Diagnosis of Cognitive Impairment using Multiple Data Modalities

Upul Senanayake

A Thesis presented for the degree of
Doctor of Philosophy

School of Computer Science and Engineering

University of New South Wales

Australia

February 27, 2022

Thesis/Dissertation Sheet

Surname/Family Name	:	**Panamaldeniya Mudiyanselage**
Given Name/s	:	**Upul Chathuranga Bandara Senanayake**
Abbreviation for degree as give in the University calendar	:	**PhD**
Faculty	:	**Engineering**
School	:	**Computer Science and Engineering**
Thesis Title	:	**Diagnosis of Cognitive Impairment using Multiple Data Modalities**

Abstract 350 words maximum: (PLEASE TYPE)

Decline in cognitive functions including memory, processing speed and executive processes, has been associated with ageing for sometime. Differentiating between cognitive decline due to a pathological process and normal ageing is an ongoing research challenge. According to the definition of the World Health Organization (WHO), dementia is an umbrella term for several diseases affecting memory and other cognitive abilities and behaviour that interfere significantly with the ability to maintain daily living activities. Although a cure for dementia has not been found yet, it is often stressed that early identification of individuals at risk of dementia can be instrumental in treatment and management. Mild Cognitive Impairment (MCI) is a prodromal condition to dementia, and patients with MCI have a higher probability of progressing to certain types of dementia, the most common being Alzheimer's Disease (AD). Therefore, accurate and early diagnosis of MCI may be useful.

Traditionally, clinicians use a few neuropsychological tests (also called NM features) to evaluate and diagnose cognitive decline in individuals. In contrast, computer aided diagnostic techniques often focus on medical imaging modalities such as magnetic resonance imaging (MRI) and positron emission tomography (PET). This thesis utilises machine learning and deep learning techniques to leverage both data modalities in a single end-to-end pipeline that is robust to missing information. Several techniques have been designed, implemented and validated to diagnose different types of cognitive impairment including mild cognitive impairment and its subtypes as well as dementia, initially directly from NM features, and then in fusion with medical imaging features.

The novel techniques proposed by this thesis build end-to-end deep learning pipelines that are capable of learning to extract features and engineering combinations of features to yield the best performance. The proposed deep fusion pipeline is capable of fusing data from multiple disparate modalities of vastly different dimensions seamlessly. Survival analysis techniques are often used to understand the progression and time till an event of interest. In this thesis, the proposed deep survival analysis techniques are used to better understand the progression to dementia. They also enable the use of imaging data seamlessly with NM features, which is the first such approach as far as is known.

FOR OFFICE USE ONLY Date of completion of requirements for Award:

ORIGINALITY STATEMENT

‘I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the award of any other degree or diploma at UNSW or any other educational institution, except where due acknowledgement is made in the thesis. Any contribution made to the research by others, with whom I have worked at UNSW or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except to the extent that assistance from others in the project's design and conception or in style, presentation and linguistic expression is acknowledged.’

Signed ……..

Date ……..

INCLUSION OF PUBLICATIONS STATEMENT

UNSW is supportive of candidates publishing their research results during their candidature as detailed in the UNSW Thesis Examination Procedure.

Publications can be used in their thesis in lieu of a Chapter if:

- The student contributed greater than 50% of the content in the publication and is the "primary author", ie. the student was responsible primarily for the planning, execution and preparation of the work for publication
- The student has approval to include the publication in their thesis in lieu of a Chapter from their supervisor and Postgraduate Coordinator.
- The publication is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in the thesis

Please indicate whether this thesis contains published material or not.

☐ *This thesis contains no publications, either published or submitted for publication*

☒ *Some of the work described in this thesis has been published and it has been documented in the relevant Chapters with acknowledgement*

☐ *This thesis has publications (either published or submitted for publication) incorporated into it in lieu of a chapter and the details are presented below*

CANDIDATE'S DECLARATION

I declare that:

- I have complied with the Thesis Examination Procedure
- where I have used a publication in lieu of a Chapter, the listed publication(s) below meet(s) the requirements to be included in the thesis.

Name	Signature	Date (dd/mm/yy)

Postgraduate Coordinator's Declaration (to be filled in where publications are used in lieu of Chapters)

I declare that:

- the information below is accurate
- where listed publication(s) have been used in lieu of Chapter(s), their use complies with the Thesis Examination Procedure
- the minimum requirements for the format of the thesis have been met.

PGC's Name	PGC's Signature	Date (dd/mm/yy)

COPYRIGHT STATEMENT

Signed …….....

Date …….....

AUTHENTICITY STATEMENT

‘I certify that the Library deposit digital copy is a direct equivalent of the final officially approved version of my thesis.’

Signed …….....

Date …….....

Abstract

Decline in cognitive functions including memory, processing speed and executive processes, has been associated with ageing for sometime. It is understood that every human will go through this process, but some will go through it faster, and for some this process starts earlier. Differentiating between cognitive decline due to a pathological process and normal ageing is an ongoing research challenge. According to the definition of the World Health Organization (WHO), dementia is an umbrella term for a number of diseases affecting memory and other cognitive abilities and behaviour that interfere significantly with the ability to maintain daily living activities. Although a cure for dementia has not been found yet, it is often stressed that early identification of individuals at risk of dementia can be instrumental in treatment and management. Mild Cognitive Impairment (MCI) is considered to be a prodromal condition to dementia, and patients with MCI have a higher probability of progressing to certain types of dementia, the most common being Alzheimer's Disease (AD). Epidemiological studies suggest that the progression rate from MCI to dementia is around 10-12% annually, while much lower in the general elderly population. Therefore, accurate and early diagnosis of MCI may be useful, as those patients can be closely monitored for progression to dementia.

Traditionally, clinicians use a number of neuropsychological tests (also called NM features) to evaluate and diagnose cognitive decline in individuals. In contrast, computer aided diagnostic techniques often focus on medical imaging modalities such as magnetic resonance imaging (MRI) and positron emission tomography (PET). This thesis utilises machine learning and deep learning techniques to leverage both of these data modalities in a single end-to-end pipeline that is robust to missing information. A number of techniques have been designed, implemented and validated to diagnose different types of cognitive impairment including mild cognitive impairment and its subtypes as well as dementia, initially directly from NM features, and then

in fusion with medical imaging features.

The novel techniques proposed by this thesis build end-to-end deep learning pipelines that are capable of learning to extract features and engineering combinations of features to yield the best performance. The proposed deep fusion pipeline is capable of fusing data from multiple disparate modalities of vastly different dimensions seamlessly. Survival analysis techniques are often used to understand the progression and time till an event of interest. In this thesis, the proposed deep survival analysis techniques are used to better understand the progression to dementia. They also enable the use of imaging data seamlessly with NM features, which is the first such approach as far as is known. The techniques are designed, implemented and validated across two datasets; an in-house dataset and a publicly available dataset adding an extra layer of cross validation. The proposed techniques can be used to differentiate between cognitively impaired and cognitively normal individuals and gain better insights on their subsequent progression to dementia.

Publications Arising from Thesis

1. Senanayake, U., Sowmya, A., Dawes, L., Kochan, N. A., Wen, W., Sachdev, P. (2016). Classification of mild cognitive impairment subtypes using neuropsychological data, in ICPRAM 2016: Proceedings of the 5th International Conference on Pattern Recognition Applications and Methods, pp. 620-629

2. Senanayake U;Sowmya A;Dawes L;Kochan NA;Wen W;Sachdev P, 2017, Deep learning approach for classification of mild cognitive impairment subtypes, in ICPRAM 2017 - Proceedings of the 6th International Conference on Pattern Recognition Applications and Methods, pp. 655 - 662

3. U. Senanayake, A. Sowmya and L. Dawes, "Deep fusion pipeline for mild cognitive impairment diagnosis," 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), Washington, DC, 2018, pp. 1394-1997. doi: 10.1109/ISBI.2018.8363832

4. Senanayake, U., Sowmya, A., Dawes, L., Kochan, N. A., Wen, W., Sachdev, P., "Diagnosis of Mild Cognitive Impairment with Neuropsychological Measures: Does Deep Learning Help?" (submitted)

5. Senanayake, U., Sowmya, A., Dawes, L., Kochan, N. A., Wen, W., Sachdev, P., "A Robust Deep Fusion Pipeline for Mild Cognitive Impairment Diagnosis" (under preparation)

6. Senanayake, U., Sowmya, A., Dawes, L., Kochan, N. A., Wen, W., Sachdev, P., "Deep Survival Analysis for Cognitive Impairment" (under preparation)

Acknowledgements

The culmination of a PhD thesis is the result of an arduous journey with many ups and downs and a few moments of enlightenment. It is these moments of enlightenment that decorates the thesis and elevates it to a PhD. I'm humbled and proud to have achieved such moments to include in my PhD thesis.

First and foremost, I would like to express my heartfelt gratitude to my supervisor, Professor Arcot Sowmya for her enduring guidance and continuous support throughout my candidature. I cannot emphasize how valuable her contributions of time, ideas, guidance and knowledge has been to my research.

I would also like to thank my collaborators, Dr. Laughlin Dawes, Professor Perminder Sachdev, Dr. Wei Wen and everyone from Centre for Healthy Brain Ageing, UNSW. I'm thankful to Alzheimer's Disease Neuroimaging Initiative for allowing me to use their dataset and my gratitude goes to everyone who was involved in curating the dataset.

The working environment at UNSW was productive and collaborative. I am thankful to my dear colleagues, Gihan Samarasinghe, Thamali Lekamge, Matt Gibson, Jian Kang, Manna Philips, Annette Spooner and Sankaran Iyer for their support and advises. I would also like to thank Sailesh Conjeti for the creative advises and insightful chats.

I was also fortunate to be part of Westpac STEM PhD program during my PhD candidature and I am grateful to Westpac and the wonderful colleagues I've met there. My sincere gratitude goes to Robert Mangum, Barry French, Sandra Casinader, Jessica Hallam, Karina Mak, Pierre Naeyaert, Mark Hunter, Bree Lloyd and all of my colleagues who I couldn't mention by name.

When you are away from your family, your friends become the cornerstone of your existence. I couldn't have been happier with the friends I have had in Sydney. My dear friends Harith, Harini, Anudhi, Uvin, Kasun, Chathu, Panduka, Umanga, Chiranjan, Sureka, Sarith, Kalpa, Milinda, thanks for putting up with me, you all are like siblings for me. My apologies for anyone who I couldn't mention by name for there are many, but know that I'm forever grateful to you.

Last but not least, I would like to thank my family. I owe everything I have accomplished to my parents for their relentless and unwavering love, support and confidence in me. I'm grateful to my sister who has always been there for me thick and through and for all the sacrifices she has made. I'm thankful to my loving wife for I couldn't have done this without her. I'm amazed by how understanding and supportive she is for I know it couldn't have been easy with my eccentric ways. I would also like to thank my extended family including my grandparents and my in-laws. Finally, I would like to acknowledge everyone who has supported me in any way, shape or form during my PhD and I'm eternally grateful for each and every one of you. Thank you for being patient with me and supporting me.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

According to the definition set forth by the World Health Organization (WHO), dementia is "an umbrella term for several diseases affecting memory, other cognitive abilities and behaviour that interfere significantly with the ability to maintain daily living activities. Although age is its strongest known risk factor, dementia is not a normal part of ageing" [Org19b]. Dementia can often cause long-term and gradual decrease in cognitive abilities, emotional problems, language difficulties and decreased motivation. A number of different diseases can cause dementia, including Alzheimer's disease (AD), frontotemporal dementia (FTD), Lewy body dementia (LBD), vascular dementia (VD), syphilitic dementia (SD), mixed dementia (MD), senility dementia (SD) or the combined effect of two or more dementia types and even stroke. However, the primary cause of dementia may not be identified as it is not a specific disease. As Fymat [Fym18] points out, the greatest shortcoming is the inability to pinpoint the root cause of the condition, making it difficult to treat. Currently available medications can treat some of the symptoms but not the condition itself and hence there is no known cure for dementia [Fym18].

According to Global Health Estimate 2016 report by WHO [Org19a], AD

Year	2016		2010	
Cause	Deaths	% of total deaths	Deaths	% of total deaths
Ischaemic heart disease	9433.22	16.58	7028.95	13.43
Stroke	5780.64	10.16	5169.67	9.88
Chronic obstructive pulmonary disease	3041.44	5.34	2972.06	5.68
Lower respiratory infections	2957.13	5.19	3324.51	6.35
Alzheimer disease and other dementias	1991.70	3.50	803.56	1.53
Trachea, bronchus, lung cancers	1707.74	3.00	1256.70	2.40
Diabetes mellitus	1598.52	2.81	944.24	1.80
Road injury	1402.30	2.46	1136.38	2.17
Diarrhoeal diseases	1382.70	2.43	2246.25	4.29
Tuberculosis	1292.90	2.27	1684.21	3.22
Cirrhosis of the liver	1254.07	2.20	987.84	1.88
Kidney diseases	1179.83	2.07	726.81	1.39
Preterm birth complications	1013.33	1.782	0	0
HIV/AIDS	1011.99	1.77	1469.30	2.80
Hypertensive heart disease	897.68	1.57	0	0

Table 1.1: Top 10 global causes of deaths in 2000 and 2016. The statistics used to create the figure are from Global Health Estimate 2016. Number of deaths are in 1000s.

and other dementias are listed as the fifth major cause of deaths. As Table 1.1 shows, this is a significant increase compared to year 2000. Cases of dementia have increased from 35.6 million in 2010 to 46 million in 2015 and around 50 million in 2017 while projections suggest it will increase to 82 million in 2030 and 152 million by 2050 [Fym18]. There are no known cures for dementia currently, therefore, it is important to identify individuals who are at risk of progressing to dementia before they become demented. This thesis examine computer aided diagnosis methods to identify individuals who are at risk of dementia from available data and better understand the progression to dementia.

Mild cognitive impairment is considered as a prodromal stage to dementia and it is characterized by cognitive decline that is greater than normal for the age without significantly impairing daily functions [HMH+13]. It can be considered as a criterion to predict progression to dementia. Previous studies have found that individuals with MCI progress to dementia at a rate of 6-15% per year, whereas the progression to dementia in the general older population is about 1-2% [MSF09]. In clinical terms, MCI is divided into several subtypes, with the main categories being amnestic MCI (aMCI) and non-amnestic MCI (naMCI). In aMCI, memory is impaired and is considered to have resemblance to Alzheimer's disease (AD). naMCI is characterized as the impairment of one or more non-memory cognitive domains such as frontal-executive function, language or visuospatial ability. naMCI is found to have an increased risk of progressing to non-Alzheimer's dementias. These subtypes can be further categorized with respect to the domain of impairment. aMCI is divided into single domain aMCI (sd-aMCI) and multiple domain aMCI (md-aMCI) while naMCI is divided into single domain naMCI (sd-naMCI) and multiple domain naMCI (md-naMCI). It is useful to identify the subtypes of MCI, because the subtypes have differential rates of conversion to dementia [KSB+10].

Cognitive decline in individuals is diagnosed mainly using neuropsycholog-

ical test measures (called NM features) in clinical settings. These are a battery of standardized tests administered to evaluate different cognitive functions including memory, executive functions and visuospatial ability. A comprehensive description of such tests are included in Chapter 2.

Recent focus in this area of research has been on using advanced medical imaging modalities such as magnetic resonance imaging (MRI) and positron emission tomography (PET) to differentiate between AD and subtypes of MCI. Some early work on the use of imaging modalities for this purpose [ALLF07, CLE+05, CWSS08, CWC+09,HMH+13,HSXJ11,RKB+13,RWK+14,RZW+14,SZBW13,SLC+13,TWZ+12] are described in detail in Chapter 3. Understanding the progression to cognitive impairment and dementia is also considered important. Survival analysis techniques may be used for this purpose effectively. Survival analysis is a set of techniques used to analyze available variables to determine the time until the occurrence of an event of interest, which in this case is the onset of any type of cognitive impairment. In past studies, mostly classifier based approaches have been considered for such progression tracking.

The focus of this thesis is on leveraging the research already performed on NM measures and MRI based features for MCI and dementia diagnosis to propose robust techniques for differentiating dementia and MCI from cognitively normal (CN) individuals and, predicting the progression to MCI and dementia over time by leveraging on the features already derived for the diagnosis task. The methods are based on conventional machine learning, recent developments in deep learning and deep survival analysis, all of which are reviewed in Chapter 3.

The thesis goals are stated in section 1.1 while the scope is defined in section 1.2. An overview of the thesis is presented in section 1.3 and contributions are described in section 1.4. Finally, the organization of this thesis is described in section 1.5.

1.1 Thesis Goals

The goal of this thesis is to present computer aided diagnostic techniques that are capable of differentiating between CN individuals and individuals with cognitive decline, and methods to track progression to dementia starting from CN and MCI using survival analysis techniques. Two datasets that contain multi-modal data are used in this study. The thesis proposes techniques that leverage multi-modal data including neuropsychological measure based features and neuroimaging based features for the diagnosis task. It also utilizes the fused data in survival analysis to achieve progression tracking of the disease states. Multiple feature extraction, classification, deep fusion and survival analysis techniques are proposed, implemented and validated for this purpose.

1.2 Thesis Scope

The methods in this thesis are based on computer vision, machine learning and deep learning. Two datasets are used in this work: an in-house dataset from Sydney Memory and Ageing Study (MAS) [SBR+10] and an open dataset from Alzheimer's Disease Neuroimaging Initiative (ADNI) [Ini05]. The latter is used as a benchmark dataset, and the results reported in this thesis will serve as a benchmark for MAS in future studies. A detailed description of these datasets are included in Chapter 2. This thesis does not attempt to differentiate the types of dementias and the datasets used focus AD exclusively. Therefore, AD and dementia are used interchangeably in this thesis. Algorithm accuracy and other performance measures are measured using appropriate metrics that are commonly used for evaluating similar work in the literature. The experiments are performed at the population level and personalised machine learning models are not in the scope of this thesis.

1.3 Thesis Overview

The thesis proposes multiple techniques to differentiate between dementia and its subtypes from cognitively normal individuals, and track progression over time between the disease states. In order to achieve the goal of differentiating between CN, MCI and dementia, NM based features were explored first. As far as is known, NM based features have not been used for automated diagnosis so far, even though they are the basis for diagnosis by physicians. The correlations between these features were explored and conventional machine learning algorithms designed and implemented to classify individuals into CN, dementia and MCI subtypes. This was used as a baseline to evaluate more optimized techniques, including deep learning techniques. The resulting pipeline was validated using cross-validation and shown to be generalizable.

While NM based features are a great starting point, acquiring them is challenging as it is both time consuming and expensive. Imaging modalities such as MRI can alleviate the difficulties in acquiring NM based features, therefore structural MR images were explored for the same task. Traditionally, the medical imaging pipeline consists of feature extraction and feature engineering, which may introduce bias and requires significant domain expertise. In order to address this challenge and create an end-to-end pipeline, a deep neural network architecture was designed. As structural MR images comprise multiple horizontal slices stitched together, a 2-dimensional (2D) deep network as well as a 3-dimensional (3D) deep network were designed.

When designing computer aided diagnostic systems, the challenges include data paucity and the complexity of leveraging all the available data modalities. To address these issues, a novel deep fusion pipeline that can seamlessly fuse information from multiple disparate modalities such as MR images and NM features has been proposed. It has also been demonstrated that a model trained on one dataset can effectively be used as a starting point for another dataset (called transfer learning)

to address the issue of data paucity. The deep fusion architecture is also resilient to missing information, which is an added advantage. It can also be readily extended to fuse information from other imaging modalities.

Finally, a novel survival analysis based deep learning technique has been proposed to better understand the progression from CN and MCI to dementia from NM and MR data. As far as is known, survival analysis has not been applied to dementia progression analysis nor have medical images been used for the purpose. The survival analysis method proposed in this thesis is capable of analyzing a patient's survival using NM features as well as MR image based features without requiring significant domain knowledge or an additional feature extraction and engineering step.

1.4 Thesis Contributions

This thesis makes contributions to the following fields: computer vision, machine learning, deep learning, medical image analysis and cognitive impairment diagnosis. Conventional machine learning based methods and novel deep learning techniques for classification of dementia and its subtypes have been designed, implemented and demonstrated on two datasets. Transfer learning has been explored to show that models trained on one dataset can be used as a starting point to train models on another dataset. Survival analysis based techniques are proposed to better understand the progression from cognitively normal and MCI to dementia.

1.5 Thesis Organization

The remaining chapters of this thesis are organized as follows. A background on brain anatomy, cognitive impairment and available datasets is included in Chapter 2. Related work on classification of cognitive impairment is provided in Chapter 3.

The focus of Chapter 4 is the diagnosis of mild cognitive impairment subtypes and dementia using NM based features. The features and the conventional machine learning algorithms used are presented and the results are compared against those of the deep learning architectures. A discussion of the utility of deep learning architectures ensues.

In Chapter 5, MR image based deep learning networks are proposed, followed by a novel deep fusion pipeline for NM and MR features. A comprehensive set of experiments is described that are used to demonstrate the efficacy of the deep fusion pipeline.

Survival analysis based techniques are proposed to better understand the progression of dementia in Chapter 6. The proposed approach is novel in the field of cognitive impairment diagnosis as it uses MR images in addition to NM features.

Finally, Chapter 7 summarizes the thesis and presents the contributions made to computer vision, machine learning, deep learning, medical image analysis and cognitive impairment diagnosis. The limitations of the proposed methods and potential future works are described in Chapter 7 followed by concluding remarks.

Chapter 2

Background

In differentiating between cognitively normal individuals from cognitively impaired individuals and predicting their risk of progression to a cognitively impaired state, neuropsychological measures are most popular with advanced medical imaging techniques such as magnetic resonance imaging (MRI) providing supplementary information.

Mild cognitive impairment (MCI) is considered as a prodromal stage to dementia, with the former characterized by cognitive decline that is greater than normal for age without significantly impairing daily functions. Previous studies have found that individuals with MCI progress to dementia at a rate of 6-15% per year, whereas the progression to dementia in the general older population is about 1-2%. From a clinical perspective, MCI is divided into several subtypes. The main categories can be categorized as amnestic MCI and non-amnestic MCI (aMCI and naMCI respectively). In aMCI, memory is impaired and is considered to have resemblance to Alzheimer's disease (AD). naMCI is characterized by impairment of one or more non-memory cognitive domains such as frontal-executive function, language or visuospatial ability. naMCI is found to have an increased risk of progressing to

non-Alzheimer's dementias. These subtypes can be further categorized with respect to the domain of impairment. Consequently, aMCI is divided into single domain aMCI and multiple domain aMCI (sd-aMCI, md-aMCI) while naMCI is divided into single domain naMCI and multiple domain naMCI (sd-naMCI, md-naMCI).

Since MCI is considered a prodromal stage of dementia, it may be considered as a criterion to predict progression to dementia. However, many patients with MCI will not necessarily progress to dementia. Further analysis of imaging data may help refine and improve such predictive efforts, and facilitate the easy and early identification of those who are at risk of dementia. Also, a combination of image and non-image data may assist in better optimizing and improving the performance of existing classification and progression analysis studies. Therefore, in this thesis, novel methods based on computer vision and machine learning techniques are developed in order to better analyze and understand brain MR images and neuropsychological measures for the purpose of MCI and dementia diagnosis. Although the focus is on structural MRI images of the brain, the methods are equally applicable to other types of imaging, and in fact other human soft tissue organs for diagnosis of other diseases. Existing algorithms are carefully examined and improved, and new algorithms are developed. The algorithms are tested and validated on the MR scans acquired as a part of Sydney Memory and Aging Study (MAS) and Alzheimer's Disease Neuroimaging Initiative (ADNI).

Another motivation for using these datasets is that relatively little work has been carried out using a combination of image and non-image features. Significantly, in routine medical practice, diagnosis is mostly based on neuropsychological measures, leading to tagging of labels arrived at by observing non-image features to the image features, which is an additional theoretical constraint.

The remainder of this chapter is organized as follows. The anatomy of the brain relevant to cognitive impairment diagnosis is briefly described in section 2.1.

The types of cognitive impairment are discussed in section 2.2 followed by a description of available datasets in section 2.3. Neuropsychological tests used to evaluate cognitive impairment are included in section 2.4 while relevant medical imaging techniques are covered in section 2.5. Labelling of the datasets are described in section 2.6 while section 2.7 summarizes this chapter.

2.1 Brain Anatomy

The human brain consists of a large mass of nerve tissues that is protected by a skull. The basic anatomy of the brain is depicted in Figure 2.1. The largest part of the brain is the cerebrum, which is divided into two hemispheres separated by a groove called the longitudinal fissure. Each hemisphere is broken into broad regions known as lobes that are associated with different functions. The frontal lobes are the largest and are located in the front part of the brain. They are responsible for coordinating high-level behaviours such as motor skills, problem solving, judgement, planning, attention, impulse control and management of emotions. The parietal lobes are responsible for organizing and interpreting sensory information from other parts of the brain. They are located behind the frontal lobes. The temporal lobes that are located on either side of the head at ear level are responsible for coordinating specific functions including visual memory, verbal memory and the interpretation of emotions and reactions of others. Finally, the occipital lobes found at the back of the brain are involved in the ability to read and recognize printed words and other aspects of vision [SSH19]. Other parts of the brain are not of direct interest in this work, and are not discussed. A more in-depth description of brain anatomy can be found [CMMK05].

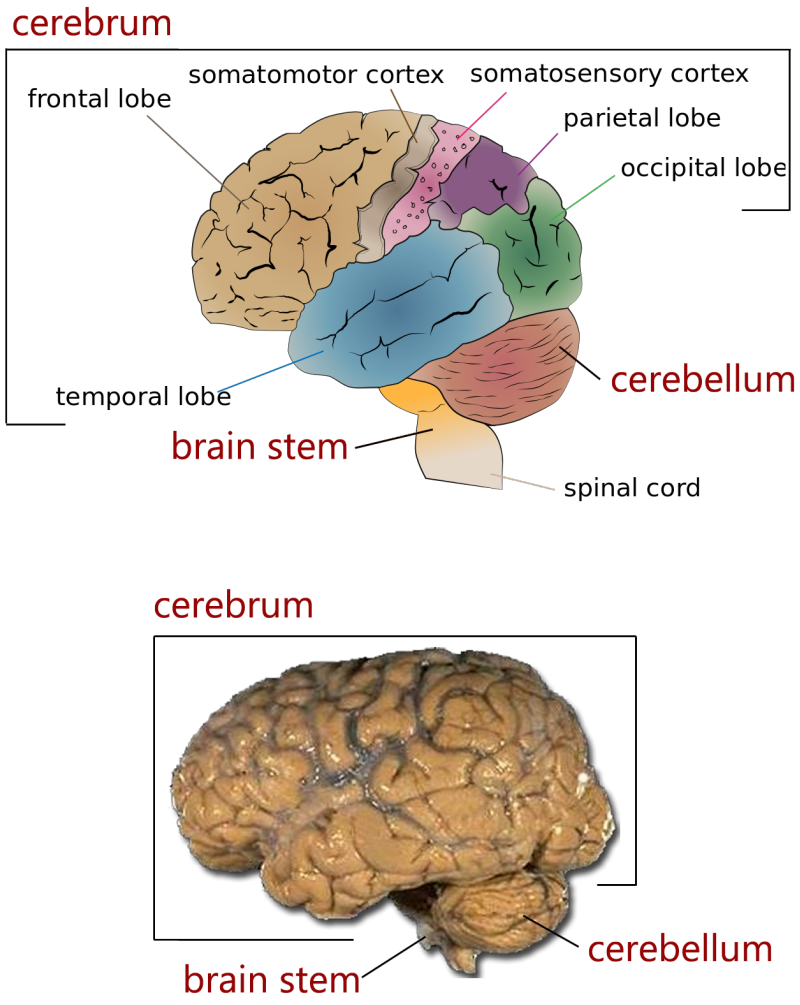


Figure 2.1: Basic Anatomy of the Brain [adapted from [Lea19]]

2.2 Cognitive Impairment

Cognitive impairment is suspected when a person has trouble remembering, leaning new things, concentrating or making decisions that affect their everyday life [fDCP19]. There are different kinds of cognitive impairments depending on the stage that the patient is at and the functions that are affected. The common types of cognitive impairments and the datasets used are discussed below.

