.10 A summary of "Detection of Changes Between Two Groups" step in the workflow (MRI images from ADNI adapted by Yuyang Liu)

Two-sample T-test

The two-sample t-test is a statistical method used to test whether the unknown population means of two groups of data are equal (Lachin, 1981). It is a commonly used method to test for ROIs and abnormalities in MRI scans (Leandrou, Lamnisos, Kyriacou, Constanti, & Pattichis, 2020; Seyedi et al., 2020; Beheshti & Demirel, 2016). Although there are other techniques, e.g., convolutional neural network architectures or attention mechanism, they need a large volume of labelled data for backpropagation. However, as mentioned in Chapter 1 and Chapter 2, medical imaging data are not easy to acquire, which leads to relatively low volumes, and it is also hard to label them because it requires professionally-trained clinical practitioners and consumes a great amount of human power. Therefore, in this study, it was decided that the two-sample t-test analysis would be utilised. The two-sample t-test was applied to 30 scans randomly selected from each of AD and CN groups, respectively, to obtain changes and ROIs between the two groups. The reason why 30 scans in each group were selected is that the t-test is generally used for small sizes of data, which is often less than 30 (Gerald, 2018). In addition, 30 scans only account for around 12% to 15% of total scans in each of the two groups, which is not a large proportion. Before the two-sample t-test was employed, three assumptions of it should be fulfilled (Rasch, Kubinger, & Moder, 2011):

1) The data must be continuous;

2) The samples must follow a normal distribution;

3) The two groups of samples must be independent and randomly selected.

In this study, the data met the first assumption because what the two-sample t-test is applied to is the pixels (voxels) in each slice of the MRI images and the intensity of the pixel is a finite region with a continuous value. The MRI images collected from the ADNI were baseline scans from different participants so they are therefore independent and randomly selected. Therefore, the third assumption is valid as well. As for the second assumption, Srivastava (1958) showed that, for practical purposes, the power of the t-test is not seriously affected even if samples are not from a normal distribution. In other words, the t-test is robust to non-normally-distributed samples. However, for a better explanation in this study, a random voxel was selected from each MRI scan of the groups and was histogrammed to indicate whether it was normally distributed or followed a nearly normal distribution. This is further discussed in Section 4.9.2.

3.5.2. Stable Mild Cognitive Impairment vs. Progressive Mild Cognitive Impairment

Based on the previous research (Elahifasaee, Li, & Yang, 2019; Y. Huang et al., 2019; Zeng et al., 2021; X. Zhang et al., 2021; Zheng et al., 2022), discriminating between sMCI and pMCI scans is more difficult than that between CN and AD scans, because both sMCI and pMCI are subtypes of MCI and the difference between them is whether they develop to AD within 36 months. This means that the visible differences between sMCI and pMCI scans are much smaller than that between CN and AD scans. Therefore, it is quite possible that sMCI and pMCI scans may be similar and they may have no obvious ROI.

To test whether there were obvious ROIs between them, a two-sample t-test was also applied to the sMCI and pMCI scans. However, due to smaller volumes of them compared with CN and AD scans, fewer scans of sMCI and pMCI were used for the ROI detection. In this part of the study, a two-sample t-test analysis was applied to 15 sMCI and 15 pMCI scans to detect ROIs between them. If there was no obvious ROIs between the sMCI and pMCI scans, more scans (30 sMCI and 30 pMCI scans) would be used for the two-sample t-test; if the ROIs were still not obvious, the ROIs detected between the CN and AD scans would be used for the discrimination. This is because MCI is an intermediate stage between CN and AD and, as described in Section 1.4.2, MCI shows minor memory problems but not as severe as AD, and this suggests that MCI is very likely to have similar pathological regions as AD but the pathological changes of the regions may be not large as in AD. In addition, pMCI will develop to AD within 36 months, thus pMCI is more similar to AD compared with sMCI; sMCI will

remain beyond 36 months and it also probable that it will return to CN, thus sMCI is more similar to CN. As a result, it is very likely that the relationship between sMCI and pMCI is similar to that between CN and AD, and thus the ROIs between sMCI and pMCI scans may be similar to those between CN and AD scans, respectively.

3.5.3. Cognitively Normal vs. Progressive Mild Cognitive Impairment

Due to the similarity between sMCI and pMCI, distinguishing them is a harder task than distinguishing between CN and AD. Therefore, a task, discriminating between CN and pMCI scans, whose difficulty is between the two was also conducted. In this task, randomly selected 25 CN and 15 pMCI scans were used for detecting ROIs between them by using a two-sample t-test analysis. Similar to the discrimination between the sMCI and pMCI scans (Section 3.5.2), if no obvious ROI was detected between CN and pMCI scans, the ROIs discovered in Section 3.5.1 would be used in the later steps of the workflow in this part of study.

To detect ROIs described above (Sections 3.5.1 to 3.5.3), the toolbox CAT12 in SPM12 was employed because it can provide the two-sample t-test estimation. The viewer tool, Xjview, can show the ROIs between the two groups in the form of a heat map. Xjview was developed by MRI researchers from leading universities, and it can be downloaded from https://www.alivelearn.net/xjview/download/. The heat map is a type of graphical representation of data where values are shown as colours (Zhao, Guo, Sheng, & Shyr, 2014). The interpretation of the heat map in Studies 1 to 3 are presented in Sections 4.9.2.2, 5.9.1, and 6.9.1, respectively. An example of Xjview showing ROIs is shown as Figure 3.11. Before the feature extraction, the detected ROIs were cropped from the original MRI scans. Figure 3.12 demonstrates an example of ROI cropping.
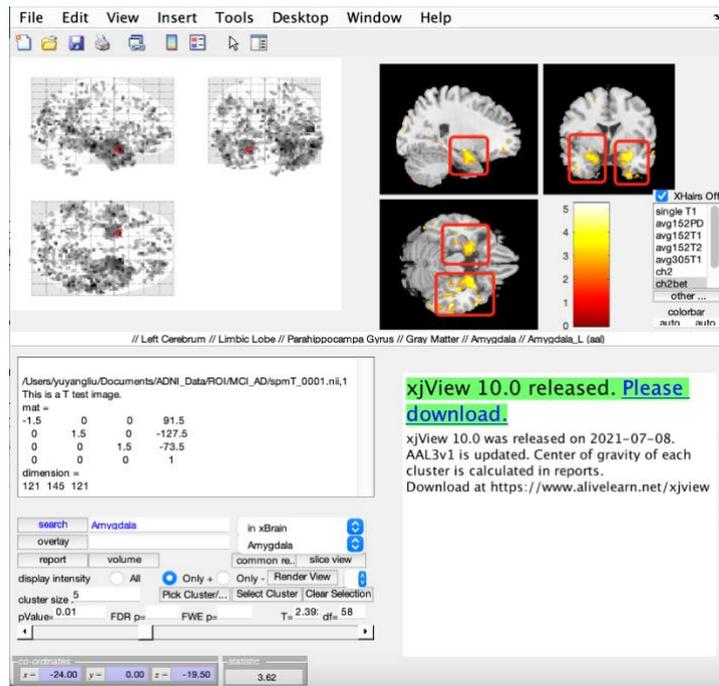
Figure 3.11 an example of Xjview showing ROIs and ROI examples have been circle in red (Screenshot by Yuyang Liu)
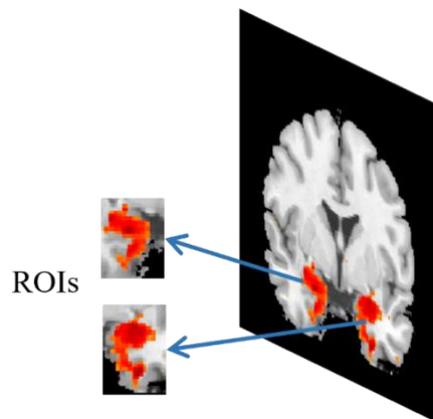
Figure 3.12 an example of ROI cropping (MRI images from ADNI adapted by Yuyang Liu)

To crop the detected ROIs, coordinates of the voxels within the ROIs need to be known. However, it was hard to acquire the exact coordinates of the voxels within the ROIs only by viewing them in Matlab. The coordinates of voxels could be known by looking at the "co-ordinates" area at the bottom of Figure 3.11.

3.6. Feature Extraction

The source code of the feature extraction step can be found in https://doi.org/10.5281/zenodo.8316114.

After the detection of the ROIs between the two groups, the computation of the voxels from the ROIs was performed. The purpose of this was to map the features that indicate the existence and non-existence of AD at a low level (voxel level) to a higher level (more abstract semantics of the data). Although ROIs can indicate the existence of AD, it is still difficult to accurately cluster data using their voxel values directly because, in some cases, changes of a small number of voxels can lead to AD, but such tiny changes may not much contribute to the discrimination between AD and CN (also between sMCI and pMCI; between CN and pMCI). In other words, it is likely that directly using the ROIs recognised in Section 3.5 may not be sufficient to separate the two groups of data well in the original vector space, especially the medical imaging scans, because the differences between AD and non-AD scans might not have large visible differences (i.e., they may be close to each other in the space). Consequently, the voxels within the ROIs should be mapped to another space and, in that space, the two groups should be easily separable.

According to Section 2.5.1.1, supervised deep learning methods can separate the two groups of data well by updating the weight of the ROIs in the process of backpropagation from labels of data. However, as mentioned earlier, this process needs a considerable amount of labelled data; thus, in this study, an unsupervised deep learning model, the PCANet, was used because it has previously shown good performance in feature extraction of MRI scans (Bi et al., 2019).

The PCANet is an unsupervised neural network structure developed by Chan et al. (2015) based on convolutional neural network (CNN) architectures. It is a neural network that combines a Principal Component Analysis (PCA) method and a Convolutional Neural Network (CNN) architecture. CNN was introduced in Section 2.5.1.1. PCA is a commonly used method that reduces the dimensions of the data by projecting it into a subspace whilst retaining the essence of the original data. A PCANet has three stages: the first two stages are similar, which learn convolution filters by PCA method for feature mapping, the final stage is to binarise and generate a block-wise histogram for feature output. The structure of PCANet is shown in Figure 3.13. The details of the PCANet will be introduced in Sections 3.5.1 to 3.5.3.
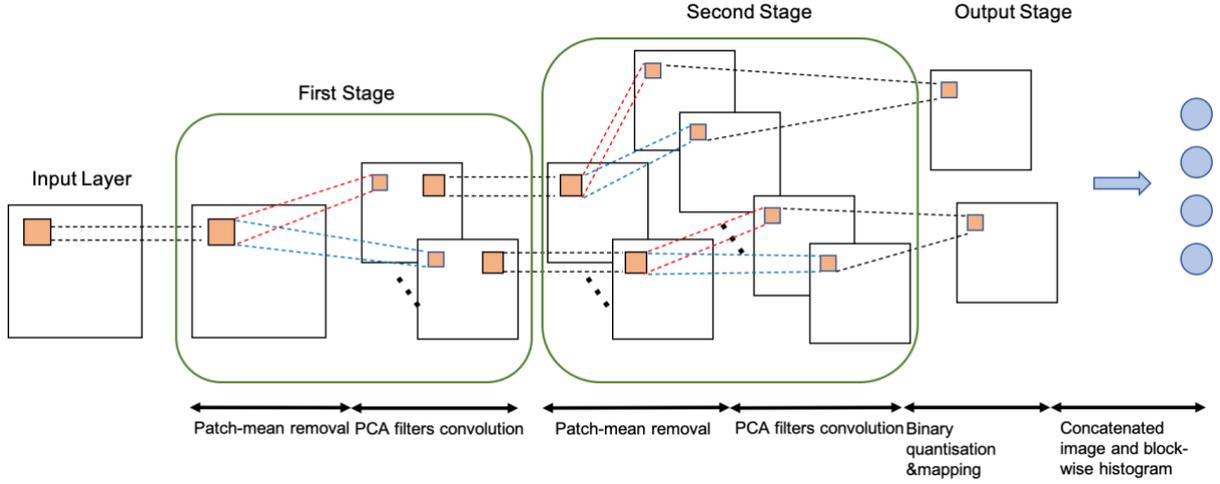
Figure 3.13 The structure of a PCANet (Developed by Yuyang Liu)

3.6.1. First Stage of the PCANet

Given N input images $\{I_i\}_{i=1}^N$ of size $m \times n$ and assuming the size of the patch is $k_1 \times k_2$ at all stages. All overlapping patches within the i_{th} image are collected and there are $m \times n$ vectorised patches in each image. We then subtract the patch mean from each patch and obtain $\bar{X}_i = [\bar{x}_{i,1}, \bar{x}_{i,2}, \ldots, \bar{x}_{i,mn}]$, where $\bar{x}_{i,j}$ is a mean-removed patch. Therefore, operating all images in the same way and we obtain $\boldsymbol{X} = [\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_N] \in \mathbb{R}^{k_1 k_2 \times Nmn}$.

Assuming that the number of filters in layer is L_1, and a PCA was used to learn the filter bank in this stage. The solution is known as the L_1 principal eigenvectors of $\boldsymbol{X}\boldsymbol{X}^T$. Therefore, the PCA filters are expressed as

$$W_l^1 = mat_{k_1,k_2}(q_l(\boldsymbol{X}\boldsymbol{X}^T)) \in \mathbb{R}^{k_1 k_2}, l = 1, 2, \ldots, L_1, \text{(Equation 3.1)}$$

where $mat_{k_1,k_2}(\boldsymbol{v})$ is a function that maps $\boldsymbol{v} \in \mathbb{R}^{k_1 k_2}$ to a matrix $\boldsymbol{W} \in \mathbb{R}^{k_1 \times k_2}$, and $q_l(\boldsymbol{X}\boldsymbol{X}^T)$ denotes the l_{th} principal eigenvector of $\boldsymbol{X}\boldsymbol{X}^T$. This captures the main variation of all the mean-removed patches.

3.6.2. Second Stage of the PCANet

The process of the second stage is almost a repeat of the same process in the first stage. Let the l_{th} filter output of the first stage be:

$$I_i^l = I_i * \boldsymbol{W}_l^1, i = 1, 2, \ldots, N, \text{(Equation 3.2)}$$

where $*$ denotes 2D convolution and the boundary of I_i is zero-padded before convolving with $\boldsymbol{W}_l^1$ in order to make the I_i^l having the same size of I_i. Similar to the first stage, all the overlapping patches of I_i^l are collected, and the patch mean is subtracted from each patch. This

91

forms $\bar{Y}_i^l = [\bar{y}_{i,l,1}, \bar{y}_{i,l,2}, \dots, \bar{y}_{i,l,mn}] \in \mathbb{R}^{k_1 k_2 \times mn}$, where $\bar{y}_{i,l,j}$ is the j_{th} mean-removed patch in I_i. We further defined $Y^l = [\bar{Y}_1^l, \bar{Y}_2^l, \dots, \bar{Y}_N^l] \in \mathbb{R}^{k_1 k_2 \times Nmn}$ for the matrix that collects all mean-removed patches of the l_{th} filter output and concatenate Y^l for the outputs of all filters:

$$Y = [Y^1, Y^2, \dots, Y^{L_1}] \in \mathbb{R}^{k_1 k_2 \times L_1 Nmn}.$$

The PCA filters of the second stage are obtained as:

$$W_l^2 = mat_{k_1, k_2}(q_l(YY^T)) \in \mathbb{R}^{k_1 k_2}, l = 1, 2, \dots, L_2, \text{ (Equation 3.3)}$$

where L_2 is the number of filters in the second stage. For each input I_i^l of the second stage, there are L_2 outputs, and each convolves W_l^2 for $l = 1, 2, \dots, L_2$:

$$\mathcal{O}_i^l = \{I_i^l * W_l^2\}_{l=1}^{L_2}. \text{ (Equation 3.4)}$$

The number of outputs in the second stage is $L_1 L_2$.

3.6.3. Output Stage of the PCANet

The final stage of PCANet employs hashing and generates a histogram. Using hashing, L_2 outputs in $\mathcal{O}_i^l$ are converted into a single integer-valued "image":

$$\mathcal{T}_i^l = \sum_{l=1}^{L_2} 2^{l-1} H(\mathcal{O}_i^l), \text{ (Equation 3.5)}$$

whose every pixel is an integer in the range $[0, 2^{l-1} - 1]$. $H(\cdot)$ is a Heaviside step function whose value is one for positive and zero otherwise.

Then for each of the L_1 images $\mathcal{T}_i^l, l = 1, 2, \dots, L_1$, we partition it into B blocks. We compute the histogram (with 2^{L_2} bins) of the decimal values in each block. Followed by concatenating all the B histograms into one vector and denote as $Bhist(\mathcal{T}_i^l)$, the feature vector of the input image I_i is then defined to be the set of block-wise histograms:

$$f_i = [Bhist(\mathcal{T}_i^1), Bhist(\mathcal{T}_i^2), \dots, Bhist(\mathcal{T}_i^{L_1})]^T \in \mathbb{R}^{(2^{L_2})L_1 B}.$$

In addition, the local blocks can be either overlapping or non-overlapping.

3.6.4. PCANet in This Research

The source code of the PCANet was downloaded from the author's web page (Chan, 2015). In addition, to run this source code, XCode on Mac was installed beforehand if MacOS operating system was used because one of the functions, i.e., "im2colstep()", was programmed by C language. Furthermore, the environment configuration using the command line "mex -setup" in the command window of Matlab was necessary to choose the suitable complier (XCode) for the mex file. The "make" command was then executed to build a group of programs in the PCANet package.

To use the PCANet in this research, the parameters of it needed to be set beforehand. The number of stages was set to two because two-stage PCANet is sufficient to achieve a good performance (Chan et al., 2015). Based on multiple trials and the default values provided by Chan et al. (2015), other parameters were properly set as follows:

- The size of the patches $k_1 \times k_2$ was set to 3×3 in both two stages;
- The number of filters L_1 and L_2 were both set to eight, respectively, which indicated that there were eight principal eigenvectors in each PCA process of the two stages;
- The block size for the histograms was set to 15×15;
- The overlap ratio of the local blocks was set to zero in this study for reducing the number of features.

The parameter settings were applied to all the three tasks: CN vs. AD; sMCI vs. pMCI; CN vs. pMCI). Figure 3.14 shows a summary of the feature extraction step in the workflow.
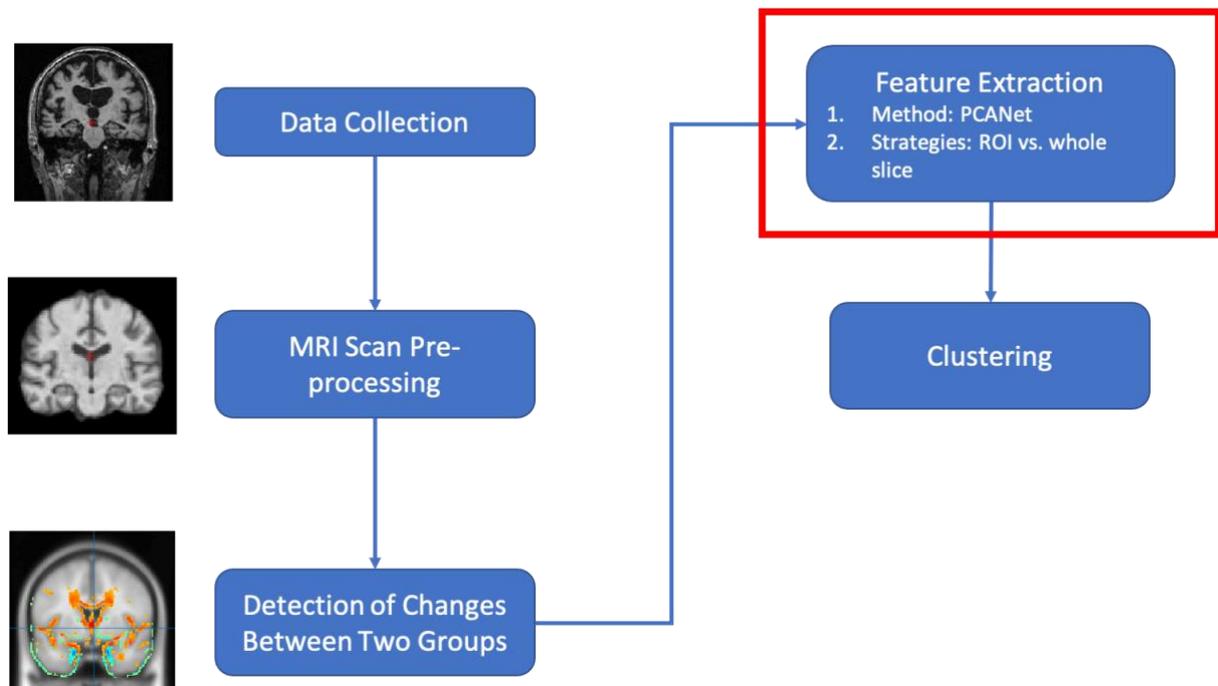
Figure 3.14 A summary of the "Feature Extraction" step in the workflow (MRI images from ADNI adapted by Yuyang Liu)

3.7. Clustering

The source code of the clustering step can be found in https://doi.org/10.5281/zenodo.8316114.

According to the study by Alashwal et al. (2019), unsupervised clustering methods applied to datasets of Alzheimer's disease in prior studies can be summarised as three types: k-means or k-means-mode, multi-layer clustering, and hierarchical agglomerative clustering. However,

previous studies employed multi-layer clustering and hierarchical agglomerative clustering to mine useful information from the dataset or were used for predictive problems, instead of diagnostic problems (Alashwal et al., 2019). Furthermore, the number of clusters in previous studies were unknown when using the two types of clustering methods because the results of them would depend on the data itself. However, in this study, data should be divided into two groups in each task: AD vs. CN, sMCI vs. pMCI, and CN vs. pMCI, which means in each of clustering tasks the data should be split into two clusters. In addition, the data in this study is MRI imagery, which has high dimensions, thus it is difficult to estimate the distribution of the data intuitively. Therefore, the clustering methods based on the density or distribution of the data, e.g., Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm and mixture of Gaussian method, were not applicable in this context. In conclusion, k-means-mode algorithms are suitable to the study because the number of clusters can be directly specified using the algorithm. As described in Section 2.5.1.3, k-means is one of the most popular clustering methods. It divides a collection of data into k number of disjointed cluster. The basic k-means algorithm is as the follows:

(1) "Select k points as initial centroids randomly;

(2) repeat

(3)   Form k clusters by assigning each point to its closest centroid;

(4)   Recompute the centroid of each cluster;

(5) Until Centroids do not change"

<div align="right">(Tan, Steinbach, Karpatne, & Kumar, 2019, p. 535)</div>

Generally, the k-means-mode algorithm includes several algorithms that are developed based on the basic k-means, i.e., k-mode algorithm, fuzzy-c-means, k-means++, etc. However, the k-mode algorithm is often used for the data with categorical features, which is not suitable for the situation in this study. The fuzzy-c-means algorithm is a soft clustering algorithm, which means it allows one piece of data to belong to two or more clusters, thus it also does not apply to this study. In this study, k-means++ (Arthur & Vassilvitskii, 2006) was employed to discriminate between AD and CN, between sMCI and pMCI, and between CN and pMCI. K-means++ is an optimised version of the basic k-means algorithms, and its algorithm is as the follows:

(1) Select the first centroid in a random way;

(2) repeat

(3) Calculate the distance from each data point x_i to the nearest, previously chosen centroid, and we note it as $dist(x_i)$;

(4) Choose the new data point from the n data points as the next centroid with the help of the maximum probability: $\frac{dist(x_i)^2}{\sum_{i=1}^{n} dist(x_i)^2}$;

(5) until k centroids are selected;

(6) repeat

(7)   Form k clusters by assigning each point to its closest centroid;

(8)   Recompute the centroid of each cluster;

(9) Until Centroids do not change

The k-means++ was applied to all three tasks (clustering CN vs. AD; sMCI vs. pMCI; CN vs. pMCI) in this study. However, different clustering strategies were implemented in these three different tasks. The strategy of each task will be presented in the corresponding sections in their own chapters (Chapters 4 to 6).

Clustering is often unstable because ambiguity may exist in data and there are various different ways to cluster data. It is therefore important to check for cluster stability using, e.g., internal and external cluster validity methods. However, due to the limited accessibility to MRI scans and limited time, the internal and external cluster validity methods were not used in this study. Nevertheless, AD-related regions were identified and used in this study and it is reasonable that the clustering is clear and the data should be clustered based on AD-related characteristics. Therefore, it is not unreasonable to expect that the clustering would be stable in this study. The internal and external cluster validity methods could be implemented in future work.

However, even though the k-means++ optimises the problem of centroid selection in basic k-means algorithms, the selection of the first centroid is still random, which means that the algorithm has different results for different centroids. Therefore, in this study, the k-means++ algorithm was repeated ten times in each task, and mean values of the ten runs were taken as the final results.

Figure 3.15 shows a summary of the clustering in the workflow.
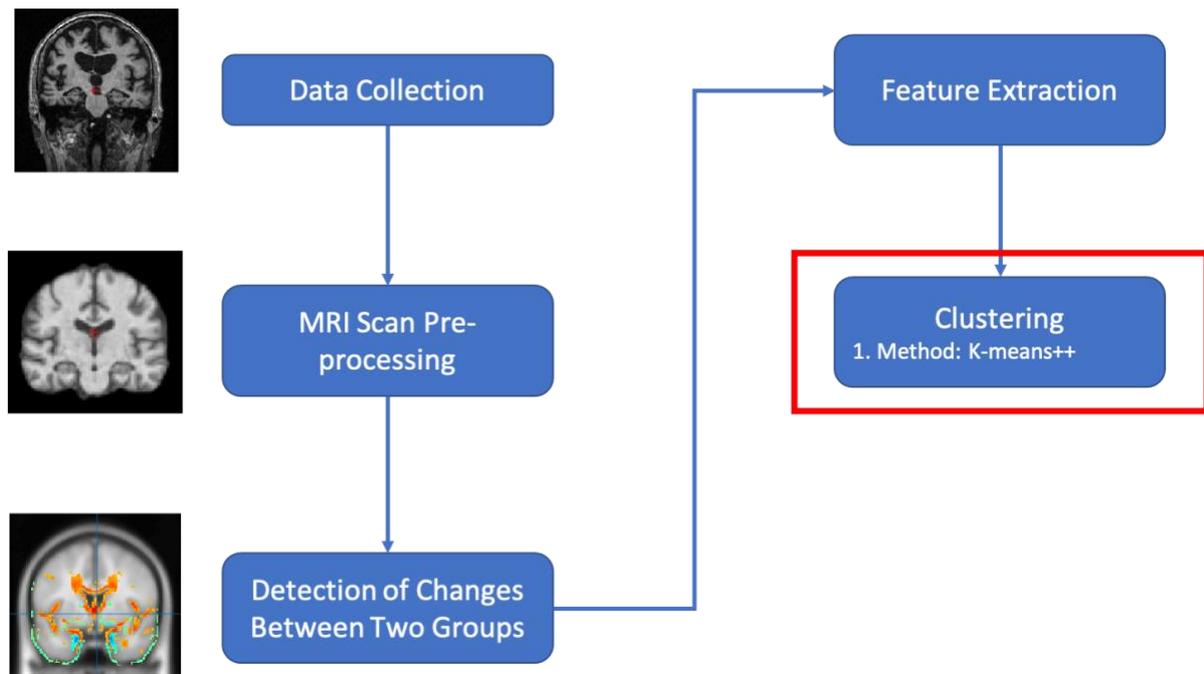
Figure 3.15 A summary of "Clustering" step in the workflow (MRI images from ADNI adapted by Yuyang Liu)

3.8. Performance Evaluation

3.8.1. Recognising Clusters

The k-means++ only labelled the clusters as 1 or 2 in this study instead of diagnostic states (CN, sMCI, pMCI, or AD), therefore the diagnostic states corresponding to the labels in each task needed to be recognised. For example, in discriminating between CN and AD scans, the k-means++ divided the data into cluster 1 and cluster 2 but it does not indicate which cluster is CN and which is AD. Therefore, such correspondence and the clusters need to be clarified.

However, the three tasks (CN vs. AD; sMCI vs. pMCI; CN vs. pMCI) used different strategies to cluster the data, thus their ways to recognise the correspondence and clusters in their tasks are presented in Section 4.8, 5.8, and 6.8, respectively. However, they used the same measures to evaluate the performance in the clustering as presented in Section 3.8.2.

3.8.2. Evaluation Measures

In each task of clustering, after recognising correspondence between diagnostic states (CN, sMCI, pMCI, and AD) and labels (k-means++ labelled clusters as 1 or 2 in this study) in each run, five metrics – accuracy, specificity, sensitivity, positive predictive value (PPV), and negative predictive value (NPV) – could be calculated. In this process, a confusion matrix was employed. Confusion matrix is a matrix used for evaluating the performance of a classification

model (clustering model in this study). For the binary clustering, an example of the confusion matrix is shown in Table 3.2.

Table 3.2 An example of the confusion matrix

		Actual Values	
		Positive	**Negative**
Predicted Values	**Positive**	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

In Table 3.2, some terms need to be explained. True positive (TP) means that both the actual value and predicted value calculated by the model are positive. False positive (FP) means that the actual value is negative while the predicted value from the model is positive; False Negative (FN) means that the actual value is positive but the predicted value from the model is negative; True negative (TN) means that the actual value is negative and the predicted value is negative as well. In the three studies (CN vs. AD; sMCI vs. pMCI; CN vs. pMCI) in this research, positive and negative results refer to positive (abnormal) and negative (normal) diagnosis results, which mean as follows:

(1) Study 1 (CN vs. AD): CN is diagnosed as negative and AD is diagnosed as positive;

(2) Study 2 (sMCI vs. pMCI): although neither sMCI or pMCI is negative (normal), pMCI is more severe because it develops to AD within 36 months. Therefore, pMCI is defined as positive results and sMCI is defined as negative results in Study 2;

(3) Study 3 (CN vs. pMCI): CN is diagnosed negative and pMCI is diagnosed as positive.

Based on this, the four values (i.e., TP, FP, FN, and TN) could be obtained in the three studies, and then five measures used to evaluate performances of the three studies could be calculated. Accuracy is the ratio of the number of data that are correctly clustered to the total number of data. Thus, accuracy is calculated by:

$$accuracy = \frac{TP+TN}{TP+FP+TN+FN} \text{ (Equation 3.6)}$$

Sensitivity is the proportion of subjects who test positive (AD/pMCI) among all those who actually are positive (AD/pMCI). It is calculated by:

$$sensitivity = \frac{TP}{TP+FN} \text{ (Equation 3.7)}$$

In contrast, specificity is the proportion of subjects who test negative (CN/sMCI) among all those who actually do not have AD/pMCI. It can be calculated by:

$$specificity = \frac{TN}{TN+FP} \text{ (Equation 3.8)}$$

The PPV is the probability that following a positive test, that subject will truly have AD/pMCI, and NPV is the probability that following a negative test, that subject will truly be CN/sMCI. They two can be calculated by:

$$PPV = \frac{TP}{TP+FP} \text{ (Equation 3.9)}$$

$$NPV = \frac{TN}{TN+FN} \text{ (Equation 3.10)}$$

In each run of the clustering algorithm, the five measures above were obtained and the final result was the mean value of each measure in ten runs. In addition, the 95% confidence interval (95% CI) of each measure was calculated and estimated as well.

3.9. Clinical Workflows in Practice

In practice, the workflow of the whole diagnosis process is shown as Figure 3.16: a new patient receives MRI scanning and a doctor obtained the patients MRI scan; the scan would have to be pre-processed by SPM12 and then it needs to be fed into the model presented in this thesis; finally, the doctor and the patient receive the results.

As for using the model, some pre-labelled scans (consisting of both CN and AD scans) should be prepared beforehand. The new scan and the pre-labelled scans are fed into the model and clustered into two groups. The results of the new scan can be identified based on the approach described in Sections 4.8, 5.8, and 6.8.
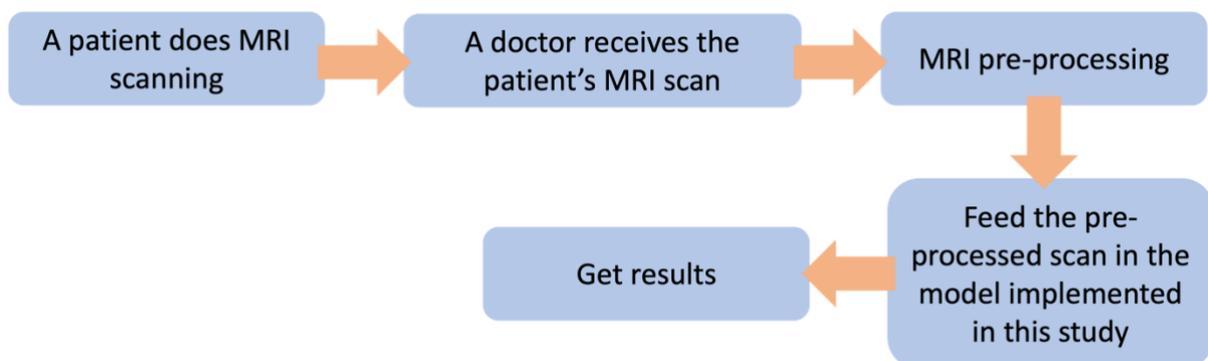
Figure 3.16 A workflow of the whole diagnosis process of a patient

3.10. Conclusion

In this chapter, the methodology of three clustering tasks: AD vs. CN, sMCI vs. pMCI, and CN vs. pMCI, was introduced. It includes the data source and data collection, the data pre-processing, the detection of ROIs, feature extraction, clustering, and performance evaluation. The data used in this study were 1.5T T1-weighted MRI scans collected from ADNI database; the performance of using SPM12 and FreeSurfer in pre-processing were compared and the better one became the pre-processing tool; a two-sample t-test was applied to detect ROIs between the two groups in each task and the ROIs were fed into a PCANet to extract features; the k-means++ algorithm was used as the clustering algorithm in each task; finally, the mean values and 95% CI of five measures – accuracy, sensitivity, specificity, PPV, and NPV, were used to evaluate the performance of the proposed method. The detailed information of the three tasks is presented in Chapters 4, 5, and 6, respectively.

In Chapter 4, detailed information of discriminating between AD and CN data, including explicit operations and use of tools of each step, the settings of the models and algorithms, the result estimation, and the discussion, are presented.

Appendix A

A.1 Research ethics approval letter

The University Of Sheffield.

Yuyang Liu
Registration number: 190198343
Information School
Programme: N/A

Dear Yuyang

PROJECT TITLE: Using machine learning techniques and brain MRI scans for detection of Alzheimer's dieseas
APPLICATION: Reference Number 034919

This letter confirms that you have signed a University Research Ethics Committee-approved self-declaration to confirm that your research will involve only existing research, clinical or other data that has been robustly anonymised. You have judged it to be unlikely that this project would cause offence to those who originally provided the data, should they become aware of it.

As such, on behalf of the University Research Ethics Committee, I can confirm that your project can go ahead on the basis of this self-declaration.

If during the course of the project you need to deviate significantly from the above-approved documentation please inform me since full ethical review may be required.

Yours sincerely

Claire Du Puget
Departmental Ethics Administrator

Script A.1

```
Path = "/Users/yuyangliu/Documents/ADNI_Data/sample";
Folder = dir(Path);
for k=4:length(Folder)
    FileName = Folder(k).name;

    set_env_cmd_1 = "export
FREESURFER_HOME=/Applications/freesurfer/7.1.1";
    set_env_cmd_2 = "source $FREESURFER_HOME/SetUpFreeSurfer.sh";
```

```matlab
    sub_dir = "export
SUBJECTS_DIR=/Users/yuyangliu/Documents/ADNI_Data/sample";
    pre_process_cmd = "recon-all -s "+FileName+num2str(k-3)+" -i
"+Path+"/"+FileName+" -autorecon1";
    full_cmd =
set_env_cmd_1+newline+set_env_cmd_2+newline+sub_dir+newline+pre_process_cmd
;
    status = system(full_cmd);
    %disp(full_cmd);
end
```

Chapter 4 Differentiating Between Alzheimer's Disease and Cognitively Normal Scans

4.1. Introduction

Section 1.3 (Aims and Objectives) presented that the research presented in this thesis includes three studies (clustering tasks): (1) discriminating between cognitively normal (CN) and Alzheimer's disease (AD) magnetic resonance imaging (MRI) scans; (2) discriminating between stable mild cognitive impairment (sMCI) and progressive cognitive impairment (pMCI) scans; (3) discriminating between CN and pMCI scans. Chapter 3 provided an overview of the methodology employed in this research. However, the three studies have differences when conducted. These three studies are presented in detail in three separate chapters.

In this chapter, the study discriminating between AD and CN scans is introduced and undertaken, and the following two chapters will present the work clustering sMCI and pMCI, and clustering CN and pMCI, respectively. This chapter provides the detailed information and operations of each phase in the workflow, including the software, tools, and programming languages used in this task (Section 4.2), how the data were collected from the Alzheimer's Disease Neuroimaging Initiative (ADNI) (Section 4.3), how the pre-processing tools were used (Section 4.4), how the two-sample t-test analysis was applied (Section 4.5), and how the PCANet (Section 4.6) and k-means++ were implemented in Matlab (Section 4.7), and how the performance of this task was evaluated (Section 4.8). In addition, the clustering results will be demonstrated and estimated in Section 4.9. Discussion of discriminating between AD and CN is presented in Section 4.10. Section 4.11 concludes this chapter.

4.2. Software, Tools, and Programming Languages

In this study, and as described in more detail in Section 3.4, two sets of software, i.e., FreeSurfer and Statistical Parametric Mapping 12 (SPM12), were employed to pre-process the magnetic resonance imaging (MRI) scans and compare their performance on pre-processing. As explained in Section 3.5, a two-sample t-test was applied to the data and the computational anatomy toolbox (CAT 12) of SPM12 was also used to undertake the two-sample t-test to detect differences of the scans between the two groups. From the feature extraction phase to the clustering phase, programming was necessary and the Matlab R2020b was chosen as the tool to program.

4.3. Data Collection

As described in Section 3.3, in this study, 1.5T T1-weighted structural MRI scans were collected as the study data from the ADNI database (ADNI-1). In addition, ADNI used the term "subject" to indicate the participant, and each subject (participant) had multiple MRI scans because the ADNI is a longitudinal study and each participant received an MRI scanning every six months or 12 months. However, this part of the study only required the baseline scan of each subject. In ADNI-1, baseline was named as "Screening". Therefore, 198 MRI scans of AD and 231 scans of CN were collected from "Screening" of ADNI-1, and for each subject (participant), only one baseline MRI image was used. A summary of the MRI scans used to cluster AD and CN is shown in the Table 4.1.

Table 4.1 The number of MRI scans of AD and CN

States	AD	CN	Total
Number of MRI scans	198	231	429

4.4. Data Pre-processing

As described in Section 3.4, two software programmes, FreeSurfer and SPM12, were employed to pre-process the structural MRI scan. In this phase of Study 1, 198 AD scans and 231 CN scans were pre-processed by both FreeSurfer and SPM12 (its Computational Anatomy Toolbox (CAT12)) and the performances of the two pre-processing tools were compared.

4.5. Detection of Difference Between the Two Groups

As described in Section 3.5, differences and abnormalities at the voxel level resulting from AD in the brain between AD and CN groups were detected using a two-sample t-test analysis after pre-processing. First, a basic model of the two-sample t-test was applied to two groups of MRI scans (AD and CN) by running CAT12 as Figure 4.1 shows. Then an SPM format file generated from the two-sample t-test was estimated by the CAT12. Finally, the result of the estimation of the two-sample t-test was displayed in the form of heat maps in different colours and the heat maps can be viewed by Xjview. The meaning of the different colours in heat maps in this study is explained in Section 4.9.2.2. In this step, the two-sample t-test was applied to 30 CN scans and 30 AD scans to detect AD-relevant regions because the t-test is applicable to small number (generally under 30) of samples, as explained in Section 3.5.1.
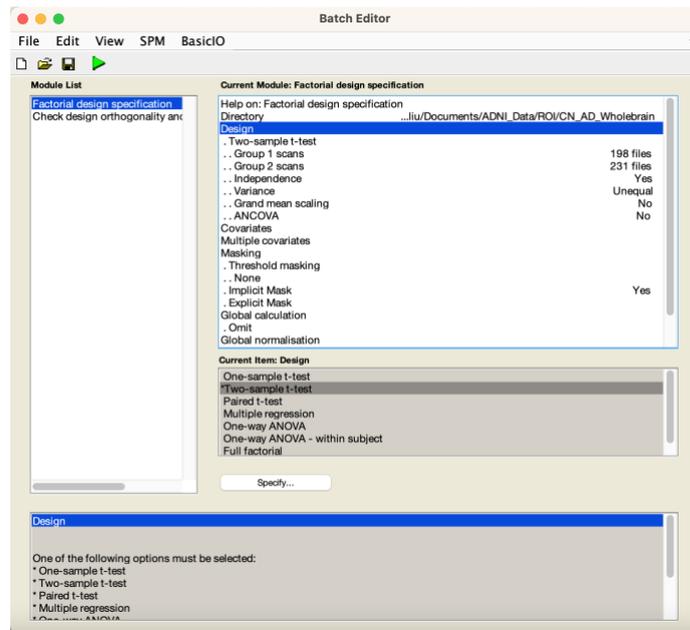
Figure 4.1 GUI of Basic Model (Screenshot by Yuyang Liu)

4.6. Feature Extraction

In Section 4.5, the ROIs between CN and AD groups were recognised and noted in a heat map, which is shown Figure 4.7 in Section 4.9.2.2. As described in Section 3.6, a PCANet was used to extract features from the detected ROIs. To demonstrate whether the ROIs and the PCANet were effective to well separate the two groups well, the distributions of the features in four scenarios were compared:

(1) features (voxel intensity values) directly from random selected slices (without PCANet);

(2) features (voxel intensity values) from multiple whole slices (as shown in Figure 4.2) containing the ROIs (without PCANet);

(3) features (voxel intensity values) directly from the ROIs (without PCANet);

(4) features from the ROIs and computed by PCANet (higher-level features) in the method reported here.
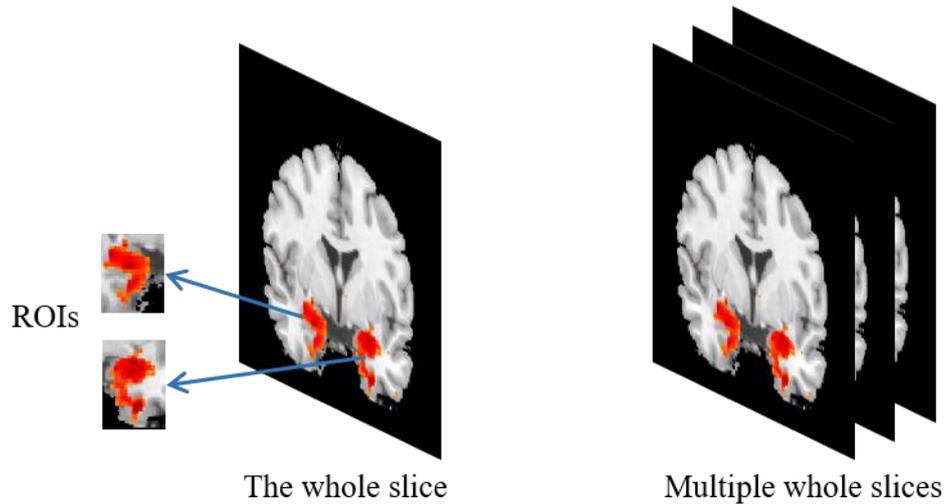
Figure 4.2 an example of a single whole slice and multiple whole slices in an MRI scan and the ROIs in them (MRI images from ADNI adapted by Yuyang Liu)

The distributions were demonstrated using the t-distributed stochastic neighbour embedding (t-SNE) technique. It is a statistical method for visualising high-dimensional data.

To use the PCANet in this study, the parameters of it needed to be set beforehand. The number of stages was set to two because two-stage PCANet is sufficient to achieve a good performance (Chan et al., 2015). Based on multiple trials and the default values provided by Chan et al. (2015), other parameters were properly set as follows:

- The size of the patches $k_1 \times k_2$ was set to 3×3 in both two stages;
- The number of filters L_1 and L_2 were both set to eight, respectively, which indicated that there were eight principal eigenvectors in each PCA process of the two stages;
- The block size for the histograms was set to 15×15;

The overlap ratio of the local blocks was set to zero in this study for reducing the number of features.

4.7. Clustering

As mentioned in Section 3.7, data of CN and AD were clustered by using a k-means++ algorithm. In the algorithm of k-means++, parameters need to be set. First, the number of clusters, k, was set to 2 in this study, because the data were to be clustered into two groups (CN and AD). In addition to the k value, the selection of distance metrics is importance in the k-means++ as well. The basic Euclidean distance was selected. The Euclidean distance is a simple and common dissimilarity measure. The Euclidean distance, $d(x, y)$, between two data points, x and y, is calculated by:

$$d(x, y) = \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2} \text{ (Equation 4.1)},$$

where n is the number of dimensions, and x_k and y_k are the kth features of x and y, respectively. In addition, cosine similarity had a better performance in Study 2, thus it was also used in this study. However, it achieved an almost the same performance as using the Euclidean distance in this study, thus the result of using cosine similarity is not reported here.

In order to demonstrate the effectiveness of PCANet and the inclusion of two ROIs (two obvious ROIs were detected in Section 4.5, which are shown in Figure 4.7), in this study, two clustering experiments were conducted:

(1) Comparison of clustering performances between using one ROI and using two ROIs with PCANet;

(2) Comparison of clustering performances between two ROIs with PCANet and two ROIs without PCANet.

In Matlab 2021b, the function "*kmeans()*" uses the k-means++ algorithm directly by default. Therefore, the function *kmeans(X,2)* was used to partition the observation of the data matrix (*X*) into two clusters using the Euclidean distance, where the *X* was the matrix of the features extracted from the PCANet. In addition, according to Section 3.7, the k-means++ algorithm was repeated ten times due to its different results in different runs.

4.8. Performance Evaluation

As mentioned in Section 3.8.1, after clustering, the data were divided into two clusters, but because of the nature of the unsupervised learning method, where labels are not typically used to cluster the data, it cannot recognise which cluster is AD data and which is CN data directly. The way to refer to AD and CN data is as follows: if, in a cluster labelled as 1, more than 50% of the data are AD data, the cluster with label 1 is regarded as AD cluster, which means that the data in this group (cluster) are expected to be AD; the other cluster (labelled as 2) is CN cluster and the data in the CN cluster are expected to be CN; and *vice versa*. For example, if in a cluster with label 1, 70% of the data are AD and 30% of the data are CN, this cluster is regarded as AD cluster and therefore those 30% of the data (CN) are thought to be mis-clustered data; the other cluster (with label 2) is regarded as CN cluster and the mis-clustered data can be known. In this way, true positive (TP), false positive (FP), true negative (TF), and false negative (FN) were obtained.

After recognising the two clusters, as described in Section 3.8, mean values of five measures: the accuracy, sensitivity, specificity, positive predictive value (PPV), and the negative predictive value (NPV), were calculated and were used to evaluate the overall performance of the method presented in this chapter. 95% confidence intervals (95% CI) of the five measures were also used to describe the stability and the variance of the performance of the method.

4.9. Results

4.9.1. Pre-processing Results

4.9.1.1. Pre-processing Duration

In the pre-processing phase, FreeSurfer and SPM12 were both employed to pre-process the MRI scans. The overall pre-processing time for a single MRI scan using the FreeSurfer was longer (35-40 minutes) than SPM12 (25-35 minutes).

4.9.1.2. Pre-processing Quality

To compare the pre-processing quality of the FreeSurfer and SPM12, two MRI scans were randomly selected as examples: the scans from subjects (participants) 033_S_0888 and 057_S_1379. The standard of comparison was whether the images were well registered in a common space (template) and whether the skull was stripped properly. The pre-processed MRI scans from the two subjects are shown in Figures 4.3 and 4.4. Figure 4.3 (a) and (b) are the brain MRI images from subjects (patients) 033_S_0888 and 057_S_1379 respectively pre-processed by SPM12, and Figure 4.4 (a) and (b) are MRI images from the two subjects after being pre-processed by FreeSurfer. From the point of view of the horizontal comparison, the image pre-processed by FreeSufer has better clarity. However, when comparing Figure 4.4 (a) and (b), it can be discovered that the two axial planes are totally different when they are not shown in a common coordinate system, which means that the image registration was not good.

(a)   (b)
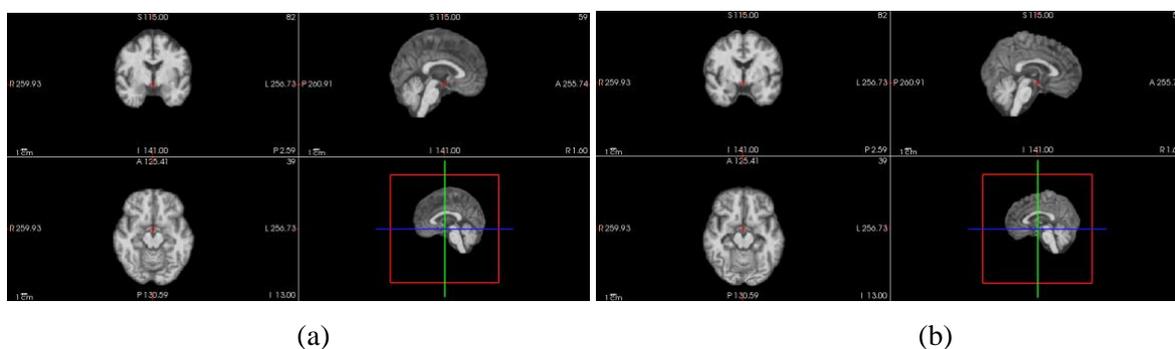
Figure 4.3 MRI scans of two patients randomly selected (033_S_0888 (a) and 057_S_1379 (b)) pre-processed by SPM (Original images from ADNI adapted by Yuyang Liu)

(a)   (b)
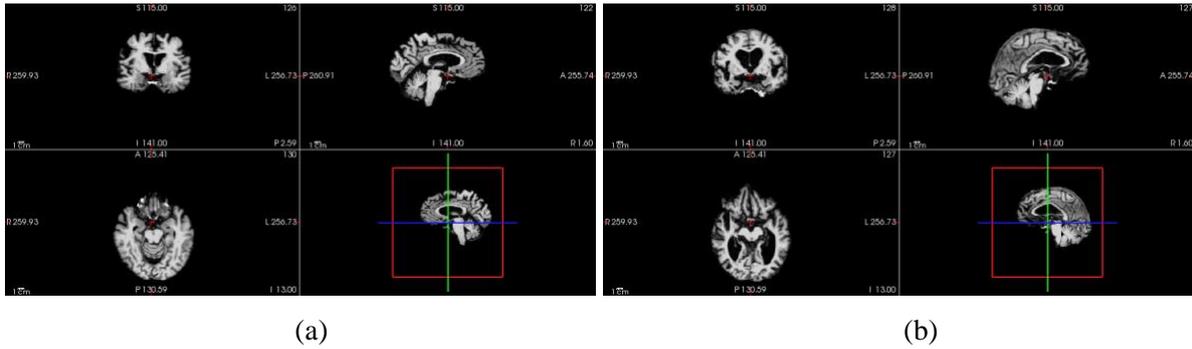
Figure 4.4 MRI scans of two patients randomly selected (033_S_0888 (a) and 057_S_1379 (b))
pre-processed by FreeSurfer (Original images from ADNI adapted by Yuyang Liu)

As for the performance of the skull stripping, SPM preserved the brain tissues to the maximum extent possible, but as can be seen in Figure 4.4 (a) and (b), images pre-processed by FreeSurfer might lose some visual information close to the edge of the brain, which means that using FreeSurfer might lead to over-stripping (Laboratory for Computational Neuroimaging at the Athinoula A. Martinos Center for Biomedical Imaging, n.d.-b), or it may result in some remains of the brain skull. In addition to these two subjects, skull over-stripping occurred in the scans of other subjects when using FreeSurfer.

In summary, considering the two factors stated above, the overall performance of SPM is better than FreeSurfer in general, because this research involves identifying visible distinguishing characteristics in brain structures but the visible impact on the images in FreeSurfer is more than SPM. Furthermore, FreeSurfer has no GUI, which is not user-friendly, thus SPM was chosen as the pre-processing software in this study, including the other two tasks (sMCI vs. pMCI; CN vs. pMCI).

After the MRI scans were pre-processed using the SPM12, the dimension of each scan was 121×145×121 voxels. In other words, there were 121 slices with size 145×121 voxels from the sagittal view, 145 slices with size 121×121 voxels from the coronal view, and 121 slices with size 145×121 voxels from the axial view. As a result, in total, each MRI scan contained 387 (145+121+121) slices from all three view planes.

4.9.1.3.   White Matter Hyperintensity

In pre-processing step, SPM reported whether each MRI scan has distinguishable areas named white matter hyperintensities (WMHs). A hyperintensity is an area of high intensity on the MRI scans of the human brains that reflect lesions, and WMHs in the brain are the consequence of cerebral small vessel disease (Prins & Scheltens, 2015). According to the Debette et al.'s (2010)

study, WMHs are associated with an growing risk of stroke, dementia, and death. In addition, they reported an association of WMHs with a fast decline in global cognitive performance, executive function, and processing speed (Debette & Markus, 2010). In the pre-processing phase of this study, as shown in Table 4.2, there were 83 participants (subjects) detected by SPM to have WMHs in their brains, of which the number of CN people was 25 and the number of AD patients was 58. This means that, in this study, if WMHs were discovered in a participant's brain, the participant would have a chance of 70% (58 out of 83) of having AD within this dataset, which highlights the potential of presence of WMHs as a possible indicator of AD.

Table 4.2 The number of subjects (participants) whose brains have been tested having WMHs

Pathological Stages	Number of Subjects (patients)
CN	25
AD	58
Total	**83**

4.9.2. Detection of Differences Between AD and CN

4.9.2.1.   Normality of the Distribution

As described in Section 3.5, the normality of the distribution of the dataset (30 AD scans and 30 CN scans) for using the two-sample t-test in this study was estimated before a two-sample t-test was applied. Figure 4.5 shows a histogram of the intensity value of a randomly-selected voxel in AD MRI scans (within the brain area); Figure 4.6 shows a histogram of the intensity value of the voxel in CN scans whose location was the same as that of the voxel selected for Figure 4.5.

From Figure 4.5, it can be seen that the pattern was approximated to a nearly-normal distribution. However, from Figure 4.6, it seems hard to judge whether the histogram is a normal-like distribution or not. However, it was shown to follow a normal distribution by using a Kolmogorov-Smirnov test. As a result, according to the Srivastava's (1958) proof mentioned in Section 3.5, a two-sample t-test could be applied to the AD and CN MRI scans in the study.
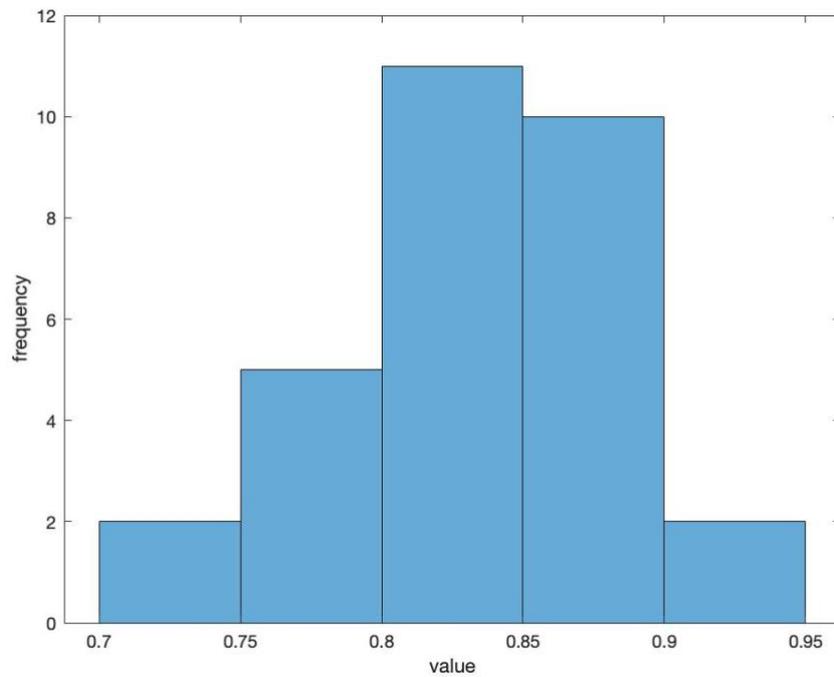
Figure 4.5 Histogram of voxels in AD MRI scans. X-axis shows the normalised intensity value of the voxel selected from 30 AD MRI scans; y-axis shows the number of occurrence of the corresponding intensity value. (Developed by Yuyang Liu)
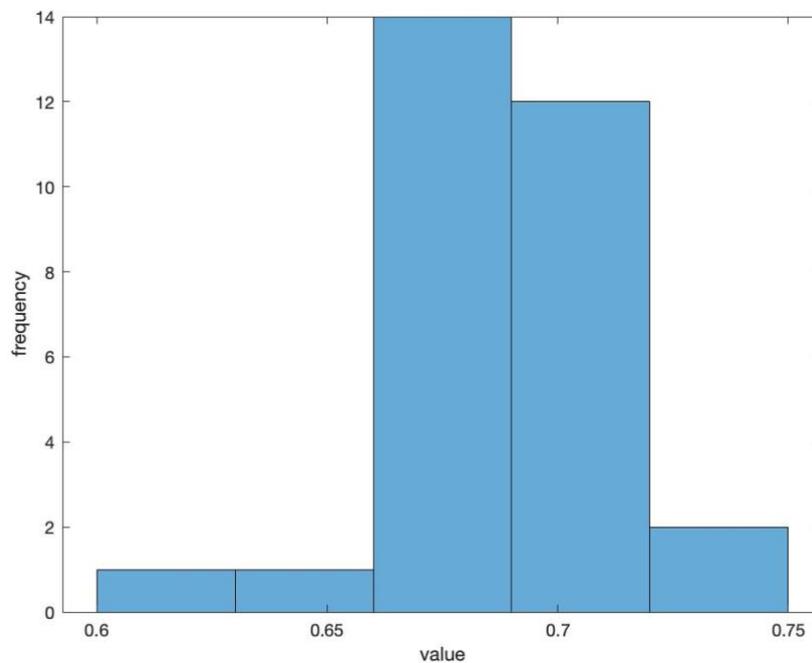
Figure 4.6 Histogram of voxels in CN MRI scans. X-axis shows the normalised intensity value of the voxel selected from 30 CN MRI scans; y-axis shows the number of occurrences of the corresponding intensity value. (Developed by Yuyang Liu)

Figure 4.7 shows the AD-relevant regions at voxel level between the 30 AD scans and 30 CN scans, which have been identified as the ROIs, based on the two-sample t-test analysis. Figure 4.7 (a) – (c) shows the sagittal plane, coronal plane, and axial plane, respectively, where the ROIs were noted in a heat map. The red and orange regions indicate that the mean values of the intensity values (after being normalised) of the voxels in the AD group are greater than that in the CN group, and the blue area indicates where the mean values of the voxels in the AD group are less than the CN group.

As can be seen from the three view planes in Figure 4.7, the most conspicuous regions are the lower parts of the limbic system, which include the parahippocampal gyrus, amygdala, and hippocampus, in both the left and right cerebra. In addition, there are some fragmentary regions including small parts of the thalamus and frontal lobe (including some tiny parts of the frontal gyrus). In this part of the study, the two most obvious regions (in orange and red) in the lower parts of the limbic system were used for the discrimination between the CN and AD scans, because other regions are so small and diffuse that they might lack generality.

| (a) sagittal plane | (b) coronal plane | (c) axial plane |
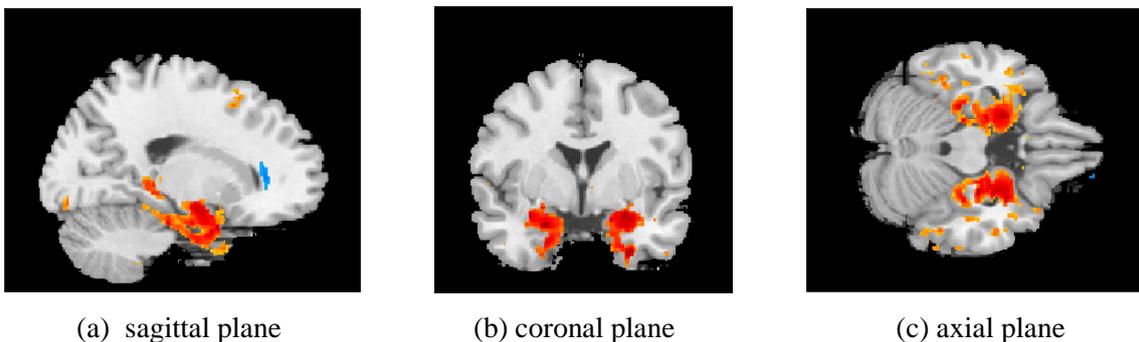
Figure 4.7 A heat map of the ROIs noted in the three view planes of the MRI scan. (a) is the sagittal plane; (b) is the coronal plane; (c) is the axial plane. (MRI images from Xjview adapted by Yuyang Liu)

4.9.3. Distributions of Features

As mentioned in Section 4.6, to demonstrate whether the ROIs and the PCANet were effective in separating the two groups, the distributions of the features in four scenarios were compared using distribution maps:

   (1) features (voxel intensity values) directly from random selected slices (without PCANet);

   (2) features (voxel intensity values) from the whole slices containing the ROIs (without PCANet);

(3) features (voxel intensity values) directly from the ROIs (without PCANet);

(4) features computed by the PCANet (higher-level features) using the voxel intensity values of the ROIs.

(a)

(b)

(c)

(d)
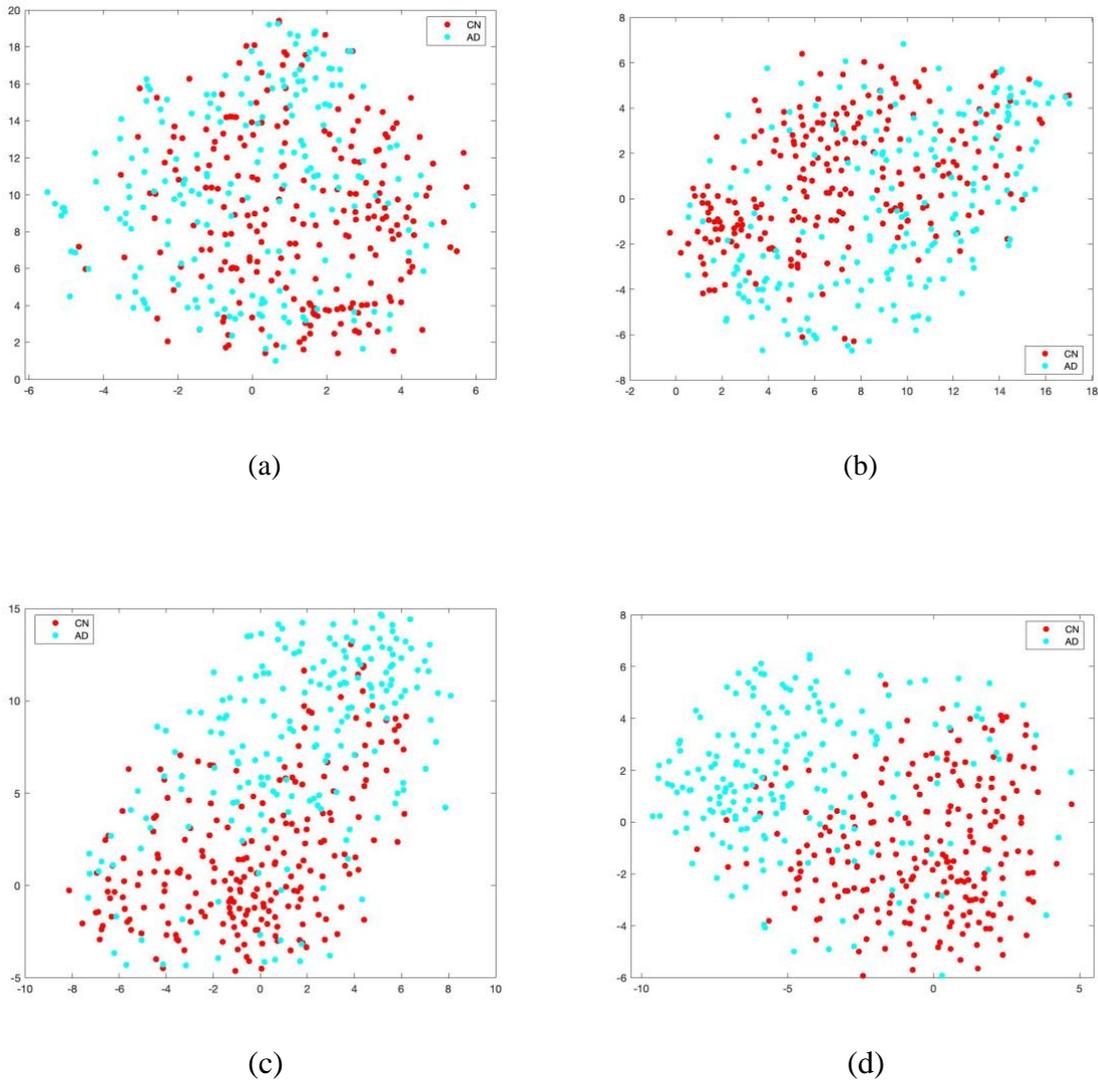
Figure 4.8 Distributions of the features from four scenarios. (a) illustrates the distribution of features (voxel intensity values) from randomly selected slices; (b) is the distribution of features (voxel intensity values) from the slices including ROIs; (c) is the distribution of features (voxel intensity values) directly from the ROIs; (d) is the distribution of the features computed by PCANet using the voxel intensity values of the ROIs. (Developed by Yuyang Liu)

Figure 4.8 shows the distributions of the features from the four scenarios using the t-SNE technique. However, the x- and y-axis in the t-SNE distributions are not easily interpretable and are different in different runs. Figure 4.8 (a) shows the distribution of the features from the randomly-selected slices of the MRI scans. It can be seen that the data were distributed irregularly and the two classes were completely visually inseparable and were therefore hard

to split. Figure 4.8 (b) shows the distribution of the features from the whole slices that contain the ROIs. It can be seen that, although the two classes are still mixed together, most AD points gather towards the lower right part of the graph and most CN points gather towards the upper left area of the graph. This means that the ROIs do help the two classes of data to split.

Figure 4.8 (c) presents the distribution of the features directly from the ROIs without PCANet. It can be seen that the data further tend to separate compared with Figure 4.8 (a) and (b). This means that the ROIs selected by the two-sample t-test were truly AD-relevant although there was still room for improvement. When examining the distribution of features computed by a PCANet using the ROIs (Figure 4.8 (d)), it can be seen that, although the two groups are still adjacent to each other, fewer deviants are mixed in the wrong group, compared with Figure 4.8 (c), and there is improved separation of the groups. The results in Table 4.3 (in Section 4.9.4) and Figure 4.8 (a) – (d) suggest a basic that the ROIs detected by the two-sample t-test are truly related to AD and that the PCANet truly helps the ROIs map to higher-level features that can be clustered accurately.

4.9.4. Clustering Results

As described in Section 3.7, k-means++ was employed to cluster the CN and AD scans into two groups (clusters) using the features extracted from the PCANet. As mentioned in Section 3.8, five measures: the accuracy, sensitivity, specificity, PPV, and NPV, were calculated based on the equations in Section 3.8.

As presented in Section 3.7, the k-means++ algorithm was run 10 times and all the runs are listed in the Appendix B (Tables A.1 to B.4). Table 4.3 and Figure 4.9 shows the mean values and 95% CI of each measure over the 10 runs using one ROI and two ROIs with using the PCANet.

Table 4.3 Measures of Clustering using one ROI and two ROIs using PCANet

Measures		K-means++	
		One ROI	**Two ROIs**
Accuracy	**Mean**	0.773	0.842
	95% CI	(0.764, 0.782)	(0.833, 0.851)
Sensitivity	**Mean**	0.695	0.797
	95% CI	(0.640, 0.750)	(0.768, 0.826)
Specificity	**Mean**	0.840	0.881

	95% CI	(0.792, 0.888)	(0.853, 0.909)
PPV	**Mean**	0.798	0.855
	95% CI	(0.764, 0.832)	(0.829, 0.881)
NPV	**Mean**	0.767	0.836
	95% CI	(0.747, 0.787)	(0.819, 0.853)
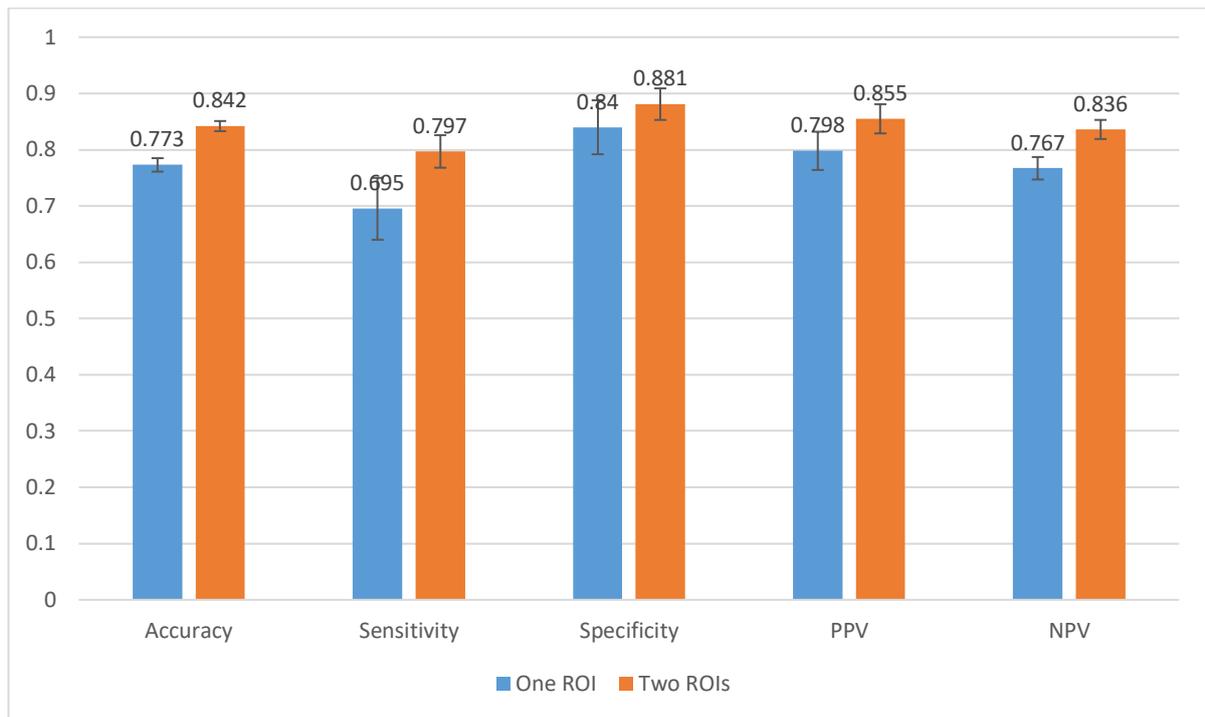
Figure 4.9 A bar graph indicating the mean values and 95% CI of each measure across the 10 runs using one ROI and two ROIs (using PCANet) *(Developed by Yuyang Liu)*

In Section 4.9.2.2, two (obvious) AD-relevant regions were detected and were therefore used in this part of the study. Therefore, the performance in discriminating between CN and AD using one of the two regions (the ROI on the left in Figure 4.7 (b)) and both regions (two ROIs) were tested respectively. As can be seen from Table 4.3, for all five measures, two ROIs performed better than when using one ROI: the accuracy for two ROIs was 0.06 higher than using one ROI and the sensitivity of using one ROI and two ROIs shows a huge gap: using two ROI achieved more than 0.1 sensitivity (i.e., 10%) greater than using one ROI. Sensitivity indicates the ability of a test to correctly identify people with AD, thus the result indicates that using two ROIs approach is more successful at recognising people with AD. The respective 95% CI (of using one ROI and two ROIs) indicates that these differences are statistically significant, i.e., the mean sensitivity for one method is outside the 95% CI for the other hand.

From another point of view, the additional ROI is meaningful to AD, or namely the two ROIs indicate two diseased areas associated with AD. Additionally, the wider 95% CI of the sensitivity when using one ROI (0.640, 0.750) indicates a more variable performance. The specificities for one (0.840) and two (0.881) ROIs indicated that they both performed well in identifying non-AD cases (CN cases in this part of the study), although the specificity for one ROI was significantly lower than that for two ROIs.

When using two ROIs, a PPV of 0.855 was achieved, which is 0.06 greater than that of using one ROI (0.798). The PPV indicates the probability that a person who is identified as having AD (i.e., a positive prediction) actually has the disease. As a result, the significantly higher PPV when using two ROIs (0.855; 95% CI=0.829-0.881) versus one ROI (0.798; 95% CI=0.764-0.832) indicates that results using two ROIs could be trusted more that a person identified as having AD actually has the disease, i.e., its diagnosing AD is fairly credible. This has important implications for clinicians in treating people diagnosed with the condition. Nevertheless, the NPV of using two ROIs (0.836) is more than 0.07 greater than that of using only one ROI (0.767). In contrast to PPV, NPV is the probability that a person who receives a negative test result actually does not have the disease. As a consequence, the significantly higher NPV for two ROIs (0.836; 95% CI=0.819-0.853) gives greater confidence that someone who receives a negative test result actually does not have the disease. This means that clinicians are less likely to miss an actual case when using two ROIs compared to using a single ROI.

Furthermore, to estimate whether the PCANet is effective in this study, the results of the clustering using two ROIs (because two ROIs have been shown to have a better performance than one ROI above) with and without PCANet were compared. The results of these analyses are presented in Table 4.4 and Figure 4.10. It can be concluded that, when using PCANet, all mean values of the five measures were much greater than when not using it, which highlights the effective contribution of PCANet to the clustering process between the CN and AD scans. The measures and figures also demonstrate that the upper and lower bounds of 95% CI of the five measures were the same, which indicates that the performance of clustering without PCANet was relatively more stable than using PCANet. Although the performance was stable when the PCANet was not used, the upper bounds of the 95% CI of all five measures were all lower than the lower bounds of those when the PCANet was used. It suggests that, although the results fluctuated when using the PCANet, the worst results were still better than the results yielded by not using the PCANet.

Table 4.4 Measures of Clustering with PCANet and without PCANet using two ROIs

Measures		K-means++	
		Without PCANet	**With PCANet**
Accuracy	**Mean**	0.709	0.842
	95% CI	(0.709, 0.709)	(0.833, 0.851)
Sensitivity	**Mean**	0.697	0.797
	95% CI	(0.697, 0.697)	(0.768, 0.826)
Specificity	**Mean**	0.719	0.881
	95% CI	(0.719, 0.719)	(0.853, 0.909)
PPV	**Mean**	0.680	0.855
	95% CI	(0.680, 0.680)	(0.829, 0.881)
NPV	**Mean**	0.735	0.836
	95% CI	(0.735, 0.735)	(0.819, 0.853)
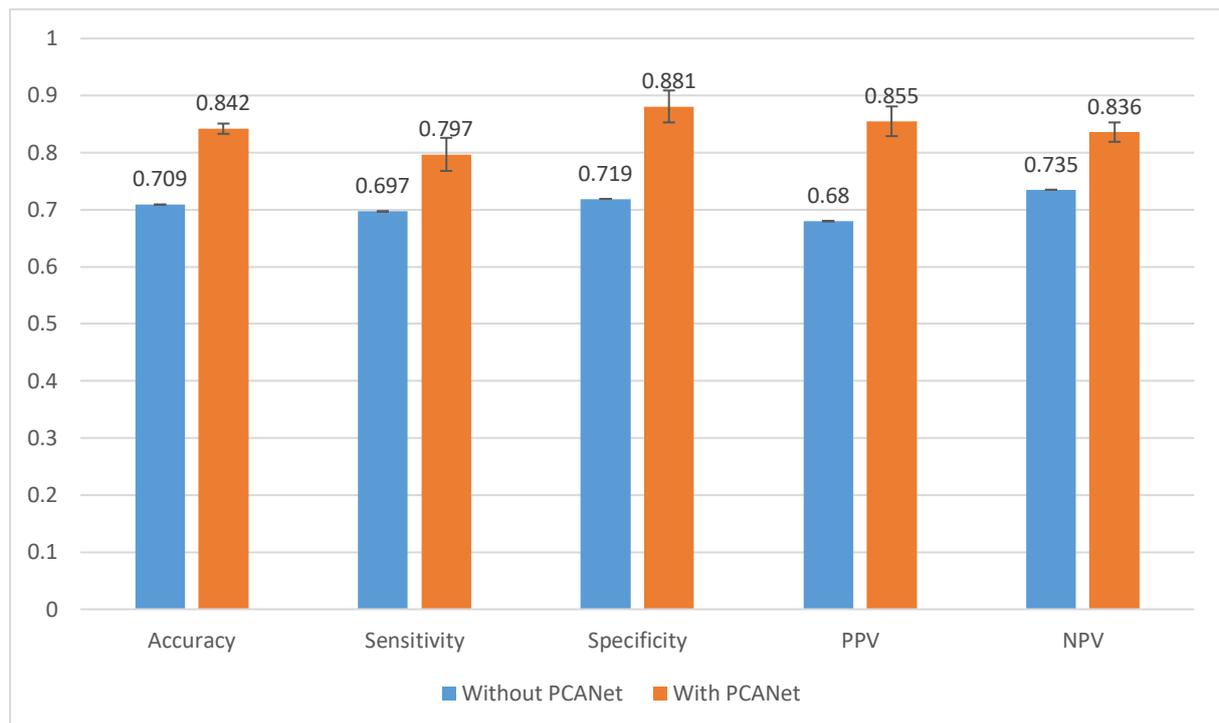
Figure 4.10 A bar chart indicating the mean values and 95% CI of each measure across the 10 runs not using PCANet and not using PCANet (using two ROIs) (Developed by Yuyang Liu)

4.9.5. Wrongly-clustered Cases

In order to understand better why some cases were mistakenly discriminated into AD at baseline, when they were actually CN, how these patients progressed was studied, as

summarised in Table 4.5. Table B.5 in Appendix B lists all the participants who were CN at baseline and then developed to MCI or developed to AD in later phases, and those who were mistakenly clustered into AD are noted in red. It can be observed that 15 participants progressed to MCI or developed to AD from CN (excluding those who developed to MCI and returned to CN later) before the 36th month and 6 out of the 15 participants were clustered into the AD cluster. If the participants who developed before the 48th month were looked at, 8 out of 25 were clustered into AD cluster; if those who developed before the 60th month were looked at, 8 out of 31 were clustered into AD; 37 participants developed to MCI or AD before the 72nd month and 9 of them were clustered into AD; if the conversion time was prolonged to 108th month, 12 out of 52 participants were clustered into AD. The records for the 84th month and 94th month are not displayed in Table 4.5 because they are the same as the record for 72nd month; the record for 120th month is the same as 108th month thus it is not shown in the table as well.

Table 4.5 A summary of the scans that are CN but were mis-clustered (using two ROIs with PCANet) into AD. The first column shows the month when the participants developed to MCI or AD from CN; the second column indicates the number of patients who developed to MCI or AD from CN in corresponding months; the third column is the number of participants who are CN at baseline but were mistakenly clustered into AD group by the method used here; the fourth column indicates the ratio of the third column to the second column.

MCI/AD detected in (months)	No. of patients	Misclassified as AD (No. of patients)	Misclassified as AD (%)
36 months	15	6	40
48 months	25	8	32
60 months	31	8	25
72 months	37	9	24
108 months	52	12	23

From Table 4.5, it can be conjectured that the method reported in this study can potentially predict the progression of the participants. The reason is that the unsupervised learning methods are based on the features themselves instead of feedback from the label of the data (which occurs using supervised learning methods). Therefore, even though some participants were diagnosed as CN at baseline, if their brains more or less showed pathological changes related to AD, naturally they might possibly be clustered into AD based on the changes.

However, the table also highlights that the prediction decreases with the time of conversion delaying. It is probably because the proposed method only used the baseline scans but if the time of conversion (to AD) is long after the baseline, the structure of the brain at the conversion time is more likely to have dramatic changes to that at the baseline, which means that the scan of the baseline still shows a totally healthy status even if it develops into AD after a long time.