2.2.1 Cognitively Normal Individuals

Cognitively normal (CN) individuals exhibit no apparent impairment to their cognitive abilities. If a patient is not diagnosed with either Mild Cognitive Impairment or dementia and no subjective complaints were made by the participant or their escort, they are diagnosed as CN individuals.

2.2.2 Mild Cognitive Impairment

Mild cognitive impairment (MCI) is considered to be an intermediate stage between normal aging and dementia. It is defined as cognitive decline greater than expected for an individual's age and education level but that does not interfere notably with activities of daily life [GRZ+06]. Epidemiological studies suggest that this is prevalent in 3-19% of adults older than 65 years [GRZ+06]. More than 50% of that population progresses to dementia within 5 years but there have been reports on stable MCI individuals as well as individuals who have reverted to a cognitively normal state. This makes MCI an interesting phenomenon to study, as there are three likely stages of progression to it.

Since more than 50% individuals with MCI progress to dementia, MCI is

considered as a prodromal stage to dementia. Recent advances in dementia treatment suggest that treatments are far more effective when performed before developing dementia. Therefore, it is worthwhile to understand the diagnosis, progression and tracking of MCI in order to reliably predict progression to dementia.

2.2.3 Dementia

Dementia is characterized by deterioration of mental function in its cognitive, emotional aspects [Gus96]. A declining memory that is evident in learning, retention and recall of new information and the remote past is considered obligatory for diagnosis of dementia. In addition, at least one of the following symptoms are usually required to diagnose dementia: spatial disorientation (increased difficulty in finding one's way); language disturbances (increased difficulties in understanding or in expressing oneself verbally or in writing); reduced practical abilities (increased difficulty in maintaining learned skills or managing everyday activities and personality changes resulting in lack of judgment, increased sentimentality, emotional bluntness, aggressiveness or lack of insight) [Gus96]. Alzheimer's disease (AD) is the most common form of dementia accounting for up to 70% of dementia cases in United States [PLF+07].

2.3 Available Datasets

The solutions proposed by this thesis are trained and validated on two datasets which are described briefly in this section, while their characteristics are elaborated in Chapters 4, 5 and 6 at time of use.

2.3.1 Sydney Memory and Ageing Study Dataset

The Sydney Memory and Ageing Study (MAS) dataset is based on 1037 community-dwelling, non-demented individuals, who were recruited randomly from two electorates of East Sydney, Australia [SBR+10]. The baseline age of these individuals was 70-90 and each participant was administered a comprehensive neuropsychological test battery. Only 52% of the study population underwent an MRI scan. Individuals were excluded if they had a Mini-Mental State Examination (MMSE) score $<$ 24 (adjusted for age, years of education and non-English-speaking background), a diagnosis of dementia, mental retardation, psychotic disorder (including schizophrenia and bipolar disorder), multiple sclerosis, motor neuron disease and progressive malignancy or inadequate English to complete assessments. Seven repetitive waves after the baseline assessment have been carried out to date at a frequency of 2 years, and the first four waves were made available for this work. Details of the sampling methodology have been published previously [SBR+10]. This study was approved by the Human Research Ethics Committees of the University of New South Wales and the South Eastern Sydney and Illawarra Area Health Service, and all participants gave written informed consent.

2.3.2 Alzheimer's Disease Neuroimaging Initiative Dataset

The second dataset was obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database [Ini05]. ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging(MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). Up-to-date information on

ADNI is available [Ini05]. A complete listing of ADNI investigators can be found at:[1] *.

2.4 Cognitive Assessments

A subset of prominent cognitive assessment measures that are used in both ADNI and MAS datasets are described herein. A number of cognitive assessment scales and protocols were selected by the designers of the MAS and ADNI studies to represent the patient population and adequately sample cognitive domains of interest in subjects who are normal, or have MCI or dementia. These studies can also be used to measure change over the duration of a longitudinal study and are reasonably efficient and meet the practical demands of the MAS and ADNI studies [Ini05]. The following subsections briefly discuss the different measures of cognitive assessment employed by the two studies.

2.4.1 Mini-mental State Exam (MMSE)

MMSE is frequently used in AD drug studies and is a fully structured screening instrument. Orientation to place, orientation to time, registration (immediate repetition of three words), attention and concentration (e.g. serially subtracting seven beginning with 100), recall (recalling the previously repeated three words), language (naming, repetition, reading, writing, comprehension) and visual construction (copy two intersecting pentagons) are evaluated by this test. It is scored as the number of correctly completed items with lower scores indicative of poorer performance and greater cognitive impairment. The total score ranges from 0 to 30 [Ini05, FFM75].

*1. http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

2.4.2 Alzheimer's Disease Assessment Scale-Cognitive (ADAS-COG)

ADAS-COG evaluates memory (word recall, word recognition), reasoning (following commands), language (naming, comprehension), orientation, ideational praxis (placing letter in envelope) and constructional praxis (copying geometric designs). It also obtains ratings of spoken language, language comprehension, ability to remember test instructions and word finding difficulty. As the results are represented in errors, higher scores reflect poorer performance where scores can range from 0 (best) to 70 (worse) [Ini05, ada84].

2.4.3 Logical Memory Test (Delayed Paragraph Recall)

This test used in the ADNI study is a modification of the episodic memory measure from Wechsler Memory Scale-Revised (WMS-R). Free recall of one short story (story A) that consists of 25 bits of information is elicited immediately after it is read aloud to the subject, and again after a thirty minute delay in this modified version. The total bits of information from the story that are recalled immediately (maximum score = 25) and after the delay interval (maximum score = 25) are recorded. A retention or "savings" score can be computed by dividing the score achieved during delayed recall by the score achieved during immediate recall [Ini05, Elw91].

2.4.4 Boston Naming Test

The Boston naming test measures visual confrontation and requires the subject to name objects depicted in outline drawings. In the ADNI study, a modified version of the original test was used where only 30 items are presented (either the odd or even numbered items from the full 60-item test). The drawings get increasingly difficult

and the easiest ones are presented first. If a naming difficulty is encountered, a stimulus cue and/or a phonemic cue is provided. The number of spontaneous correct responses (maximum score = 30) and spontaneous plus semantically-cued correct responses (maximum score = 30) are recorded. The number of perceptual errors, circumlocutions, paraphasic errors, and perseverations can also be used to evaluate the subjects' language performance [Ini05, KGW83].

2.4.5 Category Fluency Test

This test measures verbal fluency in which the subject is asked to generate examples from each of two semantic categories (animals and vegetables) in successive one-minute trials. The number of correct, unique examples generated for the two categories is considered the primary performance measure. Repetitions of a correct item (perseveration) and non-category items (intrusion) errors are also taken into consideration [Ini05, BGS+87].

2.4.6 Clock Drawing Test

This test is a visuo perceptual constructional task where the subject is given a blank sheet of 8 1/2" x 11" paper and instructed to "Draw a clock, put in all the numbers and set the hands for 10 after 11". Thereafter, a copy condition ensues where the subject attempts to copy a drawing of a clock with the hands set at ten past eleven. A maximum total score of 10 is derived for each drawing by adding the scores of three separate features: a maximum of 2 points is given for the integrity of the clock face; a maximum of 4 points for the presence and sequencing of the numbers; a maximum of 4 points for the presence and placement of hands [Ini05, Nas84]. This test has been proven effective in discriminating between patients with AD and cognitively normal adults [CSM+96].

2.4.7 Digit Span Test

The subject is expected to repeat sequences of single digit numbers which are read aloud by the examiner. The subject must repeat the sequence in the same order and then digits should be repeated in reverse order. The lengths of the sequences increase progressively from three to nine digits in the forward direction and from two to eight digits in the backward direction. Two trials are presented for each sequence length. When the subject misses both trials at a given sequence length, testing is terminated. One point is awarded for each sequence correctly produced and the maximum score for each condition is 14 points.

2.4.8 American National Adult Reading Test (ANART)

Estimating premorbid verbal intelligence is the objective of this test where this skill is thought to remain relatively preserved until the later stages of Alzheimer's disease. It requires patients to read and correctly pronounce fifty "irregular" words that do not follow common rules of phonography and orthography. Pronouncing such words depends solely on previous familiarity and cannot be accomplished by applying common grammatical rules (eg: the word 'naive' might be pronounced 'nave' if common English grammatical rules were employed). A large premorbid vacabulary is correlated with a high premorbid verbal intelligence and this is characterized by the ability to correctly pronounce progressively less common irregular words [Ini05, NO78]. Premorbid verbal intelligence can be estimated from the formula derived by Grober and Sliwinski [GSRK08].

2.4.9 Rey Auditory Verbal Learning Test

Multiple cognitive parameters associated with learning and memory are assessed by this learning task. There are five learning trials where 15 unrelated words (all nouns) are presented orally at the rate of one word per second and immediate free recall of the words is elicited. The number of correctly recalled words are recorded. After a 20-minute delay filled with unrelated testing, free recall of the original 15 word list is elicited. Finally, a yes/no recognition test is administered which consists of the original 15 words and 15 randomly interspersed distracter words [Ini05, Rey64].

2.4.10 Trail Making Test: Parts A and B

There are two parts to this test: part A consists of 25 circles numbered 1 through 25 distributed over a white sheet of 8 1/2" x 11" paper while part B also consists of 25 circles but these circles are either numbered (1 through 13) or contain letters (A through L). In part A, the subjects are instructed to connect the circles with a drawn line as quickly as possible in ascending numerical order. In part B, subject should connect the circles while alternating between numbers and letters in an ascending order (eg: A to 1; 1 to B; B to 2; 2 to C). The time required for a subject to complete each trial and the number of errors of commission and omission are used as the assessing criteria [Ini05, MR58].

2.4.11 Digit Symbol Substitution Test

This test has 110 small blank squares presented in seven rows, each randomly paired with one of nine numbers (1 to 9) printed directly above it. There is a "key" printed above the row of blank squares that pairs each of the numbers 1 through 9 with an unfamiliar symbol. A short series of practice trials are given and then the subjects

must use the key to fill in the blank squares in order (working left to right across the rows) with the symbol that is pared with the number above it, working as quickly as possible for 90 seconds. The correct number of blank squares filled within the time limit is the measure of interest where the maximum raw score is 110 [Ini05, Wec81].

2.4.12 Clinical Dementia Rating (CDR)

Five degrees of impairment in performance on each of 6 categories of cognitive functioning including memory, orientation, judgment and problem solving, community affairs, home and hobbies and personal care are described by this test. The degree of impairment obtained on each of the six categories of function are synthesized into one global rating of dementia (ranging from 0 to 3), with a more refined measure of change available by use of the sum of boxes. It is used as a global measure of severity of dementia, as its reliability and validity has been established as well as the high inter-rater reliability [Ini05, Ber84].

2.4.13 Functional Activities Questionnaire (FAQ)

This test is based on an interview with a caregiver or a qualified partner. A subject is rated on their ability to carry out ten complex activities of daily living: 1) manage finances, 2) complete forms, 3) shop, 4) perform games of skill or hobbies, 5) prepare hot beverages, 6) prepare a balanced meal, 7) follow current events, 8) attend to television programs, books or magazines, 9) remember appointments, 10) travel out of the neighborhood. The score has three scales, 0 (does without difficulty), 1 (needs frequent advice or assistance) and 2 (someone has taken over the activity). Scores are then summed across items to provide a total disability score with a maximum score of 20, with a higher score indicating greater impairment [Ini05, PKH$^+$82].

2.4.14 Neuropsychiatric Inventory Q (NPIQ)

This test is also performed as an interview with a caregiver or a qualified partner. It is a well-validated, reliable,multi-term instrument to assess psychopathology in AD. It is a brief interview of 15 minutes and carries a maximum score of 36 [Ini05, KCK+00].

2.4.15 Geriatric Depression Scale

The objective of this test is to identify symptoms of depression in the elderly in a self report. It consists of 15 printed questions that the subject is asked to answer by circling yes or no on the basis of how they felt over the past week. More benign items are presented first. Five items are negatively oriented for depression (ex: do you feel full of energy) while ten are oriented positively (ex: do you often feel helpless). Each appropriate positive or negative answer indicative of a symptom of depression is given one point. Scores ranging from 0-5 are considered normal while 6-15 are considered depressed [Ini05, SY86].

2.5 Medical Imaging

There are a number of advanced medical imaging techniques such as X-ray radiography, X-ray computer tomography (CT), magnetic resonance imaging (MRI), ultrasonography, elastography, optical imaging, positron emission tomography (PET) and thermography [KEbS15]. However, the focus here is on MRI, as it is the primary imaging data source used in the experiments. Other imaging modalities such as PET and diffusion tensor imaging (DTI) are also described briefly, as they are used in cognitive impairment diagnosis.

MRI is frequently divided into structural MRI and functional MRI (fMRI).

Nuclear magnetic resonance (NMR) was first observed in 1945 and was subsequently developed over the next three decades to obtain the first human in-vivo MR images [SJSY04]. MR images of the head provided excellent anatomical detail and strong grey / white matter contrast compared to imaging modalities extant at the time. The core protocols of clinical MRI include T2- and T1-weighted sequences, which are described below. Therefore, most pathological processes are described in terms of T1 or T2 signal behavior. In recent times, fluid attenuated inversion recovery (FLAIR) has been introduced as a complement of the conventional T2-weighted sequence. Newer techniques have been introduced over the years to increase the spatial resolution and sensitivity in T2 imaging.

In MRI acquisition, no ionizing radiation is used. A magnetic field is used that causes protons in the body to align and then pulsed radio waves are directed at the patient, causing a disturbance of the proton alignment. Atoms then realign by emitting absorbed radio-frequency. The time it takes the protons to regain their equilibrium state is known as the relaxation time. There are two types of relaxation times, T1 (longitudinal: parallel to the magnetic field) and T2 (transverse: perpendicular to the magnetic field). Relaxation time and proton density can be considered as the main determinants of signal strength. The main determinants of contrast (also known as weighting) are the repetition time (TR), which is the time between two successive radio frequency (RF) pulses, and the echo time (TE), which is the time between the arrival of the RF pulse that excites and the arrival of the return signal at the detector.

There are two main governing characteristics of image quality; spatial resolution and signal to noise ratio (SNR). Resolution is mostly determined by the number of picture elements (pixels) in the frequency, and phase encoding directions and the through-plane resolution by the slice thickness. Pixel size, slice thickness, scan time and the sequence used govern the SNR. Any motion while capturing MRI

Figure 2.2: A sample T1w brain MR image [Ins19]

can cause motion artifacts [SJSY04].

T1 weighted MRI (T1w) is one of the basic MR image techniques that demonstrates the T1 relaxation times of tissues. In T1w, the image relies on the longitudinal relaxation of a tissue's net magnetization vector such that spins aligned to an external field are put into the transverse plane by an RF pulse. These can then slide back toward the original equilibrium of the external field. As all tissues do not get back to equilibrium equally quickly, a tissue's T1 relaxation time reflects the amount of time that its protons' spins take to realign with the main magnetic field [MJ19]. Essentially, T1 weighted MRI is acquired when TR and TE are short. A sample T1w MR image of the brain is shown in figure 2.2.

The difference between T1w and T2w is that T2w MR images rely on the transverse magnetization as opposed to the longitudinal magnetization used in T1w. The method of acquisition is similar, however with long TR and long TE. A sample

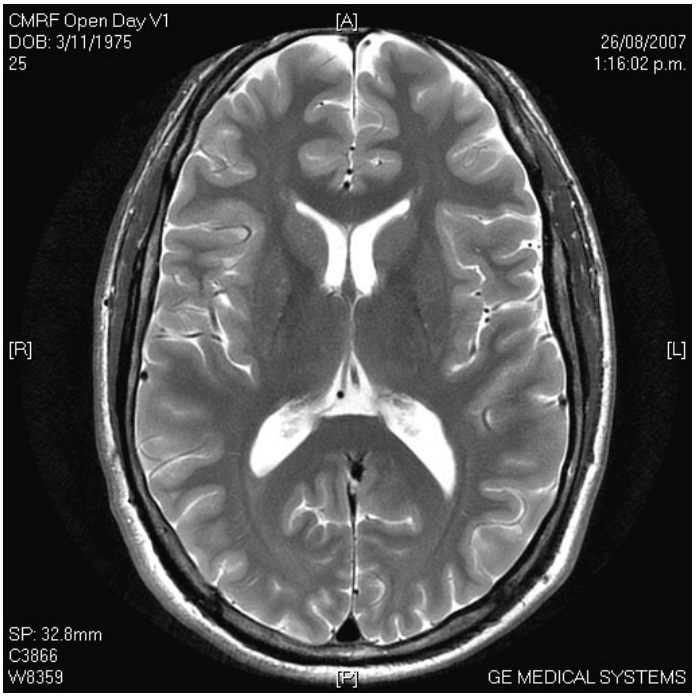


Figure 2.3: A sample T2w brain MR image [Ins19]

T2w MR image of the brain is shown in figure 2.3.

Diffusion tensor imaging (DTI) is a modality for characterizing micro-structural changes or differences. DTI can be used to map and characterize the three-dimensional diffusion of water as a function of spatial location. The diffusion tensor elaborates the magnitude, degree of anisotropy and orientation of diffusion anisotropy. The most commonly used diffusion-weighted imaging (DWI) approach is the pulsed-gradient spin echo (PGSE) pulse sequence with a single-shot, echo planar imaging (EPI) readout. At least six non-collinear diffusion encoding directions are needed to measure the full diffusion tensor [ALLF07]. Owing to the way image acquisitions are performed, DTI images are considered as a temporal sequence and hence cannot be represented by a single image. However values derived from DTI images such as fractional anisotropy (FA) or mean diffusivity (MD) can be displayed, as shown in Figure 2.4.

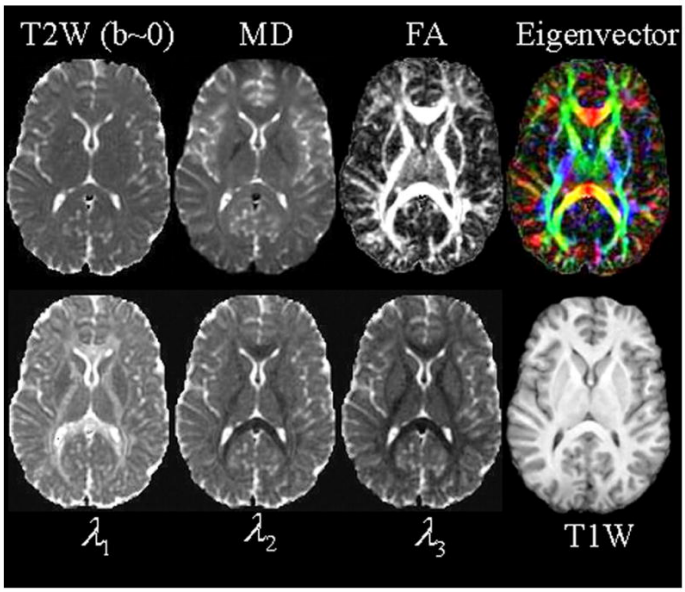


Figure 2.4: A sample MD and FA image derived from DTI sequence. T1w and T2w images are also displayed [ALLF07].

PET imaging is based on detecting two time-coincident high-energy photos from the emission of a positron-emitting radioisotope [VK15]. It is a non-invasive imaging modality that provides physiological information through the injection of radioactive compounds, detection of radiation and reconstruction of the distribution of the radiotracer [VK15]. It is a prominent method used in diagnosis and staging in oncology as well as for neurological and cardiovascular indications [VK15].

2.6 Labelling of Datasets

The characteristics used to diagnose and label each patient in the MAS and ADNI datasets are described here. The labelling procedure for MCI is discussed first followed by those for AD and Dementia.

2.6.1 Labelling of MCI

The participants of MAS dataset were labelled with MCI if they exhibited the following criteria as outlined in international consensus criteria [Pet04].

i A subjective complaint of decline in memory or other cognitive function (from the participant and/or their informant).

ii Cognitive impairment as shown by performance at 1.5 standard deviations (or equivalent) below published normative values (matched for age and education where available) on a neuropsychological test measure.

iii Normal or minimally impaired functional activities as determined by informant ratings on the Bayer-ADL scale [Hin98]. This is a questionnaire completed by the participant's informant that rates the participant's level of difficulty in instrumental activities of daily living.

iv Not demented i.e.no diagnosis of DSM-IV dementia [APATFoNaAPACoNa00], as determined by a consensus diagnosis from an expert team comprising psychogeriatrician, neuropsychiatrists and neuropsychologists.

The inclusion criteria for MCI participants in the ADNI study are outlined in Table 2.1.

Table 2.1: Inclusion criteria for MCI in ADNI study

Item	Description
Memory complaints	Memory complaint by subject or study partner that is verified by a study partner.

Memory Function	Abnormal memory function documented by scoring below the education adjusted cutoff on the Logical Memory II subscale (Delayed Paragraph Recall) from the Wechsler Memory Scale – Revised (the maximum score is 25): a) less than or equal to 8 for 16 or more years of education b) less than or equal to 4 for 8-15 years of education c) less than or equal to 2 for 0-7 years of education.
MMSE	Mini-Mental State Exam score between 24 and 30 (inclusive) (Exceptions may be made for subjects with less than 8 years of education at the discretion of the project director).
CDR	Clinical Dementia Rating = 0.5. Memory Box score must be at least 0.5.
General Cognition	General cognition and functional performance sufficiently preserved such that a diagnosis of Alzheimer's disease cannot be made by the site physician at the time of the screening visit.
Hachinski	Modified Hachinski score of less than or equal to 4.
Age	Age between 55 and 90 (inclusive).

Stability of Permitted medications	Permitted medications stable for at least 4 weeks prior to screening. In particular: a) Subjects may take stable doses of antidepressants lacking significant anticholinergic side effects (if they are not currently depressed and do not have a history of major depression within the past 1 year) b) Estrogen replacement therapy is permissible c) Gingko biloba is permissible, but discouraged d) Washout from psychoactive medication (e.g., excluded antidepressants, neuroleptics, chronic anxiolytics or sedative hypnotics, etc.) for at least 4 weeks prior to screening. e) Cholinesterase inhibitors and memantine are allowable if stable for 4 weeks prior to screen
Geriatric Depression Scale	Geriatric Depression Scale score of < 6
Study partner	Study partner is available who has frequent contact with the subject (e.g. an average of 10 hours per week or more), and can accompany the subject to all clinic visits for the duration of the protocol.
Visual and auditory acuity	Adequate visual and auditory acuity to allow neuropsychological testing.
General Health	Good general health with no additional diseases expected to interfere with the study.

Pregnancy/Childbearing Potential	Subject is not pregnant, lactating, or of childbearing potential (i.e. women must be two years post-menopausal or surgically sterile).
Testability	Willing and able to complete all baseline assessments. Willing and able to participate in a 3-year protocol.
Commitment to neuroimaging and providing study samples	Willing to undergo MRI 1.5 Tesla neuroimaging (PET and MRI 3Tesla are optional) and provide DNA for ApoE assessments and banking as well as plasma samples at protocol specified time points.
Commitment to provide CSF samples	Willing to provide CSF for biomarker studies at protocol specified time points (optional).
Education	Completed 6 grades of education (or had a good work history sufficient to exclude mental retardation).
Language	Fluent in English or Spanish.

In the MAS study, MCI individuals are labelled by four subtypes in turn depending on the affected region and function [Pet04] according to their neuropsychological test performances. The four subtypes are amnestic single domain MCI(only memory domain impaired)[sd-aMCI], amnestic multiple domain MCI(memory plus at least one non-memory domain impaired)[md-aMCI], non-amnestic single domain MCI(one non-memory domain impaired)[sd-naMCI] and non-amnestic multiple domain MCI(more than one non-memory domain impaired)[md-naMCI]. The subclas-

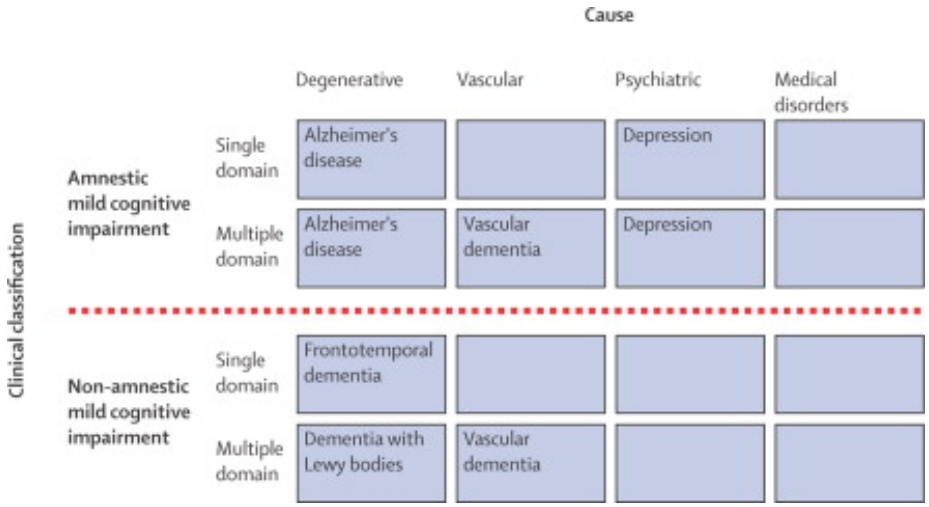


Figure 2.5: The sub-classification of MCI and their causes [Pet04].

sification is better represented in figure 2.5.

In the ADNI study, MCI is more generally subdivided into two categories, early MCI (EMCI) and late MCI. Therefore, ADNI does not have the same level of resolution in the subtypes of MCI as MAS does.

A variety of neuropsychological tests were used to evaluate and diagnose individuals in the MAS study and these tests are listed in Table 2.2.

A number of these overlap the tests used in ADNI. However, for completeness, the tests used by ADNI are also listed in Table 2.3. It should be noted that some tests are regional variants of the same tests as MAS, as ADNI was conducted in the Americas while MAS was conducted in Australia.

2.6.2 Labelling of Dementia

For the MAS study, diagnosis of dementia was based on DSM-IV criteria [APATFoNaAPACoNa00] that takes into consideration factors such as the presence of multiple cognitive deficits that represent a decline from a previous level of functioning and in-

Table 2.2: Neuropsychological Test Battery and Normative Data used for Diagnostic Classification for MAS [SBR+10]

Cognitive Domain	Test	Normative Data & Demographic adjustments
Premorbid Intelligence	National Adult Reading Test (NART)*	
Attention/ Processing speed	Digit Symbol-Coding [Axe01] Trail Making Test(TMT) A [SSS06]	Age [Axe01] Age & Education [Tom04]
Memory	Logical Memory Story A delayed recall [Axe01] Rey Auditory Verbal Learning Test (RAVLT) [SSS06] RAVLT total learning: sum of trials 1-5 RAVLT short-term delayed recall:trial 6 RAVLT long-term delayed recall: trial 7 Benton Visual Retention Test Recognition [Ste05]	Education [MRSea04] Age [IMS+92a] Age & Education
Language	Boston Naming Test - 30 items [Rot11] Semantic Fluency (Animals) [SSS06]	Age [FDM98] Age & Education [TKR99]
Visuo-spatial	Block Design [Axe01]	Age [IMS+92b]
Executive Function	Controlled Oral Word Association Test [SSS06] Trail Making Test(TMT) B [SSS06]	Age & Education [TKR99] Age & Education [Tom04]

Table 2.3: Neuropsychological Tests conducted for ADNI [Ini05]

Tests	Cognitive domain
American National Adult Reading Test	Premorbid Intelligence
Mini Mental State Examination	Multiple
Logical Memory I and II	Memory
Digit Span	Attention / Processing speed
Category Fluency	Language
Trails A & B	Executive Function
Digit symbol	Attention / Processing speed
Boston Naming Test	Language
Auditory Verbal Learning Test	Memory
Geriatric Depression Scale	Multiple
Clock drawing	Visua-spatial
Neuropsychiatric Inventory Q	Executive function
ADAS-Cog	Multiple
Clinical Dementia Rating Scale	Multiple
Activities of Daily Living(FAQ)	Multiple

clude memory impairment and at least one other cognitive disturbance as described above. Participants who were diagnosed with dementia at baseline were excluded from the study.