In addition, from Table B.5 in Appendix B, it can be seen that seven participants only eventually developed to MCI instead of to AD but they were still clustered into AD. As well as the mistaken clustering by the proposed methods, it is also likely that the participants developed to MCI and their brain had some AD-related changes but still showed symptoms of MCI. Moreover, these participants, such as 020_S_0899, 127_S_0622, and 130_S_1200, have no further record after they developed to MCI (this may be because of their end of life or they were awaiting further scans). Although the findings highlight potential applications of the method in understanding the progression of AD, there is a need to further study this in more detail.

4.9.6. Runtime of the Proposed Method

In addition to the performance of diagnosis, the efficiency of the proposed method was also reported. To show the advantage in term of efficiency of the method reported in this chapter (using two ROIs with PCANet), the whole slices of the MRI scans were used to compare (with the ROIs detected in Section 4.7.2.2). Due to the limited memory space (a limitation deposit the use of a high performance computer), 77 out of 387 slices (approximate 1/5 of all the slices) were selected in each MRI scan. The measures used for comparison are maximum memory usage in the running process, the final memory usage (size of the final features processed by PCANet), and the run time. Specifically, only run time of the processing by PCANet was calculated because it occupied the most time in the whole workflow.

The two methods were run in a same high performance computing environment. The specification of the high performance computer is as follows:

- Machine: Dell PowerEdge C6420
- CPUs: 2 × Intel Xeon Gold 6138
- RAM: 192 GB (i.e. 4.8 GiB / core)
- Local storage: 1 TiB SATA III HDD

The results of the comparison are shown in Table 4.6. It can be seen that, even though only 1/5 of all the slices in each MRI scan were used, the maximum memory usage was approximately

100G. In other words, it is estimated that a very high amount of memory (i.e., around 500G) would be required if all 387 slices of each scan were fed into the PCANet. Similarly, if the whole scan (387 slices) was to be used, the runtime of it would likely be longer than that of our method because more slices need to be processed by PCANet. As can be seen from Table 4.6, the runtime for 77 slices is already higher than the method in this chapter and any increase in the slices would likely be significantly higher.

Table 4.6 Efficiency of the two methods

	Maximum memory usage	Final memory usage	Runtime
77 whole slices (1/5 whole scan)	100G	20G	7 hours
Method in this chapter	7G	0.480G (480M)	6.5 hours

4.10. Discussion

4.10.1. Comparison with State-of-the-art Methods

Ideally, the methods used in this study (i.e., in Studies 1 to 3) should be run on the same dataset when compared with previous studies. However, some images from the same database that did not meet the requirement of the study or did not meet the quality standards may have been removed in these studies. In addition, due to the policy of the ADNI, the data cannot be directly shared. A consequence of this is that, even if many studies collected data from ADNI-1, they used different datasets from this study. Therefore, it is not feasible to compare the performance with previous studies using exactly the same datasets.

Compared with some state-of-the-art methods that detected regions pathologically related to AD, the method reported here located similar regions to theirs. Zhang et al. (2022) noted the left side of hippocampus from slice 72 to slice 74 of the brain MRI scan as the parts with obvious changes in their study. Nigri et al. (2020) highlighted the region within the left hippocampus and ventricles in form of a heat map using Occlusion Test from their proposed neural network, AlexNet 2D plus Channel. It can be seen that their highlighted regions are similar to the red region in Figure 4.7 (a). Furthermore, Nigri et al. (2020) also used another method, The Swap Test, to interpret their network: the Swap Test highlighted the hippocampus in the coronal plane, which is also similar to the ROI in Figure 4.7 (b).

In addition to the detection of AD-regions, the performance in clustering using the method reported here compared favourably with other state-of-the-art unsupervised learning methods (shown in Table 4.7). Bi et al. (2019) used similar methods, PCANet and k-means, to cluster CN and AD data,

and achieved an accuracy of 0.8915. Although the accuracy in their study was 0.05 greater, the accuracy alone is not sufficient to comprehensively evaluate the performance of a method, because it only measures its overall capability to discriminate between AD and CN. However, it cannot evaluate the performance in correctly clustering CN data or AD data, respectively. For example, if the number of the data that are correctly clustered into CN is extremely large but the number of the data correctly clustered into AD is small, the overall accuracy still can be high, especially when the two classes of data are unbalanced, i.e., the sensitivity, specificity, PPV, and NPV, can offer a more comprehensive insight into how different methods compare. As presented in Table 4.3, this study provided a more rounded evaluation of the performance of a method.

Farouk et al. (2020) used the k-means method alone based on the selected ROIs and the whole brain, respectively. They obtained the best performance when using the whole brain, where the accuracy was 0.763. However, Farouk et al. (2020) also used accuracy as the only measure in their study, and the aforementioned measures are needed to comprehensively evaluate the results. Escudero et al. (2011) also utilised k-means to cluster CN and AD base on the normalised left and right hippocampus volumes from MRI scans, and 69% (0.69) of AD cases were clustered correctly. The method reported here achieved a sensitivity of 0.797, namely, 79.7% (0.797) of AD data were clustered into the correct cluster, which is more than 0.1 greater than Escudero et al.'s (2011) method. In addition, Escudero et al. (2011) did not provide the reason why they chose left and right hippocampus; however, the two-sample t-test was used to show that the left and right hippocampus are important regions related to AD. Cabreza et al. (Cabreza et al., 2022) applied a generative adversarial network (GAN) to MRI scans and achieved an accuracy of 0.7444, a sensitivity of 0.7386, and a PPV of 0.8228, all of which were lower than our method. The results from these studies are summarised in Table 4.7

Table 4.7 The studies of CN/AD clustering (unsupervised learning methods) based on structural MRI

Study	ACC	SEN	SPE	PPV	NPV	Dataset	Method
Method in this chapter	**0.842**	**0.797**	**0.881**	**0.855**	**0.836**	**231 CN, 198 AD**	**Two-sample t-test, PCANet, k-means++**
(Farouk & Rady, 2020)	0.763	-	-	-	-	113 CN, 162 AD	K-means
(Bi et al., 2019)	0.8915	-	-	-	-	307 CN, 243 AD	PCANet, k-means

(Escudero et al., 2011)	-	0.69	0.94	-	-	200 CN, 200 AD	K-means
(Cabreza et al., 2022)	0.7744	0.7486	-	0.8228	-	Not reported	GAN

In addition to the studies that employed unsupervised learning methods listed in Table 4.7, the performance of the method reported in this chapter is also better than some state-of-the-art supervised learning methods. These studies are listed in Table 4.8. The method reported in this chapter achieved a higher accuracy and also provided other measures (SEN, SPE, PPV, and NPV) compared with some of them, which means that the results in this way offers a more comprehensive picture of the performance. As Table 4.8 also demonstrates, the performance of the method reported here was lower than some studies that have used supervised learning (e.g., J. S. Kim et al., 2022; Kumari et al., 2022; Zeng et al., 2021). It is likely because of the impact of training datasets on the models (as labelled data are used in supervised learning methods) or additional contextual data contribute to improved performances (e.g., Zeng et al. (2021) included the mini mental state examination (MMSE) scores and the AD assessment scale). However, this requires a considerable amount of labelled MRI scans and MMSE scores, which is time-consuming and requires more clinical input or training.

As for the semi-supervised learning methods in Table 4.8, Saravanakumar et al. (2022), S. Liu et al. (2014), and D. Zhang & Shen (2011) all achieved good performances, where the accuracies were above 0.85. However, Saravanakumar et al. (2022) only reported the maximum accuracy (0.97) and it is hard to evaluate the performance of their method only based on the maximum accuracy because the minimum accuracy might be low and their other measures (e.g., sensitivity, specificity, etc.) might be unbalanced. S. Liu et al. (2014) and D. Zhang & Shen (2011) achieved high and balanced sensitivities and specificities. It is likely because they chose the most AD-related regions (ROIs) and the pseudo labels generated in the process of the semi-supervised learning methods were proper. However, the sizes of their datasets were both small, thus the evaluation to performance might not be convincible.

Both of the supervised and semi-supervised learning methods require labelled data to train, especially supervised learning, which needs a large amount of labelled data. The method reported in this study only required a limited volume of labelled data to detect AD-related regions and did not require training. Therefore, a trade-off can be seen from Table 4.8, in that

the simplicity (limited reliance on the labelled data) of the unsupervised learning methods is sometimes at the sacrifice of minor performance. However, the method reported in this study obtained a good balance between the performance and the simplicity of the method. In other words, the unsupervised learning method in this chapter achieved a relatively advantageous performance.

Table 4.8 The studies of CN/AD classification (semi-supervised and supervised learning) based on structural MRI

Learning type	Study	ACC	SEN	SPE	PPV	NPV	Dataset	Method
Unsupervised learning	**Method in this chapter**	**0.842**	**0.797**	**0.881**	**0.855**	**0.836**	**231 CN, 198 AD**	**Two-sample t-test, PCANet, k-means++**
Supervised learning	(Kumari et al., 2022)	0.969	0.73	0.88	0.767	0.82	36 CN, 38 AD	HPT-TSVM
	(Hengxin Li, Shi, Zhu, Wang, & Zhang, 2022)	0.844	0.836	0.859	-	-	226 CN, 186 AD	FSNet
	(J. S. Kim et al., 2022)	0.871	0.933	0.855	-	-	61 CN, 37 AD	DBAD CNN based model
	(Zeng et al., 2021)	0.986	-	-	-	-	92 CN, 92 AD	DBN-based multi-task learning
	(W. Lin et al., 2018)	0.888	-	-	-	-	229 CN, 188 AD	CNN, PCA, extreme learning machine
	(Aderghal et al., 2017)	0.837	0.792	0.872	-	-	228 CN, 188 AD	CNN
	(S. Korolev et al., 2017)	0.80	-	-	-	-	61 CN, 50 AD	3D-CNN
	(Valliani & Soni, 2017)	0.813	-	-	-	-	233 CN, 188 AD	ResNet
	(Zeng et al., 2018)	0.813	-	-	-	-	92 CN, 92 AD	SDPSO-SVM
Semi-supervised learning	(Saravanakumar et al., 2022)	Maximum 0.97	-	-	-	-	1000 scans	GAN
	(S. Liu et al., 2014)	0.8776	0.8857	0.8722	-	-	77 CN, 65 AD	Auto-encoder
	(D. Zhang & Shen, 2011)	0.95	0.95	0.95	-	-	52 CN, 51 AD	mLapRLS

From some other studies, which are listed in Table 4.9, they were proved to have data leakage by the investigation of Wen et al. (2020). Data leakage refers that any part of training data involves test data, or vice versa (Wen et al., 2019). However, data leakage may affect the performance and may make it unreliable, and such a biased result could potentially lead to a bad performance and serious misdiagnoses after the model deployment. Hence, data leakage should be avoided in studies. In the study reported here, an unsupervised learning method was used, which means that there were neither training data nor test data, and therefore no data leakage.

Table 4.9 The accuracy of the related studies based on structural MCI scans with data leakage

Study	Performance				
	AD vs. CN	sMCI vs. pMCI	MCI vs. CN	AD vs. MCI	Multi-class
(Asl et al., 2018)	0.99	-	0.94	1.00	0.95
(Backstrom et al., 2018)	0.9837	-	-	-	-
(Backstrom et al., 2018)	0.9883	-	0.9842	0.9361	0.9752
(Basaia et al., 2019)	0.992	0.751	-	-	
(Farooq et al., 2017)	-	-	-	-	0.9888

4.10.2. Memory Usage and Runtime Problems

From Table 4.6, it can be seen that the PCANet has a drawback when it is applied with MRI scans: from an algorithmic point of view, it is a process that increases the dimensions (i.e., the number of features). PCANet was developed initially for recognising general pictures (2-D images) but the MRI scan is a 3-D image, and this will cause much memory usage and long executing time. In this study, the dimension of a pre-processed MRI scan is $145 \times 121 \times 121$, which means that there are 387 slices (2-D images) in each scan. As can be seen from Table 4.6, even if only 77 out of 387 slices in each scan were used, the memory usage was around 100GB in this study. Therefore, the resultant features processed by the PCANet using all slices (i.e., 387 slices) of all the 429 MRI scans, will occupy a much larger volume of memory space. However, the method reported here only required less than 8GB of memory usage. Hence, when the programme is executed on a public high performance computer, the former method (that uses all slices of all 429 scans) involves a considerable of allocated memory is required and, as a result, less memory can be requested by other users. This may therefore require dedicated computing facilities.

In addition to the memory usage, feeding all slices of the MRI scans into the PCANet indicates more processing time than the method reported here (which only feeds AD-relevant regions).

From results in Table 4.6, it can be seen that, in the same executive environment, the methods that only involves 77 out of 387 slices in each scan had a close runtime to the method reported here. Therefore, if all slices of each scan were used, the runtime could be expected to be multiple times longer than the method reported here. Although parallel computing can be considered and it may potentially shorten the executing time in this case, the required memory might be larger than the serial computing based on the mechanism of the memory usage in the process of parallel computing. In a word, this will be a vicious cycle in this case: the method requires an amount of huge memory when running and it leads to a long runtime; if the runtime is shortened via parallel computing, it may require more memory. Again, this is a trade-off for this kind of computational task.

Moreover, the waiting time in queue when the programme is submitted also needs to be considered. The waiting time might be very long when the programme requires a large volume of memory space and is submitted at rush hours. For example, in a certain run of the method that requires 100G of memory, the waiting time in queue was 20 hours. In practice, long waiting time could be a problem because it will delay the diagnosis.

In summary, detecting AD-relevant regions is necessary when the PCANet is used in this study, because it can significantly reduce the need for memory space and can significantly reduce the runtime of the program. Lower memory usage and shorter runtime also mean low energy consumption and naturally low carbon environment.

4.11. Conclusion

In this chapter, detailed information of the method of discrimination between CN and AD was provided and the results of the discrimination were discussed and concluded. The method indicated that the lower parts of the limbic system in brain were the ROIs related to AD, and using both two ROIs could achieve a better performance than using one in the discrimination. In addition, in the comparison of the two pre-processing tools, SPM was a relatively easily used and was a better pre-processing tool compared with FreeSurfer, thus it was used in the next two studies reported in Chapters 5 and 6 for pre-processing MRI scans. Finally, the method yielded an accuracy of 0.842, a sensitivity of 0.797, a specificity of 0.881, a PPV of 0.855, and a NPV of 0.836. The results indicated that the proposed method could diagnose AD well and Table 4.5 indicated that it could also potentially predict the progress of patients, especially those who developed to AD or who developed to MCI within 36 months.

In addition to diagnosing AD, it is also important to discriminate between sMCI and pMCI scans because they two are crucial stages for potentially delaying the development to AD. Based on the predictive capability of the proposed method, it is worth considering whether sMCI and pMCI can be clustered using CN and AD data. In the following chapter, Chapter 5, the discrimination between sMCI and pMCI will be explicitly presented and discussed. The detailed information of the discrimination will be provided in the chapter.

Appendix B

Table B.1 Number of MRI scans clustered correctly using one ROI from two-sample t-test

	Run 1	Run 2	Run 3	Run 4	Run 5	Run 6	Run 7	Run 8	Run 9	Run 10	Mean	Standard Deviation
Negative (out of 231)	197	214	174	196	154	188	208	203	199	208	194.1	18.0951
Positive (out of 198)	134	108	154	144	172	146	130	136	126	125	137.5	17.6588

Table B.2 Metrics of results using one ROI calculated from two-sample t-test

	Run 1	Run 2	Run 3	Run 4	Run 5	Run 6	Run 7	Run 8	Run 9	Run 10	Mean	Standard Deviation
Accuracy	0.7716	0.7506	0.7646	0.7925	0.7599	0.7786	0.7879	0.7902	0.7576	0.7762	0.7730	0.0146
Sensitivity	0.6768	0.5455	0.7778	0.7273	0.8687	0.7374	0.6566	0.6868	0.6364	0.6313	0.6945	0.0892
Specificity	0.8528	0.9264	0.7532	0.8485	0.6667	0.8139	0.9004	0.8788	0.8615	0.9004	0.8403	0.0783
PPV	0.7976	0.864	0.7299	0.8045	0.6908	0.7725	0.8496	0.8292	0.7974	0.8446	0.7980	0.0548
NPV	0.7548	0.7039	0.7982	0.784	0.7662	0.7833	0.7536	0.766	0.7343	0.8221	0.7666	0.0332

Table B.3 Number of MRI scans clustered correctly using two ROIs from two-sample t-test

	Run 1	Run 2	Run 3	Run 4	Run 5	Run 6	Run 7	Run 8	Run 9	Run 10	Mean	Standard Deviation
Negative (out of 231)	216	208	217	192	202	202	198	202	213	184	203.4	10.5325
Positive (out of 198)	155	160	141	171	157	153	158	153	155	174	157.7	10.3458

Table B.4 Metrics of results using two ROIs calculated from two-sample t-test

	Run 1	Run 2	Run 3	Run 4	Run 5	Run 6	Run 7	Run 8	Run 9	Run 10	Mean	Standard Deviation
Accuracy	0.8648	0.8578	0.8345	0.8462	0.8368	0.8275	0.8298	0.8275	0.8578	0.8345	0.8417	0.0139
Sensitivity	0.7828	0.8081	0.7121	0.8636	0.7929	0.7727	0.7980	0.7727	0.7828	0.8788	0.7965	0.0472
Specificity	0.9351	0.9004	0.9394	0.8312	0.8745	0.8745	0.8571	0.8745	0.9221	0.7965	0.8805	0.0456
PPV	0.9117	0.8743	0.9097	0.8143	0.8441	0.8407	0.8272	0.8407	0.8960	0.7873	0.8546	0.0419
NPV	0.8339	0.8455	0.7920	0.8767	0.8313	0.8178	0.8319	0.8178	0.8320	0.8846	0.8364	0.0274

Table B.5 Subjects in this study who were CN at baseline and developed MCI or AD in the future visits. The rows in red are the progression of the subjects who were wrongly diagnosed as AD using the method in this chapter.

Subjects	BL	6	12	24	36	48	60	72	84	96	108	120
005_S_0223	CN	MCI	MCI	MCI	AD	-	-	-	-	-	-	-
009_S_0842	CN	CN	CN	CN	CN	CN	CN	MCI	MCI	MCI	CN	-
010_S_0420	CN	CN	CN	CN	CN	MCI	CN	MCI	MCI	MCI	MCI	-
011_S_0008	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	MCI
011_S_0022	CN	CN	CN	MCI	-	-	-	-	-	-	-	-
012_S_1009	CN	CN	MCI	CN	-	-	-	-	-	-	-	-
014_S_0520	CN	CN	CN	CN	CN	CN	CN	MCI	-	-	-	-
014_S_0548	CN	CN	CN	CN	MCI	MCI	MCI	AD	-	-	-	-
014_S_0558	CN	CN	CN	CN	MCI	MCI	MCI	AD	-	-	-	-
018_S_0055	CN	CN	CN	CN	CN	MCI	MCI	MCI	MCI	MCI	MCI	-
020_S_0883	CN	CN	CN	CN	MCI	-	-	-	-	-	-	-
020_S_0899	CN	CN	CN	CN	MCI	-	-	-	-	-	-	-
021_S_0984	CN	CN	CN	CN	CN	CN	CN	CN	AD	AD	AD	-
023_S_0061	CN	CN	CN	CN	CN	CN	CN	MCI	AD	AD	AD	
023_S_1190	CN	CN	CN	CN	CN	CN	CN	CN	MCI	MCI	MCI	-
024_S_0985	CN	CN	CN	CN	CN	CN	MCI	AD	AD	-	-	-
024_S_1063	CN	CN	CN	MCI	MCI	MCI	-	-	-	-	-	-
029_S_0843	CN	CN	CN	MCI	MCI	MCI	-	-	-	-	-	-
032_S_1169	CN	CN	CN	CN	CN	CN	MCI	MCI	MCI	MCI	MCI	-
033_S_0741	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	MCI	MCI
033_S_0920	CN	CN	CN	CN	CN	CN	CN	CN	MCI	MCI	AD	AD
033_S_1098	CN	CN	CN	CN	CN	CN	CN	CN	CN	MCI	MCI	AD
035_S_0156	CN	CN	CN	CN	CN	CN	CN	CN	MCI	CN->MCI	MCI	MCI
035_S_0555	CN	CN	CN	CN	CN	MCI	MCI	MCI	MCI	MCI	AD	MCI
037_S_0454	CN	CN	CN	CN	CN	CN	CN	MCI	MCI	MCI	MCI	NC->MCI
037_S_0467	CN	CN	CN	CN	CN	MCI	MCI	AD	AD	AD	AD	AD
041_S_0898	CN	CN	CN	CN	CN	AD	-	-	-	-	-	-
051_S_1123	CN	CN	CN	CN	CN	MCI	MCI	MCI	MCI	AD	AD	-
057_S_0779	CN	CN	CN	MCI	MCI	-	-	-	-	-	-	-
067_S_0056	CN	CN	CN	CN	CN	CN	CN	CN	MCI	CN	CN	CN
068_S_0127	CN	CN	CN	CN	CN	CN	MCI	MCI	MCI	MCI	MCI	MCI
068_S_0210	CN	CN	CN	CN	MCI	MCI	MCI	MCI	MCI	MCI	MCI	CN
082_S_0363	CN	CN	CN	CN	CN	CN	CN	MCI	-	-	-	-
094_S_1241	CN	CN	MCI	CN	CN	-	-	-	-	-	-	-
098_S_0171	CN	CN	CN	CN	CN	CN	CN	CN	MCI	-	-	-
098_S_0896	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	MCI	MCI
100_S_0015	CN	CN	CN	CN	CN	CN	CN	MCI	-	-	-	-
100_S_0035	CN	CN	CN	CN	CN	CN	MCI	-	-	-	-	-
100_S_0069	CN	CN	CN	CN	CN	CN	CN	MCI	CN	CN	CN	CN
114_S_0166	CN	CN	CN	CN	CN	CN	AD	MCI	AD	AD	AD	AD
114_S_0173	CN	CN	CN	CN	CN	CN	CN	CN	CN	MCI	MCI	MCI
116_S_1249	CN	CN	CN	CN	CN	CN	CN	CN	MCI	AD	AD	-
123_S_0106	CN	CN	CN	CN	CN	CN	MCI	MCI	MCI	MCI	AD	AD
126_S_0605	CN	CN	CN	CN	CN	CN	CN	CN	MCI	MCI	MCI	MCI
126_S_0680	CN	CN	CN	CN	CN	CN	CN	CN	MCI	MCI	MCI	MCI
127_S_0259	CN	CN	CN	CN	CN	CN	CN	CN	CN	MCI	AD	AD

127_S_0622	CN	CN	CN	CN	CN	MCI	-	-	-	-	-	-
128_S_0229	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	CN	MCI
128_S_0230	CN	CN	CN	CN	AD	AD	AD	AD	AD	AD	AD	-
128_S_0522	CN	CN	CN	CN	CN	CN	CN	MCI	MCI	CN	CN	MCI
128_S_0545	CN	CN	CN	CN	CN	CN	CN	CN	CN	MCI	MCI	MCI
129_S_0778	CN	CN	CN	CN	CN	MCI	MCI	AD	AD	AD	AD	-
130_S_0232	CN	CN	CN	CN	MCI	MCI	MCI	MCI	-	-	-	
130_S_1200	CN	CN	CN	CN	MCI	-	-	-	-	-	-	-
131_S_0123	CN	CN	CN	CN	CN	CN	CN	MCI	MCI	AD	AD	AD
137_S_0459	CN	CN	CN	CN	CN	MCI	MCI	-	-	-	-	-
137_S_0972	CN	CN	CN	CN	CN	CN	MCI	MCI	MCI	AD	-	-
141_S_0717	CN	CN	CN	CN	MCI	MCI	MCI	MCI	-	-	-	-
941_S_1194	CN	CN	CN	CN	CN	MCI	-	-	-	-	-	-
941_S_1202	CN	CN	CN	MCI	MCI	MCI	MCI	-	-	-	-	-
941_S_1203	CN	CN	CN	CN	MCI	MCI	-	-	-	-	-	-

Chapter 5 Differentiating Between Stable Mild Cognitive Impairment Scans and Progressive Mild Cognitive Impairment Scans

5.1. Introduction

In Chapter 4, the detailed information, results, and discussions from the work discriminating between cognitively normal (CN) and Alzheimer's disease (AD) scans was presented. The discrimination between CN and AD yielded an accuracy of 0.842, a sensitivity of 0.797, a specificity of 0.881, a positive predictive value (PPV) of 0.855, a negative predictive value (NPV) of 0.836. In addition, it also indicated that the proposed method has potential to predict the development of AD and has potential to help in the discrimination between stable mild cognitive impairment (sMCI) and progressive mild cognitive impairment (pMCI).

This chapter seeks to explore the use of unsupervised learning methods in discriminating between sMCI and pMCI. The chapter presents detailed information and an exhaustive description of the operations of each phase in the workflow of the discrimination between sMCI and pMCI scans are presented. This chapter discusses the software, tools, and programming languages that were used in this part of the study to cluster sMCI and pMCI (Section 5.2), how the sMCI and pMCI scans were collected from the ADNI (Section 5.3), how the pre-processing tools were used (Section 5.4), how the two-sample t-test analysis was applied (Section 5.5), and how the PCANet (Section 5.6) and the k-means++ were implemented in Matlab (Section 5.7), and how the performance of this part of the study was estimated (Section 5.8). Finally, the chapter presents the findings of the discrimination and discussions on the results.

5.2. Software, Tools, and Programming Languages

In Section 4.9.1, the performance of Statistical Parametric Mapping (version 12, SPM12) and FreeSurfer was compared: the SPM12 had a better performance in pre-processing the scans. As a result, SPM12 was used to pre-process the sMCI and pMCI MRI scans. Similar to previous phase of the study (as presented in Chapter 4), Matlab 2020b was also used in this phase as the SPM12 (including its computational anatomy toolbox) is based on it and the feature extraction (PCANet) and clustering (k-means++) were also implemented on the Matlab.

5.3. Data Collection

As with the CN and AD MRI scans, described in Section 4.3, sMCI and pMCI scans were collected from the ADNI-1 and only the baseline scans (i.e., the scan from the first visit of each

patient) were used in this study. However, unlike CN and AD, ADNI does not directly provide sMCI and pMCI data. In other words, baseline MCI scans that were directly obtained from ADNI-1 needed to be manually divided into sMCI and pMCI beforehand, as explained below.

To divide the MCI scans into sMCI and pMCI, it is necessary to know the diagnostic state (i.e., CN, AD, or MCI) of each visit of the participants used in this study, because the sMCI and pMCI classification is based on whether the patients later developed to AD or not within the following 36 months. A diagnostic summary form was downloaded from the ADNI through the account, and a part of the diagnostic summary form is shown in Figure 5.1. The column D in the form indicates the identifiers of patients, the column F indicates the order of the visit (e.g., "bl" means the baseline or first visit, "m06" means the visit 6 months after the baseline, etc.), and the column L indicates the current diagnostic states (i.e., CN, MCI, or AD).

According to the definition of sMCI and pMCI mentioned in Section 1.4.1, those participants who were diagnosed as MCI at baseline and then developed to AD within later 36 months were considered as having pMCI, and otherwise they were considered as having sMCI.

Phase	ID	RID	PTID	SITEID	VISCODE	VISCODE2	USERDATE	USERDATE2	EXAMDATE	DXCHANGE	DXCURREN	DXCONV	DXCONTYP	DXREV	DXNORM	DXNODEP	DXMCI	DXMDES
ADNI1	2	2	011_S_0002	107	bl	bl	10/1/05		9/29/05		1	0	-4	-4	1	-4	-4	-4
ADNI1	336	2	011_S_0002	107	m06	m06	4/27/06		3/6/06		1	0	-4	-4	1	-4	-4	-4
ADNI1	4	3	011_S_0003	107	bl	bl	10/1/05		9/30/05		3	0	-4	-4	-4	-4	-4	-4
ADNI1	338	3	011_S_0003	107	m06	m06	4/27/06		3/13/06		3	0	-4	-4	-4	-4	-4	-4
ADNI1	990	3	011_S_0003	107	m12	m12	9/14/06		9/13/06		3	0	-4	-4	-4	-4	-4	-4
ADNI1	3792	3	011_S_0003	107	m24	m24	9/22/07		9/12/07		3	0	-4	-4	-4	-4	-4	-4
ADNI1	34	4	022_S_0004	10	bl	bl	11/18/05		11/8/05		2	0	-4	-4	-4	-4	1	1
ADNI1	382	4	022_S_0004	10	m06	m06	5/3/06		5/2/06		2	0	-4	-4	-4	-4	1	1
ADNI1	1406	4	022_S_0004	10	m12	m12	11/15/06		11/14/06		2	0	-4	-4	-4	-4	1	1
ADNI1	3136	4	022_S_0004	10	m18	m18	6/22/07		5/14/07		2	0	-4	-4	-4	-4	1	1
ADNI1	6	5	011_S_0005	107	bl	bl	10/1/05		9/30/05		1	0	-4	-4	1	-4	-4	-4
ADNI1	340	5	011_S_0005	107	m06	m06	4/27/06		3/9/06		1	0	-4	-4	1	-4	-4	-4
ADNI1	956	5	011_S_0005	107	m12	m12	9/7/06		9/6/06		1	0	-4	-4	1	-4	-4	-4
ADNI1	3788	5	011_S_0005	107	m24	m24	9/22/07		9/7/07		1	0	-4	-4	1	-4	-4	-4
ADNI1	38	6	100_S_0006	4	bl	bl	12/1/05		12/1/05		2	0	-4	-4	-4	-4	1	1
ADNI1	536	6	100_S_0006	4	m06	m06	6/7/06		6/7/06		2	0	-4	-4	-4	-4	1	1
ADNI1	1736	6	100_S_0006	4	m12	m12	1/10/07		1/9/07		2	0	-4	-4	-4	-4	1	1
ADNI1	3048	6	100_S_0006	4	m18	m18	6/8/07		6/8/07		2	0	-4	-4	-4	-4	1	1
ADNI1	4196	6	100_S_0006	4	m24	m24	11/19/07		11/19/07		2	0	-4	-4	-4	-4	1	1
ADNI1	10	7	022_S_0007	10	bl	bl	10/6/05		10/6/05		3	0	-4	-4	-4	-4	-4	-4
ADNI1	290	7	022_S_0007	10	m06	m06	4/17/06		4/13/06		3	0	-4	-4	-4	-4	-4	-4
ADNI1	1186	7	022_S_0007	10	m12	m12	10/12/06		10/12/06		3	0	-4	-4	-4	-4	-4	-4
ADNI1	3972	7	022_S_0007	10	m24	m24	10/11/07		10/11/07		3	0	-4	-4	-4	-4	-4	-4
ADNI1	8	8	011_S_0008	107	bl	bl	10/1/05		9/30/05		1	0	-4	-4	1	-4	-4	-4
ADNI1	342	8	011_S_0008	107	m06	m06	4/27/06		3/20/06		1	0	-4	-4	1	-4	-4	-4
ADNI1	1054	8	011_S_0008	107	m12	m12	9/20/06		9/20/06		1	0	-4	-4	1	-4	-4	-4

Figure 5.1 Diagnostic summary of the subjects in ADNI (Screenshot by Yuyang Liu)

Therefore, using the diagnostic summary and the definitions of sMCI and pMCI, 134 patients who did not progress to AD within 36 months were therefore considered as having sMCI and 152 patients progressed to AD within 36 months and were considered as having pMCI in this part of study. As mentioned in Section 3.3.1, only baseline scans of participants were used so 134 sMCI scans and 152 pMCI scans were acquired. A summary of the MRI scans used to cluster sMCI and pMCI is shown in Table 5.1. In addition, as described in later Sections 5.6 and 5.7, CN and AD scans were also used in this study, and the numbers of CN and AD scans used in this study were 231 and 198 (the same as those in Study 1), respectively.

Table 5.1 The number of MRI scans of sMCI and pMCI

Status	sMCI	pMCI	CN	AD	Total
Number of MRI scans	134	152	231	198	715

5.4.　Data Pre-processing

As described in Section 4.9.1, SPM12 is easier to use and offers better performance than FreeSurfer in pre-processing MRI scans. Therefore, the MRI scans of sMCI and pMCI were pre-processed only by SPM12 and its computational anatomy toolbox 12 (CAT12). Similar to the pre-processing step in Section 4.4, the parameters and the settings of the pre-processing ("segment" in CAT12) were the default values. The pre-processed outputs of each MRI scan via "segment" are a grey matter scan without the skull, a white matter scan without the skull, a whole brain scan without the skull, and other text files. As with Study 1, the whole brain scans without skulls were still used in this task. In addition, all output scans were motion corrected, normalised, and registered into a common space, i.e., Montreal Neurological Institute (MNI) template. The size of the output scans was 121×145×121 voxels (sagittal, coronal, and axial planes, respectively).

5.5.　Detection of Differences Between sMCI and pMCI

A two-sample t-test was used to detect visual differences (regions of interest, ROIs) between the sMCI and the pMCI scans. In Chapter 4, 30 CN and 30 AD scans were used to detect differences between them. However, the number of sMCI and pMCI scans (286 scans) was much smaller than the CN and AD scans (429 scans) and this study sought to use a limited number of labelled data, thus as described in Section 3.5.2, 15 sMCI and 15 pMCI scans was used to detect ROIs using a two-sample t-test. Similar to Section 4.5, the two-sample t-test was implemented using the CAT12 of SPM12 in Matlab.

5.6.　Feature Extraction

Due to the similarity between sMCI and pMCI scans, as shown using similarity maps by Wee, Yap, & Shen (2013), sMCI and pMCI were not easy to distinguish. In some previous studies using supervised learning methods, sMCI and pMCI scans were classified by training CN and AD scans or their other types of data (e.g., Assessment Scale-Cognitive Subscale of CN and AD) (Y. Huang et al., 2019; W. Lin et al., 2018; Luo, He, Cui, Chen, & Ward, 2023). Therefore, in this study, distributions of the features in two scenarios were compared:

(1) Features (voxel intensity values) from the ROIs and then computed by PCANet only from sMCI and pMCI scans;

(2) Features (voxel intensity values) from the ROIs and then computed by PCANet from CN, sMCI, pMCI, and AD scans.

The ROIs used in the two scenarios are discussed in Section 5.9.1. Similar to Section 4.6, the distributions were demonstrated using the t-distributed stochastic neighbour embedding (t-SNE) technique.

As Figure 5.2 (a) shows, sMCI and pMCI are considered similar and are probably difficult to separate in the feature space. In an ideal situation, as shown in Figure 5.2 (b), it is expected that the differences between sMCI and pMCI are enhanced after including CN and AD cases because the difference between CN and AD scans is greatly larger.

(a)   (b)
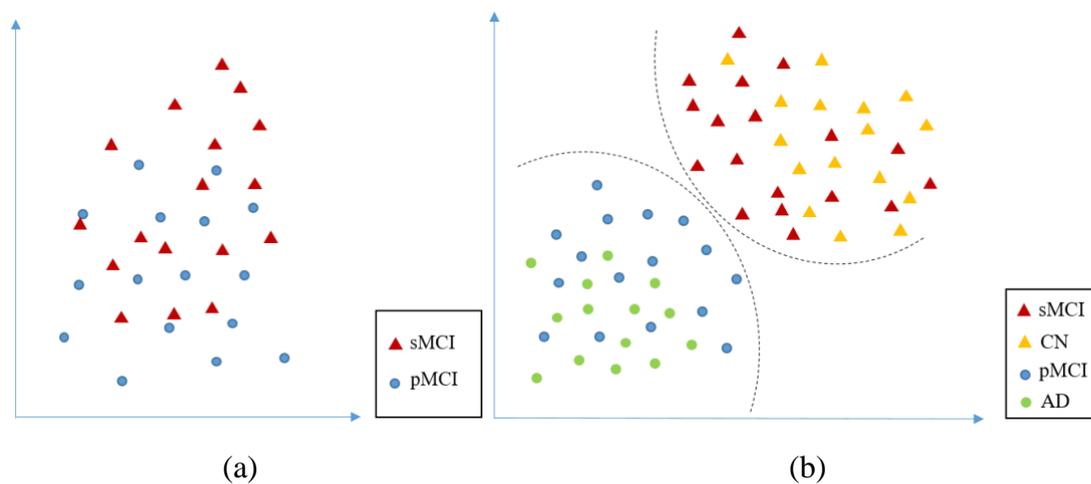
Figure 5.2 (a) shows a hypothetical distribution of sMCI and pMCI; (b) shows an ideal distribution of sMCI and pMCI after including CN and AD cases (Developed by Yuyang Liu)

In this study, the parameters of the PCANet were the same as mentioned in Section 3.6.4:

- the number of stages was set to two;
- the size of the patches $k_1 \times k_2$ was set to 3×3 in both two stages;
- the number of filters L_1 and L_2 were both set to eight, respectively;
- the block size for the histograms was set to 15×15;
- the overlap ratio of the local blocks was set to zero in this study for reducing the number of features.

The reasons for the choice of these parameters were given in Section 3.6.4., and the same applied in this study.

5.7. Clustering

Similar to Section 5.6, in this phase, two scenarios were implemented:

(1) Clustering sMCI and pMCI into two clusters.

(2) Clustering all four groups (i.e., CN, sMCI, pMCI, and AD) of scans into two clusters;

The aim of the two experiments (scenarios) was to compare the performances of the discriminations with and without the inclusion of CN and AD scans, and was to verify the effectiveness of the inclusion of the CN and AD scans.

K-means++ was used to cluster data in this study. As for the distance (similarity) used in the k-means++, using cosine similarity had a better performance than using the Euclidean distance. The performances of the two distances (similarities) are presented in Table 5.4 in Section 5.9.4. Therefore, cosine similarity was used in the clustering. The cosine similarity is calculated by:

$$Cosine\ Similarity = \frac{X \cdot Y}{\|X\| \, \|Y\|} = \frac{\sum_{i=1}^{n} X_i Y_i}{\sqrt{\sum_i^n X_i^2} \sqrt{\sum_i^n Y_i^2}}, \text{(Equation 5.1)}$$

where X_i and Y_i are components of vector X and Y

The function *kmeans(X,2,'Distance','cosine')* was used in Matlab to run the k-means++ to partition the observation of the data matrix (*X*) into two clusters with cosine similarity; the function *kmeans(X,2)* was used to run the k-means++ with Euclidean distance.

5.8. Performance Evaluation

As mentioned in Section 3.8.1, the clusters needed to be identified after using the k-means++. For the clustering of sMCI and pMCI, there are two scenarios described in Section 5.7:

(1) sMCI vs. pMCI: the approach to identifying the clusters was similar to that for identifying AD and CN clusters, described in Section 4.8: the data labelled as 1 were first check and were confirmed which diagnostic state (sMCI or pMCI) account for more than 50% of them. For example, if the sMCI data accounted for more than 50% of the data labelled as 1, sMCI is referred as 1 in this run and pMCI is referred as 2, and *vice versa*.

(2) sMCI+CN vs. pMCI+AD: the AD and CN scans were regarded as labelled data. After clustering using the k-means++, the MCI scans that were clustered into the AD cluster were identified as being pMCI, and the MCI scans that were clustered into the CN cluster were identified as being sMCI.

After recognising the clusters, the same as Section 4.8, the mean values of five measures: the accuracy, sensitivity, specificity, PPV, and the NPV, were used to evaluate the overall performance of the method presented in this chapter. The k-means++ was run 10 times due to its different centroids in different runs leading to different clustering results. 95% confidence intervals (95% CI) of the five measures over the ten runs were also used to describe the stability and the variance of the performance of the method.

5.9.   Results

5.9.1. ROIs between sMCI and pMCI

As described in Section 5.5, to demostrate whether sMCI and pMCI scans had clear ROIs, 15 sMCI and 15 pMCI scans were used to test. As can be seen from Figure 5.3, when a two-sample t-test was applied to 15 sMCI and 15 pMCI scans, there was no obvious visible difference at voxel level between the two groups.
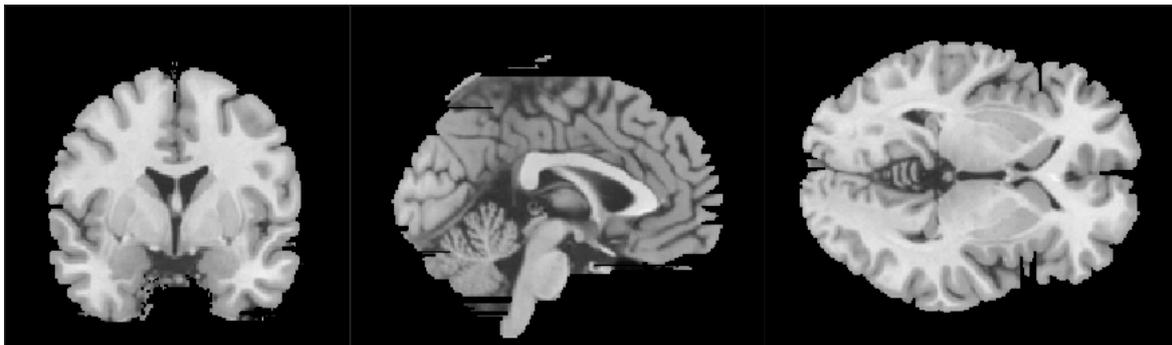
Figure 5.3 Three view planes (sagittal, coronal, and axial planes) to demonstrate the lack of ROIs detected using 15 sMCI and 15 pMCI scans (MRI images from Xjview adapted by Yuyang Liu)

The reason for this might be that the number of selected scans was low. Therefore, 30 sMCI and 30 pMCI scans (around 21% of the dataset) were then used to try and detect the ROIs, and the results are shown in Figure 5.4. As can be seen from Figure 5.4, when the two-sample t-test was applied to the 30 sMCI and 30 pMCI scans, although some regions (red and organge regions) were detected, the regions were too small and diffuse (scattered). In addition, 30 scans account for nearly 1/4 of sMCI scans and 1/5 of pMCI scans, which is a relatively large proportion of the whole dataset. However, this study aimed to use only a limited number of MRI scans to distinguish between sMCI and pMCI. Therefore, although very few ROIs were detected in Figure 5.4, it was decided to explore a different strategy, one that does not involve pre-labelled sMCI and pMCI data.
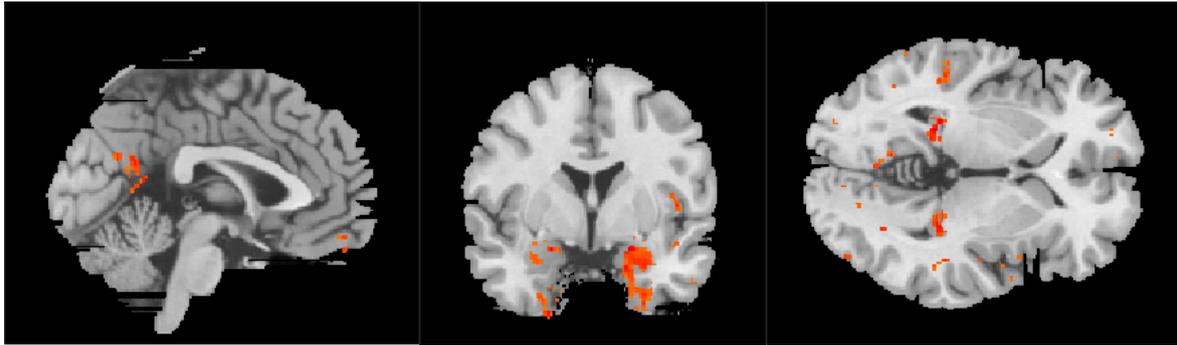
Figure 5.4 Three view planes (sagittal, coronal, and axial planes) to show the ROIs detected using 30 sMCI and 30 pMCI scans (MRI images from Xjview adapted by Yuyang Liu)

The fact that no ROI was detected from the limited number of scans (15 sMCI and 15 pMCI scans) suggests that sMCI and pMCI scans are somehow similar. Hence, such similarities made it hard to discriminate between sMCI and pMCI scans as what Study 1 (CN vs. AD) did: locating ROIs between the two groups using a limited number of sMCI and pMCI scans first and then clustering them. However, sMCI and pMCI are two prodromal stages before AD, in which the severity is less (based on the definitions of MCI). As mentioned in Section 5.6, Y. Huang et al. (2019), W. Lin et al. (2018), and Luo et al. (2023) used CN and AD data to train models and classified sMCI and pMCI data, thus it is highly likely that the ROIs between sMCI and pMCI scans are the same as those between CN and AD. Therefore, the same ROIs used in discriminating between CN and AD scans were used to discriminate between sMCI and pMCI scans.

5.9.2. Distributions of Features

The ROIs used in discriminating between sMCI and pMCI scans have been determined in Section 5.9.1. As presented in Section 5.6, distributions of the features from the determined ROIs in two scenarios were compared. The results of the distributions are shown in (a) and (b) below.

(a) sMCI vs. pMCI

To further demonstrate that sMCI and pMCI are similar, the distribution of features extracted from ROIs between sMCI and pMCI scans using the PCANet was shown in Figure 5.5. As can be seen from Figure 5.5, although the features were computed by voxels within the ROIs, the two groups of data were still mixed and scattered randomly. Therefore, as can be seen in Table 5.2 in Section 5.9.3, the performance of the clustering only using sMCI and pMCI alone was not satisfactory.
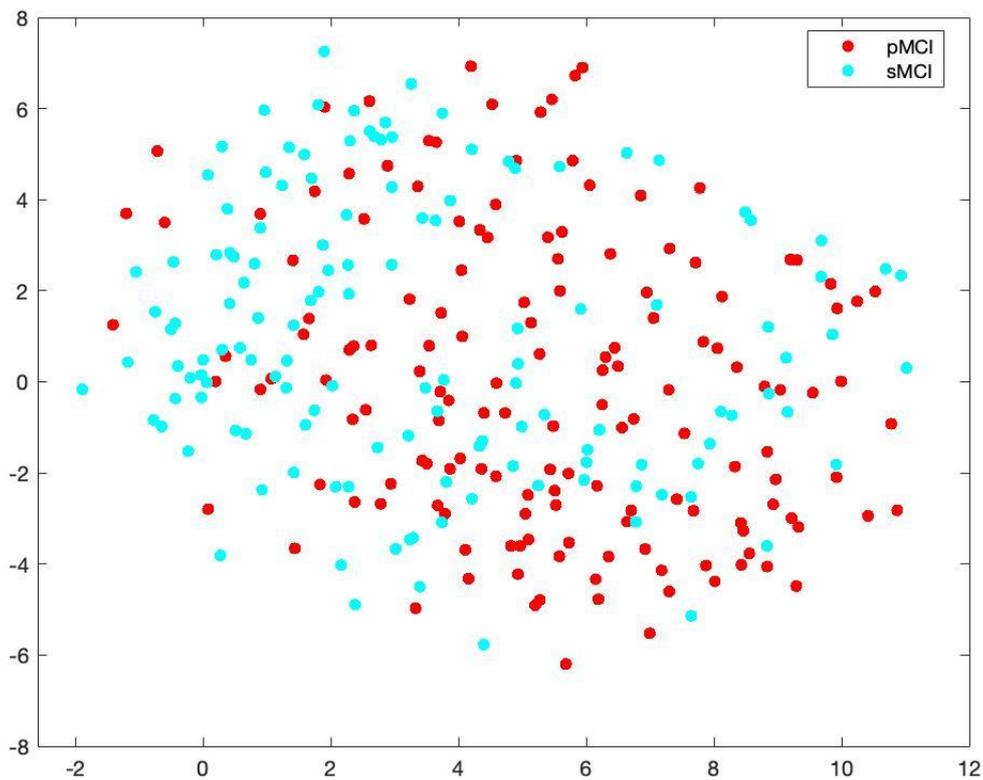
Figure 5.5 Distribution of extracted features of ROIs in sMCI and pMCI scans (Developed by Yuyang Liu)

(b) CN+sMCI vs. pMCI+AD

Figure 5.6 shows a distribution of features extracted from ROIs within sMCI and pMCI scans using the PCANet including AD and CN scans. There was improved separation of the two groups compared with Figure 5.5. For easily distinguishing the two groups (CN+sMCI and pMCI+AD) in the coordinate in Figure 5.6, features of CN and sMCI are shown in red, which are called negative data because CN is a negative diagnosis result in clinical practice; features of pMCI and AD are shown in light blue, which are called positive data because AD is a positive diagnosis result in clinical practice. As can be seen from Figure 5.6, negative data were distributed at the lower left part and positive data distributed at the upper right part though the two groups of data are still adjacent and there was overlap in the data points.
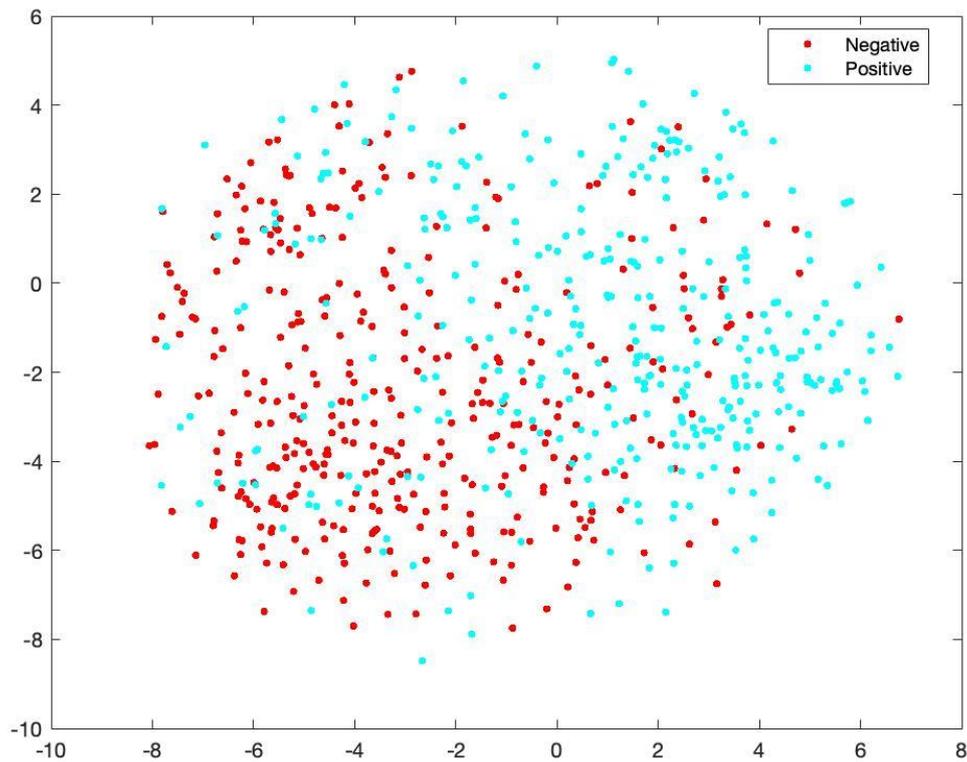
Figure 5.6 Distribution of extracted features of positive (AD and pMCI) and negative (CN and sMCI) scans (Developed by Yuyang Liu)

Figure 5.7 shows the same distribution as Figure 5.6, but all four groups of data (AD, CN, sMCI, and pMCI) are noted in four different colours. As can be seen, most pMCI data (points in light blue) were distributed together with most AD data (points in light green); most sMCI data (points in purple) were close to most CN data (points in red). This is in line with the expectation: pMCI is a late phase of MCI and it develops to AD within 36 months thus it is more similar to AD; sMCI is an early phase of MCI and it remain more than 36 months or even finally return to CN, so it is more similar to CN.
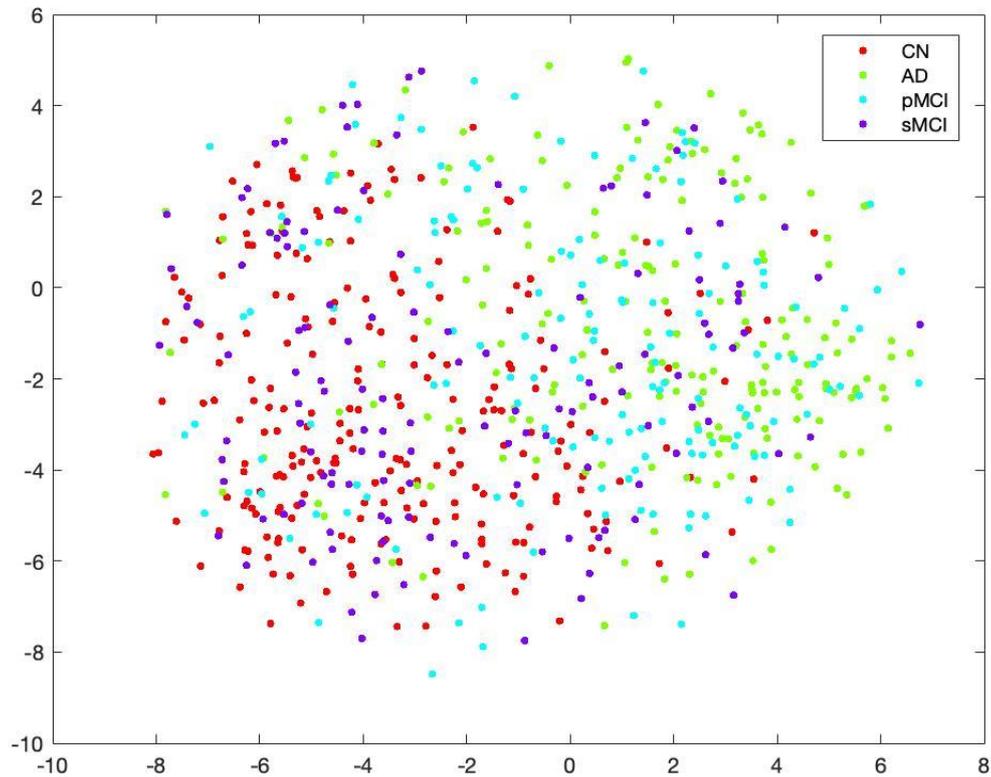
Figure 5.7 Distribution of extracted features of AD, CN, sMCI, and pMCI scans (Developed by Yuyang Liu)

In summary, based on Figures 5.5, 5.6, and 5.7, AD and CN data play important roles in the process of distinguishing MCI data: the inclusion of them could potentially enhance the performance of the discriminating between sMCI and pMCI.

5.9.3. Clustering sMCI and pMCI Using AD and CN vs. Not Using AD and CN data

As can be seen from Figures 5.5, 5.6, and 5.7, the inclusion of CN and AD scans help separate out sMCI and pMCI scans better in feature space. The performances of the k-means++ (cosine similarity used due to its better performance, and its effectiveness is shown in Section 5.9.4) applied to sMCI and pMCI alone (not including AD and CN data) and to all four groups (including AD and CN data) were compared. The results are shown in Table 5.2 and in Figure 5.8.

Table 5.2 Measures of results using k-means++ including and not including AD and CN data

Measures		K-means++	
		Not including AD and CN	**Including AD and CN**
Accuracy	**Mean**	0.557	0.672
	95% CI	(0.538, 0.576)	(0.665, 0.679)
Sensitivity	**Mean**	0.569	0.693
	95% CI	(0.405, 0.733)	(0.676, 0.71)
Specificity	**Mean**	0.543	0.660
	95% CI	(0.352, 0.734)	(0.648, 0.672)
PPV	**Mean**	0.625	0.698
	95% CI	(0.57, 0.68)	(0.691, 0.705)
NPV	**Mean**	0.536	0.653
	95% CI	(0.504, 0.569)	(0.643, 0.663)
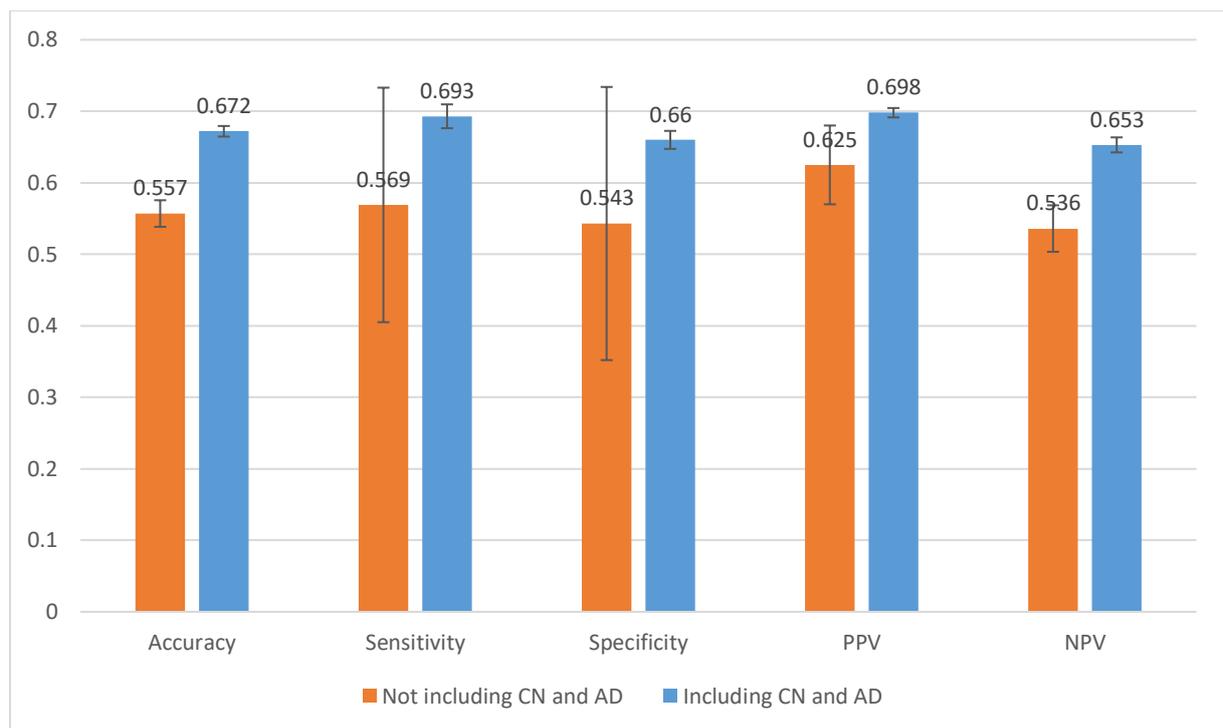
Figure 5.8 A bar chart indicating the mean values and 95% CI of each measure across the 10 runs (not including CN and AD scans and including CN and AD scans with cosine similarity) (Developed by Yuyang Liu)

From Table 5.2, it can be seen that all five measures that were obtained when including the AD and CN scans were superior to those yielded when using sMCI and pMCI scans alone. When only sMCI and pMCI scans were used, four out of the five measures (except PPV) were less

than 0.6. Furthermore, the 95% CI of the sensitivity and specificity were particular noteworthy. Although the highest values of the two measures approached more than 0.7, the lower boundaries of the 95% CI were as low as 0.4 or even 0.35. This means that, in certain runs, these two measures were extremely low and, in addition, that the variability was high and the stability correspondingly low. The 95% CI of the accuracy achieved by not using CN and AD scans was 0.538 to 0.576, which indicated a relatively more stable accuracy. Therefore, based on the stable accuracy, when one of the two measures (sensitivity and specificity) was significantly high, the other one would be very low, e.g., run 7, run 8, and run 10 in Table C.5 in Appendix C (also shown in Table 5.3), and this led to the extremely wide ranges of 95% CI for the sensitivity and specificity.

Table 5.3 Example runs of unbalanced sensitivity and specificity of discriminating between sMCI and pMCI without including CN and AD. The rest of the table is available in Table C.5 in Appendix C.

Runs	…	Run 7	Run 8	Run 9	Run 10
Negative (out of 134)	…	15	32	68	15
Positive (out of 152)	…	135	127	79	144

In a summary, the inclusion of the AD and CN data could truly enhanced the performance in discriminating between sMCI and pMCI scans using the unsupervised learning method presented in this thesis. The reason of this is likely due to the fact that sMCI is similar to CN and pMCI is similar to AD so that using the CN and AD data can help separate out sMCI and pMCI data better. This was further shown in Section 5.9.2. In addition, including AD and CN scans can help improve the stability (i.e., the range of 95% CI) of all five measures and can make the performance of the method more stable.

5.9.4. Clustering Using Euclidean Distance vs. Using Cosine Similarity

Using the k-means++ algorithm, different distances (or namely similarity measures) can result in different performances by the algorithm. Therefore, the cosine similarity and the Euclidean distance were used in the k-means++ to test which distance could achieve a better performance in this part of study. Table 5.4 and Figure 5.9 shows the five measures of the clustering results using the Euclidean distance and cosine similarity respectively in a k-means++.

Table 5.4 Measures of results using Euclidean distance and cosine similarity respectively in k-means++ (with inclusion of CN and AD scans)

Measures		K-means++	
		Euclidean Distance	**Cosine Similarity**
Accuracy	**Mean**	0.634	0.672
	95% CI	(0.622, 0.646)	(0.665, 0.679)
Sensitivity	**Mean**	0.578	0.693
	95% CI	(0.518, 0.637)	(0.676, 0.71)
Specificity	**Mean**	0.698	0.660
	95% CI	(0.652, 0.744)	(0.648, 0.672)
PPV	**Mean**	0.687	0.698
	95% CI	(0.675, 0.698)	(0.691, 0.705)
NPV	**Mean**	0.598	0.653
	95% CI	(0.575, 0.62)	(0.643, 0.663)
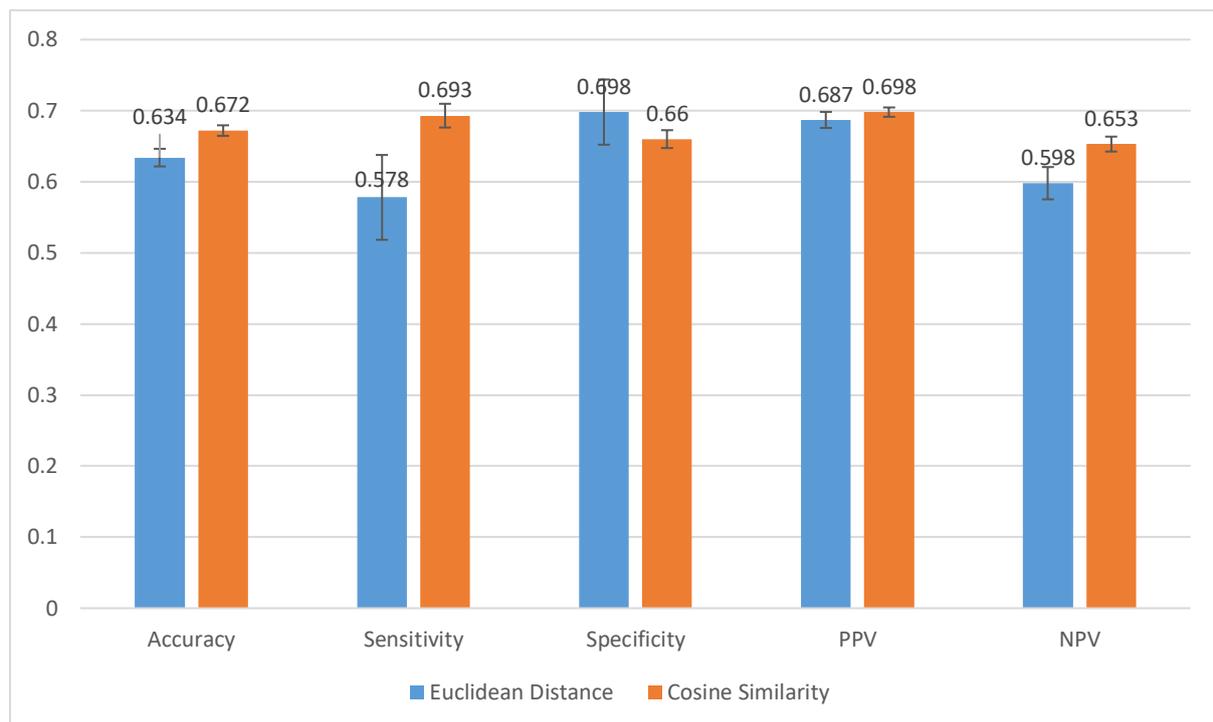
Figure 5.9 A bar graph indicating the mean values and 95% CI of each measure across the 10 runs using Euclidean distance and cosine similarity (Developed by Yuyang Liu)

As can be seen from Table 5.4, the mean values of the overall accuracies using both distances (similarity measures) were relatively stable (based on their relatively narrow ranges of 95%

CI). The cosine similarity (0.672) had nearly 0.04 overall accuracy higher than the Euclidean distance (0.634). However, compared with using the Euclidean distance (0.578), the sensitivity was considerably higher (approximately 0.12 higher) than when using the cosine similarity (0.693). In addition, using the cosine similarity achieved a more stable sensitivity (95% CI: 0.676-0.71) than when using Euclidean distance (95% CI: 0.518-0.637). Although the specificity (0.698) when using the Euclidean distance was higher, it was still variable compared with that when using the cosine similarity, based on their 95% CI. This is because the accuracies were relatively stable over all of the runs, but the sensitivity was relatively low (mean values: 0.578) and unstable (95% CI: 0.518-0.637). In other words, over a certain number of runs, the number of the correctly clustered pMCI scans was low, but to balance this (because the number of all correctly clustered scans in both pMCI and sMCI was stable), the number of correctly clustered sMCI scans was high, e.g., run 1 and run 2 in Table C.1 in Appendix C (also shown in Table 5.5), and *vice versa*. Hence, this led to some high specificities in some runs that increased the mean of specificity (to 0.6978) over the 10 runs, which was much higher than the sensitivity. This means that the ability to diagnose sMCI was much better than diagnosing pMCI when using the Euclidean distance in k-means++. In contrast, utilising the cosine similarity in the k-means++ achieved similar abilities to respectively recognise sMCI and pMCI.

Table 5.5 Example runs of unbalanced sensitivity and specificity of discriminating between sMCI and pMCI using Euclidean distance. The rest of the table is available in Table C.1 in Appendix C.

Runs	Run 1	Run 2	…
Negative (out of 134)	103	103	…
Positive (out of 152)	71	71	…

The PPV when using the Euclidean distance (0.687) and cosine similarity (0.698) were close: such levels of reliability (i.e., nearly 0.7) indicates that, when a patient is diagnosed as pMCI using them, there is 70% chance that the patient is actually positive. However, the NPV yielded when k-means++ using Euclidean distance was 0.598, which was more than 0.05 lower than that using cosine similarity (0.653). This indicates that a more reliable diagnosis is provided when a patient is diagnosed as sMCI using the k-means++ with the cosine similarity.

In summary, the cosine similarity in the k-means++ showed some stability: all of its five measures were between 0.65 and 0.70. In contrast, Euclidean distance in k-means++ yielded a

relatively high specificity and a high PPV, but the counterparts (sensitivity and NPV) of the two measures were much lower. As a consequence, when discriminating between sMCI and pMCI, the cosine similarity demonstrates a better performance than when the Euclidean distance was used in the k-means++.

5.10. Discussion

In the task of discriminating between sMCI and pMCI scans, the method proposed in this part of thesis achieved a relatively good performance, and it outperformed some previous studies. Based on the reviewed literature shown in Table 2.9 in Chapter 2, no previous study has been identified that applied unsupervised learning methods to discriminate between sMCI and pMCI; previous state-of-the-art studies used supervised and semi-supervised learning methods and these are listed in Table 5.6.

Table 5.6 Comparison of the method reported here and previous studies (supervised and semi-supervised learning) that discriminated between sMCI and pMCI

Learning Type	Study	ACC	SEN	SPE	PPV	NPV	Dataset	Models used
Unsupervised learning	**Study reported in this chapter**	**0.672**	**0.693**	**0.660**	**0.698**	**0.653**	**134 sMCI, 152 pMCI**	**PCANet, k-means++, involvement of AD and CN data**
Supervised learning	(Zheng et al., 2022)	0.8333	0.625	1	-	-	99 sMCI, 80 pMCI	3D EfficientNet
	(X. Zhang et al., 2021)	0.821	0.812	0.809	-	-	232 sMCI, 172 pMCI	Self-attention ResNet34
	(Zeng et al., 2021)	0.8778	-	-	-	-	82 sMCI, 95 pMCI	DBN-based multi-task learning algorithm
	(Bloch & Friedrich, 2021a)	0.6389	-	-	-	-	400 sMCI, 318 pMCI	RF classifier
	(Sujathakumari et al., 2019)	0.5899	0.7277	0.4110	0.618	-	191 sMCI, 146 pMCI	Gradient boosting classifier
	(Spasov et al., 2019)	0.72	0.63	0.81	-	-	409 MCI (only use structural MRI)	A neural network architecture based on AlexNet

	(Elahifasaee et al., 2019)	0.6594	0.8144	0.5469	-	-	167 pMCI, 236 sMCI	Feature decomposition, kernel discrimination analysis
	(L. Xu et al., 2019)	0.6369	0.7856	0.4539	-	-	126 pMCI, 95 sMCI	SVM
	(Zeng et al., 2018)	0.6923	-	-	-	-	95 pMCI, 82 sMCI	PCA, SDPSO-SVM
	(Sun, Qiao, Lelieveldt, & Staring, 2018)	0.654	0.676	0.642	-	-	134 sMCI, 76 pMCI	SVM integrating spatial-anatomical information
	(Salvatore et al., 2015)	0.66	-	-	-	-	134 sMCI, 76 pMCI	SVM
	(Moradi et al., 2015)	0.7474	0.8885	0.5159	-	-	100 sMCI, 164 pMCI	Random forest
	(Y. Huang et al., 2019)	0.7690	0.6815	0.8393	-	-	480 CN, 567 MCI, 465 AD	3D CNN
Semi-supervised learning	(Ye et al., 2011)	0.561	0.941	0.408	-	-	169 sMCI, 68 pMCI	LapSVM
	(Filipovych & Davatzikos, 2011)	-	0.788	0.794	-	-	174 sMCI, 68 pMCI	Transductive SVM
	(P. Zhang, Shi, Smith, & Liu, 2016)	0.7827	0.8638	0.5232	-	-	94 sMCI, 110 pMCI	CPD-LapSVM
	(Adel, Cohen, Caan, & Welling, 2017)	0.827	-	-	-	-	100 sMCI, 164 pMCI	LapSVM

Bloch et al. (2021), Sujathakumari et al. (2019)**,** L. Xu et al. (2019), Sun et al. (2018), Elahifasaee et al. (2019)**,** and Salvatore et al. (2015) achieved overall accuracies less than the study reported in this chapter. Although Zeng et al. (2021) and Zeng et al. (2018) achieved higher accuracies, which were 0.8778 and 0.6923, they did not report other measures, e.g., specificity, sensitivity, etc. However, as described in Section 4.9.3, accuracy is not sufficient on its own to comprehensively evaluate the performance of a method.

B. Zheng et al. (2022), X. Zhang et al. (2021), Moradi et al. (2015), Spasov et al. (2019)**,** and Y. Huang et al. (2019) achieved higher accuracies and reported sensitivities and specificities in their studies. Zhang et al. (2021) yielded an accuracy of 0.821, a sensitivity of 0.812, and a specificity of 0.809, which were balanced and advantageous. They employed a self-attention ResNet34 to classify

MCI scans. The self-attention mechanism can well locate the ROIs and the ResNet34 can powerfully compute the weights of each neuron (in the neural network), thus it is reasonable that they achieved a superior performance. However, the other four studies achieved unbalanced sensitivities and specificities: B. Zheng et al. (2022) yielded a sensitivity of 0.625 and a specificity of 1; Moradi et al. (2015) achieved a sensitivity of 0.8885 and a specificity of 0.5159; Y. Huang et al. (2019) achieved a sensitivity of 0.6815 and a specificity of 0.8393. This suggests that these studies were not able to well recognise sMCI and pMCI scan at the same time. Therefore, this can explain why accuracies alone cannot evaluate the performance of studies comprehensively: a study yielding a high accuracy, e.g., B. Zheng et al. (2022), may have unbalanced sensitivity and specificity.

In Table 5.6, Ye et al. (2011), Filipovych et al. (2011), P. Zhang et al. (2016), and T. Adel et al. (2017) used semi-supervised learning methods. The models they used were semi-supervised support vector machines (SVM), i.e., transductive SVM and Laplacian SVM (LapSVM). Filipovych et al. (2011) achieved a balanced and relatively good sensitivity (0.788) and specificity (0.794). It is likely because they used a different standard to define pMCI: they defined pMCI based on Clinical Dementia Rating (CDR) scores. This may lead that the difference between sMCI and pMCI cases in their study is greater than that in this study, and as a result, they achieved a better performance. As for the other studies using semi-supervised learning methods in Table 5.6, Ye et al. (2011) and P. Zhang et al. (2016) achieved unbalanced sensitivities and specificities. The possible reason could be that they trained the models on unbalanced datasets. T. Adel et al. (2017) achieved a good accuracy (0.827), but they did not provide any other measures to evaluate their method.

As can be seen from Table 5.6, none of studies provided both PPV and NPV values (only Sujathakumari et al. (2019) reported a PPV). However, as explained in Section 4.9.3, PPV and NPV indicate the reliability of the method thus it is necessary for studies to report these two measures as well. In addition, if the sensitivity and specification are not balanced (as are some studies in Table 5.6), it is very likely that either PPV or NPV is low. When the PPV is high and the NPV is low, there is a high probability that a patient has the pMCI being investigated when the patient has a positive test (tested pMCI by the model), but there is a low probability that a patient has sMCI when the patient has a negative test (tested sMCI through the model). In other words, those who are diagnosed as sMCI are very likely to be pMCI using the model with low NPV and they could well develop to AD within 3 years; however, as mentioned in the Chapter 1, so far there is no remedy for AD and pMCI is an important phase for delaying the development to AD. Hence, such misdiagnosing could

lead to delays in treatment. Therefore, including PPV and NPV is important for studies in clinical settings.

Although some studies in Table 5.6 had better performances, they used supervised and semi-supervised learning methods and their models greatly depended on labelled sMCI and pMCI data or depend on pseudo labels of sMCI and pMCI data generated in the process of semi-supervised learning methods. The study reported in this chapter applied unsupervised learning methods and it did not involve labelled sMCI and pMCI data. Thus, this study made a trade-off between the performance and dependency on labelled data: it moderately lowered the performance but entirely eliminated the reliance on labelled sMCI and pMCI scans and it also saved much time because it has no training phase. In addition, the study in this chapter also offered a stable and balanced performance compared with previous studies.

5.11. Conclusion

In this chapter, the software used in this study and the detail of each step of the methodology in this study were introduced, and the results of the clustering were presented and discussed. As can be seen from Figure 5.3 and 5.5, sMCI and pMCI scans are similar and thus have no obvious ROIs between them. Hence, based on the definition of the sMCI and pMCI, the ROIs detected from CN and AD scans were used to distinguish them. Finally, the method in this chapter achieved an accuracy of 0.672, a sensitivity of 0.693, a specificity of 0.660, a PPV of 0.698, and a NPV of 0.653. The results in Table 5.2 and 5.3 also showed that the inclusion of CN and AD scans and the cosine similarity could help in the discrimination of the sMCI and pMCI scans.

The study reported in this chapter achieved a stable performance and balanced measures compared with other studies in Table 5.6, despite not relying on training data. In addition, to the best of my knowledge, this study was the first to employ unsupervised learning methods to discriminate between sMCI and pMCI scans, and it was also the first to involve labelled CN and AD scans in unsupervised learning methods to help discriminate between sMCI and pMCI scans. In addition, this chapter recommends that studies should provide a comprehensive evaluation: providing the PPV and NPV as well as the overall accuracy, sensitivity, and specificity as measures for evaluating the performance.

The results given in this chapter indicated that distinguishing sMCI and pMCI scans is much more difficult than distinguishing CN and AD scans, because all measures (accuracy: 0.672; sensitivity:

0.693; specificity: 0.660; PPV: 0.698; NPV: 0.653) achieved by the method in this chapter are much lower than those (accuracy: 0.842; sensitivity: 0.797; specificity: 0.881; PPV: 0.855; NPV: 0.836) presented in the Chapter 4. This also indicates that the similarity between sMCI and pMCI is much greater than that between CN and AD. In other words, distinguishing between sMCI and pMCI is much more difficult than distinguishing between CN and AD. Therefore, to better distinguish pMCI, an easier, but still clinically meaningful, task, discriminating between CN and pMCI scans, was designed, and the detailed information and the results of this task will be provided in Chapter 6.

Appendix C

Table C.1 Number of MCI scans (sMCI and pMCI) correctly clustered using Euclidean distance in k-means++.

	Run 1	Run 2	Run 3	Run 4	Run 5	Run 6	Run 7	Run 8	Run 9	Run 10	Mean	Standard Deviation
Negative (out of 134)	103	103	99	86	94	73	97	102	95	83	93.5	9.914
Positive (out of 152)	71	71	82	96	87	118	84	75	95	99	87.8	14.657

Table C.2 Measures of results of discriminating between sMCI and pMCI using Euclidean distance in k-means++.

	Run 1	Run 2	Run 3	Run 4	Run 5	Run 6	Run 7	Run 8	Run 9	Run 10	Mean	Standard Deviation
Accuracy	0.6084	0.6084	0.6329	0.6363	0.6329	0.6678	0.6329	0.6189	0.6643	0.6364	0.6339	0.0200
Sensitivity	0.4671	0.4671	0.5395	0.6316	0.5724	0.7763	0.5526	0.4934	0.625	0.6513	0.5776	0.0964
Specificity	0.7687	0.7687	0.7388	0.6418	0.7015	0.5448	0.7238	0.7612	0.7090	0.6194	0.6978	0.0740
PPV	0.6960	0.6960	0.7009	0.6667	0.6850	0.6592	0.6942	0.7009	0.7090	0.66	0.6868	0.0183
NPV	0.5597	0.5597	0.5858	0.6056	0.5912	0.6822	0.5879	0.5698	0.625	0.6103	0.5977	0.0366

Table C.3 Number of MRI scans (sMCI and pMCI scans) clustered correctly using cosine similarity in k-means++.

	Run 1	Run 2	Run 3	Run 4	Run 5	Run 6	Run 7	Run 8	Run 9	Run 10	Mean	Standard Deviation
Negative (out of 134)	85	88	90	92	85	90	85	88	92	89	88.4	2.716
Positive (out of 152)	109	100	109	104	111	101	109	106	104	100	105.3	4.111

Table C.4 Measures of results of discriminating between sMCI and pMCI using cosine similarity in k-means++.

	Run 1	Run 2	Run 3	Run 4	Run 5	Run 6	Run 7	Run 8	Run 9	Run 10	Mean	Standard Deviation
Accuracy	0.6783	0.6573	0.6958	0.6853	0.6853	0.6678	0.6783	0.6783	0.6853	0.6608	0.6723	0.0120
Sensitivity	0.7171	0.6579	0.7171	0.6842	0.7303	0.6645	0.7171	0.6973	0.6842	0.6579	0.6928	0.0270
Specificity	0.6343	0.6567	0.6716	0.6866	0.6343	0.6716	0.6343	0.6567	0.6866	0.6642	0.6597	0.0203
PPV	0.6899	0.6849	0.7124	0.7123	0.6938	0.6965	0.6899	0.6973	0.7123	0.6897	0.6979	0.0106
NPV	0.6640	0.6285	0.6767	0.6571	0.6746	0.6383	0.6640	0.6567	0.6571	0.6312	0.6529	0.0169

Table C.5 Number of MRI scans (sMCI and pMCI scans) correctly clustered using cosine similarity in k-means++ without including AD and CN scans.

	Run 1	Run 2	Run 3	Run 4	Run 5	Run 6	Run 7	Run 8	Run 9	Run 10	Mean	Standard Deviation
Negative (out of 134)	59	95	106	105	103	130	15	32	68	15	72.8	41.3
Positive (out of 152)	107	66	62	63	65	17	135	127	79	144	86.5	40.3

Table C.6 Measures of results of discriminating between sMCI and pMCI using cosine similarity in k-means++ without including AD and CN scans.

	Run 1	Run 2	Run 3	Run 4	Run 5	Run 6	Run 7	Run 8	Run 9	Run 10	Mean	Standard Deviation
Accuracy	0.5804	0.5629	0.5874	0.5874	0.5874	0.5140	0.5245	0.5559	0.5140	0.5559	0.5570	0.0300
Sensitivity	0.7039	0.4342	0.4079	0.4145	0.4276	0.1118	0.8882	0.8355	0.5197	0.9474	0.5691	0.2654
Specificity	0.4403	0.7090	0.7910	0.7836	0.7686	0.9701	0.1119	0.2388	0.5075	0.1119	0.5433	0.3085
PPV	0.5879	0.6286	0.6889	0.6848	0.6708	0.8095	0.5315	0.5546	0.5483	0.5475	0.6252	0.0889
NPV	0.5673	0.5249	0.5408	0.5412	0.5421	0.4906	0.4688	0.5614	0.4823	0.6522	0.5362	0.0525

Chapter 6 Discriminating Between Cognitively Normal and Progressive Mild Cognitive Impairment Scans

6.1. Introduction

Progressive mild cognitive impairment (pMCI) is an earlier stage of Alzheimer's disease (AD), as described in Section 1.4.2, and it plays an important role for delaying the development to AD because identifying people with pMCI allows for early detection of cognitive decline, which can be crucial for interventions. Thus it is crucial to distinguish pMCI from other stages, e.g., people who are cognitively normal (CN) and those who have AD.