The inclusion criteria for dementia diagnosis in ADNI are outlined in Table 2.4. Compared to MAS, patients diagnosed with dementia at baseline were not excluded from ADNI.

Table 2.4: Inclusion criteria for Dementia in ADNI study [Ini05]

Item	Description
Memory complaints	Memory complaint by subject or study partner that is verified by a study partner.

Memory Function	Abnormal memory function documented by scoring below the education adjusted cutoff on the Logical Memory II subscale (Delayed Paragraph Recall) from the Wechsler Memory Scale – Revised (the maximum score is 25): a) less than or equal to 8 for 16 or more years of education b) less than or equal to 4 for 8-15 years of education c) less than or equal to 2 for 0-7 years of education
MMSE	MMSE between 20 and 26 (inclusive) (Exceptions may be made for subjects with less than 8 years of education at the discretion of the project director).
CDR	Clinical Dementia Rating = 0.5, 1.0
General Cognition	NINCDS/ADRDA criteria for probable AD.
Hachinski	Modified Hachinski score of less than or equal to 4.
Age	Age between 55 and 90 (inclusive).

Stability of Permitted medications	Permitted medications stable for at least 4 weeks prior to screening. In particular: a) Subjects may take stable doses of antidepressants lacking significant anticholinergic side effects (if they are not currently depressed and do not have a history of major depression within the past 1 year) b) Estrogen replacement therapy is permissible c) Gingko biloba is permissible, but discouraged d) Washout from psychoactive medication (e.g., excluded antidepressants, neuroleptics, chronic anxiolytics or sedative hypnotics, etc.) for at least 4 weeks prior to screening. e) Cholinesterase inhibitors and memantine are allowable if stable for 4 weeks prior to screen.
Geriatric Depression Scale	Geriatric Depression Scale score of ¡ 6
Study partner	Study partner is available who has frequent contact with the subject (e.g. an average of 10 hours per week or more), and can accompany the subject to all clinic visits for the duration of the protocol.
Visual and auditory acuity	Adequate visual and auditory acuity to allow neuropsychological testing.
General Health	Good general health with no additional diseases expected to interfere with the study.

Pregnancy/Childbearing Potential	Subject is not pregnant, lactating, or of childbearing potential (i.e. women must be two years post-menopausal or surgically sterile).
Testability	Willing and able to complete all baseline assessments. Willing and able to participate in a 2- year protocol.
Commitment to neuroimaging and providing study samples	Willing to undergo MRI 1.5 Tesla neuroimaging (PET and MRI 3Tesla are optional) and provide DNA for ApoE assessments and banking as well as plasma samples at protocol specified time points.
Commitment to provide CSF samples	Willing to provide CSF for biomarker studies at protocol specified time points (optional).
Education	Completed 6 grades of education (or had a good work history sufficient to exclude mental retardation).
Language	Fluent in English or Spanish.

2.7 Summary

This chapter described cognitive impairment in detail, including relevant brain anatomy and the methods currently utilized to diagnose cognitive impairment. Two data modalities from two different studies are used in this thesis, and both modalities and datasets were described in detail including the labelling protocols employed.

Chapter 3

Machine Learning and Deep Learning for Diagnosis of Cognitive Impairment and Survival Analysis

This thesis utilizes machine learning techniques as the method of choice to assign a class label of CN, MCI or dementia to patient data. Classical machine learning algorithms reason from externally supplied labelled instances to produce a general hypothesis, which is then used to make predictions about the labels of future instances, called supervised learning. Formally, the goal of supervised learning is to build a concise model of the distribution of class labels in terms of predictor features. The resulting classifier is then used to assign class labels to test instances, where the values of the predictor features are known but the class label is unknown [Kot07]. Generally, the same features that are used to arrive at class labels are used to make the predictions as well. If labels are unavailable, algorithms exist that can still produce a model for a similar purpose and is called unsupervised learning.

Recently, machine learning has bifurcated into conventional machine learn-

ing and deep learning. Deep learning is a special type of machine learning that predominantly uses deep artificial neural networks that are capable of working with structured as well as unstructured data. Conventional machine learning on the other hand almost always requires structured data and is further described in section 3.1. Deep learning techniques are discussed in section 3.2. Diagnosis of cognitive impairment is discussed in detail and prior work reviewed in section 3.3, while survival analysis for cognitive impairment and dementia is introduced in section 3.4. Finally, the motivation for this thesis is described in section 3.5 and the chapter is summarised in the final section.

3.1 Conventional Machine Learning

Machine learning may be defined as a set of techniques enabling machines to learn from data without explicit programming or handwritten rules [Cha19]. In the context of this thesis, machine learning techniques other than deep learning are considered to be conventional machine learning techniques. These require structured data and are often shallow compared to deep learning techniques. A number of conventional machine learning techniques such as decision trees, random forest (RF), support vector machine (SVM), AdaBoost (AB) and ensemble methods (ES) are discussed. These have been selected as they perform effectively with structured data such as the neuropsychological measure (NM) features described in Chapter 2. Feature subset selection algorithms are introduced briefly as well. A set of supporting terms are defined before discussing specific techniques.

i Class is a set of similar samples grouped together. For example, patients with dementia may be considered to be a class.

ii Class label is the label that indicates the class that an instance belongs to. A

patient with dementia will have a class label of dementia.

iii Classifier is a model trained on instances, labelled or unlabelled, that can then be used to classify a new instance and provide a label.

iv Classification is the process of estimating the class label of an unlabeled instance using a trained classifier.

v Supervised learning occurs when a training set consists of pairs of input features and a class label, and the objective is to learn a mapping function between the two [Sim18].

vi Unsupervised learning occurs when the training set consists of unlabelled input features, and the objective is to explore and discover properties of the mechanism generating the data [Sim18].

vii Training and test sets are used to build and validate the performance of a classifier. The dataset is usually divided into a larger training set and a smaller non-overlapping testing set. The training set is used to train the classifier which is then used to estimate the class labels of the test set to evaluate the performance of the classifier.

viii Cross validation occurs when multiple non-overlapping training and test sets are derived from a dataset to better validate the performance of a classifier.

ix n-fold cross validation is the division of the dataset into n equal sets, where nine are used for training and one for testing. This is then repeated n times by changing the test set each time, and the results are averaged to measure the performance of the classifier.

3.1.1 Logic Based Algorithms

Decision trees are one of the oldest techniques used for classification and have evolved much in the last two decades. A good overview can be found [Mur98]. Decision trees can be considered as trees that classify instances by sorting based on feature values [Kot07]. Each node in a decision tree represents a feature of an instance to be classified and each branch represents a possible value that the node can take. The classification of instances starts from the root node and sorted based on their feature values.

Constructing optimal binary decision trees is an NP-complete problem and therefore efficient heuristics have been proposed to construct near-optimal decision trees. The feature that best divides the training data is placed at the root node of the tree. A number of methods are used to find the most discriminative feature (ie: root node), including information gain and gini index. Although myopic measures estimate each attribute independently, algorithms such as RELIEFF [Kon94] estimate them in the context of other attributes. However, most studies have concluded that there is no single best method and the choice depends on the application domain [Mur98]. The same procedure is then reapplied over the created data partitions in order to create subtrees, until all the training data is divided into subsets of the same class.

A decision tree, or any learned hypothesis h for that matter, is said to overfit training data if another hypothesis $\acute{h}$ exists such that $\acute{h}$ has a larger error than h when tested on the training data, but a smaller error than h when tested on the entire dataset. Decision trees use two common approaches to avoid overfitting.

i Stop the algorithm before it reaches a point at which it perfectly fits the training data

ii Prune the induced decision tree after the algorithm has terminated

If two trees arrived at using the two approaches above respectively attain the same accuracy level, the tree with fewer leaves is then preferred. One of the most useful characteristics of decision trees is that they are human-readable. They can also be converted to a set of rules without much effort, which can be an advantage in certain instances. Decision trees tend to perform better when dealing with discrete/categorical features as the key assumption in constructing decision trees is that instances belonging to different classes have different values in at least one of their features.

3.1.2 Random Forests

A random forest (RF) is a collection of decision trees [LW02]. Classification of a new instance is obtained by majority vote over the classifications provided by individual trees included in the forest. A random bootstrap sample of data is used to train a tree [LW02]. Conventional decision trees use the best split among all variables to decide how each node is split. However, the best split among a subset of all variables is chosen in RF. Although this may appear counterintuitive, it has been pointed out that random forests perform comparably or better than a majority of classifiers such as discriminant analysis, SVM and neural networks, and are also inherently robust against overfitting [LW02].

3.1.3 Support Vector Machines

Support Vector Machines (SVM) was introduced by Vapnik et al. in 1995, which can be considered as a more recent algorithm compared to the history of other learning algorithms [CV95]. SVM is a margin based technique, where the margin is on either

side of a hyperplane that separates two data classes. Maximizing the margin creates the largest possible distance between the separating hyperplane and the instances on either side of it, and it has been proven to reduce the upper bound on the expected generalization error [Kot07].

If the training data x is linearly separable, then a pair ($\mathbf{w}$,b) exists such that

$$\begin{aligned} \mathbf{w}^T\mathbf{x}_i + b \geq 1, \text{ for all } \mathbf{x}_i \in P \\ \mathbf{w}^T\mathbf{x}_i + b \leq -1, \text{ for all } \mathbf{x}_i \in N \end{aligned} \tag{3.1}$$

with the decision rule $f_{\mathbf{w},b}$ given by

$$f_{\mathbf{w},b} = sgn(\mathbf{w}^T\mathbf{x} + b) \tag{3.2}$$

where $\mathbf{w}$ is termed the weight vector and b the bias (-b is termed as threshold). It has been shown that when there are two linearly separable classes, an optimum separating hyperplane can be found by minimizing the squared norm of the separating hyperplane. This minimization problem can be reduced to a convex quadratic programming (QP) problem:

$$\begin{aligned} \underset{\mathbf{w},b}{\text{Minimize }} \Phi(\mathbf{w}) = \frac{1}{2}||\mathbf{w}||^2 \\ \text{subject to } y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1, i = 1, .., l \end{aligned} \tag{3.3}$$

Data points that lie on its margin are known as support vector points and hence the name Support Vector Machine, as shown in figure 3.1. The solution is represented as a linear combination of support vectors while other data points are ignored. The model complexity of a SVM is not affected by the number of features encountered in the training set and SVM is considered well suited for learning tasks with a large number of features compared to the number of training instances.

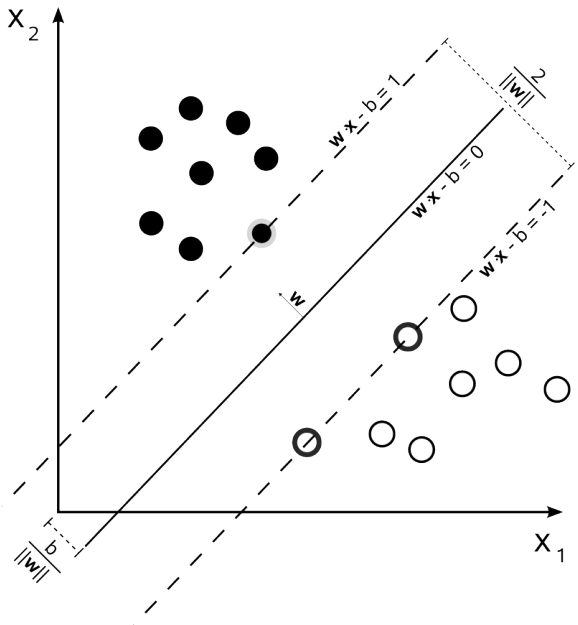


Figure 3.1: Maximum margin hyperplane and margins of a SVM trained on a two class dataset

When the classes are not linearly separable, SVM uses a kernel trick to map data onto a higher-dimensional space and define a separating hyperplane in that space. This higher-dimensional feature space is known as the transformed feature space. Conversely, a linear separation in the transformed feature space corresponds to a non-linear separation in the original input space. The mapping of the data to a Hilbert space H can be written as $\Phi : \mathbb{R}^d \rightarrow H$. Therefore the training algorithm would only depend on the data through the dot products in H, ie: on functions of the form $\Phi(x_i).\Phi(x_j)$. If there is a kernel function K such that $K(x_i, x_j) = \Phi(x_i).\Phi(x_j)$, the training algorithm can use K instead of explicitly determining Φ. Hence, kernels are considered to be a special class of functions that allow inner products to be calculated directly in feature space [Kot07, Mit97].

Training of the SVM may be considered as solving an *Nth* dimensional QP problem, where N is the number of samples in the training dataset. The Sequential Minimal Optimization (SMO) algorithm is used to solve this problem relatively quickly, and there are several optimizations available as extensions [Kot07].

3.1.4 AdaBoost

AdaBoost (AB) is a variant of boosting, with the roots of boosting going back as far as the theoretical framework of PAC (Probably Approximately Correct) learning [FS99]. Boosting builds on the concept that a 'weak' learning algorithm that performs slightly better than chance (random guessing) can be boosted into a strong learning algorithm. AdaBoost addresses the potential difficulties faced by other boosting algorithms and has become a standard in recent times.

The AdaBoost algorithm description is available [FS99] and is presented in Algorithm 3.1. Decision trees are used as the base algorithm for AdaBoost in the experiments in this thesis.

Given: $(x_1, y_1), ..., (x_m, y_m)$ where $x_i \in X, y_i \in Y = \{-1, +1\}$

Initialize $D_1(i) = 1/m$

For $t = 1, ..., T$:

i Train weak learner using distribution D_t

ii Get weak hypothesis $h_t : X \rightarrow \{-1, +1\}$ with error $\epsilon_t = Pr_{i\ D_t}[h_t(x_i) \neq y_i]$

iii Choose $\alpha_t = \frac{1}{2} ln(\frac{1-\epsilon_t}{\epsilon_t})$

iv Update $D_{t+1}(i) = \frac{D_t(i)}{Z_t}$ x $\{^{e^{-\alpha_t} \text{ if } h_t(x_i)=y_i}_{e^{\alpha_t} \text{ if } h_t(x_i)\neq y_i}$ where Z_t is a normalization factor (chosen so that D_{t+1} will be a distribution).

Output the final hypothesis: $H(x) = sign(\sum_{t=1}^{T} \alpha_t h_t(x))$

Algorithm 3.1: The AdaBoost Algorithm [FS99]

3.1.5 Ensemble Methods

The underlying concept of ensemble methods (ES) is similar to boosting. A set of weak learners, each of which performs slightly better than chance, can be combined to train a strong classifier. While many other methods of combination exist, weighted averaging and voting are commonly used. The ensemble method used is trained with multiple types of base learners, whose outputs are combined using voting. Multiple copies of base learners are trained with varying parameters and the best classifiers are determined. While some classifiers can be considered as the best reported, others yield mediocre performance. Therefore, instead of combining both the good and bad models together, a forward stepwise selection may be used to select the subset of models that, when averaged together, yield excellent performance. The basic ensemble selection procedure is summarized in Algorithm 3.2 as follows [CNMCK04].

i Start with the empty ensemble.

ii Add to the ensemble the model in the library that maximizes the ensemble's performance based on the error metric on a hillclimb validation set.

iii Repeat step 2 for a fixed number of iterations or until all the models have been used.

iv Return the ensemble from the nested set of ensembles that has maximum performance on the hillclimb validation set.

Algorithm 3.2: The Ensemble selection procedure [CNMCK04]

3.1.6 Feature Subset Selection Algorithms

The objective of feature subset selection is to eliminate redundant features from a dataset. A number of algorithms have been introduced for this purpose [Hal00] that falls into three categories: filter based, information gain based and wrapper based. The wrapper based techniques use a classifier to evaluate the features based on accuracy estimates. For a given classifier, wrapper based algorithms omit a feature at a time and test the impact of omission on the accuracy. In contrast, filter and information gain based methods use characteristics of the dataset such as correlations between features and the class label or information theoretic measures to determine and select a subset of features that are important. Filter based techniques used in this thesis include correlation based, Pearson correlation based and RELIEF subset selection, while wrapper based techniques include SVM wrapper based, RF wrapper based and cross validation based feature selection. Information gain based techniques include gain ratio based and information gain based feature selection. More detailed descriptions of these techniques can be found [Hal00, Yil15].

3.2 Deep Learning

Deep learning is a branch of machine learning that has recently became prominent [LBH15a]. Deep learning has made major advances on many problems that have been challenging for the AI community [LBH15b]. A better description of such applications can be found [LD14]. Deep learning models used in this thesis may be divided into two categories, namely discriminative models and generative models. Convolutional neural networks (CNN) and recurrent neural networks (RNN) are the discriminative models and stacked auto-encoders (SAE) is the generative model that are described and later used. Many other architectures of both categories are available in the literature [LD14].

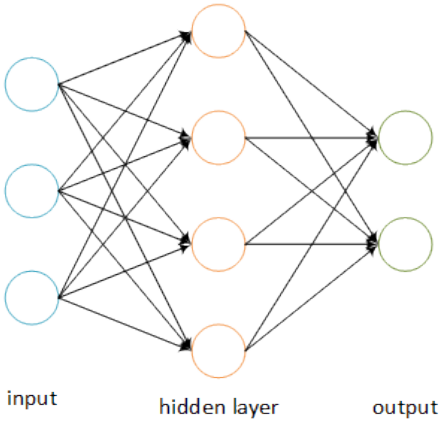


Figure 3.2: A multi layered perceptron network which is better know as ANN.

3.2.1 Convolutional Neural Networks

Convolutional Neural Networks (CNN) are an advanced type of artificial neural networks (ANN) that consists of multi layered perceptrons. The basic version has an input layer, an hidden layer and an output layer as depicted in figure 3.2. If $x_1...x_n$ are input feature values and $w_1...w_n$ are connection weights, then the perceptron computes the sum of weighted inputs: $\sum_i x_i w_i$ and the output goes through an adjustable threshold: if the sum is above the threshold, output is 1, otherwise it is -1. The most common way a perceptron is used for learning is to run the algorithm repeatedly through the training set until it finds a prediction vector that is correct on all of the training set. This prediction rule can then be used to predict labels of new instances [Kot07].

Convolutional Neural Networks (CNN) have been explored in the past and draw inspiration from typical neural networks [LBBH98]. While they were initially shown to be excellent at hand written digit recognition [LBBH98], the inability to scale CNNs to cater to larger image sizes made it impossible to use in most applications. As this was largely due to hardware and memory constraints, coupled with lack of sufficiently large datasets, recent advances in GPU computing and the curation of

large datasets such as ImageNet [Lab16] have recently made it possible to use CNNs again. The building blocks of CNNs are briefly introduced below and two popular architectures are discussed, while leaving detailed descriptions to others [SSM+16].

3.2.2 Convolution

The issue with traditional fully connected neural networks when dealing with images has always been the explosion of parameters. For example, when an image of size 100 x 100 is considered as input, the neural network would have 10,000 input nodes which in turn implies 10,000 x 10,000 = 100 million parameters, assuming that the network has 10,000 nodes in the first hidden layer. As the networks become deeper, the number of parameters grows exponentially and nearly impossible to handle even by the most advanced hardware. Therefore, instead of dealing with the whole image at one go, learning a set of convolutional filters of varying sizes is much more tractable. Convolution can be considered as a linear operation that is used for feature extraction [YNDT18]. The height and width of a convolutional filter are smaller than the input image. Each convolutional filter is slid across the width and height of the input image and the dot products between the image and the filter are computed at every spatial position, which then makes up an activation map. The output of the convolutional layer consists of the activation maps of all filters stacked along the depth dimension [KLB+18].

The added advantage of this approach is that the spatial characteristics of the image can be accounted for, unlike conventional neural networks. CNNs can be thought of as regular neural networks with two constraints [Bis07]:

Local Connectivity: In essence, each neuron is only connected to a small part of the image instead of the whole image which is the case in regular neural networks.

Parameter Sharing: Since the same convolution filter is applied across the whole image, weights are shared between these filters.

3.2.3 Pooling

A pooling operation essentially reduces the size of the activation maps for the next layer, enabling use of a smaller number of parameters progressively [YNDT18]. CNNs use different types of pooling depending on the architecture, however max-pooling is the most used pooling technique. For an n x n region, max-pooling will replace that region with its maximum value, reducing the size by a factor of n^2. Providing a small degree of spatial invariance can be considered as an added advantage of pooling [YNDT18].

3.2.4 Non-linearity

Since a cascade of linear systems such as convolutions generate another linear system, non-linearities between convolutions are added to expand their expressive power. The outputs of the convolutional layer are then passed through the non-linear activation layer. Modern CNNs typically use ReLu non-linearity which can be expressed as $ReLu(x) = max(0, x)$. CNNs with ReLu non-linearity are shown to converge faster [NH10].

3.2.5 AlexNet

Adding the building blocks together, AlexNet may be taken as an early example of a CNN network [KSH12a]. AlexNet was trained on ILSVRC 2012 training data which contained 1.2 million training images categorized into 1000 classes. AlexNet has 7

layers consisting of combinations of convolution, pooling and non-linearity as shown in Figure 3.3. Visualization of the output of layers shows that earlier layers tend to learn low level features similar to gabor-like oriented edges and blob-like features, while later layers tend to learn higher level features such as shapes and textures. Final layers also appear to learn semantic attributes such as eyes or wheels [LBH15a].

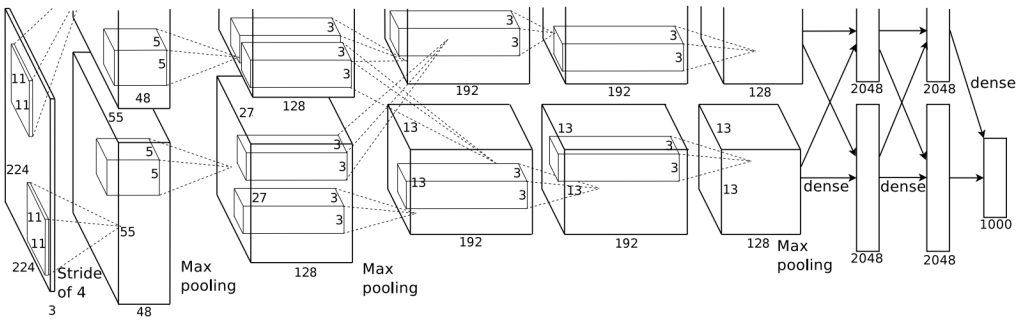


Figure 3.3: An illustration of the architecture of our CNN, explicitly showing the delineation of responsibilities between the two GPUs. One GPU runs the layer-parts at the top of the figure while the other runs the layer-parts at the bottom. The GPUs communicate only at certain layers. Illustration and description adapted [KSH12a]

3.2.6 GoogleNet

GoogleNet is another CNN trained on the ILSVRC14 dataset, and has 22 layers and the visualizations of the network are available [SLJS14]. The same group presented a new architecture called Inception that tries to use readily available dense components to approximate the optimal local sparse structure of a convolutional vision network. A better description of the layers and the rationale for their individual use can be found in the original paper.

3.2.7 RNN

Despite the power of standard neural networks, one limitation they suffer from is the assumption of independently generated samples. If the samples used to train a classifier are related in time or space, neural networks can fall short. A typical example would be time-series data such as frames from a video or snippets of audio. Another issue with these types of data is that the samples can be vectors of different lengths at different time points, whereas typical neural networks rely upon sample vectors of fixed length. Recurrent neural networks (RNN) was proposed to alleviate these issues as a connectionist model with the ability to selectively pass information across sequence steps, while processing sequential data one element at a time [Lip15].

A standard RNN computes the hidden vector sequence $h = (h_1, h_2...h_T)$ and output vector sequence $y = (y_1, y_2, ...y_T)$ given an input sequence $x = (x_1, x_2, ...x_t)$ by iterating over equation 3.4 from $t = 1$ to T [GMH13]:

$$\begin{aligned} h_t &= \mathcal{H}(W_{xh}x_t + W_{hh}h_{t-1} + b_h) \\ y_t &= softmax(W_{hy}h_t + b_y) \end{aligned} \tag{3.4}$$

Here W_{xh} is similar to the conventional weight matrix between the input and the hidden layer, while w_{hh} can be thought of as the weight matrix between the hidden layer and itself at adjacent time steps [Lip15]. This can be represented in Figure 3.4 and the dynamics of the network across time steps can also be visualized by unfolding it as shown in Figure 3.5.

RNNs are better described elsewhere [Lip15, GMH13, HS13]. A popular extension of RNN is long-short term memory networks (LSTM) that are widely used today [HS13]. RNNs are rarely used in medical imaging, although there can be interesting applications, as medical imaging applications do have sequential data.

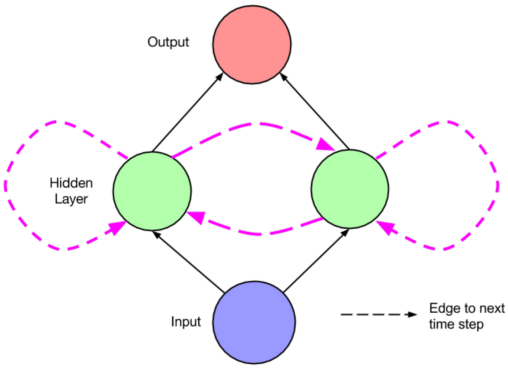


Figure 3.4: A simple recurrent network. At each time step t, activation is passed along solid edges as in a feed-forward network. Dashed edges connect a source node at each time t to a target node at each following time $t + 1$. Illustration adapted from [Lip15]

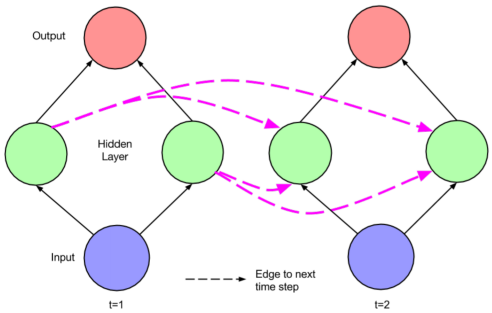


Figure 3.5: The recurrent network of Figure 3.4 unfolded across time steps. Illustration adapted from [Lip15]

3.2.8 Auto-encoders

Auto-encoders are a type of artificial neural network that can be defined with three layers: (i) input layer (ii) hidden layer and (iii) output layer. They transform inputs into outputs with the least possible amount of distortion. Auto-encoders were first introduced in the 1980s and their history and evolution are elaborated elsewhere [Bal12]. The typical architecture of an AE is shown in Figure 3.6. It is predominantly an unsupervised learning algorithm but recent advances have made it possible to use a set of auto-encoders stacked on top of each other as a supervised learning algorithm [HOT06].