In Chapter 5, the detailed information, results, and discussion of discriminating between stable mild cognitive impairment (sMCI) and pMCI were presented. The performance of discrimination between sMCI and pMCI was lower than that between scans from people who were CN and those with AD (Chapter 4), which indicated that sMCI and pMCI are more similar and they are more difficult to distinguish. Through discriminating between sMCI and pMCI, researchers can gain insights into the different degrees and patterns of cognitive decline. It is also important to differentiating between CN and pMCI because by studying this researchers can better understand the progression of cognitive decline. In this chapter, discriminating between CN and pMCI scans will be introduced and undertaken.

This chapter describes and presents detailed information on the operations of each phase in the workflow, including the software and tools used in this study to cluster CN and pMCI (Section 6.2), how the CN and pMCI Magnetic Resonance Imaging (MRI) scans were collected from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (Section 6.3), how the pre-processing tools were used (Section 6.4), how the two-sample t-test analysis was applied (Section 6.5), and how PCANet (Section 6.6) and k-means++ was implemented in Matlab (Section 6.7). The results of the experiments to discriminate between CN and pMCI MRI scans are presented (Section 6.8). The results are then discussed in Section 6.9.

6.2. Software, Tools, and Programming Languages

In Section 4.9.1, it was shown that Statistical Parametric Mapping (version 12, SPM12) outperformed in pre-processing, thus all MRI scans in this study were pre-processed using the Computational Anatomy Toolbox (version 12, CAT12) package in SPM. Similar to the previous two studies (CN vs. AD in Chapter 4; sMCI vs. pMCI in Chapter 5), the programming

language used was Matlab, because the SPM12 and its CAT12 package are based within Matlab. In addition, the MRI reading function provided by the FreeSurfer and the code of the PCANet were developed by Matlab, therefore it was convenient to read, operate MRI scans, and implement this task using them in Matlab.

6.3. Data Collection

In this study, pMCI scans were differentiated from CN scans using the equivalent dataset as those that were used in Section 5.3. The pMCI scans used in this study were the same as those used in Study 2 in Chapter 5 (sMCI vs. pMCI). The sMCI and pMCI scans were divided from MCI scans based on the diagnostic summary provide by the Alzheimer's Disease Neuroimaging Initiative (ADNI). As for the CN scans, they were the same as those used in the study 1 in Chapter 4 (CN vs. AD). Therefore, the number of CN scans used was 231 and the number of pMCI scans used was 152, which are respectively the same as the number of them in the previous two chapters. Table 6.1 shows a summary of the dataset used in this task. As described in later Sections 6.6 and 6.7, AD scans were also used in this study, and the number of the AD scans used in this study was 198 (the same as that in Study 1).

Table 6.1 The number of MRI scans of CN and pMCI

	CN	pMCI	AD	Total
Number of MRI scans	231	152	198	581

6.4. Data Pre-processing

All MRI scans used in this study were pre-processed using the CAT12 package in SPM12. The settings were the defaults and were the same as that in Section 3.4.2. The pre-processing in CAT12 package is called "segment" and the outputs of each MRI scan via "segment" are a grey matter scan without the skull, a white matter scan without the skull, a whole brain scan without the skull, and other text files. As with Studies 1 and 2, the whole brain scans without skulls were still used in this study. In addition, all output scans were motion corrected, normalised, and registered into a common space, i.e., Montreal Neurological Institute (MNI) template. The size of the output scans was 121×145×121 voxels (sagittal, coronal, and axial planes, respectively).

6.5. Detection of Differences Between CN and pMCI

Similar to the Section 4.5, SPM12 was used to apply the two-sample t-test to scans of CN and scans of pMCI to detect the differences between the two groups at the voxel level. According

to the description in Section 3.5.1, a t-test is generally applied to a small size of data, which is often less than 30 (Gerald, 2018). In addition, it also needed to ensure that the scans used for detecting differences did not account for a large proportion of the whole dataset. Therefore, a two-sample t-test was applied to 15 scans of pMCI and 25 scans of to detect differentiated regions between the two groups. This ensured that the numbers of scans of the two groups did not exceed 30 and did not account for large proportions of the two groups (15 scans of pMCI and 25 scans of CN account for approximately 10% of each group, respectively). This was because the aim of this study was to develop tools that could diagnose different stages of AD using a limited amount of data.

6.6. Feature Extraction

In this study, the PCANet was also used to extract features from the ROIs. Distributions of the features in two scenarios were compared:

(1) Features (voxel intensity values) from the ROIs and then computed by PCANet only from CN and pMCI scans;

(2) Features (voxel intensity values) from the ROIs and then computed by PCANet from CN, pMCI, and AD scans.

The inclusion of AD scans in the scenario (2) was to test whether the inclusion of AD scans can improve the separation of the CN and pMCI scans in the feature space. Similar to Sections 4.6 and 5.6, the distributions were demonstrated using the t-distributed stochastic neighbour embedding (t-SNE) technique.

In this study, the parameters of the PCANet were the same as those used in Sections 4.6 and 5.6:

- the number of stages was set to two;
- the size of the patches $k_1 \times k_2$ was set to 3×3 in both two stages;
- the number of filters L_1 and L_2 were both set to eight, respectively;
- the block size for the histograms was set to 15×15;
- the overlap ratio of the local blocks was set to zero in this study for reducing the number of features.

The reasons for the choice of these parameters were discussed in Section 3.6.4., and the same applied in this study.

6.7.   Clustering

In order to discriminate between the two groups (i.e., CN and pMCI), k-means++ clustering was used, in the same manner as described in Section 3.7. Similar to Section 6.6, the performances of two clustering experiments in two scenarios were compared:

(1) Clustering feature extracted only from CN and pMCI;

(2) Clustering features extracted from three groups: CN, pMCI, and AD.

The inclusion of AD scans in the process (clustering scenario (2)) was to test whether the inclusion of AD scans can enhance the performance of the discrimination between CN and pMCI scans.

Owing to an improved performance in Study 2 (discriminating between sMCI and pMCI) in Section 5.9.4, cosine similarity (compared to Euclidean distance) was used again in the k-means++ algorithm for discriminating between CN and pMCI. Thus, the function *kmeans(X,2,'Distance','cosine')* was used in Matlab to run the k-means++ to partition the observation of the data matrix (*X*) into two clusters with cosine similarity. Due to the different results in different runs, the k-means++ algorithm was run ten times.

6.8.   Performance Evaluation

For the clustering of CN and pMCI, when the data of CN, pMCI, and AD were used (clustering scenario (2)), the AD data were regarded as the known data (labelled data). Therefore, for the unknown scans (CN and pMCI scans) that were clustered into the same group as the AD scans, the data were identified as pMCI while the other scans were identified as CN.

If only CN and pMCI scans are used (clustering scenario (1)), the approach to identifying clusters was the same as that for identifying the AD and CN clusters in Section 4.8, where the majority of the scans in the cluster determined the appropriateness of membership.

After identifying the clusters, the same as Sections 4.8 and 5.8, mean values (across the ten runs) of five measures: the accuracy, sensitivity, specificity, PPV, and the NPV, were used to evaluate the overall performance of the method presented in this chapter. 95% confidence intervals (95% CI) of the five measures were also used to describe the stability and the variance of the performance of the method.

6.9. Results

6.9.1. ROIs Between CN and pMCI

As described in Section 6.5, to test whether there were obvious ROIs that could help differentiate between CN and pMCI scans, a two-sample t-test was applied to scans of pMCI and CN. Figure 6.1 (a) shows the differented regions (regions of interest, ROIs) at the voxel level between CN and pMCI using 25 CN scans and 15 pMCI scans. The colours in the figure have the similar meanings as those in Figure 4.7: i.e., the red and orange regions indicate that the mean of the intensity value (after being normalised) of the voxel in the pMCI group is greater than that in the CN group, and the blue areas (which are very tiny in the Figure 6.1 (a)) indicates where the mean intensity value of the voxel in the pMCI group is less than that in the CN group.

(a)
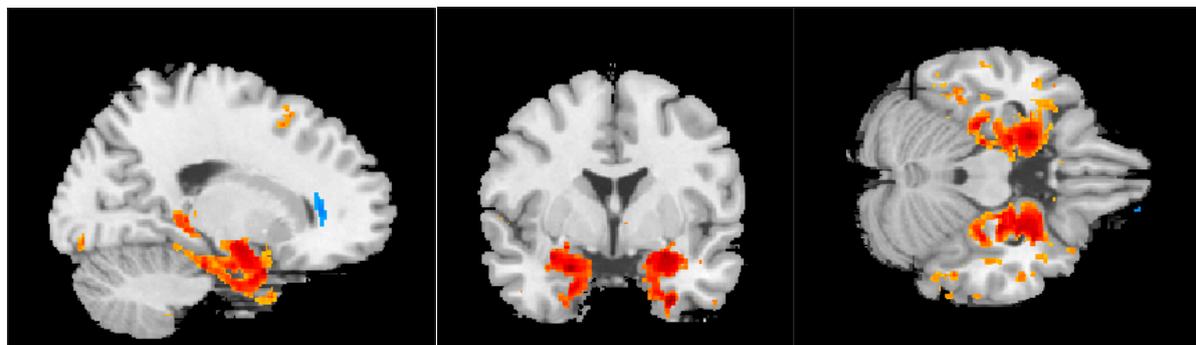
(b)

Figure 6.1 (a) Three view planes (sagittal, coronal, and axial planes) of ROIs detected using 25 CN scans and 15 pMCI scans; (b) Three view planes (sagittal, coronal, and axial planes) of ROIs detected in Study 1 (i.e., Figure 4.7) (MRI images from Xjview adapted by Yuyang Liu)

Similar to Figure 4.7 (Figure 6.1(b)), as can be seen in Figure 6.1 (a), the most obvious ROIs between pMCI and CN are also lower parts of the limbic system, which include the parahippocampal gyrus, the amygdala, and the hippocampus, in both left and right cerebra. However, the colour around these parts is mainly in orange in Figure 6.1 (b), which is lighter

than the colours in Figure 6.1 (b) (which were mainly in red and deep red). This indicates that the mean differences within these regions (ROIs) between pMCI and CN scans were less than those between AD and CN. This will be further explained and discussed in Section 6.10.1 (Results Analysis).

This study visually demonstrated that the different regions between CN and pMCI scans were similar to those between CN and AD. Since the regions were observed to be similar, it was decided that the location of the ROIs using CN and AD scans were used in discrimination between CN and pMCI. In this way, none of labelled pMCI scans were needed in this study, and this can reduce the necessity of labelled scans.

In addition to the obvious red and orange regions, there are still some extremely tiny areas in light blue in Figure 6.1 (a). However, as can be seen in the sagittal plane in Figure 6.1, the light blue area is outside of the brain. This might be because some tiny parts (e.g., skulls) outside the brain were not fully removed in the pre-processing stage. Therefore, these areas are presumed to be less relevant for this study and did not need to be considered further.

6.9.2. Distributions of Features

As mentioned in Section 6.9.1, the regions detected in the CN and pMCI groups were similar to those identified in the CN and AD groups (this will be explained and discussed in Section 6.10.1). Therefore, regions detected in the Section 4.9.2.2 were used to extract features by PCANet for discriminating between CN and pMCI scans. The distribution of the extracted features in the two scenarios (presented in Section 6.6) is presented in Figure 6.2. Figure 6.2 (a) shows a distribution of features extracted by PCANet from (voxel intensity values of) the ROIs of the CN and pMCI scans alone, and Figure 6.2 (b) shows a distribution of features computed by PCANet from (voxel intensity values of) the ROIs of three groups: CN, pMCI, and AD. In the medical field, a negative result means that the disease tested has not been identified, thus CN was named as negative in the figure; in contrast, pMCI and AD were named as positive in Figure 6.2 (b).

As can be seen in Figure 6.2 (a), CN data and pMCI data are relatively well visually separable in the space, although a few of them are mixed: the red points (CN data) are broadly gathered above and to the left of the light blue points (pMCI data). After adding AD data into the CN and pMCI data, the distribution of the two groups (negative and positive) are shown in Figure 6.2 (b). It can be seen that the two groups are slightly more visually separated compared with

Figure 6.2 (a), however more light blue points (pMCI and AD data) are mixed in with the red points (CN data).

(a)   (b)
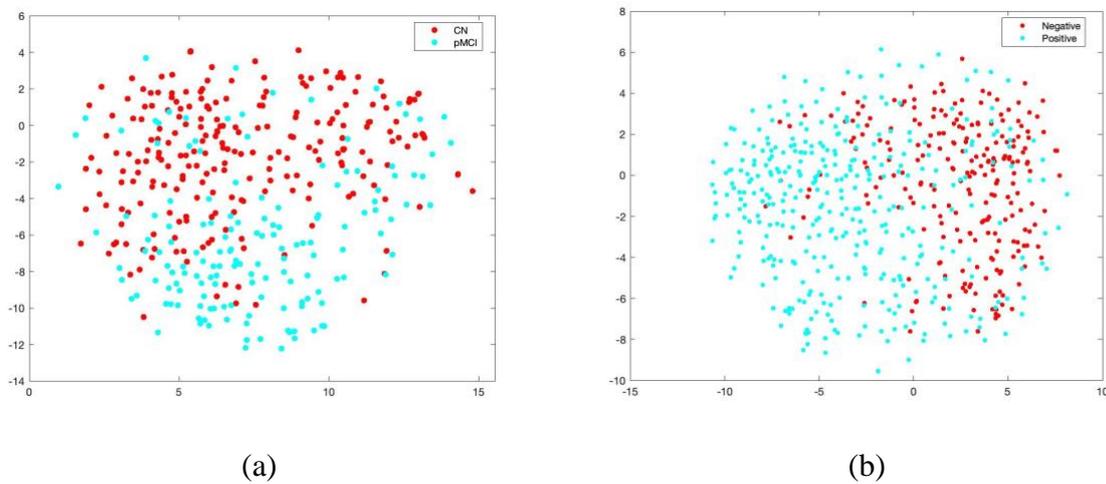
Figure 6.2 Distributions of features. (a) is the distribution of features of ROIs from the CN and pMCI scans computed using PCANet; (b) is the distribution of features of ROIs from negative (CN) and positive (AD and pMCI) scans computed using PCANet (Developed by Yuyang Liu)

6.9.3. Clustering Results

After the features were extracted from the ROIs mentioned in Section 6.6 using PCANet, the computed features were distinguished by using k-means++. As mentioned in Section 6.7, two experiments in two scenarios were conducted:

(1) Clustering feature extracted only from CN and pMCI;

(2) Clustering features extracted from three groups: CN, pMCI, and AD.

Table 6.2 shows the results of five measures in the two scenarios, and Figure 6.3 shows a bar graph indicating the results of the five measures in two situations. When the pMCI and CN scans were clustered alone, the method reported in this chapter yielded an accuracy of 0.752, a sensitivity of 0.768, a specificity of 0.742, a PPV of 0.662, and a NPV of 0.829. When scans of AD were included into the scans of pMCI and CN, three out of the five measures improved: the accuracy increased to 0.776, the specificity to 0.821, and the PPV to 0.724. However, the sensitivity and NPV were slightly reduced, i.e., to 0.714 and 0.815, respectively. This will be discussed in Section 6.10.1. However, although, from the mean values, the inclusion with the AD scans helped enhance the performance of discriminating between CN and pMCI, it can be seen that, from the 95% CI, all measures except accuracy had wider ranges when AD scans were added into the dataset. Specifically, the 95% CI of sensitivity was reduced from 0.675 to 0.752 when the AD scans were included, which was twice of the value when AD scans were

not included (ranged from 0.752 to 0.784). Apart from this, the mean value of the sensitivity (0.714) when including AD scans was also less than that when not including AD scans (0.768). Based on both of the mean value and 95% CI, it can be concluded that the inclusion with AD scans was not greatly helpful for recognising pMCI scans (due to the lower sensitivity and wide range of 95% CI) in the task of discriminating between CN and pMCI scans.

Table 6.2 Measures of results using k-means++ (using cosine similarity) including and not including AD data (with cosine similarity)

Measures		K-means++	
		Not including AD data	**Including AD data**
Accuracy	**Mean**	0.752	0.776
	95% CI	(0.742, 0.762)	(0.768, 0.783)
Sensitivity	**Mean**	0.768	0.714
	95% CI	(0.752, 0.784)	(0.675, 0.752)
Specificity	**Mean**	0.742	0.821
	95% CI	(0.73, 0.754)	(0.781, 0.86)
PPV	**Mean**	0.662	0.724
	95% CI	(0.653, 0.671)	(0.695, 0.752)
NPV	**Mean**	0.829	0.815
	95% CI	(0.82, 0.838)	(0.799, 0.83)
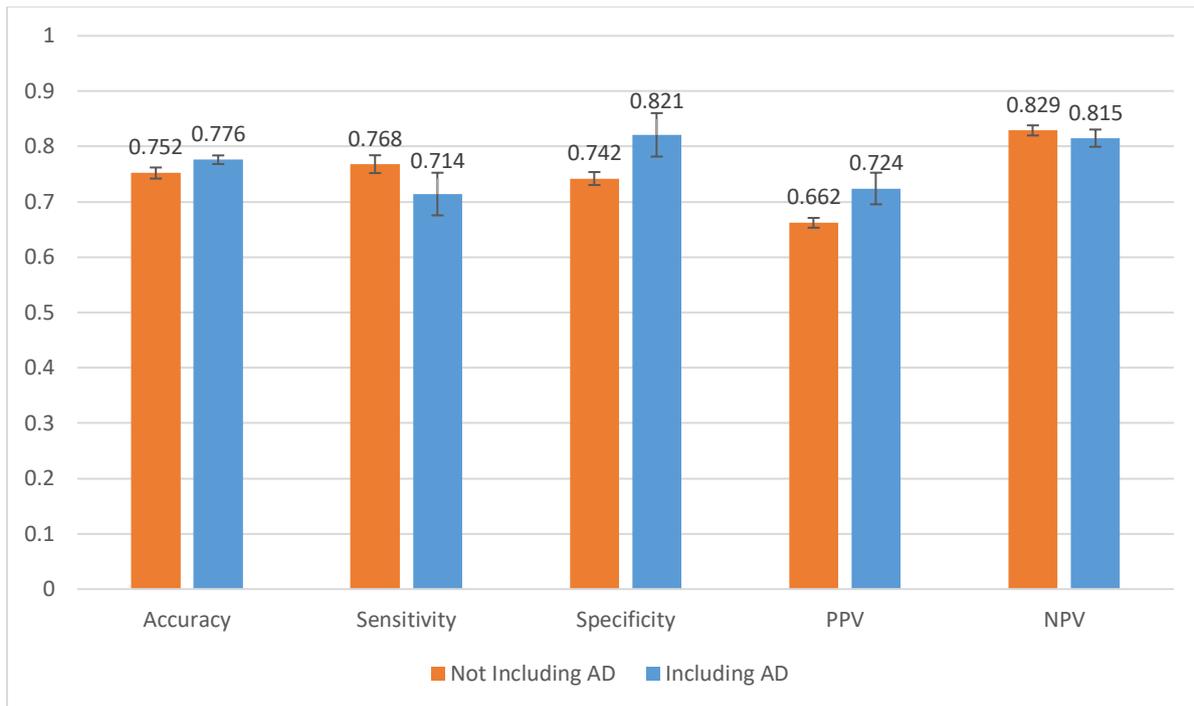
Figure 6.3 A bar graph indicating the mean values and 95% CI of each measure across the 10 runs (Developed by Yuyang Liu)

In contrast, including AD data assisted with differentiating CN scans from pMCI scans. As can be seen in Table 6.2, the specificity increased considerably (i.e., from 0.742 to 0.821) when AD data were included, although the 95% CI had a wider range. However, the lower bound of the 95% CI (0.781) was still higher than upper bound (0.754) when only using CN and pMCI scans, which indicates that these differences were statistically significant. Similarly, when including the AD scans the PPV also increased, compared with not including AD scans, i.e., the PPV increased from 0.662 to 0.724. It also had a wider 95% CI and the lower bound (0.695) was also higher than upper bound (0.671) of the 95% CI when not including AD scans, again indicating that the difference was statistically significant. This suggests that the inclusion of AD scans is helpful for increasing the reliability of the pMCI results diagnosed using the proposed method.

Finally, the mean values of the NPV in the two situations were close: these were 0.815 when AD data were included, and 0.829 when they were not included. Therefore, both scenarios achieved relatively reliable results, if patients received CN results from the proposed method in this study.

In the following section, Section 6.10, the analysis and the probable reasons for these results are discussed.

6.10. Discussion

In Section 6.9, the results of the analyses of the ROIs between CN and pMCI, the distribution of the extracted features, and the clustering results from them were presented in Figure 6.1, Figure 6.2, and Table 6.2, respectively. In this section, the results will be discussed and compared with previous state-of-the-art studies.

6.10.1. Results Analysis

From Table 6.2, it can be seen that, even if the AD data were not included, using pMCI and CN scans alone achieved a relatively good performance (an accuracy of 0.752) in the discrimination between the pMCI and CN scans, particularly when compared with the discrimination between the sMCI and pMCI reported in Chapter 5 (Section 5.9.2 and Section 5.9.3). This was demonstrated clearly in Figure 6.2. Figure 6.2 (a) shows that CN data and pMCI data are relatively well visually separable in the space even though the AD data were not included. This also indicates that the ROIs used in this study are helpful and meaningful for discriminating between CN and pMCI scans. Figure 6.2 (b) shows that, as described in Section 6.9.2, more of the light blue points (pMCI and AD cases) are mixed into the red points (CN cases), even though the two groups are more visually split. This could partially explain why the sensitivity dropped from 0.768 to 0.714 after the AD data were included: more pMCI and AD data were clustered into the CN group and thus the sensitivity decreased.

From Table 6.2, it can be seen that the accuracy of including AD data (0.776) was 0.2 higher than when not including the AD data (0.752), and the main contribution to this rise was the large increase in the number of correctly clustered CN cases, which can also be seen from the increase in specificity (increased from 0.742 to 0.821) in Table 6.2. The reason why the CN scans were discriminated into the correct cluster could be because the inclusion of AD data magnified the differences between the two groups. As introduced in Chapter 1, having pMCI is an intermediate state between a person being CN and them developing AD later on, thus the difference between scans from CN people and scans from people with pMCI is likely to be less than that between scans from people who are CN and those with AD. This can be demonstrated by the fact that the colour within the ROIs in Figure 6.1 (a) is lighter than that within the ROIs in Figure 6.1 (b) (Figure 4.7). This indicates that, as mentioned in Section 6.9.1, differences in the mean values of voxels within the ROIs in the pMCI and CN scans are less than those in the AD and CN scans. Therefore, when only CN and pMCI scans are used for discrimination, some CN data might be mis-clustered into the pMCI group due to their relatively small differences; when AD scans are included into the pMCI, some of the aforementioned mis-clustered CN data

could be correctly clustered due to the enlarged differences between the two groups, and thus the specificity was improved.

In contrast, the inclusion with AD data had a negative influence on recognising pMCI scans, causing a drop of sensitivity. This is probably because, although the inclusion of AD scans enhanced the performance in recognising CN scans, some pMCI scans might appear to be more similar to the CN scans after the AD scans were included. Some early-phase pMCI scans, whose group membership is relatively ambiguous, might be between CN and late-phase pMCI scans in the space, and were clustered into pMCI by this method before the AD scans included; however, after AD scans were included, the late-phase pMCI scans might become distant to the CN scans while the early-phase pMCI scans might remain in the same position or be closer to the CN scans in the space. As a result, they were clustered into the CN group, leading to a drop of sensitivity. After some of those ambiguous pMCI scans were clustered into the CN group, the clustering of the remaining pMCI (which are closer to AD) was more certain, and fewer CN scans were mistakenly clustered into the pMCI group. As a consequence, the results of clustering pMCI was more reliable and the PPV was higher. As for the high NPV, it indicates that the proposed method is reliable in recognising CN no matter whether the AD data are included or not. In addition, the possible reasons above may also have caused an imbalance in the sensitivity and specificity when the AD scans were included in certain runs: the extremely high specificity was with relatively low sensitivity, e.g., run 6 in Table D.4 (sensitivity=0.638, specificity=0.939) in the Appendix D. This lead to an unstable performance in the discrimination.

In summary, the inclusion of AD scans enhanced the overall performance in discriminating between CN and pMCI scans: three (accuracy, specificity, and PPV) out of the five measures increased; however, it also led to a less stable performance (based on the wider ranges of the 95% CI across all measures) compared with the method when AD scans were not included.

6.10.2. Comparison with State-of-the-art Studies

In the task of discriminating between CN and pMCI, the method reported here achieved a relatively good performance, and it (or at least some of the measures) outperformed some previous state-of-the-art studies (e.g., Samper-González et al., 2018; Tong et al., 2014; Wen et al., 2021; Zheng et al., 2022) that were mentioned and described in the literature review (Section 2.5). Based on Section 2.5.1.3 in the literature review chapter, no study was identified that has applied unsupervised learning methods (no study used semi-supervised learning method was

identified as well) for the discrimination of CN and pMCI, however, some studies using supervised learning methods were identified and are listed in Table 6.3.

Table 6.3 Comparison of our proposed study and previous studies (supervised learning) that discriminated between CN and pMCI.

Learning type	Study	ACC	SEN	SPE	PPV	NPV	Dataset	Method
Unsupervised learning	**Study in this chapter**	**0.776**	**0.714**	**0.821**	**0.724**	**0.815**	**231 CN, 152 pMCI**	**PCANet, k-means++, inclusion of AD**
		0.752	**0.768**	**0.742**	**0.662**	**0.829**		**PCANet, k-means++, no inclusion of AD**
Supervised learning	(Zheng et al., 2022)	0.8667	0.75	0.9091	-	-	218 CN, 80 pMCI	3D EfficientNet
	(Zeng et al., 2021)	0.9667	-	-	-	-	92 CN, 95 pMCI	DBN-based multi-task learning algorithm
	(Wen et al., 2021)	0.74±0.118	-	-	-	-	46 CN, 24 pMCI	Linear SVM
	(Elahifasaee et al., 2019)	0.8429	0.904	0.7985	-	-	229 CN, 167 pMCI	Feature decomposition, kernel discrimination analysis
	(Sujathakumari et al., 2019)	0.9772	0.9589	0.9902	0.986	-	205 CN, 146 pMCI	Gradient boosting classifier
	(Zeng et al., 2018)	0.8571	-	-	-	-	92 CN, 95 pMCI	SDPSO-SVM
	(Samper-González et al., 2018)	0.77	-	-	-	-	282 CN, 167 pMCI	Linear SVM
	(Tong et al., 2014)	0.829	0.689	0.931	0.878	0.805	231 CN, 167 pMCI	Multiple instance graph

As can be seen from Table 6.3, Zeng et al. (2021), Wen et al. (2021), Zeng et al. (2018), and Samper-González et al. (2018) only reported overall accuracies in their studies. However, as mentioned earlier, using accuracies alone cannot evaluate the performance of methods

comprehensively, because it cannot reflect the performance in identifying CN and pMCI scans, respectively. The study presented in this chapter provided five measures to evaluate the performance and this made the evaluation more thorough.

In addition, Wen et al. (2021) was the only study in Table 6.3 which provided the standard deviation value. It is also important to provide standard deviation values or confidence intervals because they can reflect the stability of the results across multiple runs. However, the standard deviation of the accuracy yielded by Wen et al. (2021) was 0.118 (11.8%). This means that the accuracies in Wen et al.'s (2021) study fluctuated strongly. However, the two clustering scenarios (including and not including AD scans) of the study presented in this chapter both achieved narrow range of 95% CI. This indicates that Study 3 in this research had a more stable performance than Wen et al.'s (2021).

In addition to accuracies, B. Zheng et al. (2022) and Elahifasaee et al. (2019) reported sensitivities and specificities in their studies. Both of them achieved high accuracies, which were 0.8667 and 0.8429 respectively, but their sensitivities and specificities were unbalanced: the differences between the two measures were more than 0.1 (10%) in their studies. Especially B. Zheng et al. (2022), the dataset they used was extremely unbalanced : the number of CN scans (218 scans) was nearly three times than that of pMCI scans (80 scans). Their unbalanced sensitivity (0.75) and specificity (0.9091) indicates that the higher number of CN scans contributed more to the high accuracy (0.8667) in their study. This can explain the reason why only using the accuracy cannot fully evaluate the performance. Study in this chapter provided a balanced sensitivity (0.768) and specificity (0.742) when including labelled AD scans.

In Table 6.3, only Tong et al. (2014) provided both PPV and NPV (although Sujathakumari et al. (2019) only provided PPV) in their studies. Sujathakumari et al. (2019) achieved an advantageous performance, where all four measures (accuracy, sensitivity, specificity, and PPV) were above 0.9. The possible reason is that they involved personal characteristic data into MRI scans to help the classification. Tong et al. (2014) achieved high and balanced PPV (0.878) and NPV (0.805), but their sensitivity (0.689) and specificity (0.931) were extremely unbalanced. This indicates that the capability of their method to identify pMCI scans was weak. However, identifying people with pMCI is important for clinical professionals because pMCI is a key stage to intervene the development to AD.

6.11. Conclusion

In this chapter, the detailed operations and information of the methodology for clustering CN and pMCI scans were described, and the results of the clustering were presented and discussed. In the data collection phase, pMCI scans were obtained in the same way as in Section 5.3. After using the two-sample t-test applied to 25 CN scans and 15 pMCI scans, it was found that the ROIs between the two groups were similar to those between CN and AD, only the mean values of voxels within the ROIs were different. Therefore, the clustering performance of including and not including AD data into CN and pMCI data was compared. The results showed that three (accuracy, specificity, and PPV) out of the five measures were higher when AD scans were included into the dataset, but it was less stable than not including AD scans (using CN and pMCI scans alone) based on the 95% CI values. This means that including AD data had a positive influence on the evaluation measures but had negative influence on the stability.

In the comparison with other state-of-the-art studies (Section 6.9), most studies had comparable performance to the studies using supervised and semi-supervised learning methods, which could learn well from training data. However, this study employed unsupervised learning methods to discriminate between CN and pMCI in order to reduce the need for having labelled data. Although some of them outperformed than the method reported here, they required a relatively large number of training data. In other words, the method reported here greatly decreased the dependence on the labelled data by sacrificing some performance. In addition, some of the previous studies only reported one measure (accuracy) to evaluate their methods, which is not sufficient, whereas this study reported more measures to evaluate the methods more fully. Unbalanced data were also used in some of those previous studies, and they achieved unbalanced sensitivity and specificity. In contrast, the method reported here achieved a balanced sensitivity and specificity, although the two groups in this study were not extremely well balanced.

In Chapters 4, 5, and 6, the three tasks were presented and discussed in depth, thus a horizontal discussion over the three tasks is necessary. In the next chapter, a discussion of the whole study reported in this thesis (for the three tasks described in Chapters 4, 5, and 6) will be presented. The results across the three tasks will be discussed and analysed and the research questions of this study will be answered. In addition, the overall limitations of the overall study will also be presented.

Appendix D

Table D.1 Number of MRI scans (CN and pMCI scans) correctly clustered in k-means++ without inclusion of AD scans (using cosine similarity).

	Run 1	Run 2	Run 3	Run 4	Run 5	Run 6	Run 7	Run 8	Run 9	Run 10	Mean	Standard Deviation
Negative (out of 231)	169	176	174	167	167	174	168	173	167	179	171.4	4.351
Positive (out of 152)	111	114	116	120	124	113	121	117	116	115	116.7	3.945

Table D.2 Measures of results of discriminating between CN and pMCI in k-means++ without inclusion of AD scans (using cosine similarity).

	Run 1	Run 2	Run 3	Run 4	Run 5	Run 6	Run 7	Run 8	Run 9	Run 10	Mean	Standard Deviation
Accuracy	0.7311	0.7572	0.7572	0.7493	0.7598	0.7493	0.7546	0.7572	0.7389	0.7676	0.7522	0.0162
Sensitivity	0.7303	0.75	0.7632	0.7895	0.8158	0.7434	0.7961	0.7697	0.7632	0.7566	0.7678	0.0260
Specificity	0.7316	0.7619	0.7532	0.7229	0.7229	0.7532	0.7273	0.7489	0.7229	0.7749	0.7420	0.0188
PPV	0.6416	0.6746	0.6705	0.6522	0.6596	0.6647	0.6576	0.6686	0.6444	0.6886	0.6622	0.0143
NPV	0.8048	0.8224	0.8206	0.8392	0.8564	0.8169	0.8442	0.8317	0.8227	0.8287	0.8288	0.0148

Table D.3 Number of MRI scans (CN and pMCI scans) correctly clustered in k-means++ with inclusion of AD scans (using cosine similarity).

	Run 1	Run 2	Run 3	Run 4	Run 5	Run 6	Run 7	Run 8	Run 9	Run 10	Mean	Standard Deviation
Negative (out of 231)	187	171	182	199	172	207	197	176	196	199	188.6	12.782
Positive (out of 152)	108	125	108	101	119	97	106	119	100	102	108.5	9.443

Table D.4 Measures of results of discriminating between CN and pMCI in k-means++ with inclusion of AD scans (using cosine similarity).

	Run 1	Run 2	Run 3	Run 4	Run 5	Run 6	Run 7	Run 8	Run 9	Run 10	Mean	Standard Deviation
Accuracy	0.7702	0.7728	0.7572	0.7833	0.7598	0.7937	0.7911	0.7702	0.7728	0.7859	0.7757	0.0125
Sensitivity	0.7105	0.8224	0.7105	0.6645	0.7829	0.6382	0.6974	0.7829	0.6579	0.6711	0.7138	0.0621
Specificity	0.8095	0.7403	0.7879	0.8615	0.7446	0.9394	0.8528	0.7619	0.8485	0.8615	0.8208	0.0634
PPV	0.7105	0.6757	0.68	0.7594	0.6685	0.8017	0.7571	0.6839	0.7407	0.7612	0.7239	0.0462
NPV	0.8095	0.8636	0.8053	0.796	0.8390	0.7901	0.8107	0.8421	0.7903	0.7992	0.8146	0.0251

Chapter 7 Discussion

7.1.   Introduction

Chapter 6 presented the work of discriminating between cognitively normal (CN) and progressive mild cognitive impairment (pMCI) magnetic resonance imaging (MRI) scans. All three discrimination tasks (CN vs. Alzheimer's disease, stable mild cognitive impairment vs. pMCI, and CN vs. pMCI) were presented and discussed individually in the previous three chapters. Therefore, all three discrimination studies were presented and discussed individually in the previous three chapters. This chapter brings together findings from all the three discrimination tasks, discussing the broader study as a whole. Section 7.2 answers the research questions posted in Section 1.3 based on the results in Chapter 4 to 6. Section 7.3 links and discusses the results across the three studies in this thesis. Section 7.4 summarises the limitations of the whole study. Specifically, Section 7.4.1 presents limitations of the research design; Section 7.4.2 presents the limitations of the methodology and the solutions to the limitations; Section 7.5 concludes this chapter.

7.2.   Revisiting the Research Questions

In Section 1.3, two research questions were posted. The questions with their answers are as follows:

1)   How can the proposed method discover regions that are helpful to discriminating between CN and AD; discriminating between sMCI and pMCI; discriminating CN and pMCI?

As mentioned in Section 3.5.1, a two-sample t-test compared 30 scans of CN and 30 scans of AD to detect differences at voxel level between the two groups. This identified the lower parts of the limbic system, including the parahippocampal gyrus, the amygdala, and the hippocampus, in both left and right cerebra appeared significantly different in the scans from people with AD compared to scans from healthy people. As could be seen from Figure 4.8 (a)-(c), these regions (shown in Figure 4.7) were helpful in discriminating between CN and AD.

However, when a similar approach was applied to sMCI and pMCI scans, there was no obvious detected difference at the voxel level between the two groups, as shown in Figure 5.3. Even if increasing the number of the scans for the two-sample t-test, the detected regions were small and diffuse, as shown in Figure 5.4. However, it more or less could be seen that the meaningful regions for discriminating between sMCI and pMCI were also around the lower parts of the

limbic system. Therefore, it has been realised that the regions detected in Study 1 (CN vs. AD) could be transferred to those used in discriminating between sMCI and pMCI scans.

A similar approach (applying a two-sample t-test to 25 CN scans and 15 pMCI scans) was used in discriminating between CN and pMCI scans. As could be seen from Figure 6.1 (a), this identified that the ROIs between CN and pMCI scans were similar to those between CN and AD scans , which include the parahippocampal gyrus, the amygdala, and the hippocampus, in both left and right cerebra. Therefore, the regions detected in Study 1 (CN vs. AD) were transferred to those used in Study 3 (discriminating between CN and pMCI scans). On the heat map of Figure 6.1 (a), the colours of the voxels within the ROIs between CN and pMCI scans were lighter than that between CN and AD scans in Figure 6.1 (b). It is reasonable because AD is more severe than pMCI and the difference between CN and AD scans should be larger than that between CN and pMCI scans.

The regions used in the three studies (CN vs. AD, sMCI vs. pMCI, and CN vs. pMCI) were shown to be helpful for the discriminations based on the distributions of the features and the performances; these have been reported in Sections 4.9.3, 5.9.3, and 6.9.2, respectively.

2) How is the proposed method compared with existing supervised, semi-supervised, and unsupervised learning methods?

Compared with previous studies using unsupervised learning method, the methods presented in this thesis achieved a better performance in discriminating between CN and AD scans: as can be seen from Table 4.7, the performance of the method used in Study 1 was higher than three out of four listed previous studies that used unsupervised learning methods. Although Bi et al. (2019) had a higher accuracy than the method reported in Chapter 4, they did not provide other measures, which could not comprehensively evaluate the performance. In addition, the method reported here detected the AD-relevant regions but Bi et al. (2019) did not. To the best of my knowledge, Studies 2 (sMCI vs. pMCI) and 3 (CN vs. pMCI) were the first studies to employ unsupervised learning methods to discriminate between sMCI and pMCI scans, and to discriminate between CN and pMCI scans, respectively. The results of the two studies could potentially become baseline studies that were compared with future studies using unsupervised learning methods for the two discrimination tasks (sMCI vs. pMCI; CN vs. pMCI).

When compared with previous studies using supervised and semi-supervised learning methods, the methods presented in Chapters 4 to 6 achieved favourable performances, although some supervised and semi-supervised learning methods outperformed my methods. However, the value of the methods reported in this thesis came from not needing exhaustive efforts in training models, which was timesaving. The methods made a trade-off between efficiency (e.g., collecting less labelled MRI scans, training models, etc.) and performances. In addition, it was difficult to conduct a thorough comparison with previous studies because many of the studies did not report all the measures (accuracy, sensitivity, specificity, positive predictive values, and negative predictive values).

7.3. Results in the Three Studies

In Chapters 4 to 6, the details and the results of discrimination between CN and AD, between sMCI and pMCI, and between CN and pMCI were presented, respectively. From the results, the three studies have relationships with each other. Table 7.1 summarises the best performances of the three studies (CN vs. AD, CN vs. pMCI, and sMCI vs. pMCI) from Tables 4.3, 6.2, and 5.2, respectively.