Denoting the input vector by $x \in \mathbb{R}^{D_I}$, where D_H and D_I denote the number of hidden and input units respectively, an auto-encoder creates a deterministic mapping from an input to a latent representation y such that $y = f(W_1 x + b_1)$. This is parameterized by the weight matrix $W_1 \in \mathbb{R}^{D_H x D_I}$ and the bias vector $b_1 \in \mathbb{R}^{D_H}$. This latent representation $y \in \mathbb{R}^{D_H}$ is mapped back to a vector $z \in \mathbb{R}^{D_I}$ which can be considered as an approximate reconstruction of the input vector x with the deterministic mapping $z = W_2 y + b_2 \approx x$ where $W_2 \in \mathbb{R}^{D_H x D_I}$ and $b_2 \in \mathbb{R}^{D_I}$.

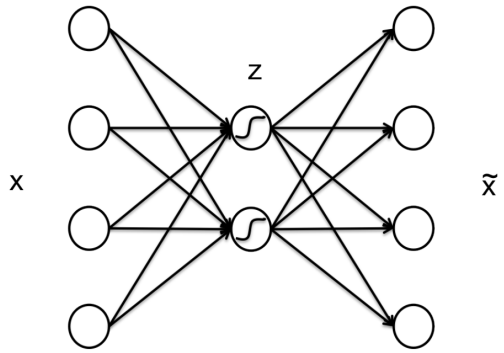


Figure 3.6: A typical AE transforms the input x to output $\widetilde{x}$ with minimum amount of distortion by encoding the input into z and decoding it back. Illustration adapted from [Le15].

3.2.9 Transfer Learning

Transfer learning was proposed as a solution to one of the prevalent problems in machine learning; namely the paucity of training data [TSK+18]. Especially in a domain like medical imaging, it becomes a challenge to find appropriate labelled training data. Transfer learning is the improvement of learning in a new task through the transfer of knowledge from a related task that has already been learned [OGS+09]. There are two prominent ways to operationalize transfer learning: further develop a trained model or fine-tune a pre-trained model [KJS18]. The former technique allows one to train a model for a specific task and then reuse all or parts of that model along

with further changes if necessary and on another similar task. The latter technique focuses on using a pre-trained network and retraining it on a different dataset which is known as fine-tuning [TSG+16]. An extension of this method specific to deep learning is to use an already trained deep neural network to extract features that are relevant and subsequently use those features to train another machine learning algorithm of choice [TSK+18, TSG+16, ARS+14]. This thesis explores transfer learning based on deep learning using discriminative models.

An important consideration when applying the transfer learning paradigm is the domain of the trained network and the new domain of application. It has been demonstrated that performance improves if both networks share the domain [ARS+14]. In natural classification problems, this has become particularly straight-forward with the availability of a plethora of pre-trained networks trained on the ImageNet dataset and other natural image datasets [KSH12a, SLJ+14, Lab16]. However, this is still a significant challenge in the medical imaging domain as the images used in the natural domain and medical imaging domains vary significantly. For instance, natural images have three channels (full colour) whereas medical images are usually grayscale. Medical images also tend to be translation and rotation invariant, whereas natural images do not exhibit such characteristics commonly [ARS+14].

3.3 Diagnosis of Cognitive Impairment

The diagnosis of cognitive impairment can be subdivided into two categories depending on the data that is used for diagnosis. The prevalent method is based on neuropsychological test scores, as already described in Section 2.4. In addition to that, medical imaging techniques as described in Section 2.5 are also used as supplementary information to diagnose cognitive impairment. The use of neuropsychological test scores are expanded in Section 3.3.1, while Section 3.3.2 describes how medical

imaging techniques are utilized. The final subsection expands on how these two data modalities are used with machine learning to diagnose cognitive impairment.

3.3.1 Diagnosing Cognitive Impairment using Neuropsychological Measures

Any patient with suspected MCI undergoes a comprehensive history and physical examination focusing on cognitive function, functional status, medications, neurological or psychiatric abnormalities, and laboratory testing [LL14]. The suggested approach for diagnosis and management of MCI is depicted in Figure 3.7.

3.3.2 Diagnosing Cognitive Impairment using Medical Imaging

National Institute on Aging and the Alzheimer's Association (NIA-AA) guidelines do not recommend routine neuroimaging in the typical clinical assessment of cognitive impairment [LL14]. They however do propose research criteria where neuroimaging may help in determining cognitive decline etiology and prognosis. There are a number of published studies that suggest the use of structural magnetic resonance imaging for identifying cognitive impairment [CSL+12, CWL+12, JSL+14, RKB+13].

A list of biomarkers based on structural MRI (T1w and T2w) were collected and are listed in Table 3.1 [CSL+12, CWL+12, JSL+14, RKB+13]. These biomarkers and their importance are further elaborated in section 3.3.3.

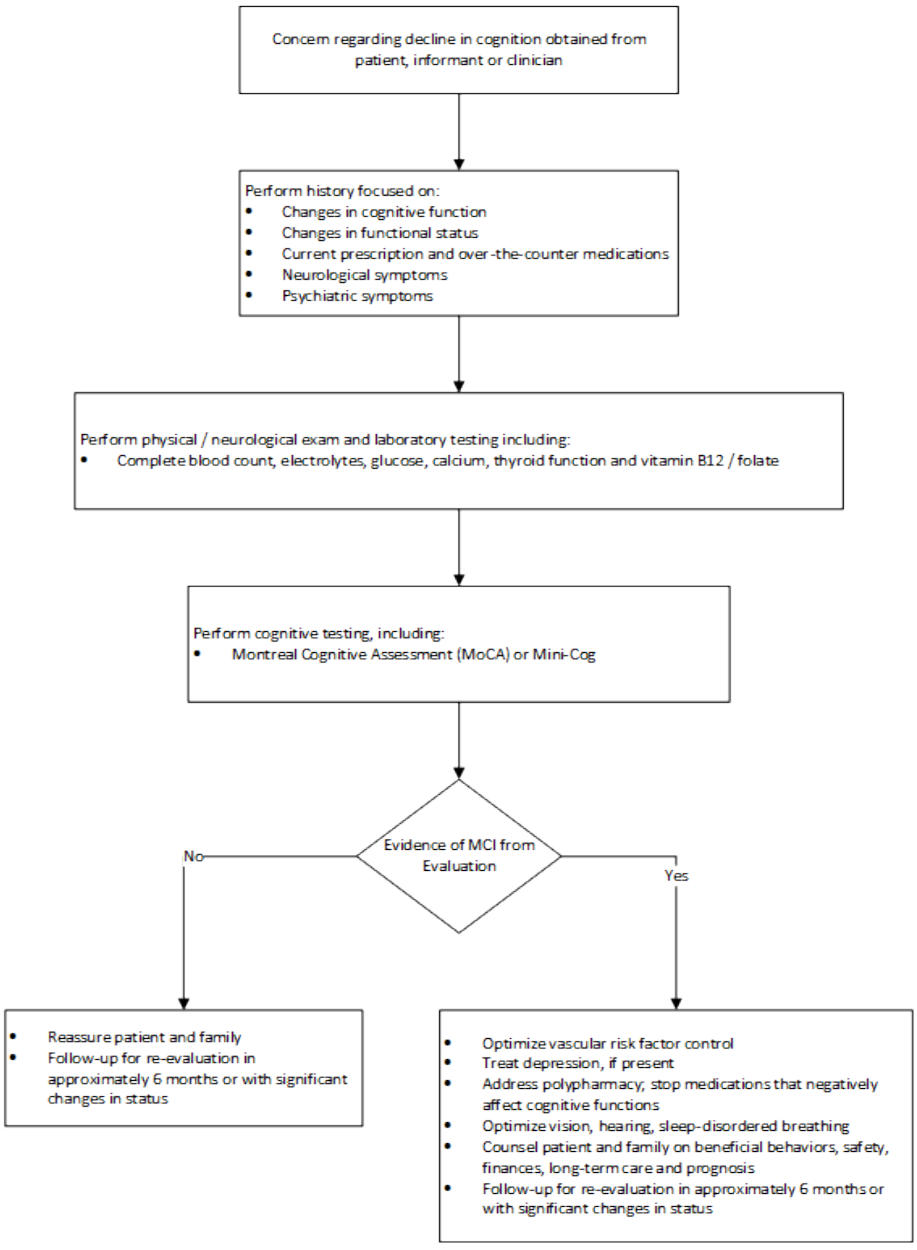


Figure 3.7: Suggested Approach to the Diagnosis and Management of MCI. The illustration is recreated from [LL14]

Table 3.1: Common biomarkers based on Structural MRI found in Literature

Biomarkers based on structural MRI
Cortical Thickness
Cortical Volume
Subcortical Volume
Grey Matter Atrophy
Hippocampal Atrophy
Entorhinal Volume

3.3.3 Diagnosing Cognitive Impairment with Machine Learning

Depending on the data modality used, diagnosis of cognitive impairment using machine learning may be divided into two; using NM features and using multi-modal MR images. However, the work presented in this thesis is the first attempt to use NM features and machine learning to diagnose cognitive impairment as far as is known. The current progress in the area may be categorized into three areas of interest:

i Brain MR image preprocessing and analysis: This includes initial preprocessing, registration, segmentation and temporal tracking of the brain MR images.

ii MCI diagnosis using multi-modal MR images: This includes the use of Diffusion Tensor Imaging and Perfusion imaging modalities in addition to structural MR images.

iii Reliable differentiation between MCI subtypes: This includes the differentiation between aMCI and naMCI as well as their single/multiple domain subtypes.

The first is a strong area of research, as it has ubiquitous utility in many other areas as well. Techniques such as registration, segmentation and tracking of brain MR images have been examined thoroughly and state-of-the-art algorithms

perform very well in these tasks [Bow14]. As brain MR preprocessing and analysis is not the focus of this thesis, they are not further reviewed.

MCI diagnosis using multi-modal MR images is a recent development, with the development of different MRI modalities in recent times. Structural MR modalities such as T1 and T2 have been in use for a long time, whereas other MR modalities such as diffusion imaging and perfusion imaging are relatively new. Scientists have begun to utilize these imaging modalities to improve the performance of existing diagnosis algorithms. A review of this area is in section 3.3.3.1.

The third area of focus is the differentiation between the individual subtypes of MCI, as different subtypes have different propensities to progress to different types of dementia. This has benefited from both structural MR image analysis and multi-modal MR image analysis and continues to improve in performance. A review is in subsection 3.3.3.2.

3.3.3.1 MCI Diagnosis using multi-modal MR images

The modalities used to diagnose MCI include T1w, T2w and DTI images. The earliest work in this regard was by Peterson et al. [JPX+99]. They reported that Hippocampal volume was associated with progression from MCI to dementia with a relative risk of 69%. Leyla et al. carried out a similar experiment to understand whether Hippocampal volume or entorhinal volume is the better predictor of conversion from MCI to AD [dMSB+XX]. A sample image is shown in Figure 3.8 highlighting both entorhinal and hippocampal formations. They reported that entorhinal volume was in fact the better predictor. They concluded that the right hemisphere entorhinal volume was the best predictor of conversion with a concordance rate of 93.5%.

Chételat et al. also tried to associate the structural changes occurring in the brain to conversion of MCI to AD [CLE+05]. They indicate that hippocampal

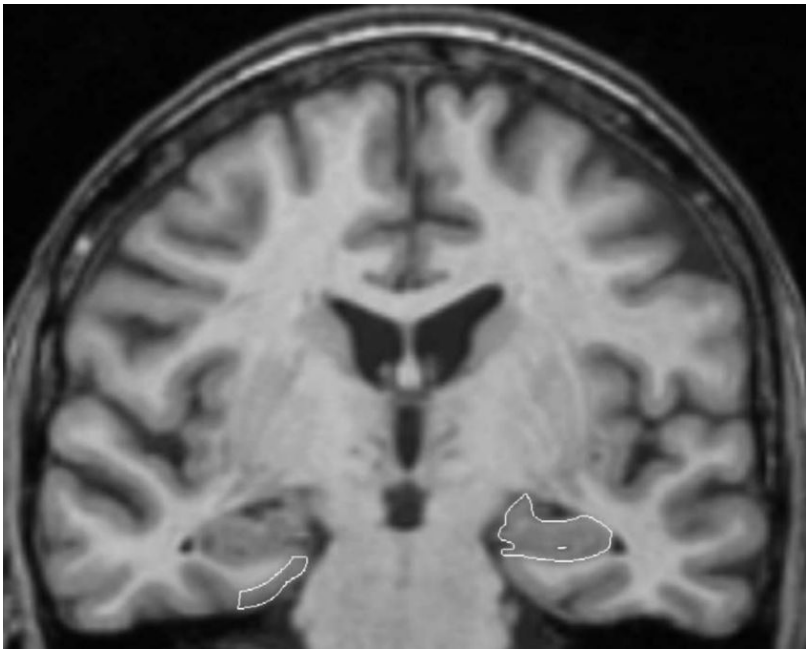


Figure 3.8: A single coronal slice illustrating the segmentation of the entorhinal cortex (left-hand side) and the hippocampal formation (right-hand side) [dMSB$^+$XX].

atrophy plays a discriminative role in this and in fact, the regions that exhibit most grey matter loss are the inferior and middle temporal gyrus, posterior cingulate, and precuneus. This can be seen in Figure 3.9. Similar approaches have been attempted for discrimination between MCI and AD, MCI and CN, CN and AD as well as MCI subtypes [CWL$^+$12, DBS$^+$11, EWT$^+$12, HSXJ11, JSL$^+$14, MFD09].

Raamana et. al took a different approach and used a derived measure instead of the volumetric measure to differentiate between the subtypes of MCI [RWK$^+$14]. They used FreeSurfer [RSRF12] to perform the initial cortical reconstruction and volumetric segmentation of the whole brain. The feature they derived from that was cortical thickness which was measured by solving a discrete approximation of Laplace's equation. It was then registered to the surface of a common atlas so that vertex-wise correspondence could be analyzed. Thereafter, they applied dimensionality reduction techniques before using SVM for classification. The results of this study are further discussed in section 3.3.3.2.

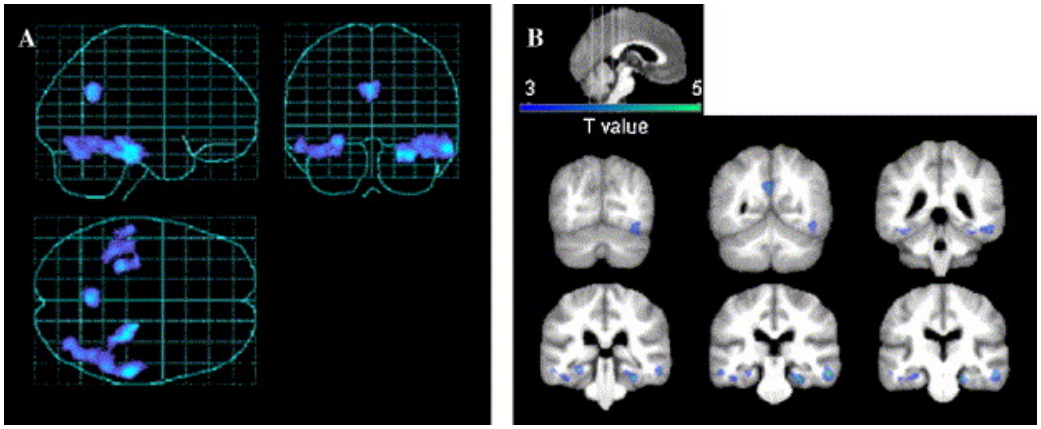


Figure 3.9: (A) 'Glass brain' representation showing significant clusters of greater gray matter loss in converters compared to non-converters. (B) The same results as projected onto coronal sections of the whole brain customized template [CLE+05].

Wolz et al. used a different approach by combining a range of features obtained from structural MR images such as hippocampal volume, tensor-based morphometry and cortical thickness [WJK+11]. They tested their techniques on the ADNI dataset and reported 90% sensitivity and 84% specificity using a LDA classifier to classify AD and CN individuals.

Another use of multi-modalities (although not MRI modalities) is the study carried out by Cui et al [CSL+12]. They used a mixture of non-image data such as neuropsychological test scores in conjunction with neuroimaging morphological measures to predict conversion from CN to MCI. This study uses the MAS dataset. The initial preprocessing was carried out with FreeSurfer and they used a Region of Interest based (74 cortical regions of interests per hemisphere) feature set that includes features such as cortical thickness average, surface area, cortical volume, mean curvature, Gaussian curvature, folding index and curvature index. Fisher Information criteria were used for feature selection and SVM for the classifier. They suggest that the combination of neuropsychological and neuroimaging features results in best performance.

So far, the discussion has been centred on T1w and T2w MR images.

There are multiple studies that use DTI images in MCI diagnosis as well. Chua et al. provide one of the first reviews that details the experiments carried out before 2008 [CWSS08]. They then proceed to identify the DTI of the posterior cingulate as a useful biomarker of MCI in a subsequent publication [CWC+09]. In this study, they found that individuals with aMCI demonstrated microstructural pathology in para-hippocampal white matter (WM), frontal WM, splenium of corpus collosum and posterior cingulate region. Individuals with naMCI on the other hand demonstrated microstructural pathology in frontal WM, internal capsule, occipital WM and posterior cingulate region. They have suggested that DTI of the left posterior cingulate is the distinguishing feature in identifying individuals with aMCI. Another study by the same group studied microstructural changes in WM of CN individuals who are at risk of aMCI [ZST+12]. It was a longitudinal study and compared 193 CN individuals out of whom 20 were diagnosed as aMCI after two years. It was reported that at baseline compared to CN-stable individuals, CN-aMCI converters had substantial reductions in WM integrity in the precuneus, parahippocampa cingulum, parahippocampal gyrus WM and fornix. It was also found that fractional anisotropy (FA) values of the precuneus is a predictor of conversion from CN to aMCI. In addition, FA values of parahippocampal gyrus WM were predictive of subsequent episodic memory decline. A sample FA map is shown in Figure 3.10.

Yet another study by the same group reports that WM changes detect early aMCI as opposed to popular observation of hippocampal atrophy [ZST+13]. They use DTI tractography and the key finding of this study is that there is a regional difference between early aMCI, late aMCI and CN individuals. Compared to CN, late aMCI had lower WM integrity in the fornix, parahippocampal cingulum and uncinated fasciculus while early aMCI showed WM damage in the fornix. Fornical measures were correlated with hippocampal atrophy in late aMCI while abnormality of the fornix in early aMCI did not correlate with hippocampal volumes because that occurred in the absence of hippocampal atrophy.

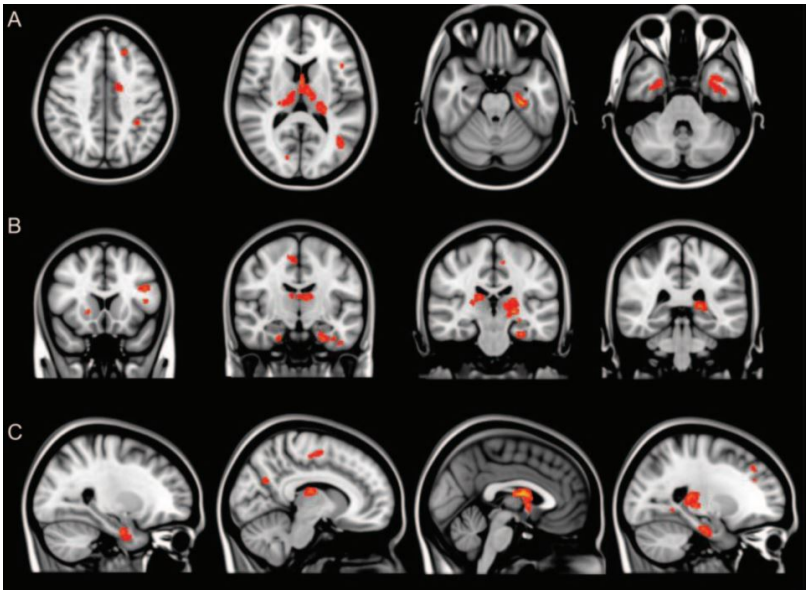


Figure 3.10: Pattern of fractional anisotropy (FA) reductions in presymptomatic individuals who eventually developed aMCI [ZST$^+$12].

Similar approaches are reported elsewhere [TWZ$^+$12, CWC$^+$09, SKF$^+$11, SM10, RZW$^+$14, SLC$^+$13].

3.3.3.2 Reliable Differentiation between MCI subtypes

Differentiation between MCI subtypes is an area that was overlooked initially but has gathered pace more recently. This is important because different MCI subtypes can progress to different types of dementia. As it has been established that any treatment for dementia may need to be initiated before the individual becomes demented, research on MCI identification and progression has come into the limelight.

Initially subtype classification was attempted using morphological images. The study carried out by Whitewell et al. gives a good base to build on [WPN$^+$07]. They report grey matter loss in both sd-aMCI and md-aMCI groups in the medial and inferior lobes compared to CN individuals. Individuals diagnosed with na-MCI were showing grey matter loss in the left inferior temporal lobe, basal forebrain and

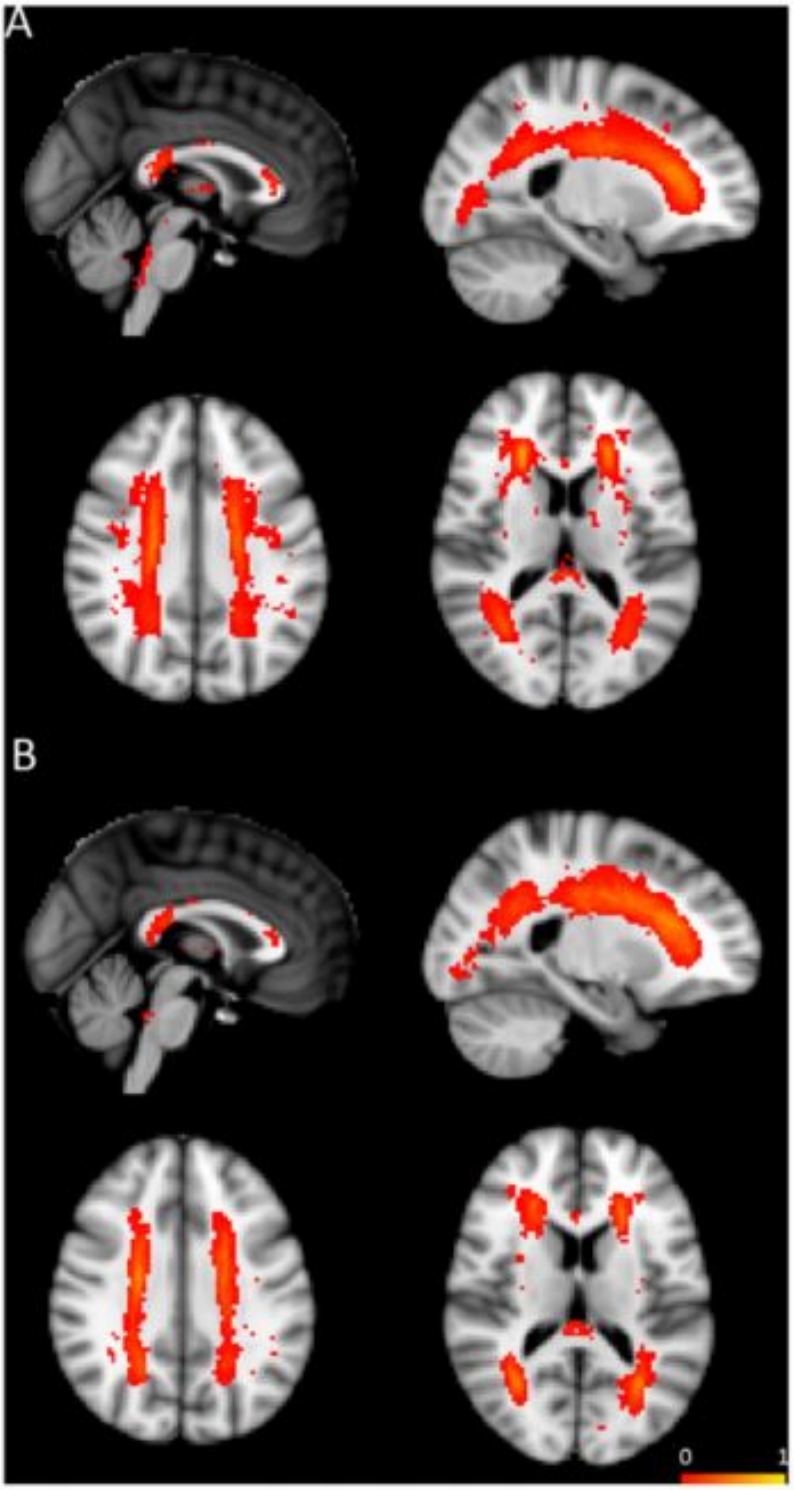


Figure 3.11: Probability maps of white matter hyperintensities (WMHs) from early (A) and late aMCI (B). WMHs were shown in the red-yellow and superimposed on the MNI T1 template. The colour bar denotes the percentage of subjects who had WMHs at each image voxel [ZST$^+$13].

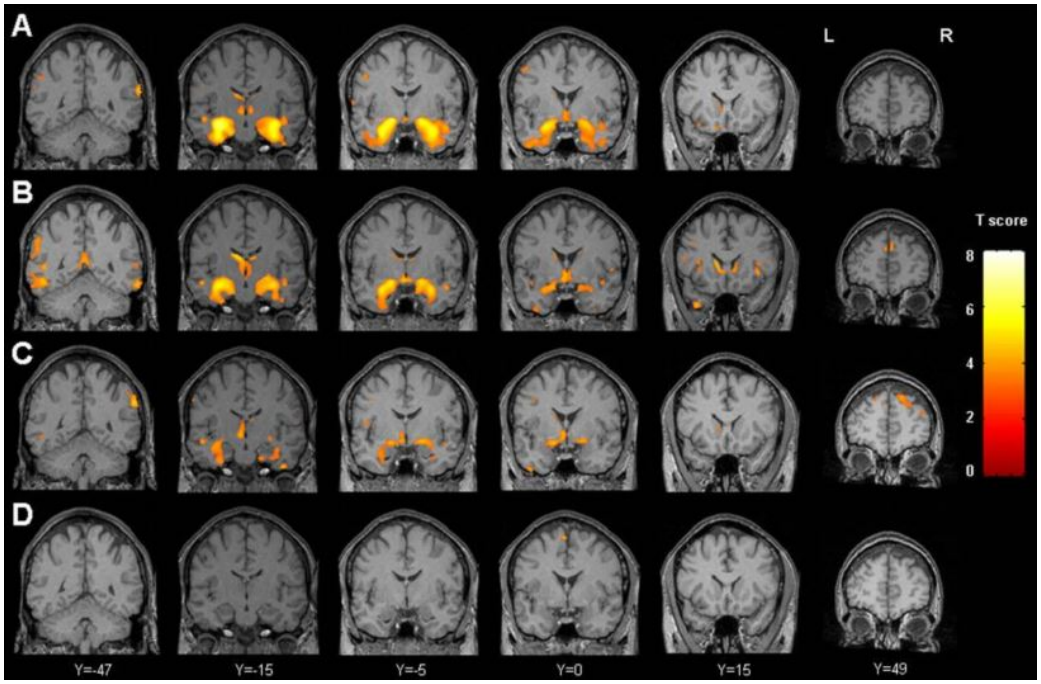


Figure 3.12: Patterns of grey matter atrophy identified by voxel-based morphometry in amnestic single-domain (A), amnestic multi-domain (B), non-amnestic single-domain (C), and non-amnestic multi-domain (D) MCI subgroups compared to controls [WPN+07].

hypothalamus.