Table 7.1 A summary of the best performances of the three studies (CN vs. AD, CN vs. pMCI, and sMCI vs. pMCI)

Measures		Three Studies			
		CN vs. AD	CN vs. pMCI		sMCI vs. pMCI
			Including AD	Not Including AD	
Accuracy	Mean	0.842	0.776	0.752	0.672
	95% CI	(0.833, 0.851)	(0.768, 0.783)	(0.742, 0.762)	(0.665, 0.679)
Sensitivity	Mean	0.797	0.714	0.768	0.693
	95% CI	(0.768, 0.826)	(0.675, 0.752)	(0.752, 0.784)	(0.676, 0.71)
Specificity	Mean	0.881	0.821	0.742	0.660
	95% CI	(0.853, 0.909)	(0.781, 0.86)	(0.73, 0.754)	(0.648, 0.672)
PPV	Mean	0.855	0.724	0.662	0.698
	95% CI	(0.829, 0.881)	(0.695, 0.752)	(0.653, 0.671)	(0.691, 0.705)
NPV	Mean	0.836	0.815	0.829	0.653
	95% CI	(0.819, 0.853)	(0.799, 0.83)	(0.82, 0.838)	(0.643, 0.663)

As can be seen from Table 7.1, the overall performances progressively decreases as the three discriminations (CN vs. AD, CN vs. pMCI, and sMCI vs. pMCI) were implemented: the mean accuracy was 0.842 for discrimination between CN and AD; the mean accuracies were 0.776

and 0.752 for discrimination between CN and pMCI using two different strategies (including and not including AD scans); the mean accuracy was 0.672 for discrimination between sMCI and pMCI. This indicates that the difficulty increased across the three studies. This is not unreasonable because, as introduced in Section 1.4.1, sMCI and pMCI are two intermediate phases between CN and AD, and as shown in Figure 7.1, the difference between CN and pMCI is not as significant between CN and AD (Bondi et al., 2017), especially in the brain structure. In Figure 4.7 (also in Figure 6.1 (b)), the coloured ROIs were mainly in red and dark red while in Figure 6.1 (a), the ROIs were in orange, which is lighter than red. In Xjview viewer, darker colour indicates the greater difference in the mean values of the voxels. Therefore, the mean values of the voxels within the ROIs between CN and pMCI were indeed less than those between CN and AD. Hence, compared with discriminating between CN and AD, it was more difficult to discriminate between CN and pMCI, and naturally the performance in distinguishing between CN and pMCI was lower than that in distinguishing between CN and AD.
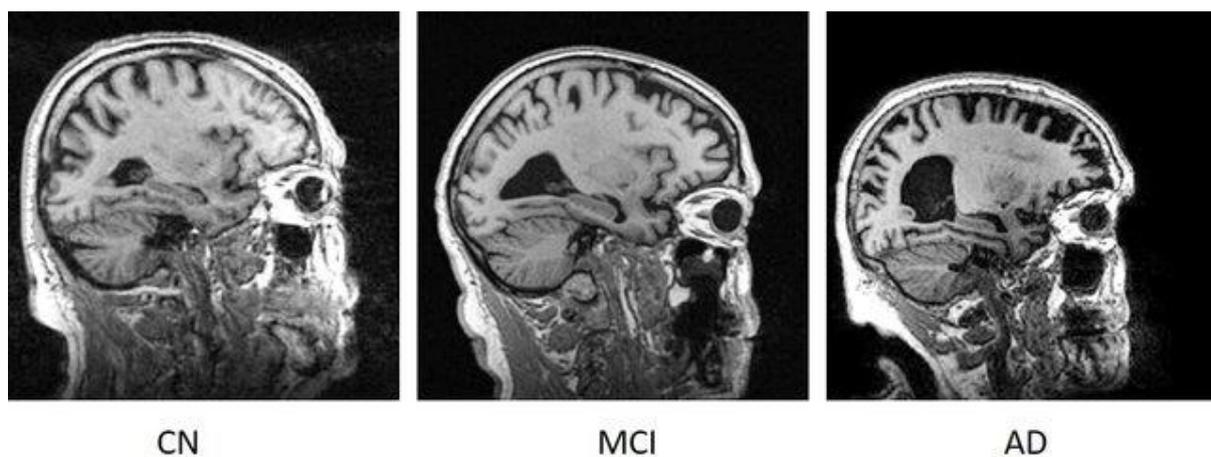
Figure 7.1 Sample brain MRI scans of a CN, MCI, and AD patient (This image is licensed under t-he Attribution 4.0 International (CC BY 4.0). [available at: https://www.researchgate.net/publication/36754680 6_An_Approach_for_Classification_of_Alzheimer%27s_Disease_Using_Deep_Neural_Network_and_Brain_M agnetic_Resonance_Imaging_MRI])

From the CN vs. AD column and the CN vs. pMCI column (including AD scans) in Table 7.1, it can be seen that sensitivities in both two columns are lower than specificities. This indicates that the method had better capabilities to recognise CN scans. In other words, AD or pMCI scans were more difficult to recognise. In addition, from Figure 4.8 (d) and Figure 6.2 (b), more AD and pMCI points are mixed into CN points. This was probably due to the following reasons:

(1) The ROIs selected and used in the two studies (CN vs. AD and CN vs. pMCI) were not sufficient to include all AD-relevant or pMCI-relevant regions;

(2) The two-sample t-test is not able to detect all ROIs between the two groups based on a limited number of scans;

(3) Voxel-based features are not able to indicate all possibilities of AD. Namely, some pathological reasons or features for AD cannot be discovered via voxels in structural MRI scans.

For the reason (1), it is reasonable that not all ROIs were detected, because only a small proportion of the whole dataset was used for detecting ROIs, and this proportion could not fully cover all possibilities leading to AD or pMCI. For the reason (2), the two-sample t-test was used to test whether the unknown population (voxels in this study) mean values of the two groups were equal. Therefore, the ROIs detected by a two-sample t-test could only indicate group-level differences of voxels between the two groups, and some individual-level differences might be ignored by the two-sample t-test. For example, if a certain area has abnormal voxel values due to AD in a very small number of AD scans (because not all AD are caused by that area), a few scans may not considerably influence the mean values of voxels within the area. The mean values of voxels within the area in AD scans may be close to those in CN scans and the two-sample t-test might not indicate the area as an AD-relevant regions. As for the reason (3), it is quite possible that AD may not fully be reflected by voxels in the structural MRI scans. As introduced in Section 1.4.2, AD might be caused by the neurofibrillary tangles or other such factors that structural MRI scans cannot capture. Therefore, AD or pMCI scans that contain such structural MRI-uncapturable characteristics may not be correctly distinguished by the method: this is a potential limitation of this method.

In addition to ROI capture, other steps might also affect the performance of the method. In Section 7.4, the influence of each step in the method will be analysis, and the limitation of the method in this study will also be discussed.

7.4. Limitations of the Study

The performance of discrimination might be influenced by each step in the workflow. In this section, factors adversely affecting the performance of the study are discussed.

7.4.1. Limitations of Research Design

The research was designed to discriminate between different stages of AD using machine learning techniques based on structural MRI scans. However, as mentioned in Section 7.3, not all AD-relevant factors can be discovered from the voxels of structural MRI scans (at the voxel level). Therefore, other modalities of medical imaging and other types of data could be considered and used in the study to facilitate this discrimination. The alternative modalities of medical imaging data that can be used are functional MRI scans, diffusion-weighted imaging (DWI) scans, and positron emission tomography (PET) scans (Guo et al., 2017; Lella et al., 2020; Samper-González et al., 2018). As for the types of data, in addition to medical imaging data, clinical biomarkers (e.g., neurological exams, cognitive assessment, medications, etc.) and genetic data could also be used in the study to diagnose AD. Moreover, the combination of the medical imaging data, clinical biomarkers, and genetic data could also be used and it has the potential to enhance the performance of the discrimination (Lu et al., 2018).

In addition, in this study, detection of AD-relevant regions and feature extraction were undertaken separately and this might make the steps cumbersome and ignore some information as mentioned in Section 7.3. Therefore, methods that can directly be applied to the whole scans to complete both detection of AD-relevant regions and feature extraction could be considered for the future research.

In the process of downloading MRI scans from ADNI, if too many scans (or all scans) were selected to download, the process would discontinue half-way through due to the long downloading time; thus, they could not be downloaded at one time. In addition, once the download was interrupted due to the network, the downloading session would restart from beginning. Therefore, the data were divided into a certain number of groups (10 scans in a group) and were downloaded group by group. This increased the manual operations and considerably increased the time of data collection.

Furthermore, it was apparent that there were some artefacts and background noises on some downloaded MRI scans, as shown in Figure 7.2. This might be because of the unstable network: some packets of the data (scans) might have been lost during the download process when the network was not stable, and the loss of the packets may have led to the poor quality of the scans. Therefore, different scans might have scratches in different download sessions. Sometimes this could be solved through re-downloading the problematical scans but this was not always effective. This may be because these images were originally damaged. In addition

to the problems of the original MRI scans, some errors also occurred during the process of the MRI scan pre-processing; however, the pre-processing tool (SPM) did not provide the information or reasons of the errors. The two problems led to low quality in the volume of the data, and this forced us to remove those low-quality MRI scans to ensure the proper running in the later phases of the workflow in this study. The removal decreased the number of usable MRI scans (from 800 to 715) and limited the access to the data obtained from ADNI. However, this was unavoidable.
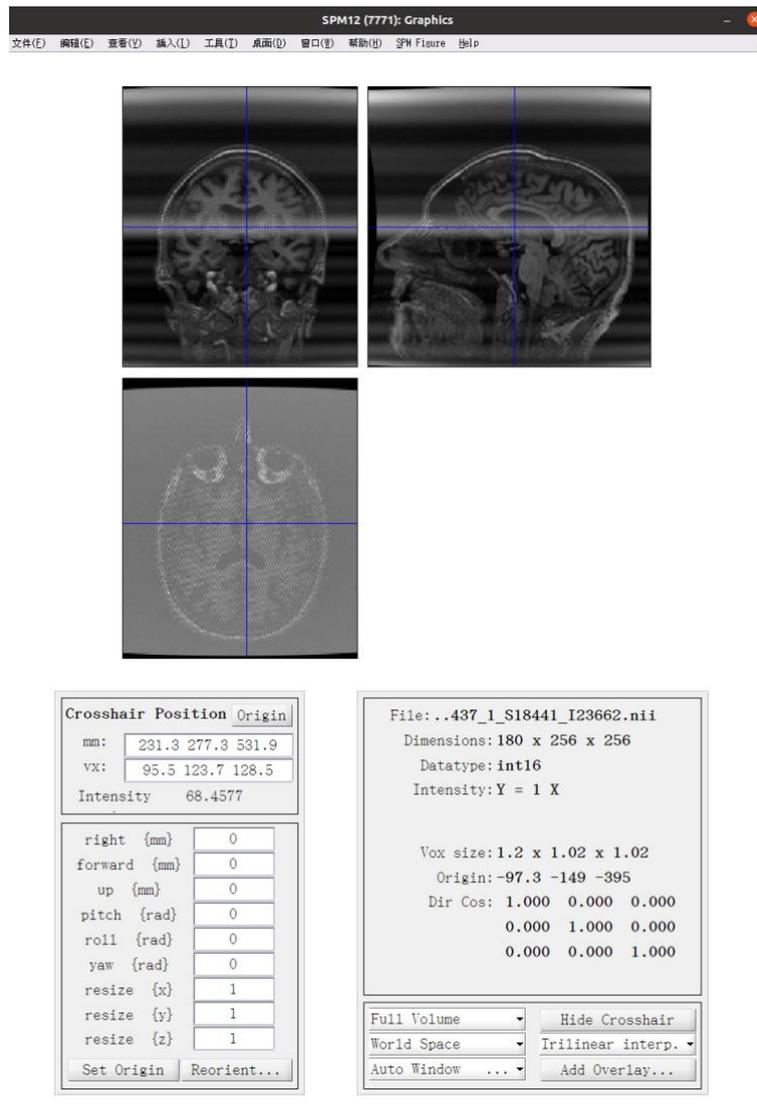
Figure 7.2 Problematic MRI scans from ADNI (Screenshot by Yuyang Liu)

It is difficult to solve the downloading problems. One solution is downloading the data from ADNI, at the location of the physical servers (the United States), with stable and high download speed. As for the problematic scans from ADNI, the practical solution is downloading the problematic scans again and the problem often could be solved by this. If the re-downloading

does not work, scans from other databases, e.g., OASIS, AIBL, etc., could potentially be used in the study to enlarge the size of the dataset. Furthermore, it could also solve the problem of external validity, which will be mentioned in Section 7.4.2. Understanding the impact of data from other datasets on performance is an area that can be explored in more detail in future.

7.4.2. Limitations of the Methodology

Limitations can be introduced in each of the phase in the methodology.

7.4.2.1.　　Data collection

Limitations: In the first phase, 1.5T T1-weighted structural MRI scans were obtained from ADNI. Although it is possible to enhance the image quality with longer sequencing 1.5T scans, this can cause a long time waiting and delay. Therefore, it is difficult to make sure that the 1.5T MRI scans provided by ADNI have high image quality because they may not have been obtained by using long sequencing. However, image quality is important to image classification and discrimination, especially in studies requiring minute detail, such as brain scans.

Potential Solutions: 3T MRI scanning is much more expensive than 1.5T scanning, thus 1.5T is more common in practical. In addition, although ADNI provided 3T MRI scans, the number of 3T scans is much smaller than 1.5T scans (only one fourth of ADNI1 subjects were scanned on 3T scanners). Therefore, if a large number of 3T MRI scans could be acquired, they could be potentially used because they can reduce noise and enhance the quality of the images.

7.4.2.2.　　Data pre-processing

Limitations: In the second phase, the Statistical Parametric Mapping (SPM) and the Computational Anatomy Toolbox (CAT) were used to pre-process MRI scans. In the process of the pre-processing, it is difficult to guarantee that there is no information loss regardless of what pre-processing tools are used. In addition, the final size of the scan after pre-processing is $121{\times}145{\times}121$ voxels as the default. If the size is much different from the original size (original sizes vary), some information loss may potentially be unintentionally introduced with the process. However, the information loss during the pre-processing is inevitable.

Potential Solutions: Information loss during data pre-processing cannot be avoided. However, different pre-processing tools may have different performances. As mentioned in Section 4.9.1.1, pre-processing a single MRI scan generally took more than 30 minutes thus it was not practical to compare all of the pre-processing tools in this study. Therefore, in addition to the SPM and FreeSurfer, which were compared and discussed in Section 4.9.1.1, other MRI pre-processing tools, e.g., FMRIB software library (FSL), Advanced Normalisation Tool (ANT),

etc., could be potentially used to conduct a comparative study of the impact of the tools on performances.

7.4.2.3. *AD-relevant regions detection*

Limitations: The limitation of the third phase, i.e., the detection of AD-relevant regions, has been mentioned and described in Section 7.3: the two-sample t-test may be not powerful enough to detect all potential AD-relevant regions.

Potential Solutions: A potential approach is to use ROIs directly based on clinical and knowledge of the pathology. The ROIs that were often used in the previous studies were hippocampus, lobes, amygdala, etc. (Y. Gupta et al., 2019; Katabathula et al., 2021; Long et al., 2017; Petrone et al., 2019). This approach does not require computation or detection of AD-regions from the scans. However, it does need professional knowledge about the pathology of AD, and the ROIs could be indicated by neuropathologists and experts in AD.

7.4.2.4. *Feature extraction*

Limitations: During the feature extraction, the PCANet was used to map the voxel-level features to higher-level features, with positive results (an increase of nearly 0.14 accuracy in discrimination between CN and AD). In this research, features were extracted from different slices by the PCANet, due to it being designed for learning features from 2D images. However, an MRI scans is a 3D image and the feature extraction based on slices can potentially ignore relationships of features between slices. Feature extraction within a slice can only build relationships between the voxels within the 2D space (within a plane) and the relationships between the voxels within 3D space (other slices) will be ignored. In addition to the features of the PCANet, its extensive memory usage is another concern. As described in Sections 4.9.6 and 4.10.2, features computed by the PCANet occupied a large proportion of the memory due to the dimension-increasing mechanism of the algorithm. This makes it difficult to extract features from the whole 3D scans.

Potential Solutions: Regarding the drawback that the PCANet cannot be used for 3D scans, a 3D-PCANet could potentially be developed to design for solving this problem. In addition, other unsupervised neural networks for 2D images are also candidates. However, the number of unsupervised neural networks is still low, thus the choice is relatively limited compared with supervised neural networks. A relatively new and alternative unsupervised neural network that could be used in a future study is anchor neighbourhood discovery (AND), which is an unsupervised deep learning structure by neighbourhood discovery developed by Huang et al.

(2019). AND could also be potentially further revised into a 3D implementation to extract features from the whole MRI scans directly.

7.4.2.5.　Clustering

Limitations: As for the clustering, a k-means++ algorithm was used, and showed positive performances in the three studies compared with previous state-of-the-art studies. K-means++ can solve the initialisation problem described in Section 3.7, but it still gives different results in different runs due to different position of the first centroid. Moreover, the values of some measures in the task of this study, e.g., the sensitivity and specificity in CN vs. pMCI, were not very stable in different runs.

Potential Solutions: The choice of the clustering algorithms also depends on the distribution of the features. Therefore, in this study, the k-means++ was applicable because the features of the two groups (CN vs. AD; sMCI vs. pMCI; CN vs. pMCI) of data were adjacent. However, other clustering algorithms, e.g., the mixture Gaussian distribution model, the fuzzy c-means, etc., did not achieve good performances in this case (through multiple experiments). Therefore, they could potentially be used to discriminate each pair of groups (CN vs. AD, sMCI vs. pMCI, and CN vs. pMCI) if other feature extraction methods were used and could lead to new feature distributions in future research.

7.4.2.6.　Results

Limitations: There are limitations in the study results, and this can be related to the data collection. The MRI scan data used in this study were collected only from ADNI and this limited the quantity and diversity of the data. The limitations of both quantity and diversity may cause reliability of the model because, if the volume of the data is not large and the data are from a single database, the model might lack generalisability to new data. In other words, the results were given by the model only based on the internal data while external validity pertains to the generalisability of results from the data in this study to the larger, target population.

Potential Solutions: Data could potentially be collected from other databases besides ADNI, e.g., the Open Access Series of Imaging Studies (OASIS), the Australian Imaging, Biomarker and Lifestyle Flagship Study of Ageing (AIBL), etc. However, collecting from multiple databases needs more time, but due to the time constraints, this could be conducted for future research.

7.5. Conclusion

This chapter discussed the three studies (CN vs. AD; sMCI vs. pMCI; CN vs. pMCI) in the previous three chapters. Section 7.2 revisited the research questions posted in Section 1.3 and it provided detailed answers to the research questions: the two-sample t-test detected the regions that were helpful to the discriminations, and the methods reported in this thesis were comparable to the existing state-of-the-art studies. Section 7.3 summarised and synthesised the results of the three studies presented in Chapters 4 to 6 and it linked them together: the difficulties of the discrimination across the three tasks in order (CN vs. AD; CN vs. pMCI; sMCI vs. pMCI) increased.

Section 7.4 presented the limitations of the whole study, including the limitations of the research design and methodology. It also provided the potential solutions to the limitations.

In the next chapter, a conclusion of the thesis will be presented. It will summarise all content in Chapter 1 to Chapter 7. It includes the summary of the thesis and the methodology, key findings in this research, contribution to the research field and the novelties of this study, and the future work. Finally, Chapter 8 will conclude and round-off the entire thesis.

Chapter 8 Conclusion

8.1.   Introduction

Chapter 7 discussed the three discrimination studies (cognitively normal vs. Alzheimer's disease; stable mild cognitive impairment vs. progressive mild cognitive impairment; cognitively normal vs. progressive mild cognitive impairment) and linked them together. Chapter 7 also answered the research questions and presented the limitations of the research. This chapter presents the wider conclusions of the research, bringing together findings from all the three studies.

The next section summarises the whole thesis; Section 8.3 provides a statement of the research ethics; Section 8.4 conclusion to the methodology used in this study and summarise the novelties of this study; the following section reviews the key findings of this study; Section 8.6 presents the contribution to the research field and clinical settings; Section 8.7 concludes with possible avenues for future research; Section 8.8 addresses the challenges during undertaking the PhD.

8.2.   Summary of the Thesis

Alzheimer's disease (AD) is the most common cause of the dementia and the number of people suffering from AD around the world is growing. However, according to the National Health Service (NHS) in UK (National Health Services (UK), n.d.-b), conventional diagnosis of AD by doctors often generally requires one to two weeks per patient. Therefore, diagnosing AD using modern computer-aided approaches, machine learning and deep learning methods, have been increasingly studied in recent years. In previous studies, most researchers employed supervised learning methods, which require a considerable volume of labelled data. However, medical imaging data are difficult to label because this needs a large number of clinical practitioners and considerable time. Semi-supervised learning and unsupervised learning methods can potentially help to overcome the problems above; however, the number of studies using these two methods is still limited.

Therefore, this research developed unsupervised learning methods to discriminate between different stages of AD using a limited amount of labelled structural MRI scans. The objectives of this research were:

(1) to extract appropriate features and regions from the brain MRI scans to enable the discrimination between cognitively normal (CN) and AD patients, between patients with stable

mild cognitive impairment (sMCI) and progressive MCI (pMCI), and between patients who are CN and with pMCI;

(2) to apply unsupervised learning approaches (including unsupervised neural networks and clustering algorithms) to discriminate between CN and AD, between sMCI and pMCI, and between CN and pMCI;

(3) to compare the performance of the methods with existing studies for discrimination of MRI scans from people with CN and AD, from people with sMCI and pMCI, and people who are CN and those with pMCI. The three objectives were implemented through the study reported in this thesis.

The three objectives were implemented through the study reported in this thesis.

In the discrimination between CN and AD, the method yielded an accuracy of 0.8417, a sensitivity of 0.7965, a specificity of 0.8805, a positive predictive value (PPV) of 0.8546, and a negative predictive value (NPV) of 0.8364 using AD-relevant regions detected by the two-sample t-test and computed by the PCANet. In Study 2, discrimination between sMCI and pMCI, the inclusion of CN and AD enhanced the performance, which yielded an accuracy of 0.672, a sensitivity of 0.693, a specificity of 0.660, a PPV of 0.698, and a NPV of 0.653. In Study 3, discrimination between CN and pMCI, the results of the discrimination between CN and pMCI scans indicated that including and not including AD scans in the dataset had their own strengths. The algorithm achieved a stable performance when AD scans were not included, which yielded an accuracy of 0.752 (95% CI: 0.742 to 0.762), a sensitivity of 0.768 (95% CI: 0.752 to 0.784), a specificity of 0.742 (95% CI: 0.73 to 0.754), a PPV of 0.662 (95% CI: 0.653 to 0.671), and a NPV of 0.829 (95% CI: 0.82 to 0.838). However, the algorithm achieved a higher performance at the cost of stability when AD scans were included, which yielded an accuracy of 0.776 (95% CI: 0.768 to 0.783), a sensitivity of 0.714 (95% CI: 0.675 to 0.752), a specificity of 0.821 (95% CI: 0.781 to 0.86), a PPV of 0.724 (95% CI: 0.695 to 0.752), and a NPV of 0.815 (95% CI: 0.799 to 0.83). The results are summarised in Table 8.1.

Table 8.1 A summary of the results of Studies 1 to 3

Measures		Three Studies			
		CN vs. AD	CN vs. pMCI		sMCI vs. pMCI
			Including AD	Not Including AD	
Accuracy	**Mean**	0.842	0.776	0.752	0.672
	95% CI	(0.833, 0.851)	(0.768, 0.783)	(0.742, 0.762)	(0.665, 0.679)
Sensitivity	**Mean**	0.797	0.714	0.768	0.693
	95% CI	(0.768, 0.826)	(0.675, 0.752)	(0.752, 0.784)	(0.676, 0.71)
Specificity	**Mean**	0.881	0.821	0.742	0.660
	95% CI	(0.853, 0.909)	(0.781, 0.86)	(0.73, 0.754)	(0.648, 0.672)
PPV	**Mean**	0.855	0.724	0.662	0.698
	95% CI	(0.829, 0.881)	(0.695, 0.752)	(0.653, 0.671)	(0.691, 0.705)
NPV	**Mean**	0.836	0.815	0.829	0.653
	95% CI	(0.819, 0.853)	(0.799, 0.83)	(0.82, 0.838)	(0.643, 0.663)

The two research questions posted in Section 1.3 were answered by the whole study in this thesis and this was discussed in Section 7.2. As can be seen from the results of the three studies (Table 8.1), the overall performances progressively decreased as the three discriminations (CN vs. AD, CN vs. pMCI, and sMCI vs. pMCI) were implemented.

8.3. Summary and Contribution of the Methodology and the Findings to New Knowledge

As shown in Figure 8.1, the overall methodology used in the whole study can be summarised as follows (further details are in Chapter 3): the data used in this study were 1.5T T1-weighted structural MRI scans, collected from the ADNI database. The collected MRI scans were pre-processed using Statistical Parametric Mapping (SPM) and Computational Anatomy Toolbox (CAT). After pre-processing, a two-sample t-test was used to detect abnormalities between each groups (CN vs. AD; sMCI vs. pMCI; CN vs. pMCI), namely regions of interest (ROIs). The ROIs were fed into an unsupervised neural network, i.e., the PCANet, to extract higher-level features. Finally, a k-means++ algorithm was applied to the extracted features from the PCANet to divide the data into two groups. The k-means++ algorithm was run ten times for each discrimination task, because k-means is non-deterministic and is like to produce different results in different runs due to its different randomly chosen centroids in each run.

To comprehensively evaluate the performance of the study, the mean values and 95% confidence intervals (95% CI) of five measures across the ten runs were used: the accuracy,

the sensitivity, the specificity, the positive predictive value (PPV), and the negative predictive value (NPV).
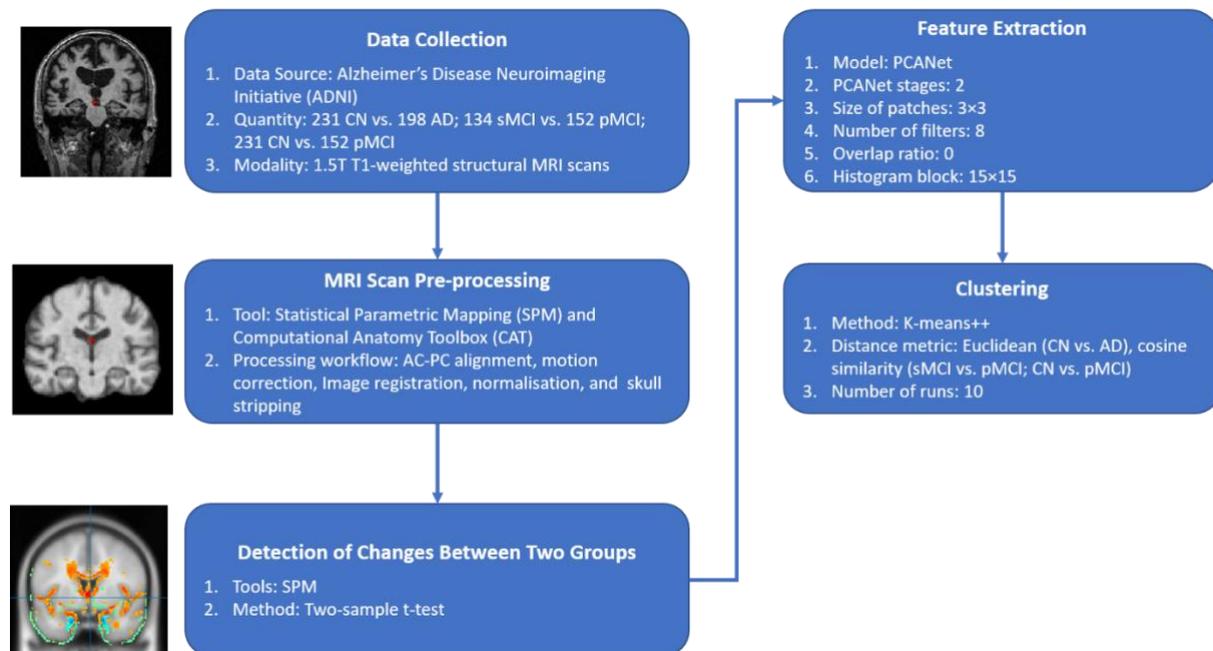
Figure 8.1 Workflow of the methodology in this thesis (MRI images from ADNI adapted Yuyang Liu)

As described in Chapter 2, in previous research, most of studies used supervised or semi-supervised learning methods to diagnose AD, including support vector machines (SVM), k-nearest neighbours (KNN) algorithms, random forests, convolutional neural networks (CNN), generative adversarial networks (GAN), etc.. These models and algorithms, primarily supervised learning models, achieved high performances (i.e., the accuracies were more than 80% for classification of CN and AD) in the previous studies. However, these methods heavily relied on the training data to achieve accurate diagnosis but MRI scans are difficult to obtain and label in practice. Hence, unsupervised learning methods need to be considered while few previous studies used unsupervised learning methods. However, as can be seen from Tables 4.7 and 4.8, most of the studies using unsupervised learning methods did not achieve comparable performance as that achieved by supervised and semi-supervised learning methods. The study in this thesis achieved an overall accuracy of 0.842 for discriminating between CN and AD using a statistical method and unsupervised learning methods based on a limited number of labelled MRI scans (30 for each group), which was also higher than some studies using supervised learning methods (as can be seen from Table 4.8).

In previous studies (Bron et al., 2021; W. Lin et al., 2018; Yu et al., 2022; Yanteng Zhang et al., 2022), the regions related to AD were detected by applying supervised neural networks (or deep learning architectures) to the whole brain MRI (or slices of the MRI) and learning from the backpropagation. However, the neural networks and deep learning architectures in these studies required nearly or more than half of the whole datasets as the training sets. In other studies (Aderghal et al., 2017; Farouk & Rady, 2020; Y. Gupta et al., 2019), they selected regions of interest (ROIs) directly from the MRI scans. However, this requires the researchers to have a considerable professional knowledge. In this study, two-sample t-tests were used to detect AD-relevant regions. The two-sample t-tests do not require so many labelled data as neural networks and deep learning architectures, where the size of the sample is generally less than 30 for each group. The regions detected by the two-sample t-test found to be helpful for the three discrimination tasks in this study.

To best of my knowledge (based on the literature review in Chapter 2), no previous study was found using unsupervised learning methods to discriminate between sMCI and pMCI scans: this is probably because the high similarity between the two groups makes this task a difficult one. This study was the first to use unsupervised learning methods to differentiate between sMCI and pMCI scans and this represents a novel contribution in terms of methods and the findings. The method achieved an accuracy of 0.672, a sensitivity of 0.693, a specificity of 0.660, a PPV of 0.698, and a NPV of 0.653. As can be seen from Table 5.6, the performance was comparable to some studies that used supervised and semi-supervised learning methods.

In this study, the CN and AD scans were included into the sMCI and pMCI scans to help the discrimination. However, no previous study that used unsupervised learning methods has done this before based on the reviewed literature in Chapter 2. This study first regarded the CN and AD scans as labelled data in unsupervised learning methods. This means that the clustering, which should not strictly have used labelled data, did include the labelled data, and this has never been conducted in previous studies that used unsupervised learning methods to discriminate between sMCI and pMCI. This represents a further novel contribution in this thesis. As described in Section 3.8.1, clustering does not give the corresponding labels to clusters. However, after including CN and AD scans and regarding them as labelled data, the two clusters computed by the k-means++ can be easily recognised: the cluster that had the most CN scans was recognised as the sMCI cluster; otherwise, it was recognised as pMCI.

Although a few studies (Y. Huang et al., 2019; Hongchao Jiang & Miao, 2022) used CN and AD scans to train the supervised learning models and tested on sMCI and pMCI scans, few of them showed the effectiveness and helpfulness of the inclusion of CN and AD scans to the discrimination between sMCI and pMCI scans. As can be seen from Figure 5.6 and Table 5.2, this study showed that the CN and AD scans truly helped the discrimination between sMCI and pMCI scans.

Similar to the discrimination between sMCI and pMCI, no previous studies have used unsupervised learning method to discriminate between CN and pMCI, and this study is the first to used this approach for these stages. The method reported in this study achieved an accuracy of 0.776, a sensitivity of 0.714, a specificity of 0.821, a PPV of 0.724, and a NPV of 0.815 when AD scans were included to help the discrimination; it also achieved an accuracy of 0.752, a sensitivity of 0.768, a specificity of 0.742, a PPV of 0.662, and a NPV of 0.829 when AD scans were not included. As can be seen from Table 6.3, the performances of the two (discriminations with and without AD scans) also were higher than some studies (Samper-González et al., 2018; Wen et al., 2021) that used supervised learning methods.

8.4. Review of the Key Findings Arising from This Thesis

This study was to develop unsupervised learning approaches to discriminate between three pairs of stages of AD (CN and AD; sMCI and pMCI; CN and pMCI) using a limited number of labelled structural MRI scans. The key findings of this study are summarised as follows:

(1) A two-sample t-test is suitable for detecting AD-relevant regions in this dataset;

(2) As indicated by the two-sample t-test analysis, lower level (i.e., voxel level) AD-relevant regions are mainly noted in the lower-parts of the limbic system, including the parahippocampal gyrus, amygdala, and hippocampus, in both the left and right cerebra;

(3) The regions detected by the two-sample t-test between sMCI and pMCI scans were not clear or obvious;

(4) The regions detected by the two-sample t-test between CN and pMCI scans were similar to those between CN and AD;

(5) PCANet was helpful for the feature extraction in the three discriminative studies (CN vs. AD; sMCI vs. pMCI; CN vs. pMCI) compared with not using it; however, it can increase the dimensionality of the features;

(6) The cosine similarity used in the k-means++ algorithm enhanced the performance in the harder discriminative task (sMCI vs. pMCI);

(7) There were limited number of previous studies using unsupervised learning method in discriminating between CN and AD, no previous study in sMCI and pMCI, no study in CN and pMCI;

(8) The method in this study had a better performance in the discrimination between CN and AD compared with previous studies using unsupervised learning methods; the whole study achieved comparable performances in the three discriminative studies (CN vs. AD; sMCI vs. pMCI; CN vs. pMCI) compared with previous studies using supervised learning methods.

(9) During the discrimination between sMCI and pMCI, the inclusion of the CN and AD scans was truly helpful for the discrimination; during the discrimination between CN and pMCI, the inclusion of AD scans was helpful for improving some measures (mean values of the accuracy, specificity, and PPV) but the stability of the measures decreased.

8.5.   Contribution of the Thesis to Clinical Settings

As described in Chapter 2, machine learning has been applied to a range of medical problems, e.g., public health, prediction of outcome, clinical decision support, etc., in recent years. The study in this thesis provided unsupervised learning methods to diagnose different stages of AD as an example of clinical decision support systems. It contributes to clinical settings for diagnosis of AD.

The regions detected by the two-sample t-test can potentially provide a guide to AD studies and can make AD researchers heed these regions more because they are demonstrated to be visibly different. In addition, based on the discovered regions, clinicians can determine the condition of AD (or pMCI) by checking these regions as a priority and this may save time in diagnosing AD, because it is very likely that most of AD cases arise as a result of the detected regions.

As described in 1.2.3, diagnosing AD requires different assessments, e.g., medical history, neuropsychological tests, etc.. However, the study in this thesis provided an effective method to distinguish patients with AD from people with CN only using MRI scans, where the overall accuracy was 84.2%. In addition, machine learning models can analyse large amounts of patient data at once, significantly reducing the time required for clinicians to analyse the results of the assessments.

Different assessments involve different types of data, e.g., text data, numerical data, medical imaging data, etc.. In addition, medical imaging data also include different modalities, e.g., MRI, computerised tomography (CT), positron emission tomography (PET), etc.. Although the study in this thesis provided approaches to diagnosing AD only using MRI scans, machine learning models have capabilities to handle the fusion of multiple data types and sources. This can not only take advantage of the complementary information contained in each data source, but also can reduce the time required for doctors to review and interpret the data, especially medical imaging data.

The study in this thesis also used a PCANet to compute voxel intensities from MRI scans. The PCANet, including other deep learning models, is capable of learning representations of the MRI scans (including other types of data) at different level of abstraction. Generally, basic features, e.g., volumes of the brain, shapes of the tissue, etc., are relatively easy for doctors to capture; however, the abstract representations often cannot be extracted by doctors. However, these representations are very likely to enhance the performance of the diagnosis of AD. The study in this thesis could provide insights for clinicians that the neural network plays an important role in the progress of diagnosing AD using artificial intelligence techniques.

The study in this thesis provided methods to discriminate between sMCI and pMCI. Combined with the predictive ability shown in Table 4.5, this could potentially provide clinicians with more advanced approaches to prognosis and AD progression studies. AD has no effective cure thus diagnosing and identifying patients with MCI, especially pMCI, is crucial. By continuously monitoring the patients with MCI (sMCI and pMCI) over time, the risk of AD progression could be assessed, predicted, and intervened by doctors.

8.6. Future Work

Based on the limitations of the research and the study (described in Chapter 7), in particular, there is considerable scope for the future work in the diagnosis of AD using machine learning and deep learning techniques.

8.6.1. Expand to Multiple Modalities

First, different modalities and types of data could be used. Functional MRI (fMRI) scans, positron emission tomography (PET) scans, and clinical biomarkers could be collected from the ADNI database and they could provide more information relevant to AD from other aspects besides voxel-level data. Guo et al. (2017) used fMRI scans to construct a hyper-network and

classify CN and AD scans. They finally achieve an average accuracy of 91.6% on average. In Tong et al.'s (2017) study, PET scans outperformed other three modalities (i.e., MRI scans, genetic features, and cerebrospinal fluid) when classifying CN and AD, which yielded an accuracy close to 0.9. Therefore, as can be seen from previous studies, using a different modality in itself has potential for future work.

In addition to using a single modality or type of data, the combination of multiple modalities or types of data could potentially be more effective in diagnosing AD, e.g., a combination of clinical biomarkers and MRI scans, a combination of PET scans and MRI scans, a combination of structural MRI scans and functional MRI scans, etc.. Tong et al. (2017) proposed a multi-modal classification using nonlinear graph fusion to diagnose AD. The modalities they used were PET scans, MRI scans, genetic features, and cerebrospinal fluid (CSF) biomarker features. Their method yielded an accuracy of 91.8% for classification of CN and AD, 79.5% for classification of MCI and CN, and 60.2% for multi-classification (AD vs. MCI vs. CN). In addition to directly using medical imaging scans, some data or features generated from medical imaging scans, e.g., brain connectivity and functional connectivity, could also be used in a future study.

8.6.2. Using Higher Quality MRI Scans

3T T1-weighted structural MRI scans can provide higher quality images than 1.5T T1-weighted structural MRI scans, and they could potentially improve the performance of the diagnosis of AD. Since the current PhD study commenced, Nigri et al. (2020) reported using 3T T1-weighted MRI scans to diagnose AD and achieved good performances in four experiments: three of the four Area under the Receiver operating characteristic (ROC) curve (AUC) were close to 0.9 and one reached 0.923. Therefore, 3T structural MRI scans could also be used in the future work.

8.6.3. Extending to Other Data Sources

In addition to ADNI, the data from other public databases could be used in future research. As described in Section 3.3, there are other commonly-used databases for AD, e.g., the Open Access Series of Imaging Studies (OASIS), Australian Imaging, Biomarkers, and Lifestyle Flagship Study of Ageing (AIBL), etc.. Li et al. (2019) trained their proposed model using the data from ADNI-1, and tested and evaluated the model using the data from AIBL, ADNI-2, and ADNI-GO. The data used in the future work could be collected from multiple of these databases and this could be used to check the external validity.

8.6.4. Other Unsupervised Learning Methods

A deep neural network applied to the whole MRI scans could also be used in future work. In this study, features were extracted by applying the PCANet to the AD-relevant regions, and this required that the AD-relevant regions were detected beforehand. In the future work, unsupervised deep neural network architectures that can extract feature directly from the whole MRI scans, e.g., Anchor Neighbourhood Discovery (AND) developed by Huang et al. (2019), could be considered. In addition, the above neural networks (i.e., AND and PCANet) were used for feature extraction of 2-D images and this may ignore some information in the third dimension because MRI scans are 3-D images; therefore, neural networks that were developed for 3-D images, e.g., 3-D convolutional neural network, or 3-D versions of PCANet and AND could be developed, and these could be employed in the future work.

Finally, other clustering algorithms, besides k-mode algorithms (e.g., k-means and k-means++), could be used in future work. Fuzzy c-means is a potential candidate, especially for discriminating between sMCI and pMCI. Fuzzy c-means is an unsupervised learning algorithm that is used to cluster data, assigning each datum a membership in each cluster from 0 to 1. MCI is a phase between CN and AD and in some cases it may be similar to CN and in other cases to AD. In other words, theoretically, the stable form of MCI, i.e., sMCI, is more similar to CN, and the more progressive form, i.e., pMCI, is more similar to AD: the fuzzy c-means method might be able to indicate such similarities between each pair (i.e., sMCI vs. CN; pMCI vs. AD) via the membership assigned to each scan. In addition to the fuzzy c-means, Gaussian mixture (GM) models could also be used for diagnosing AD. This is a probabilistic model that assumes the data points are generated from a mixture of finite number of Gaussian distributions. It is likely that the distributions of AD and CN scans are normally distributed and Gaussian mixture models may achieve an advantageous performance in distinguishing AD and CN. However, these two clustering algorithms require proper distributions of the inputs (features) (e.g., GM models require the inputs to be normally distributed) and therefore proper feature extraction methods before the clustering would be important.