Another study carried out by Thillainadesan et al. conducts a group comparison between individuals classified to MCI subtypes [TWZ+12]. They use regional FA and mean diffusivity (MD) scores to compare groups. They report that individuals with MCI subtypes had increased MD in the right putamen, right anterior limb of the internal capsule, genu and plenium of the corpus callosum, right posterior cingulate gyrus, left superior frontal gyrus and right and left corona radiate. They also localized the regional changes for different subtypes of MCI. The study by Cui et al. is aimed at identifying individuals with aMCI using a combined feature set derived from T1w images and DT images [CSL+12]. A classification accuracy of 71% was reported in this study in discriminating between CN and aMCI individuals. A subsequent study by the same group reports microstructural WM changes occurring in CN individuals at risk of aMCI [CWL+12]. They ascertained that CN-aMCI converters

had substantial reductions in WM integrity in the precuneus, parahippocampa cingulum, parahippocampal gyrus WM and fornix. It was also found that FA values of the precuneus is a predictor of conversion from CN to aMCI. In addition, FA values of parahippocampal gyrus WM were predictive of subsequent episodic memory decline.

Another study by the same group compares the microstructural WM changes between CN, early aMCI and late aMCI individuals [ZST+13]. This may be considered as a step towards progression tracking and helps the medical community to better understand the exact changes occurring. They use DTI parameters such as FA, MD, and axial diffusivity in this experiment. The key finding of this study is that there is a regional difference between early aMCI, late aMCI and CN individuals. Compared to CN, late aMCI had lower WM integrity in the fornix, parahippocampal cingulum and uncinated fasciculus, while early aMCI showed WM damage in the fornix. Fornical measures were correlated with hippocampal atrophy in late aMCI while abnormality of the fornix in early aMCI did not correlate with hippocampal volumes because that occurred in the absence of hippocampal atrophy.

A recent study on sub-classification of aMCI using MRI based cortical thickness measures uses an SVM classifier [RWK+14]. They aim to discriminate between aMCI subtypes such as sd-aMCI and md-aMCI. The highest accuracy they obtained is for the classification of md-aMCI and CN, where they report 61% accuracy. The other experiments resulted in accuracies of around 50% which is not very useful. A subsequent study using a derived feature set reported improved highest accuracy of 62% for the same classification and improved accuracy levels centered around 57% for the other experiments [RKB+13]. Although the accuracies reported are on the low side, the method they used to derive the feature set is of interest. They derived secondary measures using cortical thickness, which were then used to train a classifier. Initially FreeSurfer was used to reconstruct and segment the whole brain to obtain pial and WM/GM surfaces. Then the cortical thickness measures were

calculated in the resulting cortical parcellations by solving a discrete approximation to Laplace's equation. The surface of each subject was registered to the surface of a common atlas in order to establish vertex wise correspondence, enabling group wise analysis of difference in thickness. Each cortical label was partitioned using k-means clustering and then a mean thickness value for each partition was calculated. Each of these partitions is modelled as a node in a graph and links between nodes are established if the difference in mean thickness exceeds a certain threshold. The network built by this process is known as ThickNet and network features such as node degree, betweenness centrality and clustering coefficient are calculated and used as features in the classifier. The classifier used in this experiment is Variational Bayes probabilistic Multiple Kernel Learning (VBpMKL).

The highest performance for a classifier trying to discriminate between MCI subtypes and CN individuals was a study carried out by Haller et. al [HMH$^+$13]. They report an accuracy of around 98% for all of their experiments and they use a feature set derived from DT images such as FA and MD. However, the authors also suggest a degree of over-fitting since, based on work done so far, accuracy of that calibre with the same feature set has not been achieved.

3.4 Survival Analysis for Cognitive Impairment and Dementia

Better understanding of the progression of cognitive impairment has been a challenge. This thesis exploits survival analysis techniques and proposes solutions to better understand the progression from CN and MCI states to dementia. Survival analysis techniques are capable of evaluating the significance of prognostic variables to determine events such as death or recurrence of a disease such as cancer [YSK$^+$16, CYA13, BT04, RA13]. The objectives of survival analysis are three fold:

(i) to estimate survival and hazard functions from survival data, (ii) to compare survival and hazard functions between groups and (iii) to assess the relationship between predictor variables and survival time [CAIH$^+$16]. The focus of this thesis is on the first and the last objectives. Typically, survival analysis requires as inputs the prognostic variables (baseline or time-varying), time to an event of interest and the binary event outcome. These can then be used to train a survival analysis model capable of predicting time to an event of interest given a new patient's prognostic variables. The deep survival analysis techniques proposed in this thesis are the first approach that allows utilization of imaging data as far as can be ascertained. They also exploit the deep fusion network proposed in Chapter 5 to extract imaging and fused features to build survival analysis models. The descriptions of current survival networks fit better with the proposed models, and are explained in detail in Chapter 6.

Other than survival analysis, personalised machine learning techniques such as evolving spiking neural networks may also be re-purposed for the same task [KH10]. Personalised modelling focuses on creating an individual model from data to better predict an unknown outcome for an individual. It is different to global (group-level) modelling in that models are created to cover an individual rather than the whole problem space [KH10, KFH$^+$14]. However, due to the complexity involved and paucity of data, personalised modelling is out of scope from this thesis.

3.5 Motivation for Thesis Approach

Many methods for diagnosing cognitive impairment have been introduced in recent years. Conventional machine learning methods as well as deep learning methods have been proposed [JNS19]. However, at the time of writing of this thesis, neuropsychological measure based features with machine learning have not been used to diagnose cognitive impairment. This type of data is significantly different from neuroimaging

data, making it a worthwhile exploration. In addition, while neuroimaging data has been used with machine learning for diagnosis of cognitive impairment, only a few studies investigate the use of multi-modal neuroimages in a single pipeline. Finally, investigating the application of survival analysis techniques for cognitive impairment prediction has not explored at the time of writing this thesis. These reasons together with access to two unique datasets, namely MAS and ADNI have provided several different motivations for this research. However, this is a fast evolving field.

The main motivation of this thesis is to investigate the fusion of data from multiple modalities in a single machine learning pipeline. To acquire a baseline for this, it was important to design and propose machine learning pipelines using NM features and structural MRI based features. Another significant contribution is the design of deep learning algorithms capable of extracting features from structural MRI without having to go through the conventional hand designed feature engineering pipeline that includes registration, segmentation and feature extraction. Better understanding of when an individual experiences cognitive decline is another strong motivation for this thesis. Survival analysis using NM features and MRI based features are proposed in thesis.

3.6 Summary

This chapter described the machine learning methods that are currently utilized to diagnose cognitive impairment and introduced the use of survival analysis to better understand progression to cognitive impairment. A brief introduction to machine learning was presented first. The techniques used to diagnose cognitive impairment were described followed by an introduction to survival analysis. Finally, the motivations for the thesis approach were presented. The identified issues are investigated and solutions proposed in the ensuing chapters.

Chapter 4

Diagnosis of MCI using Machine and Deep Learning Techniques*

Ageing is frequently associated with brain pathology, especially neurodegenerative and vascular, and the distinction between normal and pathological ageing may sometimes be difficult to make [HG04]. The current understanding is that everyone goes through this process but at varying rates [CWC+09, CSL+12, GRZ+06]. Therefore, distinguishing between normal ageing and decline due to pathological processes has become increasingly important. Alzheimer's disease (AD) is one of the

Portions of this chapter appear in:

1. Senanayake, U., Sowmya, A., Dawes, L., Kochan, N. A., Wen, W., Sachdev, P. (2016). Classification of mild cognitive impairment subtypes using neuropsychological data, in ICPRAM 2016: Proceedings of the 5th International Conference on Pattern Recognition Applications and Methods, pp. 620-629

2. Senanayake U;Sowmya A;Dawes L;Kochan NA;Wen W;Sachdev P, 2017, Deep learning approach for classification of mild cognitive impairment subtypes, in ICPRAM 2017 - Proceedings of the 6th International Conference on Pattern Recognition Applications and Methods, pp. 655 - 662

best studied neurodegenerative diseases that involve dementia in this context. It is a neurodegenerative disease that causes progressive cognitive impairment with lasting effects for the patients and their families. Although drug treatment and preventative clinical trials have largely been unsuccessful, there remains a major research focus on early and accurate identification of individuals at risk of AD and other dementias who are the most appropriate target group for trials.

Mild Cognitive Impairment (MCI) is considered a prodromal condition of dementia that is characterized by a decline in cognitive functions without severely affecting the patient's daily functions [CLE+05, CWL+12, HMH+13, PKB+09]. Individuals with MCI have a 6-10% risk of progressing to dementia in a year [MSF09]. Consensus diagnosis criteria for MCI exist [WPK+04, ADD+11], but these are operationalized differently in different studies and regions, resulting in differing rates of MCI [KSB+10]. This makes it difficult to reliably separate individuals with MCI from cognitively normal individuals and to identify those at increased risk of progression to dementia. While researchers in this area focus on three distinct problems, namely: (i) differentiating between cognitively normal (CN) and MCI individuals, (ii) predicting conversion from MCI to AD and (iii) predicting the time to conversion from MCI to AD [LSG+12], the focus of this chapter is on the first, with the second and third problem dealt with in Chapter 6. In addition, classification of MCI subtypes is also of interest as each subtype has differential rates of conversion to different types of dementia [AHK+17].

There are two major subtypes of MCI; amnestic MCI and non-amnestic MCI. The amnestic subtype of MCI (aMCI) refers to impairment in memory, while the non-amnestic subtype of MCI (naMCI) refers to non-memory impairments affecting executive functions, attention, visuospatial ability or language. These two subtypes are further divided depending on the number of domains impaired, leading to four subtypes of MCI, as shown in Table 4.1 [WPK+04, ADD+11]. Recent studies

Amnestic subtype of MCI (aMCI)	**Non-amnestic subtype of MCI (naMCI)**
Single domain aMCI (sd-aMCI)	Single domain naMCI (sd-naMCI)
Multi domain aMCI (md-aMCI)	Multi domain naMCI (md-naMCI)

Table 4.1: The subtypes of MCI

point out that md-aMCI has the highest probability of progression to AD and other types of dementia [GSS+11].

Much recent effort in this area of research has been spent on the use of advanced medical imaging modalities such as magnetic resonance imaging (MRI) and positron emission tomography (PET) to differentiate between AD and subtypes of MCI [CLE+05, SM10, SKF+11, RZW+14]. This chapter, however, is focused on evaluating the efficacy of neuropsychological measures (NM) for the same purpose. This is especially important as current clinical diagnosis criteria of dementia and mild cognitive impairment are mostly derived from NM measures.

The rest of this chapter is outlined as follows. An introduction to the dataset is presented in section 4.1 while section 4.2 delves into the methodology. Results are presented in section 4.3 and the chapter is concluded with section 4.4.

4.1 Dataset

The MAS dataset used for this work was described in section 2.3.1. The demographics of the participants at baseline are given in Table 4.2. Of the sample of 1037 individuals at baseline, 164 individuals of non-English speaking background (English acquired after the age of nine years of age) were excluded as their test performance on neuropsychological tests may be disadvantaged on measures developed and normed on native English speakers [KSB+10], and also 36 others who did not have complete neuropsychological data. Dataset sizes for all four waves grouped by the individual

Sample size: 837	**Baseline (wave 1)**
Age (years)	78.57 ± 4.51 (70.29-90.67)
Sex (male/female)	43.07% / 56.92%
Education (years)	12.00 ± 3.65
MMSE (Mini-Mental State Exam)	28.77 ± 1.26
CDR (Clinical Dementia Rating)	0.066 ± 0.169

Table 4.2: Demographic characteristics of participants at baseline

classes are listed in Table 4.3.

Class	Wave 1	Wave 2	Wave 3	Wave 4
CN	504	405	378	293
sd-aMCI	103	93	44	63
md-aMCI	82	58	46	42
sd-naMCI	122	87	64	76
md-naMCI	26	22	16	14
Dementia	0	18	29	39

Table 4.3: Dataset size for different waves

4.1.1 Cognitive Assessments

The list of Neuropsychological measures (NM) used are listed in Table 4.4 [SBR+10]. MCI was diagnosed according to international consensus criteria [WPK+04] and was based on the presence of cognitive impairment (i.e., 1.5 standard deviations below published normative values for age and/or education on a neuropsychological measure), a subjective complaint of decline in memory or other cognitive function along with normal or minimally impaired instrumental activities of daily living attributable to cognitive impairment (total average score <3.0 on the Bayer Activity of Daily Living Scale [Hin98] adjusted for physical impairment). Participants with no impairments on NM were deemed to have normal cognition. Apart from this, when unusual clinical features or an indication of possible dementia were found, a panel of

psychiatrists, neuropsychiatrists and neuropsychologists were consulted. Consensus diagnosis of MCI, dementia or cognitively normal (CN) was made using all available data where necessary and the detailed methodology has been published [SBR+10]. The same tests were administered over all the waves at two yearly intervals. The relevant test for each cognitive domain is in Table 4.4. Each is then further broken down into individual tests, with the mean and standard deviation values grouped for each wave in Table 4.5. As the number of CN individuals decreases as the waves progress, an interesting trend seems to emerge for each test. For example, the mean of DSym_raw decreases from wave 1 to wave 4 while the mean of TMTA_raw appears to increase. While these are not considered as diagnostic biomarkers individually, they do indicate cognitive decline if the same cohort is considered.

4.2 Method

The proposed method uses neuropsychological test scores as input and builds models that are capable of differentiating between individuals with MCI and those that are cognitively normal. In the study, the initial focus was on identifying the relevant techniques. Once the techniques were identified, improving performance became the main focus. The rest of this section is continued as follows. Correlation between covariates are studied in section 4.2.1, while section 4.2.2 and 4.2.3 describe class breakdown and feature subset selection. In section 4.2.4, the deep learning techniques used and the results are described, and section 4.2.5 concludes by discussing validation of results.

Table 4.4: Neuropsychological Test Battery and Normative Data used for Diagnostic Classification

Cognitive Domain	Test	Normative Data & Demographic adjustments
Premorbid Intelligence	National Adult Reading Test (NART)	
Attention/ Processing speed	Digit Symbol-Coding [Axe01] Trail Making Test(TMT) A [SSS06]	Age [Axe01] Age & Education [Tom04]
Memory	Logical Memory Story A delayed recall [Axe01] Rey Auditory Verbal Learning Test (RAVLT) [SSS06] RAVLT total learning: sum of trials 1-5 RAVLT short-term delayed recall:trial 6 RAVLT long-term delayed recall: trial 7 Benton Visual Retention Test Recognition [Ste05]	Education [MRSea04] Age [IMS$^+$92a] Age & Education
Language	Boston Naming Test - 30 items [Rot11] Semantic Fluency (Animals) [SSS06]	Age [FDM98] Age & Education [TKR99]
Visuo-spatial	Block Design [Axe01]	Age [IMS$^+$92b]
Executive Function	Controlled Oral Word Association Test [SSS06] Trail Making Test(TMT) B [SSS06]	Age & Education [TKR99] Age & Education [Tom04]

Feature	Wave 1	Wave 2	Wave 3	Wave 4
DSym_raw	48.45±12.16	47.86±12.97	46.39±12.62	44.09±12.56
TMTA_raw	45.70±15.37	47.08±22.49	48.36±18.16	50.71±20.72
BNT_raw	24.97±3.48	25.00±3.55	25.00±3.68	24.95±3.85
Animal_raw	15.94±4.33	15.05±4.24	14.81±4.37	14.32±4.54
TMTB_raw	118.95±53.75	125.98±73.43	128.32±62.54	146.39±87.00
FAS_raw	37.83±12.27	36.79±12.34	37.43±12.27	37.51±12.97
RVLTtotal_raw	41.26±9.19	39.68±9.89	40.21±10.31	39.43±11.16
RVLT6_raw	8.12±3.25	7.76±3.36	7.85±3.47	7.56±3.59
RVLT7_raw	7.63±3.47	7.21±3.49	7.44±3.64	7.04±3.83
LM_delay_raw	9.42±4.05	9.55±4.11	9.61±4.01	9.01±4.39
BVRT_raw	11.85±1.80	12.06±1.75	12.07±1.82	11.86±1.94
Block_raw	21.60±8.12	21.61±8.61	22.03±8.40	20.27±9.07
LM_immed_raw	11.12±4.01	10.97±3.87	11.09±3.94	10.61±4.26
RVLT1_raw	4.78±1.48	4.78±1.71	4.89±1.59	5.02±1.81
RVLT3_raw	8.79±2.36	8.47±2.41	8.48±2.48	8.38±2.68
RVLT4_raw	9.84±2.41	9.35±2.56	9.55±2.74	9.11±2.83
RVLT5_raw	10.69±2.48	10.02±2.71	10.21±2.73	9.81±2.99
RVLT_LOT_raw	17.39±7.13	15.75±6.92	15.76±7.25	14.31±7.55
RVLT_RPC_raw	89.55±9.16	87.95±10.34	88.89±10.63	87.47±11.37
RVLT_intrusions_raw	2.94±3.40	3.24±3.49	3.16±3.40	3.30±3.41
RVLT_recTP_raw	13.54±1.71	13.21±2.03	13.53±1.92	13.21±2.34
RVLT_recFP_raw	1.67±1.89	1.83±2.00	1.88±2.00	1.97±2.09
F_raw	12.74±4.45	12.72±4.51	13.03±4.34	13.10±4.64
A_raw	11.25±4.56	10.79±4.43	10.98±4.53	10.97±4.71
S_raw	13.86±4.80	13.29±4.84	13.42±4.77	13.44±5.08

Table 4.5: Statistics of Neuropsychological Measure based Features

4.2.1 Correlation between covariates

Pearson's correlation was used to analyze the correlation between input variables before proceeding to classification experiments and the results are shown in Figure 4.1. This allows one to define the scope of machine learning experiments and exclude linear methods such as linear discriminant analysis due to the varying levels of correlation between input variables [NM01, VLBMR16]. It is interesting to observe clusters of sequential variables that are highly correlated, which is more accentuated in the earlier waves.

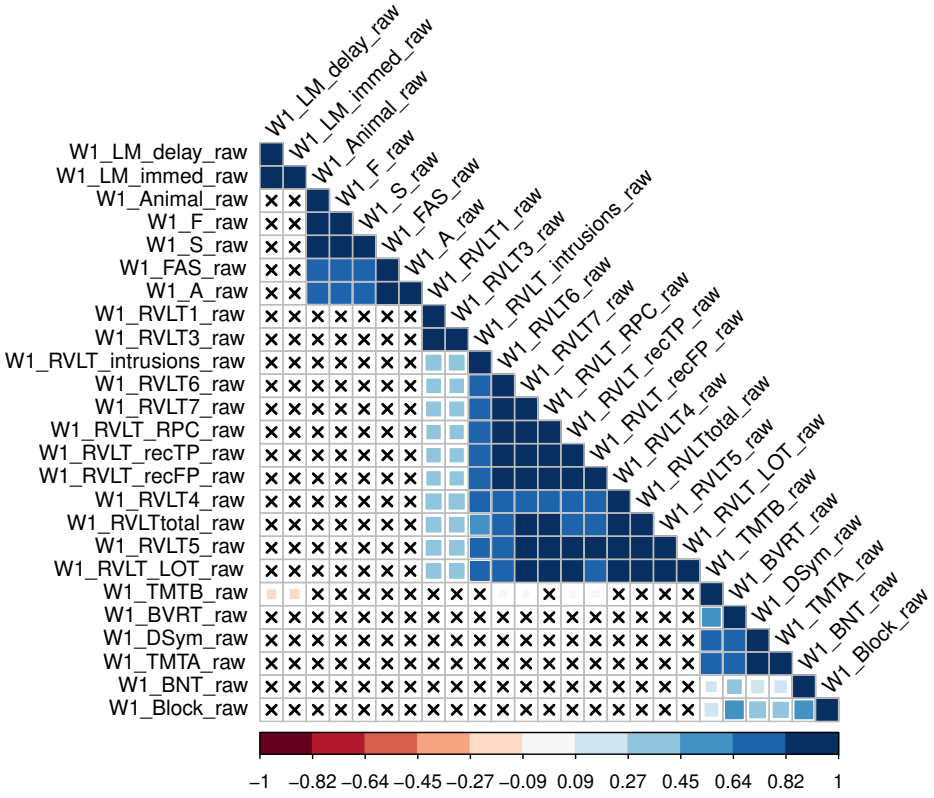
After defining the scope, a number of experiments comparing different conventional machine learning algorithms as well as deep learning methods were carried out, for the purpose of classification of MCI subtypes and CN, and the results are in Table 4.6. These machine learning algorithms were identified as the most well established algorithms in their class. In addition, both AdaBoost and ensemble methods were selected to decrease the variance and increase the robustness of the models [Bü12]. NM features were then used to train models that are capable of learning complex latent patterns from the underlying dataset.

Conventional Machine Learning	**Deep Learning**
Support Vector Machines	Stacked Auto Encoders
Random Forest	Deep Neural Networks
AdaBoost	1D Convolutional Neural Networks
Ensemble Methods	2D Convolutional Neural Networks

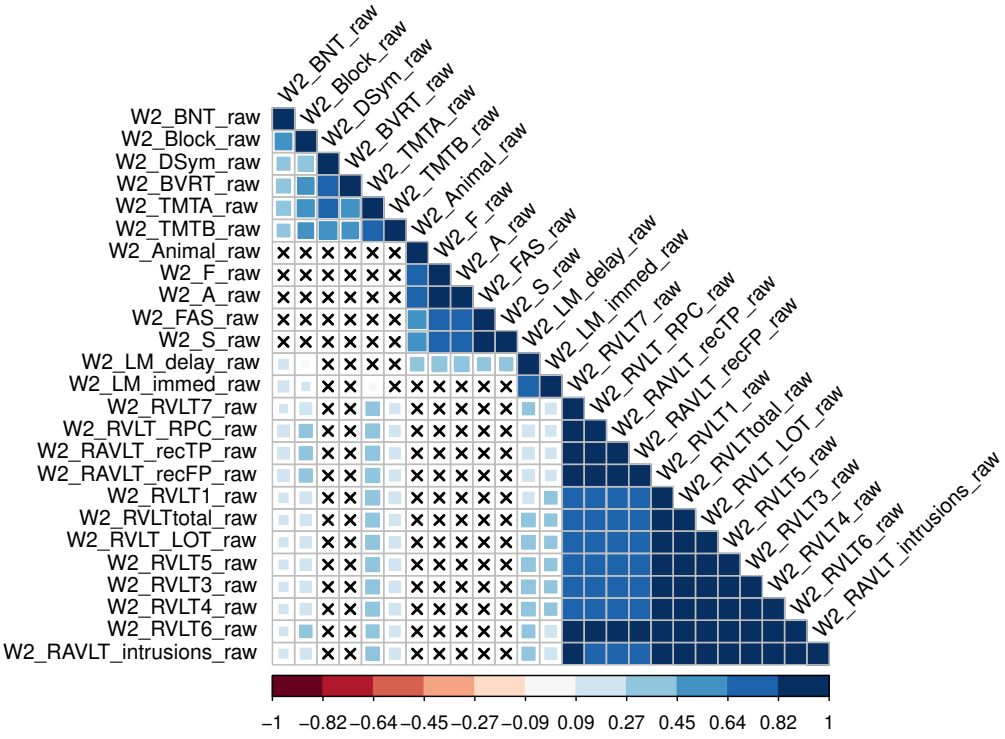
Table 4.6: Different Machine Learning Methods Compared

4.2.2 One vs One and One vs All experiments

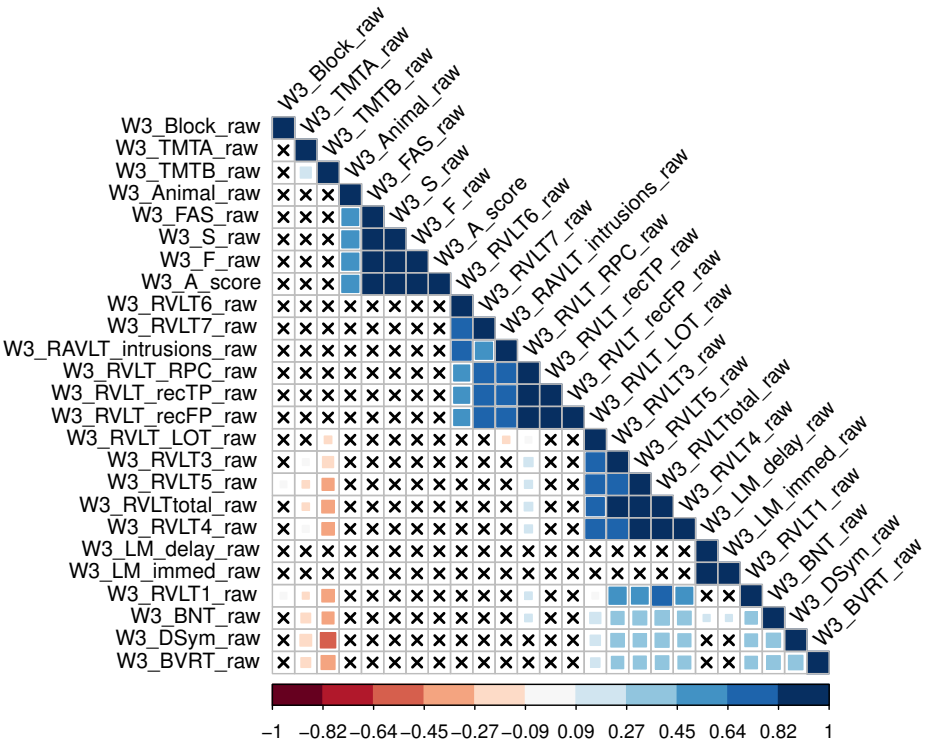
In order to evaluate the performance of the different techniques used, a number of distinct experiments were conducted. Data from four waves were used, and each wave was treated as an independent dataset, giving rise to four datasets to compare the



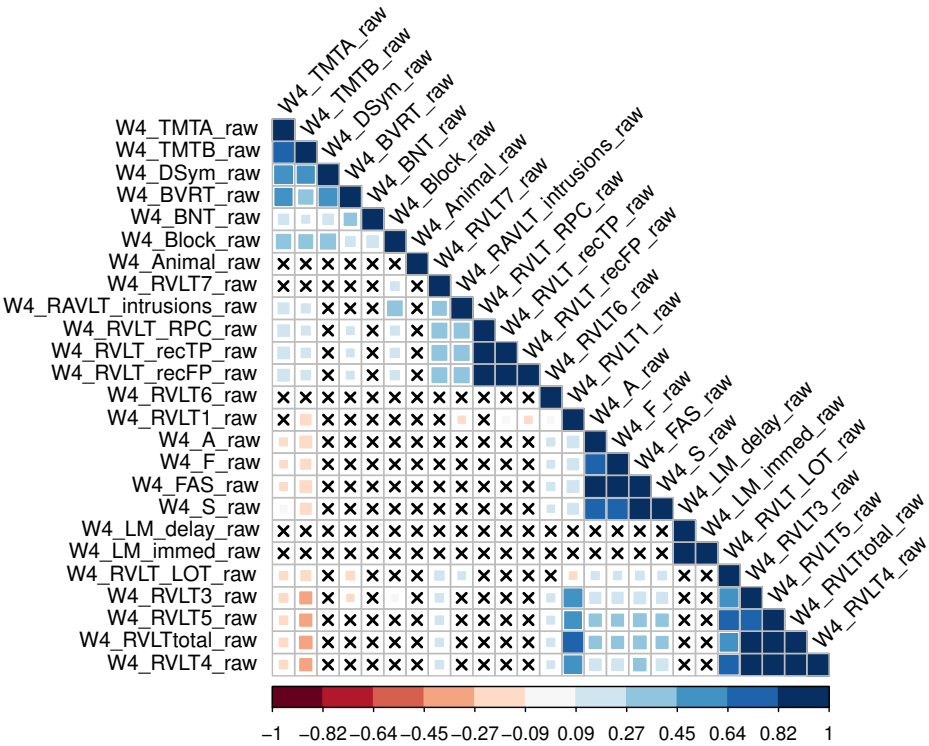
(a) Correlation of variables in Wave 1



(b) Correlation of variables in Wave 2



(c) Correlation of variables in Wave 3



(d) Correlation of variables in Wave 4

Figure 4.1: Correlation of variables used in all four waves. Insignificant correlations ($p \leq 0.05$) are indicated with 'x' marks in the graphs. The order in which features appear in the graphs corresponds to how strong the correlations are for each wave

techniques on. For each wave, two scenarios were considered for each technique in use: (i) one-vs-one experiments and (ii) one-vs-all experiments. All conventional classifiers used are binary classifiers, and in one-vs-one experiments, the goal is to discriminate between two classes, while one-vs-all experiments treat one class as positive and all other classes as negative. This creates a richer negative sample set but may also give rise to class imbalance problems. The reason for training one-vs-all classifiers was to extend them to multi-class classifiers. The classes of interest are MCI subtypes and cognitively normal individuals, and Table 4.7 describes the different models trained. For deep neural network based methods, multi-class classification experiments were conducted directly, where classification of all classes at once is performed.

One vs One	**One vs All**
MCI vs CN	aMCI — everything else
aMCI vs CN	naMCI — everything else
naMCI vs CN	sd-naMCI — everything else
aMCI vs naMCI (MCI Subtypes)	md-naMCI — everything else
sd-aMCI vs md-aMCI (aMCI Subtypes)	sd-aMCI — everything else
sd-naMCI vs md-naMCI (naMCI Subtypes)	md-aMCI — everything else

Table 4.7: The different classes used for experimentation

4.2.3 Feature Subset Selection

Feature subset selection was carried out in an effort to improve the performance of the conventional machine learning techniques, and the experimental pipeline is shown in Figure 4.2. In order to evaluate the effectiveness of feature subset selection, random forests was picked as the main technique, as it showed relatively stable results across the initial experiments. Three major categories of feature selection algorithms were used: correlation based feature selection, information gain based feature selection and wrapper based feature selection. Specific algorithms used are (i) Correlation based subset evaluation (ii) Pearson correlation based (iii) Cross validation based (iv) Gain

ratio based (v) Information gain based (vi) SVM wrapper based (vii) Random forest wrapper based and (viii) RELIEFF. The results are presented in Section 4.3.

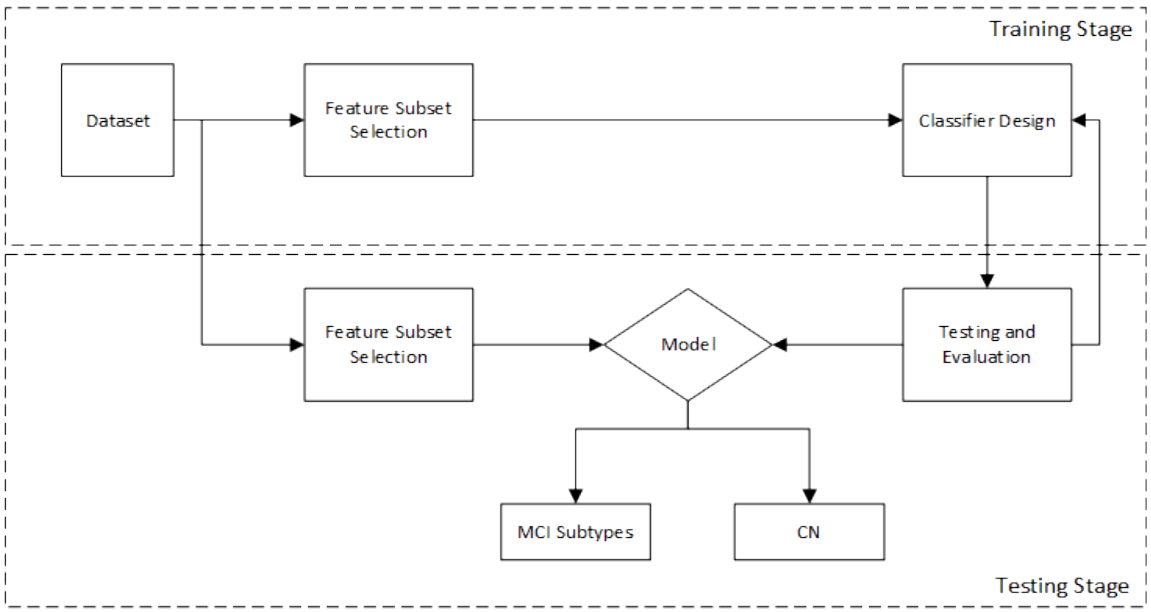


Figure 4.2: Experimental Framework for Conventional Methods

Feature selection for deep learning methods was not considered as a separate step, because deep learning methods are themselves capable of feature selection by taking into account the complex latent patterns that the dataset exhibits. In order to improve the performance of the SAE classifier, an ensemble of SAE classifiers at the model level was designed. The same training/test datasets were used to train multiple SAE classifiers with different hyper-parameters and these classifiers were then used in conjunction with a voting scheme, to come up with the final class label.

4.2.4 Deep Learning Methods

The hyper-parameters for convolutional neural networks were designed heuristically and fine-tuned after experimentation; they are included in Table 4.8. In order to utilise 2D convolutional neural networks, the dataset was put through a feature

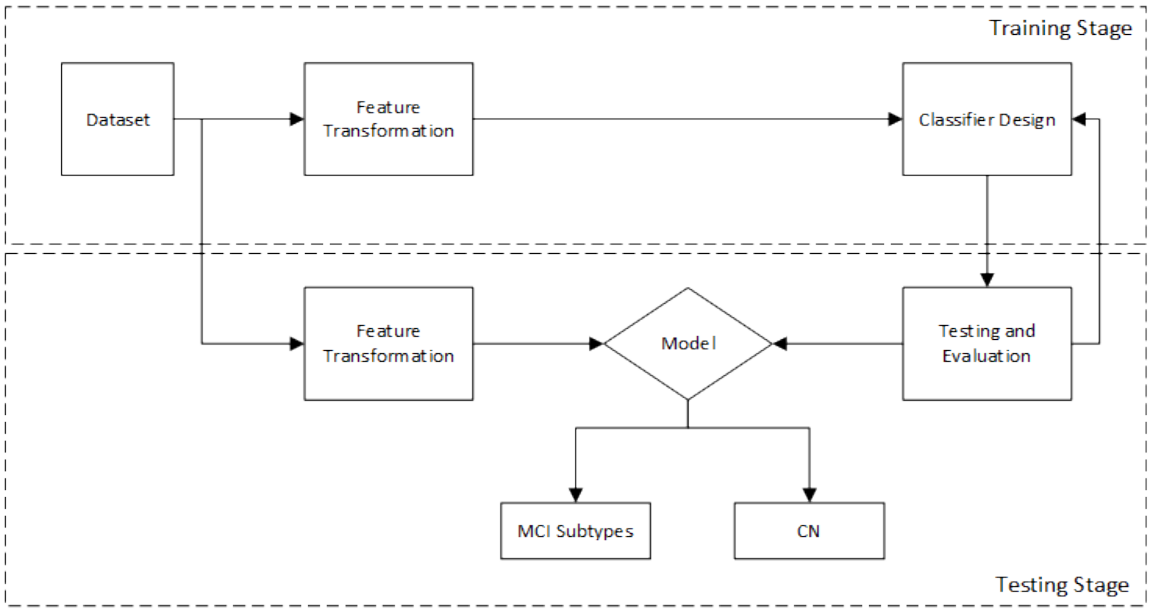


Figure 4.3: Experimental Framework for Deep Learning Methods

transformation stage as shown in Figure 4.3. In this step, the uni dimensional feature vector for each subject was rearranged into a two-dimensional feature matrix that was then used as input to the 2D CNN.

Layers	1D CNN	2D CNN
Convolution	1x3 convolution kernel, 8 activations	3x3 convolution kernel, 8 activations
BatchNormalization and ReLu	No Parameters	No Parameters
MaxPooling	1x2 max pooling	2x2 max pooling
Convolution	1x3 convolution kernel, 16 activations	3x3 convolution kernel, 16 activations
BatchNormalization and ReLu	No Parameters	No Parameters
Dropout	0.5 dropout	0.5 dropout
FullyConnected and Classification	No Parameters	No Parameters

Table 4.8: The Hyper parameters for Convolutional Neural Networks

An ensemble of SAEs at the model level was considered to improve performance. The proposed ensemble is a model level ensemble rather than a data level

ensemble, as different models with different hyper-parameters are trained on the same training set and tested on the same test set. The results of individual SAE classifiers are then taken into consideration and the majority vote is used to predict the class label. This strategy enables the use of different versions of auto-encoders including conventional auto-encoders and sparse auto-encoders together.

Instead of carrying out one-vs-all experiments with deep learning, a multi-class classification approach was taken to distinguish between all six classes. The original rationale behind training one-vs-all classifiers using conventional techniques was in fact to convert them to multi-class classifiers.

4.2.5 Validation of Results

The experiments used the same dataset consistently. Ten-fold cross validation was performed for all experiments based on conventional machine learning techniques. The same randomly picked folds were used to compare the techniques. Ten repetitions of ten-fold cross validation were carried out to reduce the variability in results. The deep learning techniques were five-fold cross validated instead of ten-fold, allowing for the higher time complexity of the models. The order of the features was shuffled for each fold as well.

4.3 Results

The results of the experiments are discussed in two main subsections. The results of the conventional machine learning methods are presented first, followed by those of deep learning methods. Both one-vs-one and one-vs-all classification scenarios are considered.

4.3.1 Conventional Machine Learning Methods

While a range of conventional machine learning methods were implemented, the results are presented for the three best classifiers, namely AdaBoost (AB), Random Forests (RF) and ensemble methods (ES). Results are presented for one-vs-one experiments in Figure 4.4 and for one-vs-all experiments in Figure 4.5 across the first wave for comparison purposes. It can be seen that AdaBoost consistently performs better than the other methods in both one-vs-one and one-vs-all experiments.

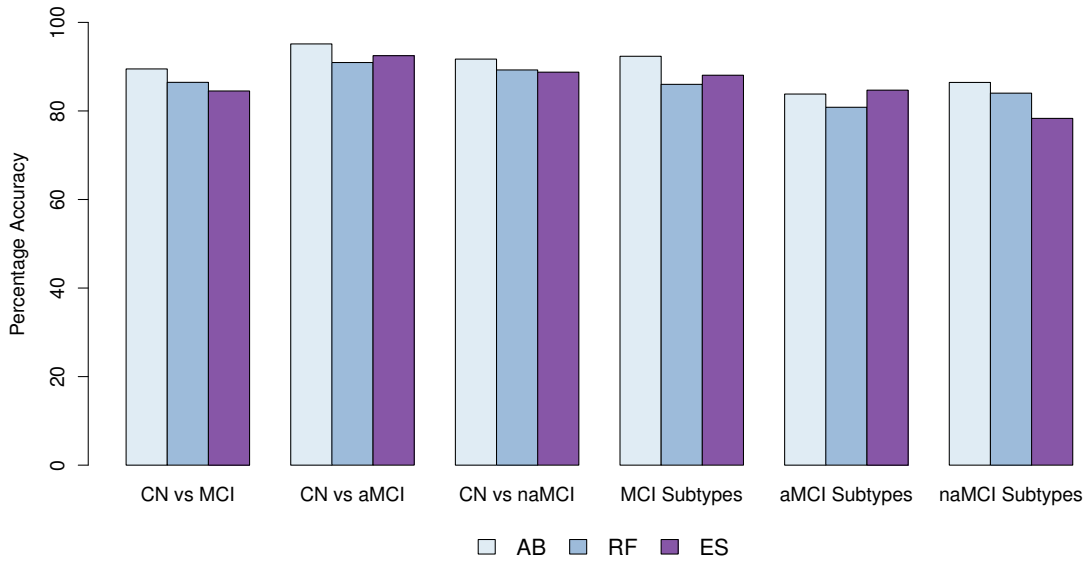


Figure 4.4: Percentage accuracy for the three best conventional machine learning methods considered in one-vs-one classification on first wave. The AUC is above 0.8 for all, except for AB in naMCI subtypes which dips to 0.74

Since data from four different time points were available (four waves, each two years apart), they are treated as four independent datasets to train classifiers on each set. Results for the best classifier (AdaBoost) are presented across all four waves for one-vs-one classification in Figure 4.6, while one-vs-all is presented in Figure 4.7. As the waves progress, the dataset gets smaller and this appears to have an adverse effect on the classification accuracy, as can be seen in Figures 4.6 and 4.7. This is

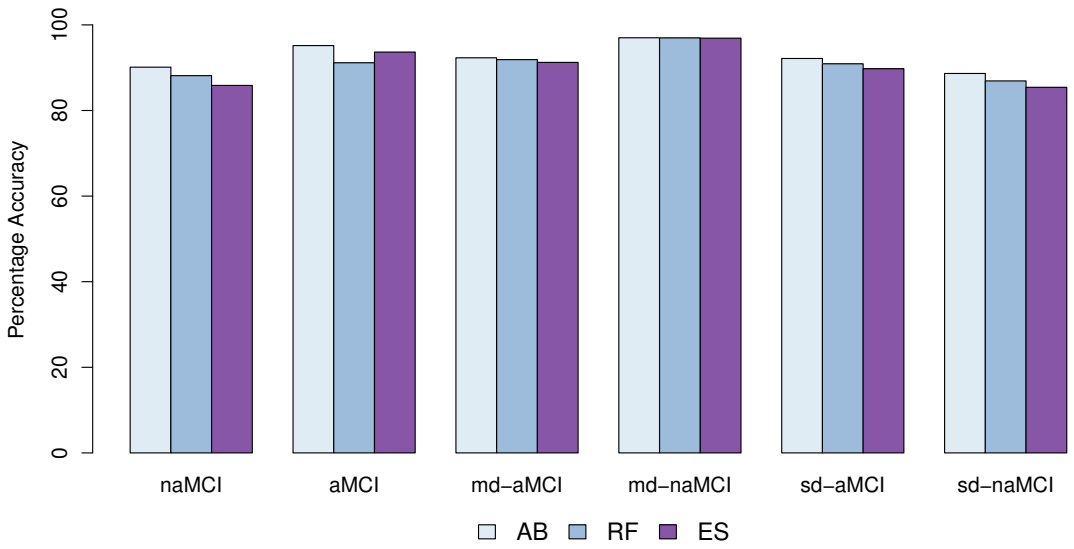


Figure 4.5: Percentage accuracy for the three best conventional machine learning methods considered in one-vs-all classification on first wave

further exacerbated when a differentiation into aMCI subtypes is attempted, as the data subset becomes increasingly small. The effects of this can be clearly seen in Figure 4.6. However, the discrimination also shows the efficacy of one-vs-all classification, as differentiation between sd-aMCI and md-aMCI in a one-vs-all scenario has improved.

Feature subset selection was attempted to improve the performance of the classifiers. The difference in accuracy across the methods is depicted in Figure 4.8. Only one model was significantly improved by feature subset selection, namely the classification of MCI subtypes where the accuracy improved to 91.27% from 86.01%. For large datasets, the best improvement in performance was obtained with random forest wrapper based feature selection. Performance was worse for CN vs naMCI and naMCI subtypes. The reason can be deduced by considering the sample sizes as listed in Table 4.3. For the naMCI subtype classifier, there are 122 instances for md-naMCI and 26 instances for sd-naMCI. With a wrapper based feature selection

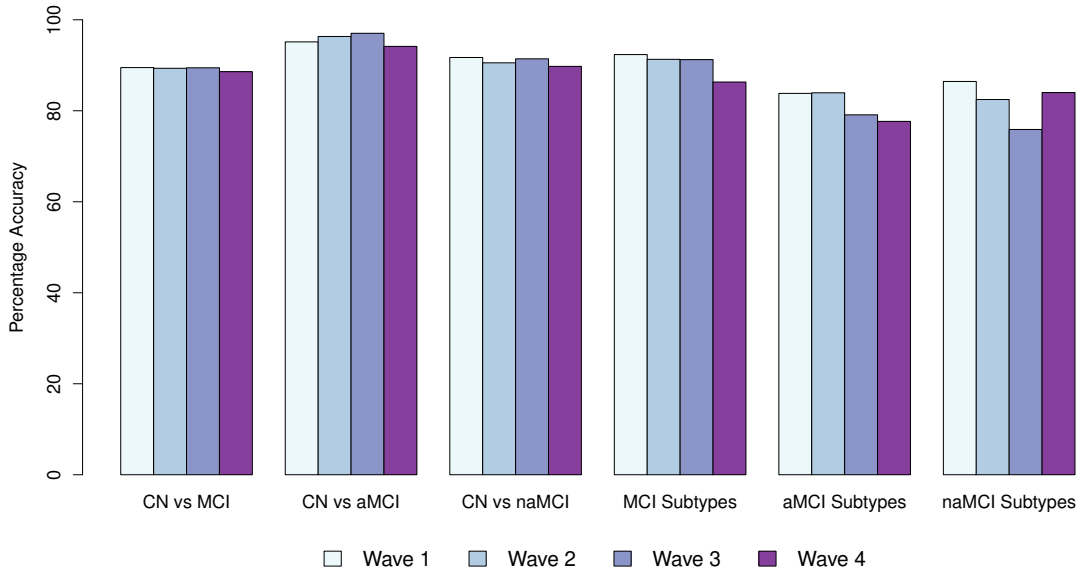


Figure 4.6: Percentage accuracy for the best conventional machine learning method (AB) in one-vs-one classification across four waves

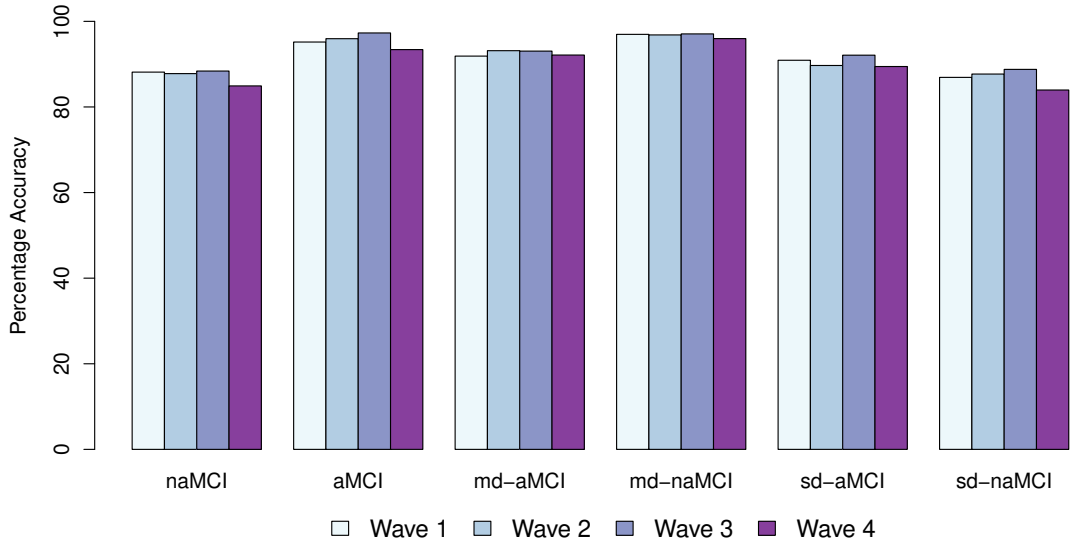


Figure 4.7: Percentage accuracy for the best conventional machine learning method (AB) in one-vs-all classification across four waves

method, the training set becomes even smaller, which explains the relative decrease in AUC of around 40% from 0.82 to 0.5. In addition, methods that improve the accuracy of at least three classifiers out of the six being tested are highlighted in Figure 4.9.

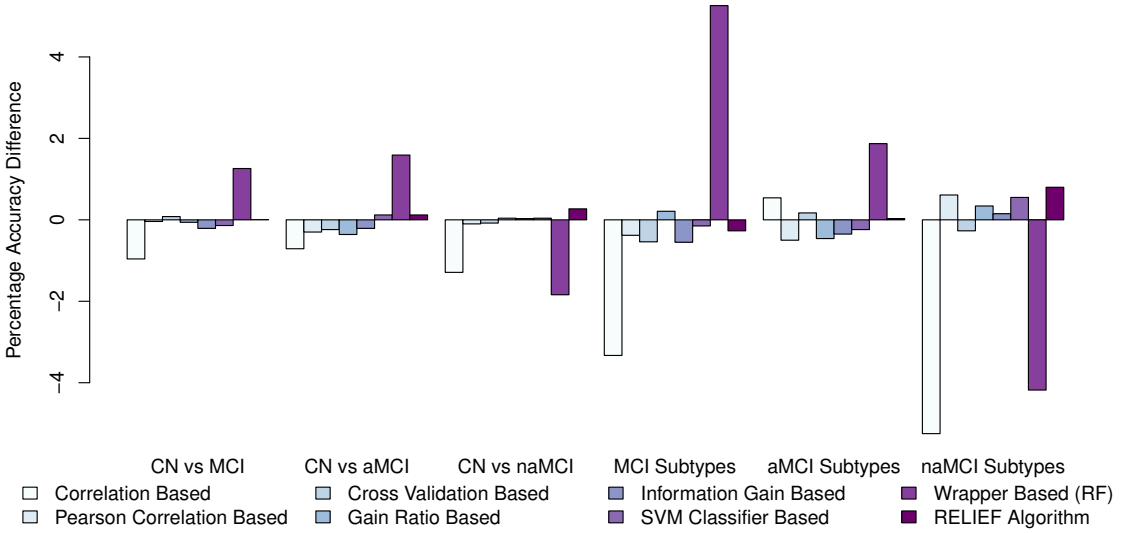


Figure 4.8: Percentage accuracy difference after feature subset selection by different methods, in comparison to original accuracy of Random Forests model for Wave 1

4.3.2 Deep Learning Methods

Experimental results for three different deep learning techniques are presented and compared against the conventional techniques in Figure 4.10 across the first wave. NM based features are transformed into a 2-dimensional input space in order to use 2D convolutional neural networks, while they are used directly as a 1D vector with 1D convolution kernel in the case of 1D CNN. Both these methods performed equally well and exceeded the performance of SAE. The results of both 1D and 2D CNNs across the four waves are presented in Figure 4.11. The results are intriguing because CNNs usually work better when there are spatial correlations between the

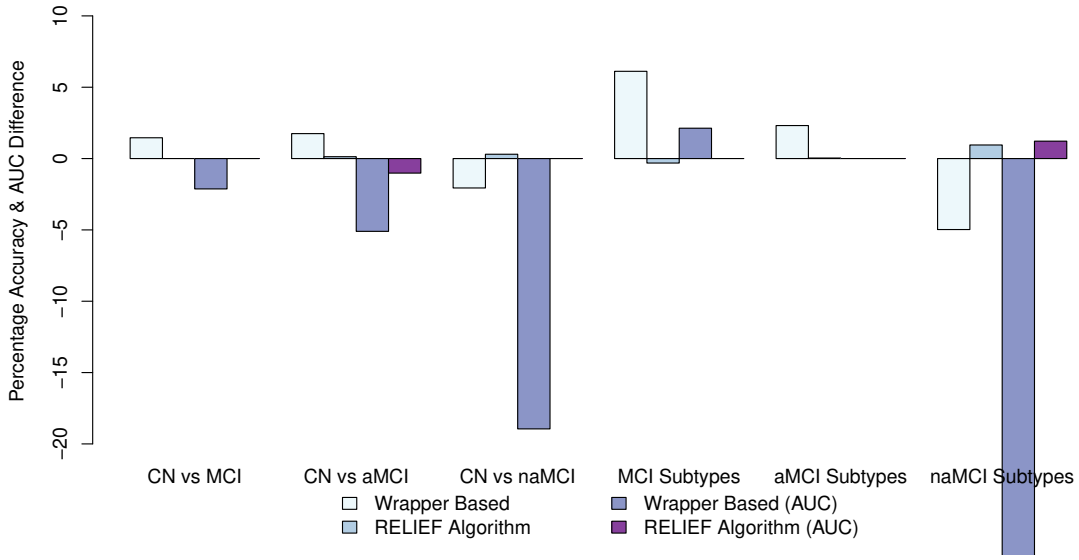


Figure 4.9: Percentage accuracy and area-under-curve (AUC) difference after feature subset selection by the two best methods, in comparison to original accuracy and AUC of Random Forests model for Wave 1. The wrapper based AUC for naMCI subtypes reaches -39.0243 and is truncated for clarity

input features, which was not the case here [RW17].

An ensemble of SAEs at the model level was built to improve performance, and results on wave 3 as an example are presented in Figure 4.12. Results of the other waves are included in Appendix A. While the accuracies almost always improved, the area under the ROC curve significantly benefited from the ensemble of classifiers. This in turn means that the classifiers trained are more generalizable and robust to noise. An optimal configuration of the ensemble was found using grid search.

Multi-class classification is inherently compatible with deep learning, and the results of the best approaches are shown in Figure 4.13.