8.6.5. Other Approaches

Diagnosing AD is a complicated process and the performance of it can be affected by multiple factors in each phase of the workflow, e.g., data modalities and types, the approaches to extract features, the parameters of the models (neural networks or classifiers), etc.. Therefore, different combinations of changing these factors could also potentially achieved a better performance,

e.g., Jabason et al. (2019) applied a semi-supervised learning method to a combination of MRI scans and clinical record data and achieved an accuracy of 98.78% for classification of CN, MCI, and AD. Thus, the combination addressed above in this section could be attempted in future work, e.g., using PET scans as the data and applying an AND to the data to extract features. The performance of these and the effect of the factors should be studied and discussed.

In addition to the above, more state-of-the-art methods could be included in future work to enhance the performance of diagnosis of AD. Continuous attention to the latest studies of machine learning and AD is the key to better study in this field. It is clear that as further AI algorithms are developed and more data become available, there will be potential ways to use these to improve the performance of diagnosing AD.

8.6.6. Deployment for Clinical Practice

This study is not enough to deploy in clinical practice, although the overall accuracy of the AD diagnosis was 84.2%. Deploying the model or the system for clinical practice requires rigorous validation, ensuring models are reliable, accurate, reproducible, and interpretable.

If the model (or system) could be used by clinicians as part of their clinical practice, it needs to be further developed and to achieve higher overall accuracies for discriminations between CN and AD, between sMCI and pMCI, and between CN and pMCI, and these would be expected to be higher than 85%. Other measures (i.e., specificity, sensitivity, PPV, and NPV) would be expected to be higher and balanced. In addition, the stability of the measures should be improved. The model is also expected to achieve similar (high) performances across different external databases. These could potentially increase the reliability of the model (system) used for diagnosing AD patients. Furthermore, the model is also expected to have capability to interpret the results for patients.

In addition to the model itself, other aspects could be considered if the model tried to be implemented in clinical practice. First, additional features, e.g., demographics, genetic data, etc., could be included to complement MRI imagery. Second, to deal with the much larger data sources, a high performance computing environment, including multiple central processing units (CPUs), graphics processing units (GPUs), etc., is necessary. Third, clinicians could be involved into the loop to verify or cross-check the ROIs identified by the model, and a user-friendly graphical user interface (GUI) could be developed for the clinicians.

8.7.   Challenges in Undertaking the PhD

In the process of undertaking the PhD, it was inevitable that challenges were encountered. These are discussed below.

The first challenge was the research itself. Although AD is becoming increasingly common in the daily life, developing comprehensive acknowledgement of it was not easy for me because it is highly related to neuroscience, and the neuroscience is complex and is not my specialty. I spent much time studying neuroscience, especially the brain anatomy (brain structures), through books and online courses, to gain a better understanding of the field.

As for the detail of the study, Alzheimer's Disease Neuroimaging Initiative (ADNI) database needed to be studied because it was used in my research and it provides several types of data and several versions of MRI scans for one patient, e.g., Magnetization Prepared-Rapid Gradient Echo (MPRAGE), B1-calibration scans, etc.; however, these terms were professional and unfamiliar to me. The methods and tools (i.e., SPM, FreeSurfer, etc.) to view them and to operate them, which are not commonly used by non-neuroscience people, were also studied, but they involve a number of parameters in the process and these parameters need to be understood because they had to be set according to the need in this study. However, these parameters are highly related to the medical profession. In addition, there was limited instruction (help, tutorials, and resources) and information from the Internet for specific tasks, e.g., preparation of sMCI and pMCI scans, thus these required me to come up with appropriate working solutions. These challenges not only required me to gain broader knowledge through limited resources, but also meant I had to communicate with other researchers (whom I went to know through blogs, WeChat, conferences, etc.) in the related research areas to overcome the difficulties in the process of doing this research.

This study as a part of the PhD research is expected to be completed within 2 years (excluding the first year for writing the confirmation review report and the final year for writing the thesis). Therefore, it is difficult to try new modalities of data, new models, and new algorithms to improve the performance of the study within this timescale. As mentioned in Section 7.3.2.1, downloading MRI scans from ADNI require many manual operations and time, thus it was difficult to further collect further functional MRI scans or DWI from ADNI. In addition to the downloading, MRI scan pre-processing also needs much time. Based on the rough observation during the pre-processing step, the pre-processing time of one MRI scan was around 30 to 60 minutes (depending on the performance of the computer used for pre-processing and the

degrees of distortion and rotation of the scans). This time does not include the manual anterior commissure - posterior commissure alignment (AC-PC alignment) but AC-PC alignment generally consumed one minute for each scan. Therefore, using new modalities suggests much longer downloading time and pre-processing time and this was not practical in this 2-year study. The new ideas that cannot be achieved during the period of PhD could be further developed in future academic work.

In addition to professional knowledge of neuroscience in this research, data preparation was also a challenge. The ADNI was founded in the United States (US) thus the servers that store the data are located in the US as well, which is far from the United Kingdom. This led difficulties in downloading MRI scans from the ADNI: the Internet was unstable and the download often interrupted. Finally, I solved the problem through asking a friend in the US to download them and to share them in the Google Drive. This saved much time for the research.

Undertaking a PhD is already a challenging endeavour, but when unexpected circumstances arose, i.e., COVID-19, it introduced additional uncertainties. Personally experiencing the impact of COVID-19 while being stuck in China for almost two years during my PhD journey undoubtedly posed unique challenges. For example, I was not able to engage with my supervisors in-person; I had to rely on online meetings with at times unstable Internet connections; connecting with other researchers was difficult without in-person events; it was difficult to be present in a research environment (e.g., research labs, university libraries, etc.) while working from home. However, in spite of the difficulties, I did not delay the process of my research. My ability to adapt and find alternative solutions allowed me to continue making progress and overall I learned a lot from having to manage the situation.

8.8. Conclusion

This chapter summarises the whole study in this thesis, including the methodology, the key findings, and the contribution to clinical settings. Furthermore, future work is also presented.

This study aimed to develop machine learning and deep learning approaches to diagnose AD based on a limited number of labelled structure MRI scans. Compared with other state-of-the-art studies, especially those using supervised learning methods, the method reported in this thesis achieved comparable performances for the three discrimination studies (CN vs. AD; sMCI vs. pMCI; CN vs. pMCI). However, in its current stage, the power of machine learning and deep learning is still not sufficient to accurately and stably diagnose AD due to the

complexity of the brain and the disease (AD). The role of artificial intelligence is currently helping neuropathologists study AD and its capabilities to diagnose AD is limited because its approaches to diagnose AD are still based on the current pathology of AD but the pathology of AD is still unclear (or, at best, partially clear). Therefore, the study needs to be developed further in future work.

This study was just the start of my research and it will lead me to further endeavours in this field. In addition to the future work presented in Section 8.6, machine learning and deep learning can be applied to broader research areas, e.g., cognitive modelling, computational neuroscience, mechanistic model, etc.. Furthermore, I will constantly stay focused on the development of artificial intelligence methods (including machine learning and deep learning), e.g., as new techniques, new models, new applications, etc., become available. In the future, with the development of new technologies, it is hopeful that more problems in medicine can be helped and be solved by the technologies, and people around the world can finally benefit from them.

References

Adel, T., Cohen, T., Caan, M., & Welling, M. (2017). 3D scattering transforms for disease classification in neuroimaging. *NeuroImage: Clinical*, *14*, 506–517. https://doi.org/10.1016/j.nicl.2017.02.004

Aderghal, K., Boissenin, M., Benois-Pineau, J., Catheline, G., & Afdel, K. (2017). Classification of sMRI for AD Diagnosis with Convolutional Neuronal Networks: A Pilot 2-D+ε Study on ADNI. *Multimedia Modeling International Conference*, *1*(January), 226–237.

Aguilar, C., Westman, E., Muehlboeck, J. S., Mecocci, P., Vellas, B., Tsolaki, M., … Wahlund, L. O. (2013). Different multivariate techniques for automated classification of MRI data in Alzheimer's disease and mild cognitive impairment. *Psychiatry Research - Neuroimaging*. https://doi.org/10.1016/j.pscychresns.2012.11.005

Ahmad, F., Almuayqil, S. N., Humayun, M., Naseem, S., Khan, W. A., & Junaid, K. (2021). Prediction of COVID-19 cases using machine learning for effective public health management. *Computers, Materials and Continua*, *66*(3), 2265–2282. https://doi.org/10.32604/cmc.2021.013067

Alam, S., Kwon, G. R., Kim, J. I., & Park, C. S. (2017). Twin SVM-Based Classification of Alzheimer's Disease Using Complex Dual-Tree Wavelet Principal Coefficients and LDA. *Journal of Healthcare Engineering*, *2017*. https://doi.org/10.1155/2017/8750506

Alanazi, A. (2022). Using machine learning for healthcare challenges and opportunities. *Informatics in Medicine Unlocked*, *30*. https://doi.org/10.1016/J.IMU.2022.100924

Alashwal, H., El Halaby, M., Crouse, J. J., Abdalla, A., & Moustafa, A. A. (2019). The application of unsupervised clustering methods to Alzheimer's disease. *Frontiers in Computational Neuroscience*, *13*(May), 1–9. https://doi.org/10.3389/fncom.2019.00031

Alexander, A. L., Lee, J. E., Lazar, M., & Field, A. S. (2007). Diffusion Tensor Imaging of the Brain. *Neurotherapeutics*, *4*(3), 316–329. https://doi.org/10.1016/j.nurt.2007.05.011

Alexander, D. C., Dyrby, T. B., Nilsson, M., & Zhang, H. (2019). Imaging brain microstructure with diffusion MRI: practicality and applications. *NMR in Biomedicine*, *32*(4), 1–26. https://doi.org/10.1002/nbm.3841

Alexander, N., Alexander, D. C., Barkhof, F., & Denaxas, S. (2020). Identifying and evaluating clinical subtypes of Alzheimer's disease in care electronic health records using unsupervised machine learning. *BMC Medical Informatics and Decision Making*, *21*, 343. https://doi.org/10.1186/s12911-021-01693-6

Ali, M. S., Islam, M. K., Haque, J., Das, A. A., Duranta, D. S., & Islam, M. A. (2021). Alzheimer's Disease Detection Using m-Random Forest Algorithm with Optimum Features Extraction. *2021 1st International Conference on Artificial Intelligence and Data Analytics, CAIDA 2021*, 1–6. https://doi.org/10.1109/CAIDA51941.2021.9425212

Almajmaie, L. K. A., Raheem, A. R., Mahmood, W. A., & Albawi, S. (2022). MRI image segmentation using machine learning networks and level set approaches. *International Journal of Electrical and Computer Engineering*, *12*(1), 793–801. https://doi.org/10.11591/ijece.v12i1.pp793-801

Almdahl, I. S., Lauridsen, C., Selnes, P., Kalheim, L. F., Coello, C., Gajdzik, B., … Fladby, T. (2017). Cerebrospinal fluid levels of amyloid beta 1-43 mirror 1-42 in relation to imaging biomarkers of Alzheimer's disease. *Frontiers in Aging Neuroscience*, *9*(FEB), 1–13. https://doi.org/10.3389/fnagi.2017.00009

AlSaeed, D., & Omar, S. F. (2022). Brain MRI Analysis for Alzheimer's Disease Diagnosis Using CNN-Based Feature Extraction and Machine Learning. *Sensors*. https://doi.org/https://doi.org/10.3390/s2082911

Alzheimer's Association. (2021). 2021 Alzheimer's Disease Facts and Figures. In *Alzheimers Dement* (Vol. 17).

Alzheimer's Disease International. (2015). World Alzheimer Report 2015: The Global Imapct of Dementia. In *Alzheimer's Disease International*. Retrieved from https://www.alz.co.uk/research/WorldAlzheimerReport2015.pdf

Alzheimer's Disease International. (2019). World Alzheimer Report 2019: Attitudes to dementia. In *Alzheimer's Disease International: World Alzheimer Report 2019* (Vol. 13). Retrieved from https://www.alz.co.uk/research/WorldAlzheimerReport2019.pdf

Alzheimer's Disease Neuroimaging Protocol (ADNI). (2008). Retrieved from https://adni.loni.usc.edu/wp-content/themes/freshnews-dev-v2/documents/clinical/ADNI-1_Protocol.pdf

Alzheimer's Society. (2016). *The progression of Alzheimer's disease and other dementias*. Retrieved from https://www.alzheimers.org.uk/sites/default/files/2019-09/458lp-the-progression-of-alzheimers-disease-and-other-dementias-190521.pdf

Alzheimer's Society. (2017). *What Is Mild Cognitive Impairment?* Retrieved from https://www.nia.nih.gov/health/what-mild-cognitive-impairment

Alzheimer's Society. (2018). *Assessment and diagnosis*. Retrieved from https://www.alzheimers.org.uk/sites/default/files/migrate/downloads/factsheet_assessment_and_diagnosis.pdf

Arabi, E. M., Ahmed, K. S., & Mohra, A. S. (2022). Advanced Diagnostic Technique for Alzheimer's Disease using MRI Top- Ranked Volume and Surface-based Features. *Journal of Biomedical Physical and Engineering*, (December 2021).

Arafa, D. A., Moustafa, H. E. D., Ali, H. A., Ali-Eldin, A. M. T., & Saraya, S. F. (2023). A deep learning framework for early diagnosis of Alzheimer's disease on MRI images. In *Multimedia Tools and Applications*. https://doi.org/10.1007/s11042-023-15738-7

Ardekani, B. A., & Bachman, A. H. (2009). Model-based automatic detection of the anterior and posterior commissures on MRI scans. *NeuroImage*, *46*(3), 677–682. https://doi.org/10.1016/j.neuroimage.2009.02.030

Ardekani, B. A., Bermudez, E., Mubeen, A. M., Bachman, A. H., & Kline, N. (2017). Prediction of Incipient Alzheimer's Disease Dementia in Patients with Mild Cognitive Impairment. *Journal of Alzheimer's Disease*, *55*, 269–281. https://doi.org/10.3233/JAD-160594

Arevalo-Rodriguez, I., Smailagic, N., Roquéi Figuls, M., Ciapponi, A., Sanchez-Perez, E., Giannakou, A., … Cullum, S. (2015). Mini-Mental State Examination (MMSE) for the

detection of Alzheimer's disease and other dementias in people with mild cognitive impairment (MCI). *Cochrane Database of Systematic Reviews*, *2015*(3). https://doi.org/10.1002/14651858.CD010783.pub2

Arthur, D., & Vassilvitskii, S. (2006). k-means++: The Advantages of Careful Seeding. *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, 1027–1035.

Aruchamy, S., Mounya, V., & Verma, A. (2020). Alzheimer's disease classification in brain MRI using modified kNN algorithm. *Proceedings - 2020 IEEE International Symposium on Sustainable Energy, Signal Processing and Cyber Security, ISSSC 2020*, 15–20. https://doi.org/10.1109/iSSSC50941.2020.9358867

Asl, E. H., Ghazal, M., Mahmoud, A., Aslantas, A., Shalaby, A., Casanova, M., … Baz, A. El. (2018). Alzheimer's disease diagnostics by a 3D deeply supervised adaptable convolutional network. *Frontiers in Bioscience - Landmark*, *23*(3), 584–596. https://doi.org/10.2741/4606

Atabo, S. M., & Umar, A. A. (2019). A review of imaging techniques in scientific research/clinical diagnosis. *MOJ Anatomy & Physiology*, *6*(5). https://doi.org/10.15406/mojap.2019.06.00269

Ayodele, T. O. (2010a). Introduction to Machine Learning. In *New Advances in Machine Learning* (p. 1). https://doi.org/10.5772/9394

Ayodele, T. O. (2010b). Types of Machine Learning Algorithms. In *New Advances in Machine Learning* (p. 19). https://doi.org/10.5772/9394

Backstrom, K., Nazari, M., Gu, I. Y. H., & Jakola, A. S. (2018). An efficient 3D deep convolutional network for Alzheimer's disease diagnosis using MR images. *Proceedings - International Symposium on Biomedical Imaging*, *2018-April*(Isbi), 149–153. https://doi.org/10.1109/ISBI.2018.8363543

Balamurugan, M., Nancy, A., & Vijaykumar, S. (2017). Alzheimer's disease diagnosis by using dimensionality reduction based on KNN Classifier. *Biomedical and Pharmacology Journal*, *10*(4), 1823–1830. https://doi.org/10.13005/bpj/1299

Baliyan, V., Das, C. J., Sharma, R., & Gupta, A. K. (2016). Diffusion weighted imaging: Technique and applications. *World Journal of Radiology*, *8*(9), 785. https://doi.org/10.4329/wjr.v8.i9.785

Bamidis, P. D., Vivas, A. B., Styliadis, C., Frantzidis, C., Klados, M., Schlee, W., … Papageorgiou, S. G. (2014). A review of physical and cognitive interventions in aging. *Neuroscience and Biobehavioral Reviews*, *44*, 206–220. https://doi.org/10.1016/j.neubiorev.2014.03.019

Bank, D., Koenigstein, N., & Giryes, R. (2020). *Autoencoders*.

Basaia, S., Agosta, F., Wagner, L., Canu, E., Magnani, G., Santangelo, R., & Filippi, M. (2019). Automated classification of Alzheimer's disease and mild cognitive impairment using a single MRI and deep neural networks. *NeuroImage: Clinical*, *21*. https://doi.org/10.1016/J.NICL.2018.101645

Baskar, D., Jayanthi, V. S., & Jayanthi, A. N. (2019). An efficient classification approach for

detection of Alzheimer's disease from biomedical imaging modalities. *Multimedia Tools and Applications*, *78*(10), 12883–12915. https://doi.org/10.1007/s11042-018-6287-8

Baskaran, K. R. (2021). Deep learning based early Diagnosis of Alzheimer's disease using Semi Supervised GAN. *Annals of the Romanian Society for Cell Biology*, *25*(3), 1583–6258. Retrieved from http://annalsofrscb.ro

Battineni, G., Chintalapudi, N., Amenta, F., & Traini, E. (2020). A Comprehensive Machine-Learning Model Applied to Magnetic Resonance Imaging (MRI) to Predict Alzheimer's Disease (AD) in Older Subjects. *Journal of Clinical Medicine*, *9*(7), 2146. https://doi.org/10.3390/jcm9072146

Bazgir, O., Zhang, R., Dhruba, S. R., Rahman, R., Ghosh, S., & Pal, R. (2020). Representation of features as images with neighborhood dependencies for compatibility with convolutional neural networks. *Nature Communications*, *11*(1). https://doi.org/10.1038/s41467-020-18197-y

Beheshti, I., & Demirel, H. (2016). Feature-ranking-based Alzheimer's disease classification from structural MRI. *Magnetic Resonance Imaging*. https://doi.org/10.1016/j.mri.2015.11.009

Bernal, J., Kushibar, K., Asfaw, D. S., Valverde, S., Oliver, A., Martí, R., & Lladó, X. (2019). Deep convolutional neural networks for brain image analysis on magnetic resonance imaging: a review. *Artificial Intelligence in Medicine*, *95*, 64–81. https://doi.org/10.1016/j.artmed.2018.08.008

Bezdan, T., & Bačanin Džakula, N. (2019). Convolutional Neural Network Layers and Architectures. *International Scienctific Conference on Information Technology and Data Related Research*, (July), 445–451. https://doi.org/10.15308/sinteza-2019-445-451

Bhargava, R., Wozniak, L., Mentzel, H.-J., & Krumbein, I. (2014). *Understanding your MRI A guide for young people*.

Bhatt, S., Gething, P. W., Brady, O. J., Messina, J. P., Farlow, A. W., Moyes, C. L., … Hay, S. I. (2013). The global distribution and burden of dengue HHS Public Access. *Nature*, *496*(7446), 504–507. https://doi.org/10.1038/nature12060

Bi, X., Li, S., Xiao, B., Li, Y., Wang, G., & Ma, X. (2019). Computer aided Alzheimer's disease diagnosis by an unsupervised deep learning technology. *Neurocomputing*. https://doi.org/10.1016/j.neucom.2018.11.111

Bigham, B., Zamanpour, S. A., & Zare, H. (2022). Features of the superficial white matter as biomarkers for the detection of Alzheimer's disease and mild cognitive impairment: A diffusion tensor imaging study. *Heliyon*, *8*(1). https://doi.org/10.1016/J.HELIYON.2022.E08725

Bloch, L., & Friedrich, C. M. (2019). Classification of Alzheimer's Disease using volumetric features of multiple MRI scans. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, 2396–2401. https://doi.org/10.1109/EMBC.2019.8857188

Bloch, L., & Friedrich, C. M. (2021a). Comparison of automated volume extraction with freesurfer and fastsurfer for early alzheimer's disease detection with machine learning.

Proceedings - IEEE Symposium on Computer-Based Medical Systems, 2021-June, 113–118. https://doi.org/10.1109/CBMS52027.2021.00096

Bloch, L., & Friedrich, C. M. (2021b). Data analysis with Shapley values for automatic subject selection in Alzheimer's disease data sets using interpretable machine learning. *Alzheimer's Research and Therapy*, *13*(1). https://doi.org/10.1186/s13195-021-00879-4

Boateng, E. Y., Otoo, J., & Abaye, D. A. (2020). Basic Tenets of Classification Algorithms K-Nearest-Neighbor, Support Vector Machine, Random Forest and Neural Network: A Review. *Journal of Data Analysis and Information Processing*, *08*(04), 341–357. https://doi.org/10.4236/jdaip.2020.84020

Bondi, M. W., Edmonds, E. C., & Salmon, D. P. (2017). Alzheimer's Disease: Past, Present, and Future HHS Public Access. *J Int Neuropsychol Soc*, *23*(9), 818–831. https://doi.org/10.1017/S135561771700100X

Bonte, S., Goethals, I., & Van Holen, R. (2018). Machine learning based brain tumour segmentation on limited data using local texture and abnormality. *Computers in Biology and Medicine*, *98*, 39–47. https://doi.org/10.1016/J.COMPBIOMED.2018.05.005

Bron, E. E., Klein, S., Papma, J. M., Jiskoot, L. C., Venkatraghavan, V., Linders, J., … van der Lugt, A. (2021). Cross-cohort generalizability of deep and conventional machine learning for MRI-based diagnosis and prediction of Alzheimer's disease. *NeuroImage: Clinical*, *31*. https://doi.org/10.1016/J.NICL.2021.102712

Cabreza, J. N., Solano, G. A., Ojeda, S. A., & Munar, V. (2022). Anomaly Detection for Alzheimer's Disease in Brain MRIs via Unsupervised Generative Adversarial Learning. *4th International Conference on Artificial Intelligence in Information and Communication, ICAIIC 2022 - Proceedings*, 230–234. https://doi.org/10.1109/ICAIIC54071.2022.9722678

Caroline dos Santos Picanço, L., Ozela, P. F., de Fátima de Brito Brito, M., Pinheiro, A. A., Padilha, E. C., Braga, F. S., … Izabel da Silva Hage-Melim, L. (2018). Alzheimer's Disease: A Review from the Pathophysiology to Diagnosis, New Perspectives for Pharmacological Treatment. *Current Medicinal Chemistry*, *25*, 3141–3159. https://doi.org/10.2174/0929867323666161213101126

Casanova, R, Hsu, F.-C., Sink, K. M., Rapp, S. R., & Williamson, J. D. (2013). Alzheimer's Disease Risk Assessment Using Large-Scale Machine Learning Methods. *PLoS ONE*, *8*(11), 77949. https://doi.org/10.1371/journal.pone.0077949

Casanova, Ramon, Barnard, R. T., Gaussoin, S. A., Saldana, S., Hayden, K. M., Manson, J. A. E., … Chen, J. C. (2018). Using high-dimensional machine learning methods to estimate an anatomical risk factor for Alzheimer's disease across imaging databases. *NeuroImage*. https://doi.org/10.1016/j.neuroimage.2018.08.040

Cercignani, M., & Bouyagoub, S. (2018). Brain microstructure by multi-modal MRI: Is the whole greater than the sum of its parts? *NeuroImage*, *182*(October 2017), 117–127. https://doi.org/10.1016/j.neuroimage.2017.10.052

Chan, T. H. (2015). PCANet. Retrieved from https://mx.nthu.edu.tw/~tsunghan/Source codes.html

Chan, T. H., Jia, K., Gao, S., Lu, J., Zeng, Z., & Ma, Y. (2015). PCANet: A Simple Deep Learning Baseline for Image Classification? *IEEE Transactions on Image Processing*, *24*(12), 5017–5032. https://doi.org/10.1109/TIP.2015.2475625

Chandra, M. A., & Bedi, S. S. (2021). Survey on SVM and their application in image classification. *International Journal of Information Technology (Singapore)*, *13*(5). https://doi.org/10.1007/s41870-017-0080-1

Chang, C. H., Lin, C. H., & Lane, H. Y. (2021). Machine learning and novel biomarkers for the diagnosis of alzheimer's disease. *International Journal of Molecular Sciences*, *22*(5), 1–12. https://doi.org/10.3390/ijms22052761

Cohen, J. P., Luck, M., & Honari, S. (2018). Distribution matching losses can hallucinate features in medical image translation. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018. MICCAI 2018.*, *11070*, 529–536. https://doi.org/10.1007/978-3-030-00928-1_60

Cortes, C., Vapnik, V., & Saitta, L. (1995). Support-Vector Networks. *Machine Learning*, *20*, 273–297.

Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehéricy, S., Habert, M. O., … Colliot, O. (2011). Automatic classification of patients with Alzheimer's disease from structural MRI: A comparison of ten methods using the ADNI database. *NeuroImage*, *56*(2), 766–781. https://doi.org/10.1016/J.NEUROIMAGE.2010.06.013

Das, N. N., Srivastav, N., & Verma, S. S. (2021). Magnetic Resonance Imaging based Feature Extraction and Selection Methods for Alzheimer Disease Prediction. *Proceedings of International Conference on Technological Advancements and Innovations, ICTAI 2021*, 454–459. https://doi.org/10.1109/ICTAI53825.2021.9673337

Davis, K. L., Mohs, R. C., Tinklenberg, J. R., Pfefferbaum, A., Hollister, L. E., & Kopell, B. S. (1978). Physostigmine: Improvement of long-term memory processes in normal humans. *Science*, *201*(4352), 272–274. https://doi.org/10.1126/science.351807

Debette, S., & Markus, H. S. (2010). The clinical importance of white matter hyperintensities on brain magnetic resonance imaging: Systematic review and meta-analysis. *BMJ*, *341*(7767), 288. https://doi.org/10.1136/bmj.c3666

Deepa, D., & Karthi, A. (2022). Knowledge data discovery on Alzheimer's images by using image enhancement techniques and KNN approach. *Indian Journal of Computer Science and Engineering*, *13*(1), 268–277. https://doi.org/10.21817/indjcse/2022/v13i1/221301202

Dhinagar, N. J., Thomopoulos, S. I., Rajagopalan, P., Stripelis, D., Ambite, J.-L., Ver Steeg, G., & Thompson, P. M. (2023). Evaluation of transfer learning methods for detecting Alzheimer's disease with brain MRI. *SPIE Conference Proceedings*, (March), 60. https://doi.org/10.1117/12.2670457

Dinu, J. A., & Manju, R. (2021). A novel modelling technique for early recognition and classification of Alzheimer's disease. *2021 3rd International Conference on Signal Processing and Communication, ICPSC 2021*, (May), 21–25. https://doi.org/10.1109/ICSPC51351.2021.9451803

Du, J., Xu, J., Song, H. Y., & Tao, C. (2017). Leveraging machine learning-based approaches to assess human papillomavirus vaccination sentiment trends with Twitter data. *BMC Medical Informatics and Decision Making*, *17*(Suppl 2). https://doi.org/10.1186/s12911-017-0469-6

Dunstan, J., Aguirre, M., Bastías, M., Nau, C., Glass, T. A., & Tobar, F. (2020). Predicting nationwide obesity from food sales using machine learning. *Health Informatics Journal*, *26*(1), 652–663. https://doi.org/10.1177/1460458219845959

Elahifasaee, F., Li, F., & Yang, M. (2019). A Classification Algorithm by Combination of Feature Decomposition and Kernel Discriminant Analysis (KDA) for Automatic MR Brain Image Classification and AD Diagnosis. *Computational and Mathematical Methods in Medicine*, *2019*. https://doi.org/10.1155/2019/1437123

Elaziz, M. A., Hosny, K. M., Salah, A., Darwish, M. M., Lu, S., & Sahlol, A. T. (2020). New machine learning method for imagebased diagnosis of COVID-19. *PLoS ONE*, *15*(6). https://doi.org/10.1371/journal.pone.0235187

Emmanuel, M., & Jabez, J. (2022). An Enhanced Fuzzy Based KNN Classification Method for Alzheimer ' s Disease Identification from SMRI Images. *Journal of Algebraic Statistics*, *13*(3), 89–103.

Erickson, B. J., Korfiatis, P., Akkus, Z., & Kline, T. L. (2017). Machine learning for medical imaging. *Radiographics*, *37*(2), 505–515. https://doi.org/10.1148/rg.2017160130

Escudero, J., Zajicek, J. P., & Ifeachor, E. (2011). Early detection and characterization of Alzheimer's disease in clinical scenarios using Bioprofile concepts and K-means. *Conf Proc IEEE Eng Med Biol Soc*, *2011*, 6470–6473. https://doi.org/10.1109/IEMBS.2011.6091597

Eskildsen, S. F., Coupé, P., Fonov, V. S., Pruessner, J. C., & Collins, D. L. (2015). Structural imaging biomarkers of Alzheimer's disease: Predicting disease progression. *Neurobiology of Aging*. https://doi.org/10.1016/j.neurobiolaging.2014.04.034

Farooq, A., Anwar, S., Awais, M., & Rehman, S. (2017). A deep CNN based multi-class classification of Alzheimer's disease using MRI. *2017 IEEE Inter- National Conference on Imaging Systems and Techniques (IST)*, 3–8.

Farouk, Y., & Rady, S. (2020). Early Diagnosis of Alzheimer's Disease using Unsupervised Clustering. *International Journal of Intelligent Computing and Information Sciences*, *20*(2), 112–124. https://doi.org/10.21608/ijicis.2021.51180.1044

Feng, J., Zhang, S. W., & Chen, L. (2022). Extracting ROI-Based Contourlet Subband Energy Feature From the sMRI Image for Alzheimer's Disease Classification. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *19*(3), 1627–1639. https://doi.org/10.1109/TCBB.2021.3051177

Filipovych, R., & Davatzikos, C. (2011). Semi-supervised pattern classification of medical images: Application to mild cognitive impairment (MCI). *NeuroImage*, *55*(3), 1109–1119. https://doi.org/10.1016/J.NEUROIMAGE.2010.12.066

Fymat, A. L. (2018). Alzheimer's Disease: A Review. *Current Opinions in Neurological Science*, *2*(2), 415–436.

Ganggayah, M. D., Taib, N. A., Har, Y. C., Lio, P., & Dhillon, S. K. (2019). Predicting factors for survival of breast cancer patients using machine learning techniques. *BMC Medical Informatics and Decision Making*, *19*(1). https://doi.org/10.1186/s12911-019-0801-4

Ganguly, D., Chakraborty, S., Balitanas, M., & Kim, T. H. (2010). Medical imaging: A review. *Communications in Computer and Information Science*, *78 CCIS*(September 2010), 504–516. https://doi.org/10.1007/978-3-642-16444-6_63

Garcia-Ptacek, S., Eriksdotter, M., Jelic, V., Porta-Etessam, J., Kåreholt, I., & Manzano Palomo, S. (2016). Subjective cognitive impairment: Towards early identification of Alzheimer disease. *Neurologia*, *31*(8), 562–571. https://doi.org/10.1016/j.nrl.2013.02.007

Gerald, B. (2018). A Brief Review of Independent, Dependent and One Sample t-test. *International Journal of Applied Mathematics and Theoretical Physics*, *4*(2), 50. https://doi.org/10.11648/j.ijamtp.20180402.13

Glenner, G. G., & Wong, C. W. (1984). Alzheimer's disease and Down's syndrome: Sharing of a unique cerebrovascular amyloid fibril protein. *Biochemical and Biophysical Research Communications*, *122*(3), 1131–1135. https://doi.org/10.1016/0006-291X(84)91209-9

Glover, G. H. (2011). Overview of functional magnetic resonance imaging. *Neurosurgery Clinics of North America*, *22*(2), 133–139. https://doi.org/10.1016/j.nec.2010.11.001

Goldgaber, D., Lerman, M. I., McBride, O. W., Saffiotti, U., & Gajdusek, D. C. (1987). Characterization and chromosomal localization of a cDNA encoding brain amyloid of Alzheimer's disease. *Science*, *235*(4791), 877–880. https://doi.org/10.1126/science.3810169

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., … Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*, *3*(January), 2672–2680. https://doi.org/10.3156/jsoft.29.5_177_2

Grover, V. P. B., Tognarelli, J. M., Crossey, M. M. E., Cox, I. J., Taylor-Robinson, S. D., & Mcphail, M. J. W. (2015). Magnetic Resonance Imaging: Principles and Techniques: Lessons for Clinicians. *Journal of Clinical and Experimental Hepatology*, *5*, 246–255. https://doi.org/10.1016/j.jceh.2015.08.001

Guo, H., Zhang, F., Chen, J., Xu, Y., & Xiang, J. (2017). Machine learning classification combining multiple features of a hyper-network of fMRI data in Alzheimer's disease. *Frontiers in Neuroscience*, *11*(NOV). https://doi.org/10.3389/fnins.2017.00615

Gupta, M., Bansal, A., Jain, B., Rochelle, J., Oak, A., & Jalali, M. S. (2021). Whether the weather will help us weather the COVID-19 pandemic: Using machine learning to measure twitter users' perceptions. *International Journal of Medical Informatics*, *145*. https://doi.org/10.1016/J.IJMEDINF.2020.104340

Gupta, Y., Lee, K. H., Choi, K. Y., Lee, J. J., Kim, B. C., & Kwon, G. R. (2019). Early diagnosis of Alzheimer's disease using combined features from voxel-based morphometry and cortical, subcortical, and hippocampus regions of MRI T1 brain images. *PLoS ONE*, *14*(10). https://doi.org/10.1371/journal.pone.0222446

Han, C., Rundo, L., Murao, K., Milacski, Z. Á., Umemoto, K., Nakayama, H., & Satoh, S. (2019). GAN-based Multiple Adjacent Brain MRI Slice Reconstruction for Unsupervised Alzheimer's Disease Diagnosis. *International Conference on Computational Intelligence Methods for Bioinformatics and Biostatistics*. Retrieved from http://arxiv.org/abs/1906.06114

Hardy, J. (2006a). A Hundred Years of Alzheimer's Disease Research. *Neuron*, *52*(1), 3–13. https://doi.org/10.1016/j.neuron.2006.09.016

Hardy, J. (2006b). Alzheimer's disease: The amyloid cascade hypothesis - An update and reappraisal. In *Journal of Alzheimer's Disease* (Vol. 9). https://doi.org/10.3233/jad-2006-9s317

Hardy, J., Cowburn, R., Barton, A., Reynolds, G., Lofdahl, E., O'Carroll, A. M., … Winblad, B. (1987). Region-specific loss of glutamate innervation in Alzheimer's disease. *Neuroscience Letters*, *73*(1), 77–80. https://doi.org/10.1016/0304-3940(87)90034-6

Hartley, K. G., Damon, B. M., Patterson, G. T., Long, J. H., & Holt, G. E. (2012). MRI techniques: A review and update for the orthopaedic surgeon. *Journal of the American Academy of Orthopaedic Surgeons*, *20*(12), 775–787. https://doi.org/10.5435/JAAOS-20-12-775

Harvey, R. J., Skelton-Robinson, M., & Rossor, M. N. (2003). The prevalence and causes of dementia in people under the age of 65 years. *J Neurol Neurosurg Psychiatry*, *37*(3), 1206–1209.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2016-Decem*, 770–778. https://doi.org/10.1109/CVPR.2016.90

Huang, J., Dong, Q., Gong, S., & Zhu, X. (2019a). Unsupervised deep learning by neighbourhood discovery. *36th International Conference on Machine Learning, ICML 2019*, *2019-June*, 5090–5099.

Huang, J., Dong, Q., Gong, S., & Zhu, X. (2019b). Unsupervised Deep Learning by Neighbourhood Discovery. *36th International Conference on Machine Learning (ICML'19)*. Retrieved from https://github.com/raymond-sci/AND.

Huang, L., Jin, Y., Gao, Y., Thung, K. H., & Shen, D. (2016). Longitudinal clinical score prediction in Alzheimer's disease with soft-split sparse regression based random forest. *Neurobiology of Aging*, *46*, 180–191. https://doi.org/10.1016/J.NEUROBIOLAGING.2016.07.005

Huang, Y., Xu, J., Zhou, Y., Tong, T., & Zhuang, X. (2019). Diagnosis of Alzheimer's disease via multi-modality 3D convolutional neural network. *Frontiers in Neuroscience*, *13*(MAY). https://doi.org/10.3389/fnins.2019.00509

Jabason, E., Omair Ahmad, M., & Swamy, M. N. S. (2019). Deep structural and clinical feature learning for semi-supervised multiclass prediction of Alzheimer's disease. *2018 IEEE 61st International Midwest Symposium on Circuits and Systems (MWSCAS)*, *2018-Augus*, 791–794. https://doi.org/10.1109/MWSCAS.2018.8623946

Jack, C. R., Bernstein, M. A., Fox, N. C., Thompson, P., Alexander, G., Harvey, D., …

Weiner, M. W. (2008). The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods. *Journal of Magnetic Resonance Imaging*, *27*(4), 685–691. https://doi.org/10.1002/jmri.21049

Jayadeva, Khemchandani, R., & Suresh, C. (2007). Twin Support Vector Machines for Pattern Classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(5), 905–910. https://doi.org/10.1016/j.eswa.2008.09.066

Jiang, Heinrich. (2017). On the consistency of quick shift. *Advances in Neural Information Processing Systems*, *2017-Decem*, 46–55.

Jiang, Hongchao, & Miao, C. (2022). Pre-Training 3D Convolutional Neural Networks for Prodromal Alzheimer's Disease Classification. *Proceedings of the International Joint Conference on Neural Networks*, *2022-July*. https://doi.org/10.1109/IJCNN55064.2022.9891966

Jiang, W., Chan, K. L., Li, M., & Zhang, H. (2005). Mapping low-level features to high-level semantic concepts in region-based image retrieval. *Proceedings - 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005*, *II*, 244–249. https://doi.org/10.1109/CVPR.2005.220

Jillings, S., Van Ombergen, A., Tomilovskaya, E., Rumshiskaya, A., Litvinova, L., Nosikova, I., … Jeurissen, B. (2020). Macro- And microstructural changes in cosmonauts' brains after long-duration spaceflight. *Science Advances*, *6*(36). https://doi.org/10.1126/sciadv.aaz9488

Jin, S., Zou, P., Han, Y., & Jiang, J. (2021). Unsupervised detection of individual atrophy in Alzheimer's disease. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, 2647–2650. https://doi.org/10.1109/EMBC46164.2021.9630103

Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement Learning: A Survey. *Journal of Artificial Intelligence Research*, *10*(19), 237–285.

Kamal, M., Pratap, A. R., Naved, M., Zamani, A. S., Nancy, P., Ritonga, M., … Sammy, F. (2022). Machine Learning and Image Processing Enabled Evolutionary Framework for Brain MRI Analysis for Alzheimer's Disease Detection. *Computational Intelligence and Neuroscience*, *2022*. https://doi.org/10.1155/2022/5261942

Kang, Y., Wang, Y., Zhang, D., & Zhou, L. (2017). The public's opinions on a new school meals policy for childhood obesity prevention in the U.S.: A social media analytics approach. *International Journal of Medical Informatics*, *103*(November 2016), 83–88. https://doi.org/10.1016/j.ijmedinf.2017.04.013

Kasban, H., El-bendary, M., & Salama, D. (2015). A Comparative Study of Brain Imaging Techniques. *International Journal of Information Science and Intelligent System*, *4*(2), 37–58.

Katabathula, S., Wang, Q., & Xu, R. (2021). Predict Alzheimer's disease using hippocampus MRI data: a lightweight 3D deep convolutional network model with visual and global shape representations. *Alzheimer's Research and Therapy*. https://doi.org/10.1186/s13195-021-00837-0

Katz, G. (2017). Vertical Ensemble Co-Training for Text Classification. *ACM Trans. Intell. Syst. Technol*, *9*. https://doi.org/10.1145/3137114

Khajehnejad, M., Saatlou, F. H., & Mohammadzade, H. (2017). Alzheimer's disease early diagnosis using manifold-based semi-supervised learning. *Brain Sciences*, *7*(8), 1–19. https://doi.org/10.3390/brainsci7080109

Khan, A., & Zubair, S. (2019). Usage of Random Forest Ensemble Classifier Based Imputation and Its Potential in the Diagnosis of Alzheimer's Disease. *INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH*, *8*. Retrieved from www.ijstr.org

Kharghanian, R., & Ahmadyfard, A. (2012). Retinal Blood Vessel Segmentation Using Gabor Wavelet and Line Operator. *International Journal of Machine Learning and Computing*, *2*(5), 593–597. https://doi.org/10.7763/ijmlc.2012.v2.196

Khedher, L., Ramírez, J., Górriz, J. M., Brahim, A., & Segovia, F. (2015). Early diagnosis of Alzheimer's disease based on partial least squares, principal component analysis and support vector machine using segmented MRI images. *Neurocomputing*. https://doi.org/10.1016/j.neucom.2014.09.072

Kim, H. E., Cosa-Linan, A., Santhanam, N., Jannesari, M., Maros, M. E., & Ganslandt, T. (2022). Transfer learning for medical image classification: a literature review. *BMC Medical Imaging*, *22*(1), 1–13. https://doi.org/10.1186/s12880-022-00793-7

Kim, J. S., Han, J. W., Bae, J. Bin, Moon, D. G., Shin, J., Kong, J. E., … Kim, K. W. (2022). Deep learning-based diagnosis of Alzheimer's disease using brain magnetic resonance images: an empirical study. *Scientific Reports*, *12*(1), 1–8. https://doi.org/10.1038/s41598-022-22917-3

Kim, J. Y., Lee, M., Lee, M. K., Wang, S. M., Kim, N. Y., Kang, D. W., … Lim, H. K. (2021). Development of random forest algorithm based prediction model of alzheimer's disease using neurodegeneration pattern. *Psychiatry Investigation*, *18*(1), 69–79. https://doi.org/10.30773/pi.2020.0304

Kim, S. J., Cho, K. J., & Oh, S. (2017). Development of machine learning models for diagnosis of glaucoma. *PLoS ONE*, *12*(5). https://doi.org/10.1371/journal.pone.0177726

Klados, M. A., Styliadis, C., Frantzidis, C. A., Paraskevopoulos, E., & Bamidis, P. D. (2016). Beta-band functional connectivity is reorganized in Mild Cognitive Impairment after combined computerized physical and cognitive training. *Frontiers in Neuroscience*, *10*(FEB), 1–12. https://doi.org/10.3389/fnins.2016.00055

Konstantinidis, E. I., Bamidis, P. D., Billis, A., Kartsidis, P., Petsani, D., & Papageorgiou, S. G. (2021). Physical training in-game metrics for cognitive assessment: Evidence from extended trials with the fitforall exergaming platform. *Sensors*, *21*(17). https://doi.org/10.3390/s21175756

Korolev, I. O. (2014). Alzheimer 's Disease : A Clinical and Basic Science Review. In *Medical Student Research Journal* (Vol. 04). Retrieved from www.msrj.org

Korolev, I. O., Symonds, L. L., & Bozoki, A. C. (2016). Predicting progression from mild cognitive impairment to Alzheimer's dementia using clinical, MRI, and plasma

biomarkers via probabilistic pattern classification. *PLoS ONE*, *11*(2). https://doi.org/10.1371/journal.pone.0138866

Korolev, S., Safiullin, A., Belyaev, M., & Dodonova, Y. (2017). Residual and plain convolutional neural networks for 3D brain MRI classification. *Proceedings - International Symposium on Biomedical Imaging*, 835–838. https://doi.org/10.1109/ISBI.2017.7950647

Kumari, R., Goel, S., & Das, S. (2022). Using SVM for Alzheimer ' s Disease detection from 3D T1MRI. *2022 IEEE 21st Mediterranean Electrotechnical Conference (MELECON)*, 600–604. https://doi.org/10.1109/MELECON53508.2022.9842935

Laboratory for Computational Neuroimaging at the Athinoula A. Martinos Center for Biomedical Imaging. (n.d.-a). FreeSurfer. Retrieved from https://surfer.nmr.mgh.harvard.edu/

Laboratory for Computational Neuroimaging at the Athinoula A. Martinos Center for Biomedical Imaging. (n.d.-b). FreeSurfer. Retrieved from https://surfer.nmr.mgh.harvard.edu/fswiki/FsTutorial/SkullStripFix_freeview

Lachin, J. M. (1981). Introduction to sample size determination and power analysis for clinical trials. *Controlled Clinical Trials*, *2*(2), 93–113. https://doi.org/10.1016/0197-2456(81)90001-5

Laino, M. E., Cancian, P., Politi, L. S., Della Porta, M. G., Saba, L., & Savevski, V. (2022). Generative Adversarial Networks in Brain Imaging: A Narrative Review. *Journal of Imaging*, *8*(4). https://doi.org/10.3390/jimaging8040083

Lama, R. K., Gwak, J., Park, J. S., & Lee, S. W. (2017). Diagnosis of Alzheimer's Disease Based on Structural MRI Images Using a Regularized Extreme Learning Machine and PCA Features. *Journal of Healthcare Engineering*, *2017*. https://doi.org/10.1155/2017/5485080

Langa, K. M., & Levine, D. A. (2009). The Diagnosis and Management of Mild Cognitive Impairment: A Clinical Review Kenneth. *Expert Rev Neurother*, *27*(4), 590–609. https://doi.org/10.1001/jama.2014.13806

Langa, K. M., & Levine, D. A. (2014). The Diagnosis and Management of Mild Cognitive Impairment: A Clinical Review NIH Public Access. *JAMA*, *312*(23), 2551–2561. https://doi.org/10.1001/jama.2014.13806

Lazli, L., Boukadoum, M., & Ait Mohamed, O. (2018). Computer-aided diagnosis system for Alzheimer's disease using fuzzy-possibilistic tissue segmentation and SVM classification. *2018 IEEE Life Sciences Conference, LSC 2018*, 33–36. https://doi.org/10.1109/LSC.2018.8572122

Leandrou, S., Lamnisos, D., Kyriacou, P. A., Constanti, S., & Pattichis, C. S. (2020). Comparison of 1.5 T and 3 T MRI hippocampus texture features in the assessment of Alzheimer's disease. *Biomedical Signal Processing and Control*, *62*(July), 3–8. https://doi.org/10.1016/j.bspc.2020.102098

Lebedev, A. V., Westman, E., Van Westen, G. J. P., Kramberger, M. G., Lundervold, A., Aarsland, D., … Simmons, A. (2014). Random Forest ensembles for detection and

prediction of Alzheimer's disease with a good between-cohort robustness. *NeuroImage: Clinical*, *6*, 115–125. https://doi.org/10.1016/J.NICL.2014.08.023

Lee, E., Choi, J.-S., Kim, M., & Suk, H.-I. (2019). Toward an interpretable Alzheimer's disease diagnostic model with regional abnormality representation via deep learning. *NeuroImage*. https://doi.org/10.1016/j.neuroimage.2019.116113

Leow, A. D., Klunder, A. D., Jack, C. R., Toga, A. W., Dale, A. M., Bernstein, M. A., … Thompson, P. M. (2006). Longitudinal stability of MRI for mapping brain change using tensor-based morphometry. *NeuroImage*, *31*(2), 627–640. https://doi.org/10.1016/j.neuroimage.2005.12.013

Li, Hengxin, Shi, X., Zhu, X., Wang, S., & Zhang, Z. (2022). FSNet: Dual Interpretable Graph Convolutional Network for Alzheimer's Disease Analysis. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 1–11. https://doi.org/10.1109/TETCI.2022.3183679

Li, Hongming, Habes, M., Wolk, D. A., & Fan, Y. (2019). A deep learning model for early prediction of Alzheimer's disease dementia based on hippocampal magnetic resonance imaging data. *Alzheimer's and Dementia*. https://doi.org/10.1016/j.jalz.2019.02.007

Lin, H., Jiang, J., Li, Z., Sheng, C., Du, W., Li, X., & Han, Y. (2022). Identification of subjective cognitive decline due to Alzheimer's disease using multimodal MRI combining with machine learning. *Cerebral Cortex*. https://doi.org/10.1093/cercor/bhac084

Lin, L., Zhang, B., & Wu, S. (2018). Hybrid CNN-SVM for Alzheimer's Disease Classification from Structural MRI and the Alzheimer's Disease Neuroimaging Initiative (ADNI). *2018 International Conference on Biomedical Engineering, Machinery and Earth Science*. https://doi.org/10.25236/bemes.2018.041

Lin, W., Tong, T., Gao, Q., Guo, D., Du, X., Yang, Y., … Qu, X. (2018). Convolutional neural networks-based MRI image analysis for the Alzheimer's disease prediction from mild cognitive impairment. *Frontiers in Neuroscience*, *12*(NOV). https://doi.org/10.3389/fnins.2018.00777

Liu, L., Zhao, S., Chen, H., & Wang, A. (2020). A new machine learning method for identifying Alzheimer's disease. *Simulation Modelling Practice and Theory*, *99*(November 2019), 102023. https://doi.org/10.1016/j.simpat.2019.102023

Liu, S., Liu, S., Cai, W., Pujol, S., Kikinis, R., & Feng, D. (2014). Early Diagnosis of Alzheimer's Disease with Deep Learning. *2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)*, 911–914. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6868019

Liu, Xialei, Weijer, J. Van De, & Bagdanov, A. D. (2019). Exploiting Unlabeled Data in CNNs by Self-Supervised Learning to Rank. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(8), 1862–1878. https://doi.org/10.1109/TPAMI.2019.2899857

Liu, Xin, Tian, S., Tao, F., & Yu, W. (2021). A review of artificial neural networks in the constitutive modeling of composite materials. *Composites Part B: Engineering*, *224*. https://doi.org/10.1016/J.COMPOSITESB.2021.109152

Liu, Xin, Tosun, D., Weiner, M. W., & Schuff, N. (2013). Locally linear embedding (LLE) for MRI based Alzheimer's disease classification. *NeuroImage*, *83*, 148–157. https://doi.org/10.1016/j.neuroimage.2013.06.033

Long, X., Chen, L., Jiang, C., & Zhang, L. (2017). Prediction and classification of Alzheimer disease based on quantification of MRI deformation. *PLoS ONE*, *12*(3). https://doi.org/10.1371/journal.pone.0173372

Long, X., & Wyatt, C. (2010). An automatic unsupervised classification of MR images in Alzheimer's disease. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2910–2917. https://doi.org/10.1109/CVPR.2010.5540031

Lopera, F., Ardilla, A., Martínez, A., Madrigal, L., & Arango-viana, J. C. (1997). Clinical Features of Early-Onset Alzheimer Disease in a Large Kindred With an E280A Presenilin-1 Mutation. *The Journal of the American Medical Association*, *277*(10), 793–799. https://doi.org/10.1001/jama.277.10.793

Lu, D., Popuri, K., Ding, G. W., Balachandar, R., Beg, M. F., Weiner, M., … Fargher, K. (2018). Multimodal and Multiscale Deep Neural Networks for the Early Diagnosis of Alzheimer's Disease using structural MR and FDG-PET images. *Scientific Reports*, *8*(1), 1–13. https://doi.org/10.1038/s41598-018-22871-z

Luo, M., He, Z., Cui, H., Chen, Y. P. P., & Ward, P. (2023). Class activation attention transfer neural networks for MCI conversion prediction. *Computers in Biology and Medicine*, *156*(August 2022). https://doi.org/10.1016/j.compbiomed.2023.106700

Ma, Z., Jing, B., Li, Y., Yan, H., Li, Z., Ma, X., … Li, H. (2020). Identifying Mild Cognitive Impairment with Random Forest by Integrating Multiple MRI Morphological Metrics / Identifying MCI with Random Forest by Integrating Multiple MRI Morphological Metrics. *Journal of Alzheimer's Disease*, *73*, 991–1002. https://doi.org/10.3233/JAD-190715

Macedonia, M. (2003). The GPU enters computing's mainstream. *IEEE Computer*, *36*(10), 106–108. https://doi.org/10.1109/MC.2003.1236476

Macqueen, J. (1967). Some methods for classification and analysis of multivariate observations.pdf. *Proc. of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, (1), 281–297.

Massalimova, A., & Varol, H. A. (2021). Input Agnostic Deep Learning for Alzheimer's Disease Classification Using Multimodal MRI Images. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, 2875–2878. https://doi.org/10.1109/EMBC46164.2021.9629807

McKhann, G., Drachman, D., Folstein, M., Katzman, R., Price, D., & Stadlan, E. (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. *Neurology*, *34*, 939–944. Retrieved from http://n.neurology.org/content/neurology/34/7/939.full.pdf

Mehmood, A., yang, S., feng, Z., wang, M., Ahmad, A. S., khan, R., … Yaqub, M. (2021). A Transfer Learning Approach for Early Diagnosis of Alzheimer's Disease on MRI

Images. *Neuroscience*, *460*, 43–52.
https://doi.org/10.1016/J.NEUROSCIENCE.2021.01.002

Mhasawade, V., Zhao, Y., & Chunara, R. (2020). Machine Learning in Population and Public Health. *In ACM Conference on Health, Inference, and Learning (ACM CHIL '20)*. Retrieved from http://arxiv.org/abs/2008.07278

Miyakawa, T., & Uehara, Y. (1979). Observations of amyloid angiopathy and senile plaques by the scanning electron microscope. *Acta Neuropathologica*, *48*(2), 153–156. https://doi.org/10.1007/BF00691158

Mohammed, M., Khan, M. B., & Bashie, E. B. M. (2006). Machine learning: algorithms and applications. In *CRC Press* (Vol. 1999).

Mohammed, M., Khan, M. B., & Bashie, E. B. M. (2016). Machine learning: Algorithms and applications. In *Machine Learning: Algorithms and Applications*. https://doi.org/10.1201/9781315371658

Moradi, E., Gaser, C., Huttunen, H., & Tohka, J. (2014). MRI based dementia classification using semi-supervised learning and domain adaptation. *Proc MICCAI Workshop Challenge on Computer-Aided Diagnosis of Dementia Based on Structural MRI Data*, 65–73. Retrieved from http://dbm.neuro.uni-jena.de/

Moradi, E., Hallikainen, I., Hänninen, T., & Tohka, J. (2017). Rey's Auditory Verbal Learning Test scores can be predicted from whole brain MRI in Alzheimer's disease. *NeuroImage: Clinical*, *13*, 415–427. https://doi.org/10.1016/j.nicl.2016.12.011

Moradi, E., Pepe, A., Gaser, C., Huttunen, H., & Tohka, J. (2015). Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects. *NeuroImage*, *104*, 398–412. https://doi.org/10.1016/j.neuroimage.2014.10.002

Muhammed Raees, P. C., & Thomas, V. (2021). Automated detection of Alzheimer's Disease using Deep Learning in MRI. *Journal of Physics: Conference Series*, 12024. https://doi.org/10.1088/1742-6596/1921/1/012024

Nanni, L., Brahnam, S., Ghidoni, S., Menegatti, E., & Barrier, T. (2013). Different Approaches for Extracting Information from the Co-Occurrence Matrix. *PLoS ONE*, *8*(12), 83554. https://doi.org/10.1371/journal.pone.0083554

National Health Services (UK). (n.d.-a). About Dementia. Retrieved from https://www.nhs.uk/conditions/dementia/about/

National Health Services (UK). (n.d.-b). MRI scans. Retrieved from https://www.nhs.uk/conditions/mri-scan/what-happens/

National Health Services (UK). (2018). MRI scan - How it's performed - NHS. Retrieved August 5, 2022, from https://www.nhs.uk/conditions/mri-scan/what-happens/

Nigri, E., Ziviani, N., Cappabianco, F., Antunes, A., & Veloso, A. (2020). Explainable Deep CNNs for MRI-Based Diagnosis of Alzheimer's Disease. *Proceedings of the International Joint Conference on Neural Networks*. https://doi.org/10.1109/IJCNN48605.2020.9206837

Ortega Hinojosa, A. M., Davies, M. M., Jarjour, S., Burnett, R. T., Mann, J. K., Hughes, E.,

… Jerrett, M. (2014). Developing small-area predictions for smoking and obesity prevalence in the United States for use in Environmental Public Health Tracking. *Environmental Research*, *134*, 435–452. https://doi.org/10.1016/j.envres.2014.07.029

Ortiz, A., Górriz, J. M., Ramírez, J., & Martínez-Murcia, F. J. (2013). LVQ-SVM based CAD tool applied to structural MRI for the diagnosis of the Alzheimer's disease. *Pattern Recognition Letters*, *34*(14), 1725–1733. https://doi.org/10.1016/J.PATREC.2013.04.014

Pang, Z., Wang, · Xiang, Wang, X., Qi, J., Zhong Zhao, ·, Yuan Gao, ·, … Yang, P. (2021). A Multi-modal Data Platform for Diagnosis and Prediction of Alzheimer's Disease Using Machine Learning Methods. *Mobile Networks and Applications*, *1*, 3. https://doi.org/10.1007/s11036-021-01834-1

Papakostas, G. A., Savio, A., Graña, M., & Kaburlasos, V. G. (2015). A lattice computing approach to Alzheimer's disease computer assisted diagnosis based on MRI data. *Neurocomputing*. https://doi.org/10.1016/j.neucom.2014.02.076

Patil, A., & Rane, M. (2021). Convolutional Neural Networks: An Overview and Its Applications in Pattern Recognition. *Smart Innovation, Systems and Technologies*, *195*, 21–30. https://doi.org/10.1007/978-981-15-7078-0_3

Petrone, P. M., Casamitjana, A., Falcon, C., Artigues, M., Operto, G., Cacciaglia, R., … Gispert, J. D. (2019). Prediction of amyloid pathology in cognitively unimpaired individuals using voxel-wise analysis of longitudinal structural brain MRI. *Alzheimer's Research and Therapy*, *11*(1), 1–13. https://doi.org/10.1186/s13195-019-0526-8

Petsani, D., Konstantinidis, E., Katsouli, A. M., Zilidou, V., Dias, S. B., Hadjileontiadis, L., & Bamidis, P. (2022). Digital Biomarkers for Well-being Through Exergame Interactions: Exploratory Study. *JMIR Serious Games*, *10*(3), 1–15. https://doi.org/10.2196/34768

Plant, C., Teipel, S. J., Oswald, A., Böhm, C., Meindl, T., Mourao-Miranda, J., … Ewers, M. (2010). Automated detection of brain atrophy patterns based on MRI for the prediction of Alzheimer's disease. *NeuroImage*, *50*(1), 162–174. https://doi.org/10.1016/j.neuroimage.2009.11.046

Prins, N. D., & Scheltens, P. (2015). White matter hyperintensities, cognitive impairment and dementia: an update. *NATURE REVIEWS | NEUROLOGY*, *11*, 157–165. https://doi.org/10.1038/nrneurol.2015.10

Qiao, S., Shen, W., Zhang, Z., Wang, B., & Yuille, A. (2018). Deep co-training for semi-supervised image recognition. *European Conference on Computer Vision*, *11219 LNCS*, 142–159. https://doi.org/10.1007/978-3-030-01267-0_9

Raghavaiah, P., & Varadarajan, S. (2022). A CAD system design for Alzheimer's disease diagnosis using temporally consistent clustering and hybrid deep learning models. *Biomedical Signal Processing and Control*, *75*. https://doi.org/10.1016/J.BSPC.2022.103571

Rasch, D., Kubinger, K. D., & Moder, K. (2011). The two-sample t test: Pre-testing its assumptions does not pay off. *Statistical Papers*, *52*(1), 219–231. https://doi.org/10.1007/s00362-009-0224-x

Raza, K., & Singh, N. K. (2021). A Tour of Unsupervised Deep Learning for Medical Image Analysis. *Current Medical Imaging Formerly Current Medical Imaging Reviews*, *17*(9), 1059–1077. https://doi.org/10.2174/1573405617666210127154257

Razavi, F., Tarokh, M. J., & Alborzi, M. (2019). An intelligent Alzheimer's disease diagnosis method using unsupervised feature learning. *Journal of Big Data*, *6*(1). https://doi.org/10.1186/s40537-019-0190-7

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1987). Learning Internal Representations by Error Propagation. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations* (pp. 318–362). https://doi.org/10.1016/B978-1-4832-1446-7.50035-2

Saim, M., & Amel, F. (2022). Classification and Diagnosis of Alzheimer's Disease based on a combination of Deep Features and Machine Learning. *2022 7th International Conference on Image and Signal Processing and Their Applications (ISPA)*, 1–6. https://doi.org/10.1109/ispa54004.2022.9786318

Salunkhe, S., Bachute, M., Gite, S., Vyas, N., Khanna, S., Modi, K., … Kotecha, K. (2021). Classification of alzheimer's disease patients using texture analysis and machine learning. *Applied System Innovation*, *4*(3). https://doi.org/10.3390/asi4030049

Salvatore, C., Cerasa, A., Battista, P., Gilardi, M. C., Quattrone, A., & Castiglioni, I. (2015). Magnetic resonance imaging biomarkers for the early diagnosis of Alzheimer's disease: A machine learning approach. *Frontiers in Neuroscience*, *9*(SEP), 1–13. https://doi.org/10.3389/fnins.2015.00307

Samper-González, J., Burgos, N., Bottani, S., Fontanella, S., Lu, P., Marcoux, A., … Colliot, O. (2018). Reproducible evaluation of classification methods in Alzheimer's disease: Framework and application to MRI and PET data. *NeuroImage*, *183*, 504–521. https://doi.org/10.1016/J.NEUROIMAGE.2018.08.042

Saratxaga, C. L., Moya, I., Picón, A., Acosta, M., Moreno-Fernandez-de-leceta, A., Garrote, E., & Bereciartua-Perez, A. (2021). Mri deep learning-based solution for alzheimer's disease prediction. *Journal of Personalized Medicine*, *11*(9). https://doi.org/10.3390/jpm11090902

Saravanakumar, S., & Saravanan, T. (2022). Early Alzheimer's Disease Detection Using SemiSupervised GAN Based on Deep Learning. *Circuit and System Conference 2022*, *1*(2), 1–24. IEEE.

Saravanakumar, S., Saravanan, T., & Thillaiarasu, N. (2022). A Deep Learning-based Semi-Supervised GAN to detect Alzheimer ' s illness efficiently. *2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, 1840–1843.

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, *2*, 160. https://doi.org/10.1007/s42979-021-00592-x

Sarraf, S., & Tofighi, G. (2016). Classification of Alzheimer's Disease Structural MRI Data by Deep Learning Convolutional Neural Networks. *ArXiv Preprint ArXiv:1603.08631*. Retrieved from http://arxiv.org/abs/1607.06583

Sathiyamoorthi, V., Ilavarasi, A. K., Murugeswari, K., Thouheed Ahmed, S., Aruna Devi, B., & Kalipindi, M. (2021). A deep convolutional neural network based computer aided diagnosis system for the prediction of Alzheimer's disease in MRI images. *Measurement: Journal of the International Measurement Confederation*, *171*. https://doi.org/10.1016/J.MEASUREMENT.2020.108838

Sethi, J. K., & Mittal, M. (2020). Monitoring the impact of air quality on the COVID-19 fatalities in Delhi, India: Using machine learning techniques. *Disaster Medicine and Public Health Preparedness*. https://doi.org/10.1017/dmp.2020.372

Seyedi, S., Jafari, R., Talaei, A., Naseri, S., Momennezhad, M., Moghaddam, M. D., & Akbari-Lalimi, H. (2020). Comparing VBM and ROI analyses for detection of gray matter abnormalities in patients with bipolar disorder using MRI. *Middle East Current Psychiatry*, *27*(1). https://doi.org/10.1186/s43045-020-00076-3

Shanmugam, J. V., Duraisamy, B., Simon, B. C., & Bhaskaran, P. (2022). Alzheimer's disease classification using pre-trained deep networks. *Biomedical Signal Processing and Control*, *71*(PB), 103217. https://doi.org/10.1016/j.bspc.2021.103217

Sharma, R., Goel, T., Tanveer, M., & Murugan, R. (2022). FDN-ADNet: Fuzzy LS-TWSVM based deep learning network for prognosis of the Alzheimer's disease using the sagittal plane of MRI scans. *Applied Soft Computing*, *115*. https://doi.org/10.1016/J.ASOC.2021.108099

Shinde, P. P., & Shah, S. (2018). A Review of Machine Learning and Deep Learning Applications. *Proceedings - 2018 4th International Conference on Computing, Communication Control and Automation, ICCUBEA 2018*. https://doi.org/10.1109/ICCUBEA.2018.8697857

Shurrab, S., & Duwairi, R. (2022). Self-supervised learning methods and applications in medical imaging analysis: a survey. *PeerJ Computer Science*, *8*. https://doi.org/10.7717/PEERJ-CS.1045

Sontakke, S., Lohokare, J., & Dani, R. (2017). Diagnosis of liver diseases using machine learning. *2017 International Conference on Emerging Trends and Innovation in ICT, ICEI 2017*, 129–133. https://doi.org/10.1109/ETIICT.2017.7977023

Sordo, M. (2002). Introduction to Neural Networks in Healthcare. *Knowledge Management for Medical Care*, *73*(2), 17. Retrieved from http://www.openclinical.org/docs/int/neuralnetworks011.pdf

Spasov, S., Passamonti, L., Duggento, A., Liò, P., & Toschi, N. (2019). A parameter-efficient deep learning approach to predict conversion from mild cognitive impairment to Alzheimer's disease. *NeuroImage*, *189*, 276–287. https://doi.org/10.1016/j.neuroimage.2019.01.031

Srivastava, A. A. B. L. (1958). Effect of Non-Normality on the Power Function of t-Test. *Biomerika*, *45*(3), 421–430.

Strittmatter, W. J., Saunders, A. M., Schmechel, D., Pericak-Vance, M., Enghild, J., Salvesen, G. S., & Roses, A. D. (1993). Apolipoprotein E: High-avidity binding to β-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer disease. *Proceedings of the National Academy of Sciences of the United States of*

America, *90*(5), 1977–1981. https://doi.org/10.1073/pnas.90.5.1977

Suetens, P. (2017). Fundamentals of medical imaging. *Fundamentals of Medical Imaging*, (February), 1–257. https://doi.org/10.1017/9781316671849

Sujathakumari, B. A., Abhishek, M. H., Darshan Singh, S., Aneesh, K. N., Rakesh, D. S., & Mahanand, B. S. (2019). Detection of MCI from MRI using Gradient Boosting Classifier. *1st IEEE International Conference on Advances in Information Technology, ICAIT 2019 - Proceedings*, 70–75. https://doi.org/10.1109/ICAIT47043.2019.8987413

Sun, Z., Qiao, Y., Lelieveldt, B. P. F., & Staring, M. (2018). Integrating spatial-anatomical regularization and structure sparsity into SVM: Improving interpretation of Alzheimer's disease classification. *NeuroImage*, *178*(April 2018), 445–460. https://doi.org/10.1016/j.neuroimage.2018.05.051

Symms, M., Jäger, H. R., Schmierer, K., & Yousry, T. A. (2004). A review of structural magnetic resonance neuroimaging. *Journal of Neurology, Neurosurgery and Psychiatry*, *75*(9), 1235–1244. https://doi.org/10.1136/jnnp.2003.032714

Tan, P.-N., Steinbach, M., Karpatne, A., & Kumar, V. (2019). *Introduction to Data Mining* (Second). New York, NY: Person Education.

Tang, X., & Liu, J. (2021). Comparing different algorithms for the course of alzheimer's disease using machine learning. *Annals of Palliative Medicine*, *10*(9), 9715–9724. https://doi.org/10.21037/apm-21-2013

Teplan, M. (2002). Fundamentals of EEG Measurement. *Measurement Science Review*, *2*.

Tong, T., Gray, K., Gao, Q., Chen, L., & Rueckert, D. (2017). Multi-modal classification of Alzheimer's disease using nonlinear graph fusion. *Pattern Recognition*, *63*, 171–181. https://doi.org/10.1016/J.PATCOG.2016.10.009

Tong, T., Wolz, R., Gao, Q., Guerrero, R., Hajnal, J. V., & Rueckert, D. (2014). Multiple instance learning for classification of dementia in brain MRI. *Medical Image Analysis*, *18*(5), 808–818. https://doi.org/10.1016/J.MEDIA.2014.04.006

Vaka, A. R., Soni, B., & K., S. R. (2020). Breast cancer detection by leveraging Machine Learning. *ICT Express*, *6*(4), 320–324. https://doi.org/10.1016/J.ICTE.2020.04.009

Valliani, A., & Soni, A. (2017). Deep residual nets for improved Alzheimer's diagnosis. *ACM-BCB 2017 - Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, 615. https://doi.org/10.1145/3107411.3108224

Vichianin, Y., Khummongkol, A., Chiewvit, P., Raksthaput, A., Chaichanettee, S., Aoonkaew, N., & Senanarong, V. (2021). Accuracy of Support-Vector Machines for Diagnosis of Alzheimer's Disease, Using Volume of Brain Obtained by Structural MRI at Siriraj Hospital. *Frontiers in Neurology*, *12*. https://doi.org/10.3389/FNEUR.2021.640696/FULL

Volkau, I., Puspitasari, F., Ng, T. T., Bhanu Prakash, K. N., Gupta, V., & Nowinski, W. L. (2012). A simple and fast method of 3D registration and statistical landmark localization for sparse multi-modal/time-series neuroimages based on cortex ellipse fitting. *Neuroradiology Journal*, *25*(1), 98–111. https://doi.org/10.1177/197140091202500114

Wang, Hongfei, Shen, Y., Wang, S., Xiao, T., Deng, L., Wang, X., & Zhao, X. (2019). Ensemble of 3D densely connected convolutional network for diagnosis of mild cognitive impairment and Alzheimer's disease. *Neurocomputing*, *333*, 145–156. https://doi.org/10.1016/j.neucom.2018.12.018

Wang, Huan, Sheng, L., Xu, S., Jin, Y., Jin, X., Qiao, S., … Xu, X. (2022). Develop a diagnostic tool for dementia using machine learning and non-imaging features. *Frontiers in Aging Neuroscience*, *14*(August), 1–14. https://doi.org/10.3389/fnagi.2022.945274

Wang, J., He, L., Zheng, H., & Lu, Z. L. (2014). Optimizing the Magnetization-Prepared Rapid Gradient-Echo (MP-RAGE) sequence. *PLoS ONE*, *9*(5). https://doi.org/10.1371/journal.pone.0096899

Wang, Y., Xu, C., Park, J. H., Lee, S., Stern, Y., Yoo, S., … Cha, J. (2019). Diagnosis and prognosis of Alzheimer's disease using brain morphometry and white matter connectomes. *NeuroImage: Clinical*. https://doi.org/10.1016/j.nicl.2019.101859

Wardlaw, J. M., Brindle, W., Casado, A. M., Shuler, K., Henderson, M., Thomas, B., … Hernandez, M. C. V. (2012). A systematic review of the utility of 1.5 versus 3 Tesla magnetic resonance brain imaging in clinical practice and research. *European Radiology*, *22*(11), 2295–2303. https://doi.org/10.1007/s00330-012-2500-8

Wee, C. Y., Yap, P. T., & Shen, D. (2013). Prediction of Alzheimer's disease and mild cognitive impairment using cortical morphological patterns. *Human Brain Mapping*, *34*(12), 3411–3425. https://doi.org/10.1002/hbm.22156

Wellcome Centre for Human Neuroimaging. (2020). Statistical Parametric Mapping. Retrieved from https://www.fil.ion.ucl.ac.uk/spm/

Weller, J., & Budson, A. (2018). Current understanding of Alzheimer's disease diagnosis and treatment. *F1000Research*, *7*. https://doi.org/10.12688/f1000research.14506.1

Wen, J., Samper-González, J., Bottani, S., Routier, A., Burgos, N., Jacquemont, T., … Colliot, O. (2021). Reproducible Evaluation of Diffusion MRI Features for Automatic Classification of Patients with Alzheimer's Disease. *Neuroinformatics*, *19*(1), 57–78. https://doi.org/10.1007/s12021-020-09469-5

Wen, J., Thibeau-Sutre, E., Diaz-Melo, M., Samper-González, J., Routier, A., Bottani, S., … Colliot, O. (2020). Convolutional neural networks for classification of Alzheimer's disease: Overview and reproducible evaluation. *Medical Image Analysis*, *63*, 101694. https://doi.org/10.1016/J.MEDIA.2020.101694

Wen, J., Thibeau-Sutre, E., Samper-González, J., Routier, A., Bottani, S., Dormont, D., … Burgos, N. (2019). How serious is data leakage in deep learning studies on Alzheimer's disease classification? *OHBM Annual Meeting*. Retrieved from https://hal.inria.fr/hal-02105133v2

Wolz, R., Julkunen, V., Koikkalainen, J., Niskanen, E., Zhang, D. P., Rueckert, D., … Lötjönen, J. (2011). Multi-method analysis of MRI images in early diagnostics of Alzheimer's disease. *PLoS ONE*, *6*(10), 1–9. https://doi.org/10.1371/journal.pone.0025446

World Health Organisation. (2023). Dementia. Retrieved from https://www.who.int/news-

room/fact-sheets/detail/dementia

Wu, C. C., Yeh, W. C., Hsu, W. D., Islam, M. M., Nguyen, P. A. (Alex), Poly, T. N., … (Jack) Li, Y. C. (2019). Prediction of fatty liver disease using machine learning algorithms. *Computer Methods and Programs in Biomedicine*, *170*, 23–29. https://doi.org/10.1016/J.CMPB.2018.12.032

Xie, Y., Meng, W. Y., Li, R. Z., Wang, Y. W., Qian, X., Chan, C., … Leung, E. L. H. (2021). Early lung cancer diagnostic biomarker discovery by machine learning methods. *Translational Oncology*, *14*(1). https://doi.org/10.1016/J.TRANON.2020.100907

Xu, L., Yao, Z., Li, J., Lv, C., Zhang, H., & Hu, B. (2019). Sparse Feature Learning With Label Information for Alzheimer's Disease Classification Based on Magnetic Resonance Imaging. *IEEE Access*, *7*, 26157–26167. https://doi.org/10.1109/ACCESS.2019.2894530

Xu, Y., Wang, Y., Yuan, J., Cheng, Q., Wang, X., & Carson, P. L. (2019). Medical breast ultrasound image segmentation by machine learning. *Ultrasonics*, *91*, 1–9. https://doi.org/10.1016/J.ULTRAS.2018.07.006

Xue, J., Chen, J., Hu, R., Chen, C., Zheng, C., Su, Y., & Zhu, T. (2020). Twitter discussions and emotions about the COVID-19 pandemic: Machine learning approach. *Journal of Medical Internet Research*, *22*(11). https://doi.org/10.2196/20550

Ye, D. H., Pohl, K. M., & Davatzikos, C. (2011). Semi-supervised pattern classification: Application to structural MRI of Alzheimer's disease. *Proceedings - International Workshop on Pattern Recognition in NeuroImaging, PRNI 2011*, 1–4. https://doi.org/10.1109/PRNI.2011.12

Yeo, B. T. T., Tandi, J., & Chee, M. W. L. (2015). Functional connectivity during rested wakefulness predicts vulnerability to sleep deprivation. *NeuroImage*, *111*, 147–158. https://doi.org/10.1016/J.NEUROIMAGE.2015.02.018

Yi, X., Walia, E., & Babyn, P. (2019). Generative adversarial network in medical imaging: A review. *Medical Image Analysis*, *58*. https://doi.org/10.1016/J.MEDIA.2019.101552

Yu, L., Xiang, W., Fang, J., Phoebe Chen, Y. P., & Zhu, R. (2022). A novel explainable neural network for Alzheimer's disease diagnosis. *Pattern Recognition*, *131*. https://doi.org/10.1016/J.PATCOG.2022.108876

Zeng, N., Li, H., & Peng, Y. (2021). A new deep belief network-based multi-task learning for diagnosis of Alzheimer's disease. *Neural Computing and Applications*. https://doi.org/10.1007/s00521-021-06149-6

Zeng, N., Qiu, H., Wang, Z., Liu, W., Zhang, H., & Li, Y. (2018). A new switching-delayed-PSO-based optimized SVM algorithm for diagnosis of Alzheimer's disease. *Neurocomputing*. https://doi.org/10.1016/j.neucom.2018.09.001

Zhang, D., & Shen, D. (2011). Semi-supervised multimodal classification of alzheimer's disease. *2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, 1628–1631. IEEE.

Zhang, D., Wang, Y., Zhou, L., Yuan, H., & Shen, D. (2011). Multimodal classification of Alzheimer's disease and mild cognitive impairment. *NeuroImage*, *55*(3), 856–867.

https://doi.org/10.1016/j.neuroimage.2011.01.008

Zhang, J., Li, T., Lu, X., & Cheng, Z. (2016). Semantic Classification of High-Resolution Remote-Sensing Images Based on Mid-level Features. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *9*(6), 2343–2353. https://doi.org/10.1109/JSTARS.2016.2536943

Zhang, P., Shi, B., Smith, C. D., & Liu, J. (2016). Nonlinear Metric Learning for Semi-Supervised Learning via Coherent Point Drifting. *15th IEEE International Conference on Machine Learning and Applications*, 314–319. https://doi.org/10.1109/icmla.2016.0058

Zhang, X., Han, L., Zhu, W., Sun, L., & Zhang, D. (2021). An Explainable 3D Residual Self-Attention Deep Neural Network For Joint Atrophy Localization and Alzheimer's Disease Diagnosis using Structural MRI. *IEEE Journal of Biomedical and Health Informatics*, *26*(11), 5289–5297. https://doi.org/10.1109/JBHI.2021.3066832

Zhang, Yanteng, Teng, Q., Liu, Y., Liu, Y., & He, X. (2022). Diagnosis of Alzheimer's disease based on regional attention with sMRI gray matter slices. *Journal of Neuroscience Methods*, *365*(October 2021), 109376. https://doi.org/10.1016/j.jneumeth.2021.109376

Zhang, Yudong, Dong, Z., Phillips, P., Wang, S., Ji, G., Yang, J., & Yuan, T. F. (2015). Detection of subjects and brain regions related to Alzheimer's disease using 3D MRI scans based on eigenbrain and machine learning. *Frontiers in Computational Neuroscience*, *9*(JUNE), 1–15. https://doi.org/10.3389/fncom.2015.0006

Zhang, Yudong, & Wang, S. (2015). Detection of Alzheimer's disease by displacement field and machine learning. *PeerJ*, *2015*(9). https://doi.org/10.7717/peerj.1251

Zhang, Z. M., Tan, J. X., Wang, F., Dao, F. Y., Zhang, Z. Y., & Lin, H. (2020). Early Diagnosis of Hepatocellular Carcinoma Using Machine Learning Method. *Frontiers in Bioengineering and Biotechnology*, *8*. https://doi.org/10.3389/FBIOE.2020.00254/FULL

Zhao, S., Guo, Y., Sheng, Q., & Shyr, Y. (2014). Advanced Heat Map and Clustering Analysis Using Heatmap3. *BioMed Research International*, *2014*. https://doi.org/10.1155/2014/986048

Zheng, B., Gao, A., Huang, X., Li, Y., Liang, D., & Long, X. (2022). A modified 3D EfficientNet for the classification of Alzheimer's disease using structural magnetic resonance images. *IET Image Processing*. https://doi.org/10.1049/IPR2.12618

Zhu, A., Lee, D., & Shim, H. (2011). Metabolic positron emission tomography imaging in cancer detection and therapy response. *Seminars in Oncology*, *38*(1), 55–69. https://doi.org/10.1053/j.seminoncol.2010.11.012

Zhu, J., Shi, J., Liu, X., & Chen, X. (2014). Co-Training Based Semi-supervised Classification of Alzheimer ' s Disease. *Proceedings of the 19th International Conference on Digital Signal Processing*. IEEE.

Zoabi, Y., Deri-Rozov, S., & Shomron, N. (2021). Machine learning-based prediction of COVID-19 diagnosis based on symptoms. *Npj Digital Medicine*, *4*(1). https://doi.org/10.1038/s41746-020-00372-6