From the array of experiments conducted, the best three techniques identified are 1D CNN, 2D CNN and AdaBoost. As can be seen from Table 4.9, all three techniques perform very well, with different techniques excelling in different

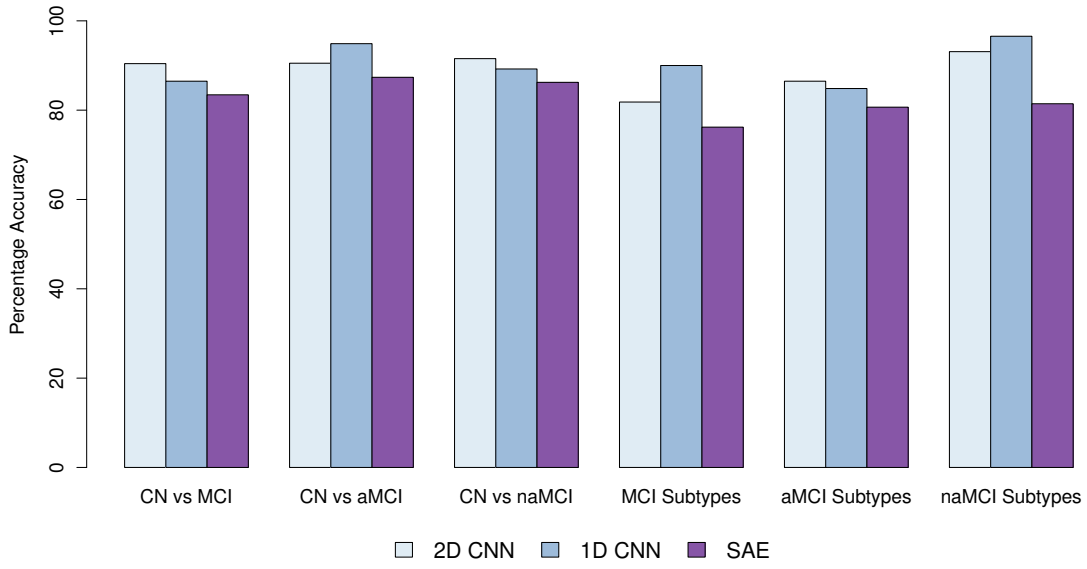
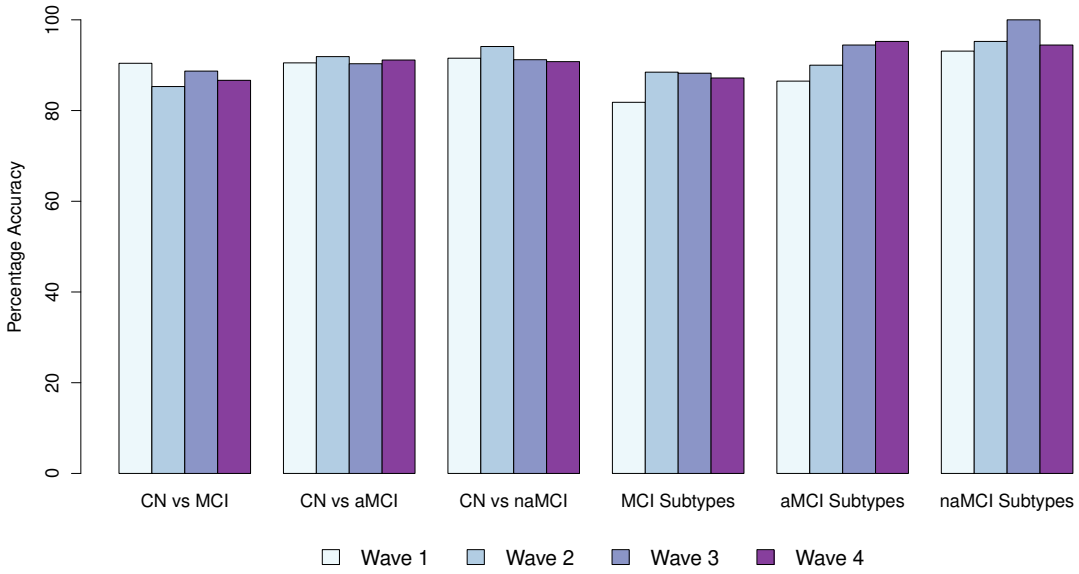


Figure 4.10: Percentage Accuracy for three deep learning models on first wave

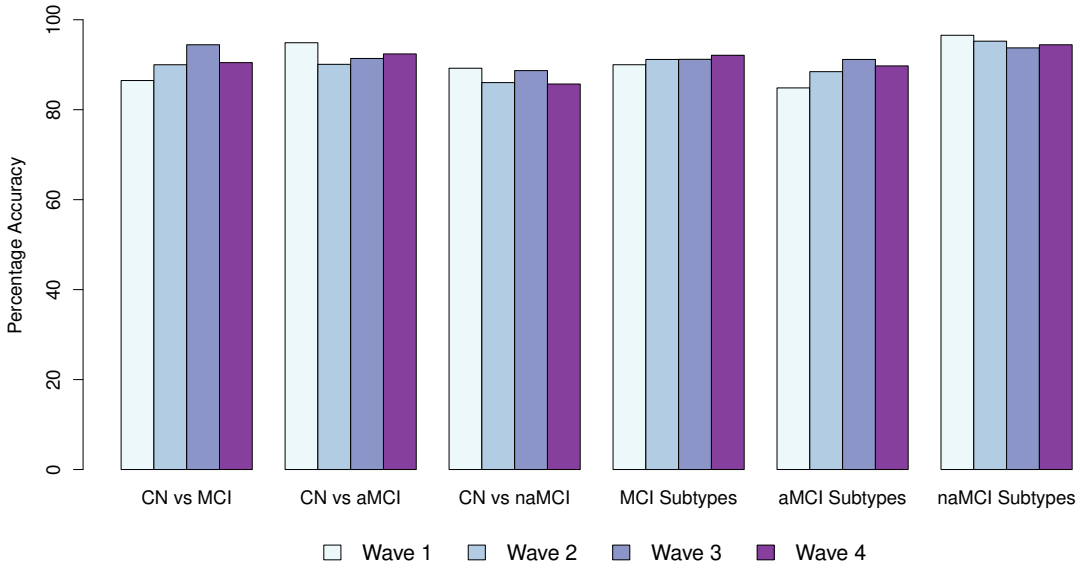
circumstances. Deep learning techniques appear to have an edge in most cases, while conventional techniques dominate in CN vs aMCI and MCI Subtypes classification, albeit not significantly.

4.4 Summary

In this chapter, the diagnostic value of NM features was studied using both deep learning techniques and conventional machine learning techniques. As far as is known, this is the first study that uses both 1D and 2D CNNs to distinguish between MCI subtypes using NM features. This is particularly interesting as CNNs are usually used with images where there is an obvious spatial correlation [RW17]. In this case, neuropsychological measures represented as a 2D matrix do not have any obvious spatial correlation. Similarly, 1D CNNs are usually used for sequence prediction problems in the literature [LZZW17]. Although NM features are used as a vector input to 1D CNNs, there is usually no relationship between consecutive elements in

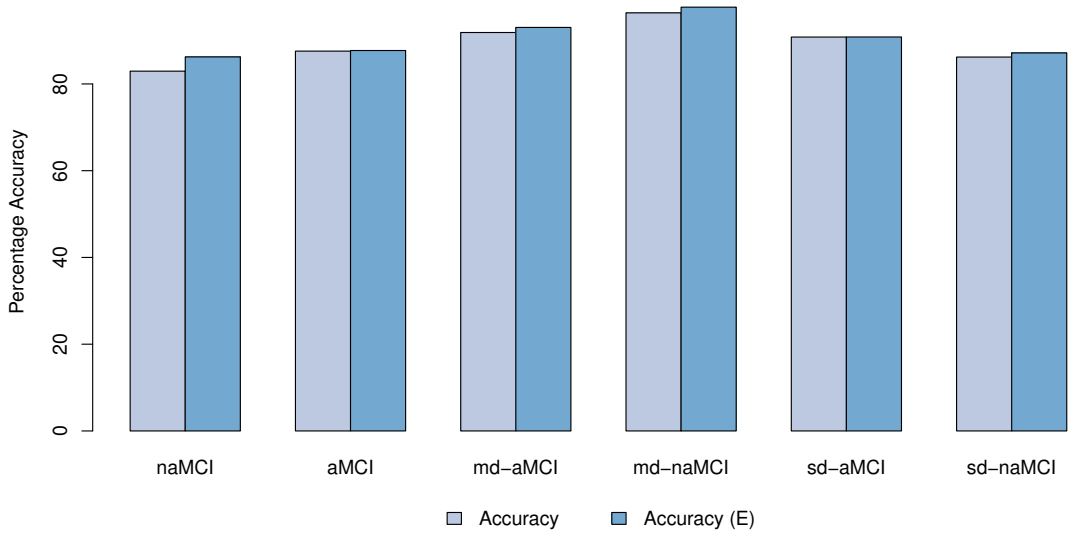


(a) Percentage Accuracy of 2D CNN across four waves

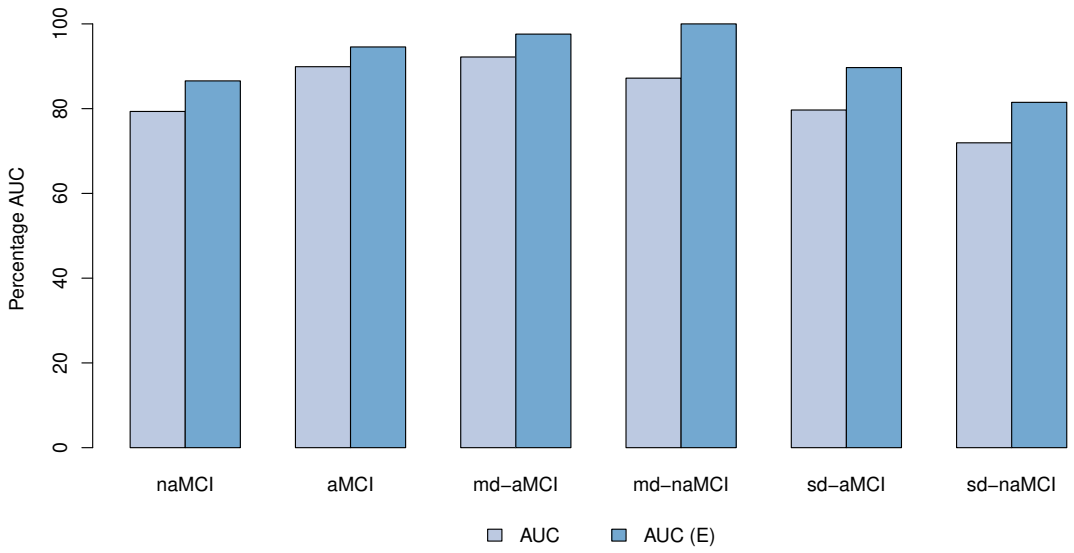


(b) Percentage Accuracy of 1D CNN across four waves

Figure 4.11: Comparison between 1D CNN and 2D CNN



(a) Percentage Accuracy of SAE classifier and SAE Ensemble classifier



(b) Percentage AUC of SAE classifier and SAE Ensemble classifier

Figure 4.12: Comparison of best SAE classifier results against SAE Ensemble classifier results for Wave 3. Results for the other waves are included in the Appendix A. Accuracy (E) and AUC (E) stands for the accuracy and AUC of the SAE ensemble classifier

		CN vs MCI	CN vs aMCI	CN vs naMCI	MCI Subtypes	aMCI Subtypes	naMCI Subtypes
2D CNN	Wave 1	**90.4192**	90.5109	**91.5385**	81.8182	**86.4865**	93.1034
	Wave 2	85.2941	91.8919	**94.1176**	88.4615	**90**	**95.2381**
	Wave 3	88.6957	90.3226	**91.2088**	88.2353	**94.4444**	**98**
	Wave 4	86.6667	91.1392	**90.7895**	87.1795	**95.2381**	**94.4444**
1D CNN	Wave 1	86.4865	94.8905	89.2216	90	84.8485	**96.5517**
	Wave 2	**90**	90.0901	86.0294	**91.1765**	88.4615	**95.2381**
	Wave 3	**94.4444**	91.3978	88.6957	**91.2088**	91.1765	93.75
	Wave 4	**90.4762**	92.4051	85.7143	**92.1053**	89.7436	**94.4444**
AdaBoost	Wave 1	89.49	**95.14**	**91.71**	**92.35**	83.82	86.44
	Wave 2	89.34	**96.33**	90.54	**91.31**	83.95	82.47
	Wave 3	89.45	**97.03**	**91.42**	**91.24**	79.11	75.89
	Wave 4	88.59	**94.15**	89.76	86.32	77.67	84

Table 4.9: Comparison of Results for 1D CNN, 2D CNN and AdaBoost across multiple waves. The best accuracy for each wave for each class is indicated in bold text

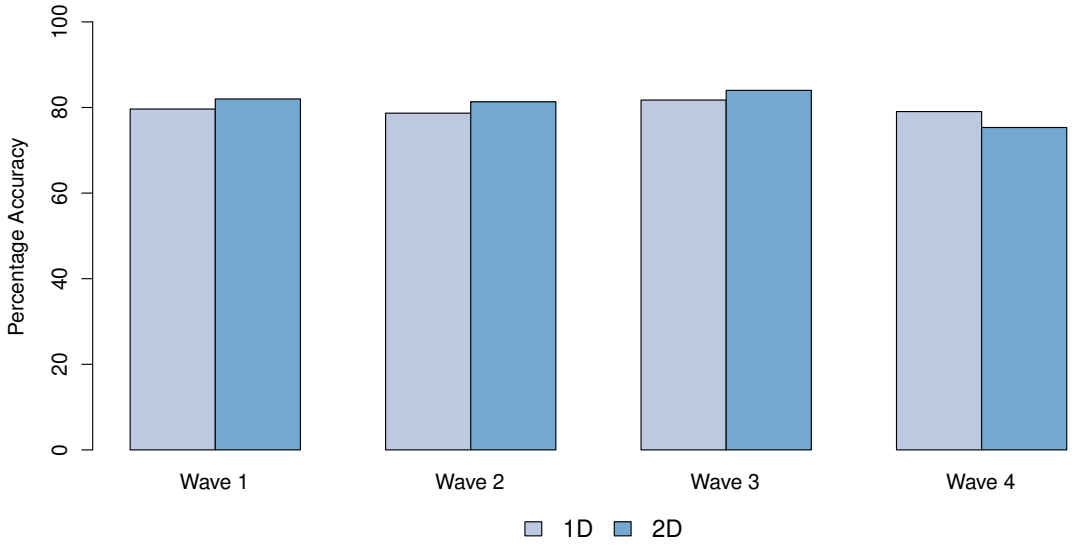


Figure 4.13: Multi-class classification using 1D CNN and 2D CNN

the NM feature vector. However, the CNNs have outperformed simple deep networks such as SAE in all cases and traditional techniques in most cases as well. An intuitive explanation for this phenomenon could be that CNNs are capable of capturing local correlations even if they are not spatially correlated, as demonstrated in the experiments. This study confirms that it is worthwhile to invest the time in exploring CNN based solutions, even when the dataset in question is not correlated spatially.

As the results presented in Table 4.9 show, conventional techniques perform well in most cases. However, an additional reason to train a deep learning network is to subsequently leverage its power and adaptability so that features from different modalities may be fused seamlessly, as will be demonstrated in Chapter 5. Typically in an MCI diagnosis scenario, either NM features or MR images (or both) are acquired for clinical diagnosis. Therefore, the dataset usually contains both NM features and multi-modal MR images, which facilitates the design of a classification system enriched with multi-modal input data. In fact, as deep learning models are capable of extracting features without intervention, this would also make the training process

end-to-end.

Chapter 5

Feature Fusion using Convolutional Neural Networks*

While Chapter 4 focused on differentiating mild cognitive impairment and dementia from cognitively normal behaviour using neuropsychological measure (NM) features, this chapter deals with the same problem using medical imaging based features. This is then further extended to treat both neuropsychological features and imaging based features seamlessly for the same task. The use of medical imaging based features is particularly important in this context because conducting a battery of neuropsychological tests can take a considerable amount of time and can also be prohibitively expensive. Medical imaging can be considered as an alternative, as it can be both relatively less expensive and also take less time. Therefore, it is worth

Portions of this chapter appear in:

1. U. Senanayake, A. Sowmya and L. Dawes, "Deep fusion pipeline for mild cognitive impairment diagnosis," 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), Washington, DC, 2018, pp. 1394-1997. doi: 10.1109/ISBI.2018.8363832

exploring the diagnostic value of medical imaging either on its own or along with neuropsychological tests.

A novel deep learning based pipeline is designed and employed in this study to combine information from multiple modalities of data seamlessly. The two modalities initially considered are 3D structural magnetic resonance images (MRI) and uni-dimensional NM features. Until recently, a mix of handcrafted features from MR images were used to train popular classifiers such as support vector machines (SVM) [HMH+13,CWL+12]. In contrast, this chapter presents a complete deep learning pipeline that encompasses feature generation, extraction, fusion and classification that can significantly simplify the classification process and reduce the expertise needed for interpretation. The architecture was designed and tested to address three issues: (i) addressing paucity of data, (ii) leveraging the characteristics unique to medical images and (iii) creating a pipeline that can meaningfully fuse data from different modalities. The basis of the proposed model is convolutional neural networks (CNN), which are a category of artificial neural networks that became prominent with the advent of deep learning techniques [KSH12b]. They are typically used for image classification problems and became well-known for solving a class of computer vision problems that computer scientists have struggled to solve in the past [SLJ+14].

The underlying concepts of ResNet, DenseNet and GoogLeNet architectures, already discussed in Chapter 3, inspired the proposed model. The model classifies MRI scans of subjects diagnosed as AD, MCI and CN. A number of previous studies using deep learning for AD classification are based on two main deep learning techniques: generative models, specifically, auto-encoder (AE) based models, and discriminative models, specifically, convolutional neural network (CNN) based models. The efforts by Suk et. al [SS13] and Li et. al [LTT+14] are a good starting point for AE based techniques. They use MRI and Positron Emission Tomography (PET) images in their classifier. The work presented in Chapter 4 is also directly

relevant as it uses deep learning in discriminating MCI subtypes using NM. Korolev et. al [KSBD17] report their approach on using a discriminative 3D CNN, which is used as a baseline in this work. A more comprehensive review can be found elsewhere [VPM17].

While many approaches have been proposed for AD classification, the main difference between the proposed approach and the closest [KSBD17] is the use of a fusion pipeline where information from multiple modalities are fused seamlessly through a single deep learning pipeline. The main components of the proposed architecture is a 3D CNN and a deep neural network that are fused and co-optimized together, which allows the use of 3D structural MR images and uni-dimensional feature vectors such as NM based features in a unified framework. As far as can be ascertained, this is the first effort that treats 3D structural MR images and NM based features in a single end-to-end deep learning pipeline. The proposed approach is also distinctive in that it incorporates a multi-class classification approach. Other than the baseline [KSBD17], classifiers built using single data modalities are also used for comparison.

The rest of the chapter is organized as follows. In section 5.1, the characteristics of the datasets used in this work, including MAS and **A**lzheimer's **D**isease **N**euroimaging **I**nitiative (ADNI) datasets, are described. The proposed methods are described in section 5.2. Results are presented in section 5.3, and the final section presents a summary of the work.

5.1 Dataset

Two datasets are used in the experiments of this work. The MRI based subset of data in the MAS dataset is described in section 5.1.1 while section 5.1.2 describes

the ADNI dataset including MRI data and NM data.

5.1.1 MRI Data of MAS Dataset

The MAS dataset was described in section 2.3.1, including a description of the NM features. The primary objective of the MAS study is to examine the clinical characteristics and prevalence of MCI and related syndromes in non-demented older Australians, and to determine the rate of change in cognitive function over time [SBR+10]. In this subsection, MRI based data acquired by the MAS study is described.

Participants at the baseline were invited to undergo an MRI scan and were scanned using a Philips 3T Achieva Quasar Dual scanner located at the Prince of Wales Medical Research Institute, Sydney [SBR+10]. The standard protocol for acquisition of 3D T1-weighted structural MRI images is described elsewhere [SBR+10]. The slice thickness was 1mm with no gap, yielding $1x1x1mm^2$ isotropic voxels. Deidentified images were made available for this work. These were then further post-processed for the experiments, to obtain sixty 2D image slices that are representative of each volume and the same sixty slices were used to create the 3D MRI volume as well. A subset of these sixty MRI image slices are depicted in Figure 5.1.

5.1.2 ADNI Dataset

The ADNI dataset was briefly introduced in section 2.3.2. The characteristics of the dataset including the data modalities and demographic information are discussed. This subsection is divided into two, focusing on NM based features and MRI based features acquired under the ADNI study respectively.

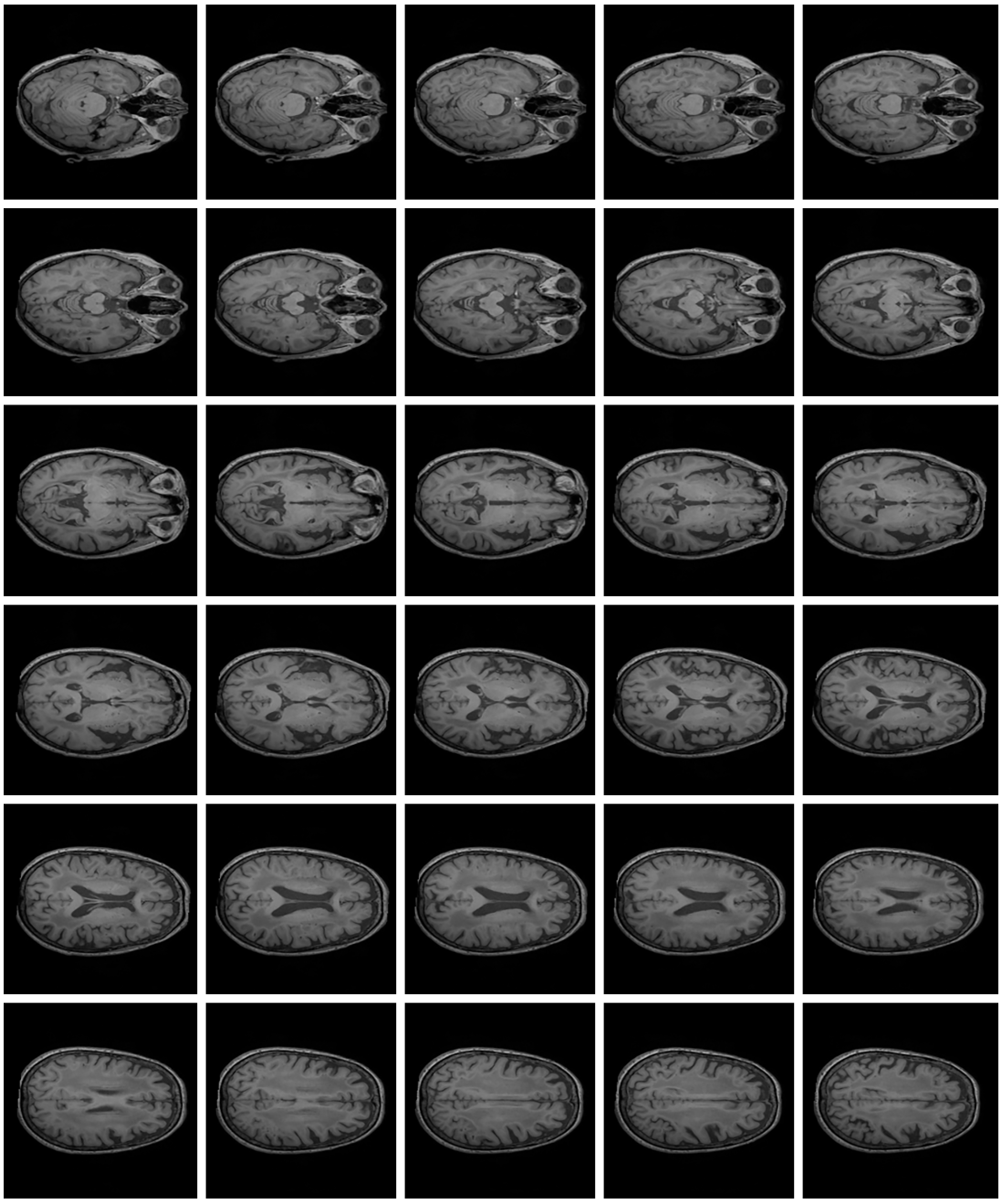


Figure 5.1: Slices of MRI volume of patient 0033A, diagnosed with multi-domain amnestic mild cognitive impairment. MRI volume considered for the experiments consists of 60 slices, out of which every other slice is shown in this figure

5.1.2.1 ADNI Neuropsychological Measures based features

There are a number of overlapping neuropsychological features in the ADNI and MAS datasets. However, all the features used in this work are presented here for completeness. The demographics of the cohort at the baseline is presented in Table 5.1.

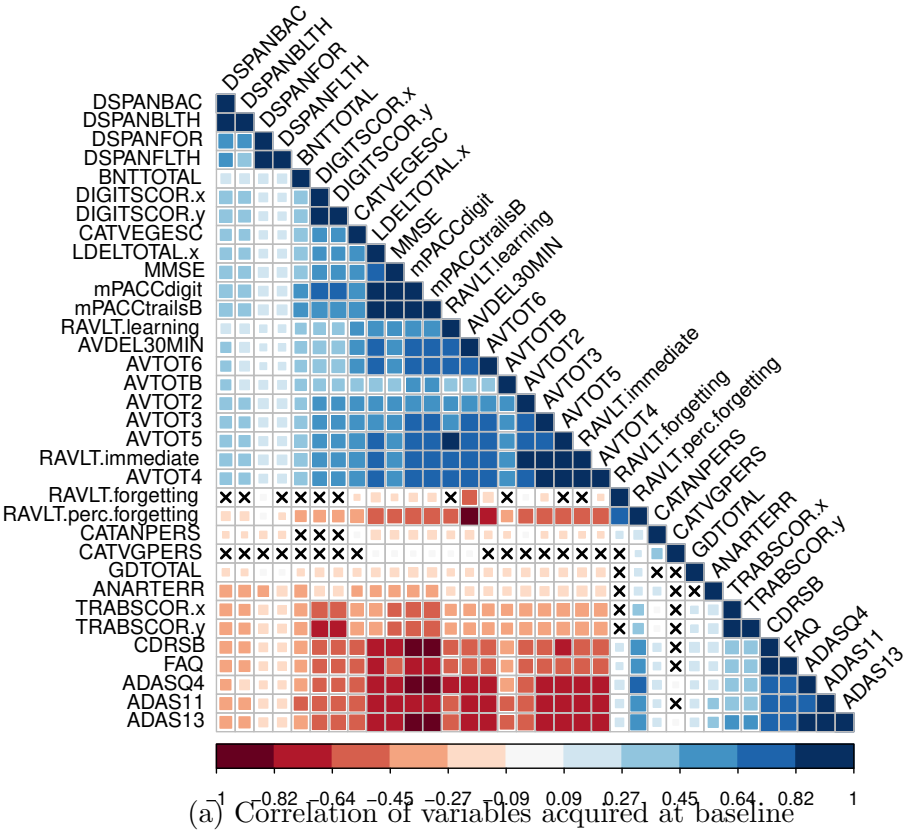
Sample size: 819	Baseline
Age (years)	75.185 ± 6.638 (54.4 - 90.9)
Sex (male/female)	59.90 / 40.1
Education (years)	15.534 ± 3.048
MMSE (Mini-Mental State Exam)	26.743 ± 2.673
CDR (Clinical Dementia Rating)	1.796 ± 1.837

Table 5.1: Demographic characteristics of participants at baseline [WHC+13]

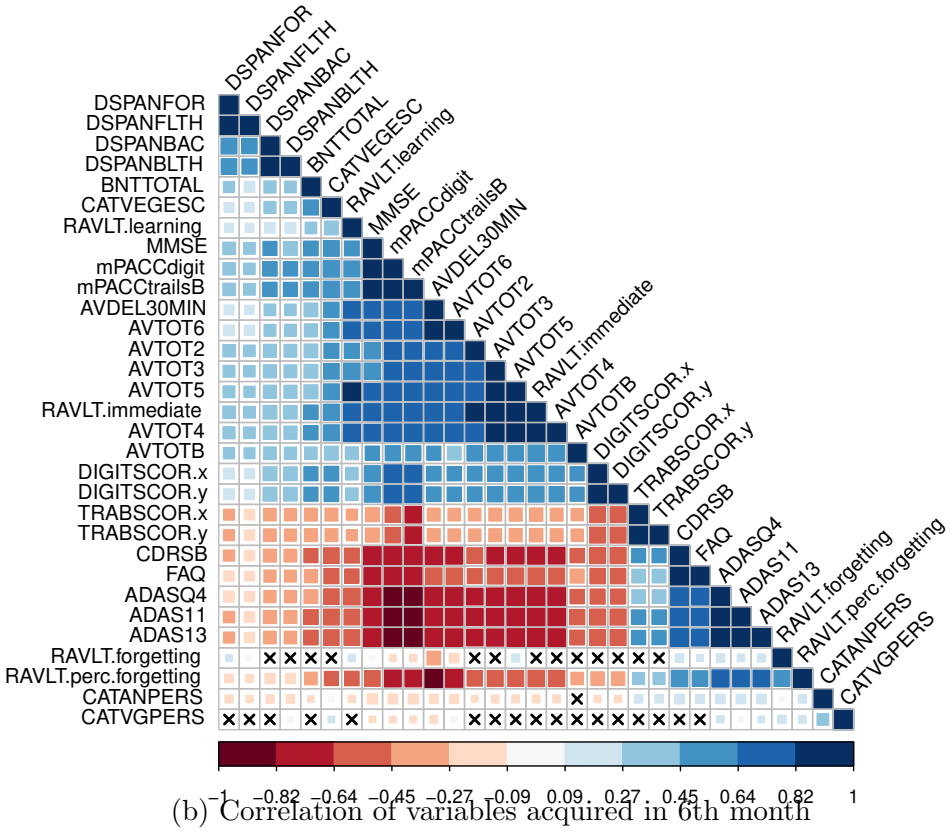
The baseline diagnosis for the patients included 188 patients with AD, 402 patients with late MCI and 229 CN individuals. The NM used are listed in Table 5.2. The pairwise correlations between variables are presented in Figure 5.2. The correlations were calculated at baseline, 6 months, 12 months and 24 months from baseline. It is evident that the correlations between variables are higher at baseline and gradually evolve and decrease over time in subsequent acquisitions. This makes it an ideal dataset to understand the longitudinal effects of the onset of Alzheimer's disease and its progression with time. While data from four time points are presented, only data from two time points, namely baseline and 12 months, are considered in this work as overlapping MRI acquisition along with the NM tests were a requirement for the experiments.

5.1.2.2 ADNI MRI image based features

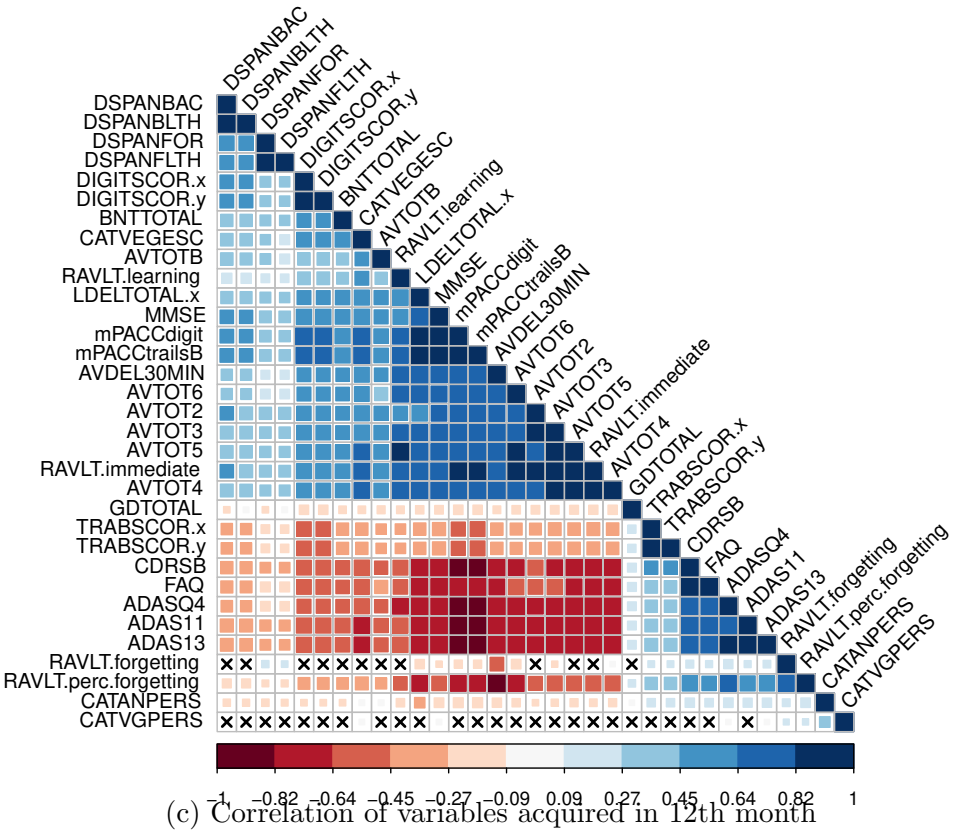
A subset of ADNI data from ADNI I phase (baseline and 1 year scans) that has already been preprocessed [WHC+13] was used in this work. The resulting dataset



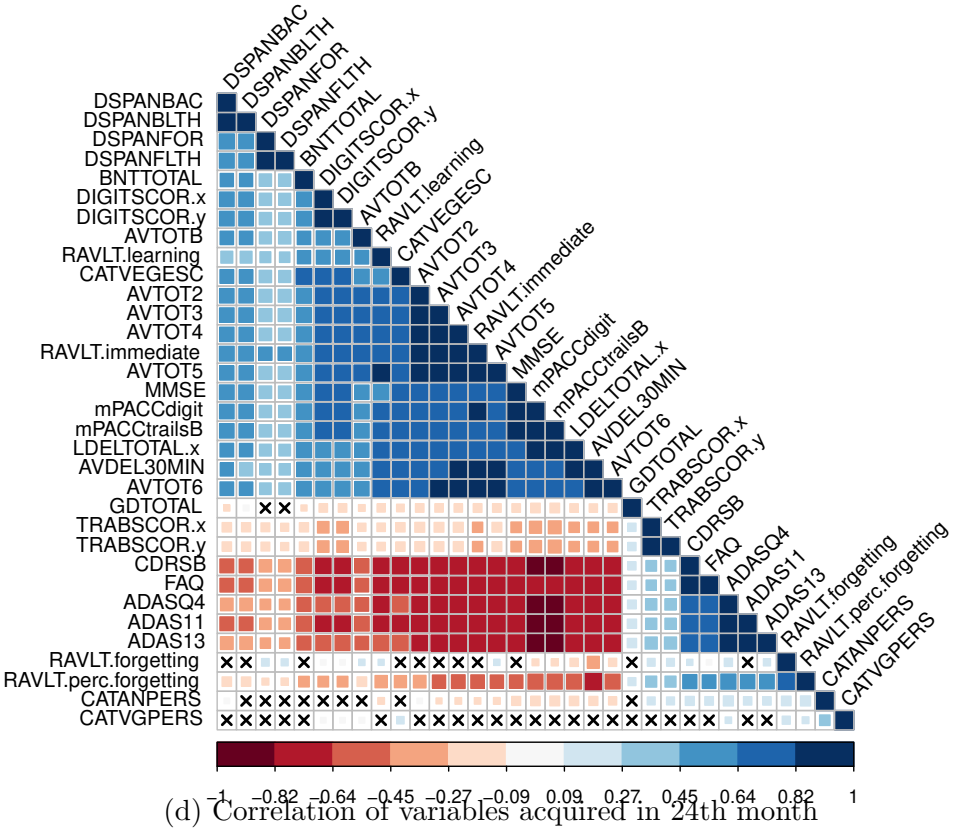
(a) Correlation of variables acquired at baseline



(b) Correlation of variables acquired in 6th month



(c) Correlation of variables acquired in 12th month



(d) Correlation of variables acquired in 24th month

Figure 5.2: Correlation of variables used in four measurement windows. Insignificant correlations ($p \leq 0.05$) are indicated with 'x' marks in the graphs. The order in which features appear in the graphs corresponds to how strong the correlations are for each measurement window

Tests	Cognitive domain
American National Adult Reading Test	Premorbid Intelligence
Mini Mental State Examination	Multiple
Logical Memory I and II	Memory
Digit Span	Attention / Processing speed
Category Fluency	Language
Trails A & B	Executive Function
Digit symbol	Attention / Processing speed
Boston Naming Test	Language
Auditory Verbal Learning Test	Memory
Geriatric Depression Scale	Multiple
Clock drawing	Visua-spatial
Neuropsychiatric Inventory Q	Executive function
ADAS-Cog	Multiple
Clinical Dementia Rating Scale	Multiple
Activities of Daily Living(FAQ)	Multiple

Table 5.2: Neuropsychological Tests conducted for ADNI [Ini05]

has 515 MR volumes that belong to three classes: 161 AD volumes, 193 MCI volumes and 161 CN volumes. Only MRI volumes acquired at a resolution of 3T were used in this work. The MRI images were broken down into two input feature spaces: two dimensional (2D) image slices and three dimensional (3D) image volumes. Each MRI image volume was sliced at the preprocessed spatial resolution of $1.0x1.0x1.2mm^3$. A representative subset of these slices (n=60) was selected as the input. In the two dimensional scenario, each patient has 60 of these selected slices, whereas in three dimensional scenario, each patient had one MRI volume comprising these 60 selected slices. A subset of these 60 slices is presented in Figure 5.3.

5.2 Method

This work proposes three progressively more complex methods to classify CN, AD, and MCI individuals. The basis of the proposed methods is CNNs. As the structural MRI images are 3D, the convolutional filters used need to be 3D as well. However, if

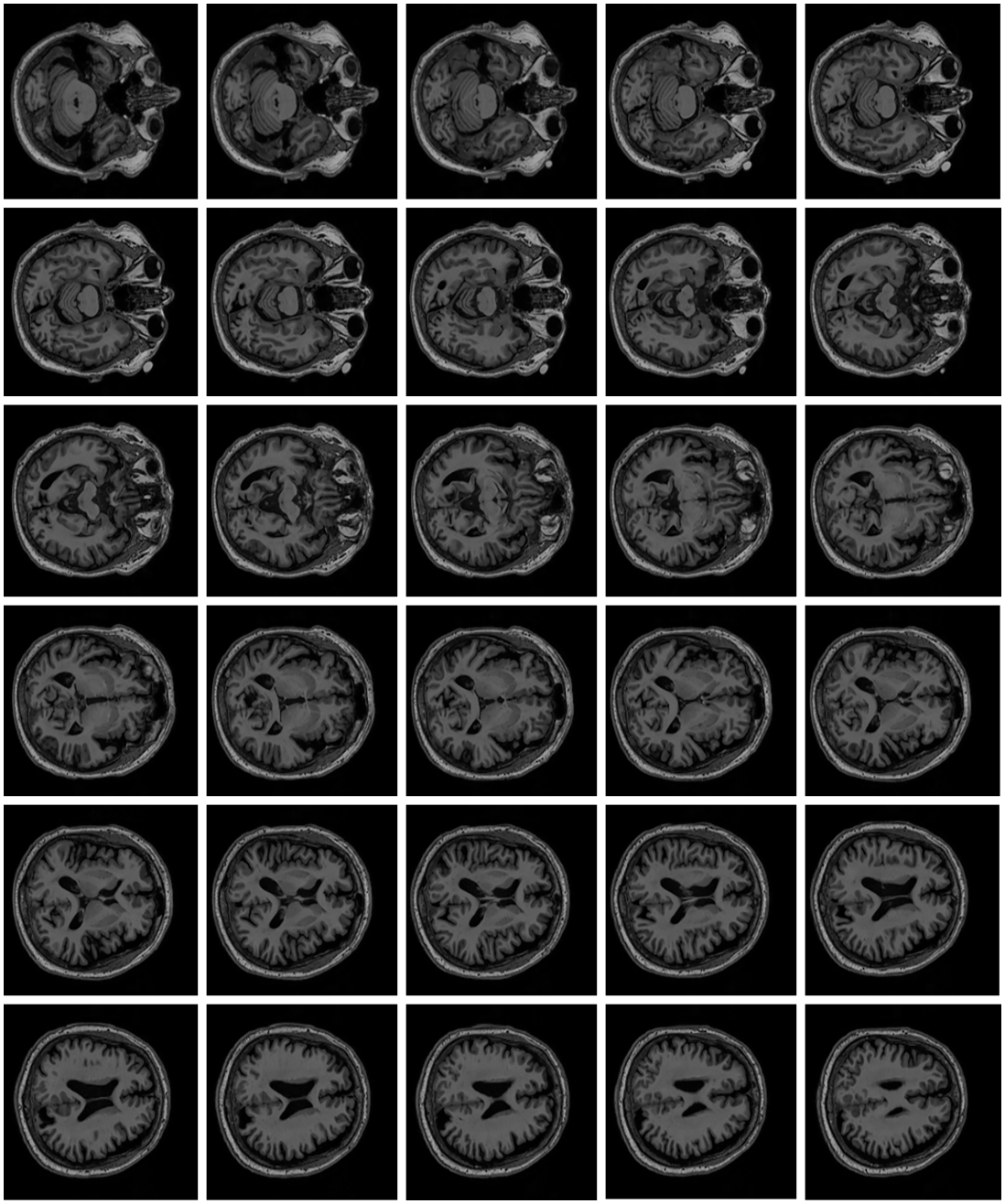


Figure 5.3: Slices of MRI volume of patient 002˙S˙0816, diagnosed with Alzheimer's disease. MRI volume used in the experiments consists of 60 slices, out of which every other slice is shown in this figure

the 3D structure of MRI images is broken down to a collection of multiple 2D slices, 2D convolution can be used. Therefore, section 5.2.1 describes the use of 2D CNNs and the dataset preparation for training and testing. This is then extended to 3D

CNNs in section 5.2.2 while section 5.2.3 presents a novel deep fusion network. In section 5.2.4, the transfer learning process is adapted to use classifiers trained on the ADNI dataset to fine-tune the training on the MAS dataset. This is followed by section 5.2.5 elaborating the behaviour of the model when dealing with missing information. In section 5.2.6, the experimental setup is presented.

5.2.1 2D CNN

A 2D CNN is the most popular form of convolutional neural networks [RW17]. As the advent of CNN occurred with image classification tasks and images in general are two-dimensional, a 2D CNN is viewed as the standard form of convolutional neural networks. The proposed architecture for the 2D CNN used in this work is depicted in Figure 5.4.

Compared to the formulation of a CNN for traditional image classification, the proposed architecture takes a set of 2D slices extracted from volumetric MRI images as the input. Each set of 2D slices corresponding to a patient is fed into the network together, and has the same label across the slices. This ensures that the classification for a patient is performed by the majority classification assigned by the network for the whole set of 2D slices. Sixty slices containing the brain were extracted for each patient from the preprocessed 3D MRI volumes. These were used to train the network and a validation set was used to evaluate the network performance.

5.2.2 3D CNN

While simultaneously feeding a set of 2D image slices to a 2D CNN allows the network to understand the sequential properties of the image slices, a 3D CNN can make better use of the inherently volumetric nature of MRI images. This is because instead of

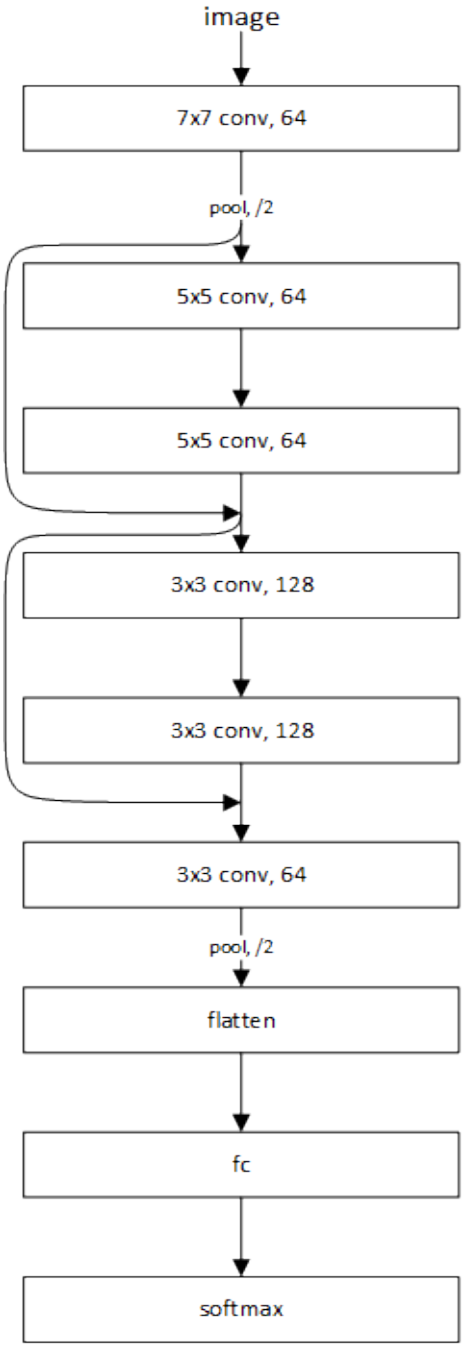


Figure 5.4: The architecture of 2D CNN

feeding 2D image slices, the 3D volumes can be fed directly into the network and 3D convolutions can convolve across the image volumetrically. This directly benefits a problem like Alzheimer's disease classification as most known MRI based biomarkers are volumetric in nature.

The architecture of a 3D CNN network is depicted in Figure 5.5. It has been sufficiently expanded to cater to the complexity of volumetric images compared to the 2D CNN. The input to this network is the 3D volumetric image per patient and to keep the results comparable, the same volume comprising the 60 slices fed into the 2D CNN were used.

5.2.3 Deep Fusion

A novel deep fusion architecture that can seamlessly fuse data from multiple modalities is proposed herein. This work uses two data modalities, namely 3D MR volumes and NM based feature vectors. The challenge in combining these two disparate data sources is in matching the dimensions. The NM based feature vector has 35 features, whereas the 3D MR volume would have slightly more than ten million features if it were flattened. The pipeline proposed meaningfully reduces the dimensions of the 3D MR volumes to a comparable dimension so that the two feature vectors can be merged together. A combination of several techniques is used for this purpose, which are explained below. The architecture of the proposed CNN is depicted in Figure 5.6. Other than the concepts discussed below, conventional components of CNNs such as batch normalization [LD14] and dropout [LD14] are also employed in designing the architecture, in order to reduce the internal covariate shift in the network and to reduce overfitting.

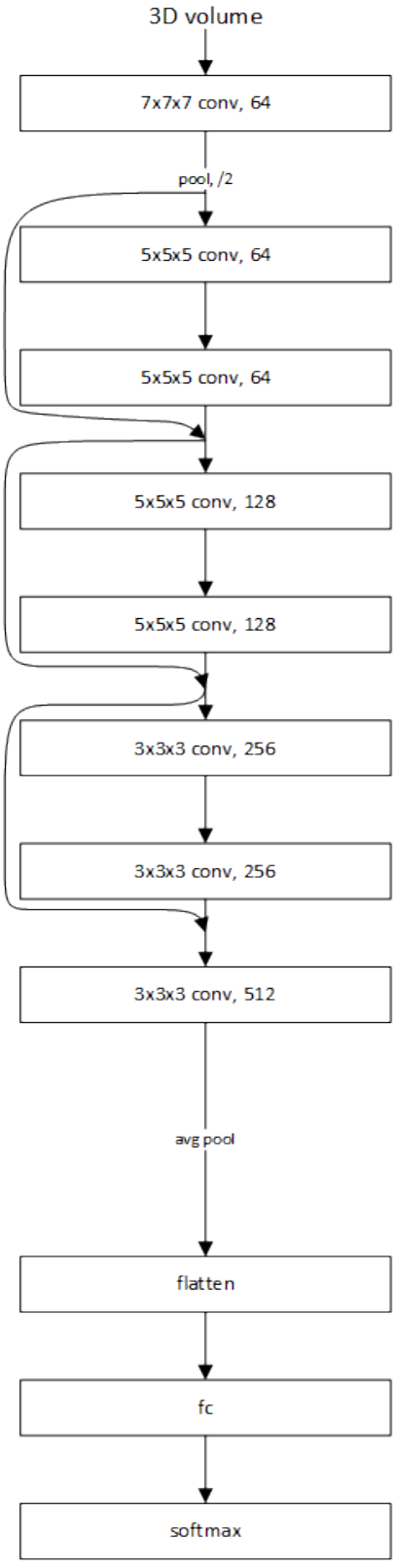


Figure 5.5: The architecture of 3D CNN

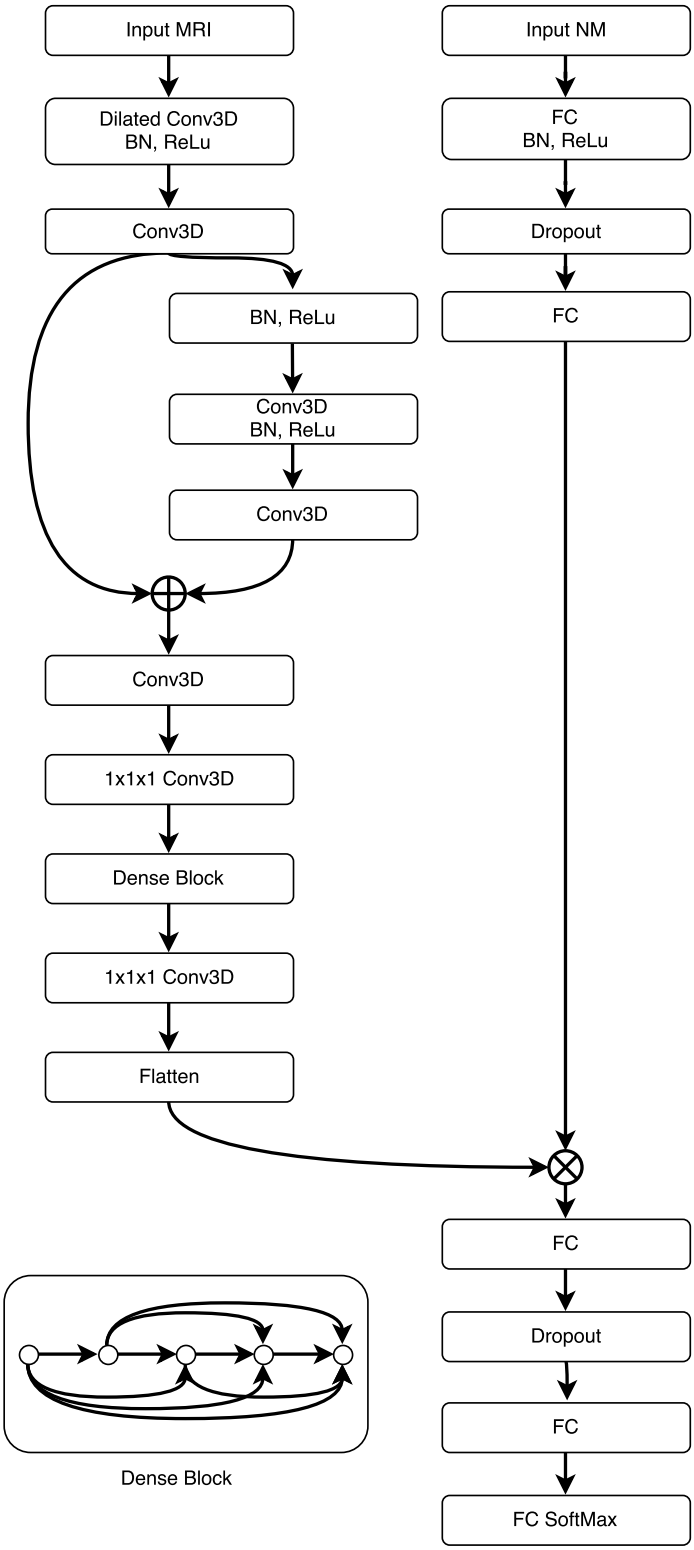


Figure 5.6: Architecture of the deep fusion network. Conv3D denotes a 3D convolutional layer, BN stands for batch normalization, ReLu stands for rectified linear units while FC stands for fully connected layer

5.2.3.1 Dilated Convolution

Dilated convolution was proposed by Yu and Koltun [YK15] for dense predictions. It can be considered as a way to incorporate global context into the convolution operation, as systematic dilation allows exponential expansion of the receptive field without loss of resolution or coverage, and the original article contains a better description.

Let $F : \mathbb{Z}^2 \to \mathbb{R}$ be a discrete function, $\Omega_r = [-r, r]^2 \cap \mathbb{Z}^2$ and $k : \Omega_r \to \mathbb{R}$ be a discrete filter of size $(2r + 1)^2$. The discrete convolution operation $*$ is defined as

$$(F * k)(p) = \sum_{s+t=p} F(s)k(t) \tag{5.1}$$

Generalizing this, let l be a dilation factor and let $*_l$ be defined as

$$(F *_l k)(p) = \sum_{s+lt=p} F(s)k(t) \tag{5.2}$$

Dilated convolution with a dilation factor l is denoted by $*_l$. The traditional convolution is simply the 1-dilated convolution. A more detailed description of dilated convolution can be found [YK15]. The proposed approach uses 2-dilated convolution. The advantage of using dilated convolution is that it addresses the multi-scale problem efficiently without increasing the number of parameters significantly. It integrates knowledge of the wider, global context by dilating the receptive field.

5.2.3.2 Residual Connections

Residual connections proposed by He et. al [HZRS16] allow easier training of deeper networks. The core idea in residual networks is to explicitly allow layers to fit a residual mapping instead of directly fitting a desired underlying mapping to a stack of layers. A better description of residual connections and ResNet architecture can be found [HZRS16]. The proposed method specifically requires a deeper network in order to meaningfully reduce the complexity of volumetric features in the fusion pipeline and therefore, residual connections are integral to maintaining ease of training of the whole network. As the proposed network has an increased number of layers, an individual layer may not modify the signal significantly. The residual connections eliminate these singularities by breaking the permutation symmetry of nodes and by reducing the possibility of node elimination [HZRS16].

5.2.3.3 Dense Connections

Dense connections were introduced by Huang et al. [HLW16], taking a step further from residual connections. A better description of dense connections and DenseNet architecture can be found [HLW16]. The core idea is to connect every layer to every other layer. This ensures ease of error propagation throughout the network regardless of its depth. Rather than using the addition operator for skip connections, dense connections consist of stacking layers. Besides better parameter efficiency, this also achieves improved flow of information and gradients throughout the network making the network easy to train. Each layer has direct access to the gradients from the loss function and the original input signal, leading to implicit deep supervision [HLW16]. The authors also suggested that dense connections have a regularizing effect that reduces over-fitting when the dataset size is small, which further justifies the use of dense connections in this proposed pipeline.

The caveat here is that only layers with the same height and width can be stacked. This is why both dense connections and residual connections are employed, as they are necessary to reduce the complexity of 3D MR volumes in a meaningful manner while maintaining trainability of the network.

5.2.3.4 1x1 Convolutions

The proposed method also employs 1x1 convolutions (or specifically, 1x1x1 convolutions for 3D CNN) as these can be used for dimensionality reduction in feature space while still maintaining spatial relevancy. This may be considered as a coordinate dependent feature transformation stage in filter space and was popularized by GoogLeNet architecture [SLJ+14].

5.2.4 Transfer Learning

In order to evaluate the generalizability of the trained models, the models trained on one dataset were fine-tuned on the second dataset. This was made possible because both MAS and ADNI datasets were used for these experiments. The models were first trained using the ADNI dataset for the different categories of experiments and the trained models were further fine-tuned and tested using the MAS dataset. This method is identified as transfer learning [KJS18].

Although the imaging modalities have the same dimensions, the NM features have different dimensions on the ADNI and MAS datasets. Therefore, when NM features from the MAS dataset were fed into the trained model to fine-tune it, the feature vector was zero-padded to match the dimension of the ADNI NM feature vector.

5.2.5 Evaluating the resilience of deep fusion network to missing information

Given that the proposed deep fusion network fuses information from two modalities, it stands to reason that it should be resilient to missing information from any one modality. The objective of this experiment was to understand the minimum number of NM based features required for a trained network to make the correct discrimination, as that could manifest in a reduced number of tests necessary to be performed on the patients. Therefore, one feature at a time was removed from the validation set and the performance of the model measured. Zero-padding was used to restore the size of the NM feature vector to the original size as the trained network expects a fixed size feature vector.

Three techniques of removing features were used. A random clipping was considered as a baseline first called 'random' hereinafter. The NM features were then ranked from the most important to the least important using the feature importance matrix derived from the deep learning model, and features were clipped both in ascending and descending orders of importance (called 'ascending' and 'descending'), in order to evaluate the robustness of the network for missing information.

5.2.6 Experimental setup

Six different classification scenarios are possible with the three classes considered: three binary (one vs one) and three multi-class (one vs all). The breakdown of the sample sizes in ADNI and MAS datasets for each of these scenarios is presented in Table 5.3. A 15% validation set was identified using stratified sampling and used to evaluate the performance of the trained model, as n-fold cross validation was not possible due to the increased computation time necessary for training a single fold.

Class	Sample size ADNI	Sample size MAS
Normal vs All	161 — 354	293 — 234
MCI vs All	193 — 322	195 — 332
AD vs All	161 — 354	39 — 488
Normal vs MCI	161 — 193	293 — 195
MCI vs AD	193 — 161	195 — 39
Normal vs AD	161 — 161	293 — 39

Table 5.3: The ADNI and MAS dataset sample size of different scenarios

5.3 Results

The results of the experiments carried out are presented in this section. As two datasets were available, it was decided to use the more stable ADNI dataset to train and test the 2D CNN, 3D CNN and deep fusion models, and then use these pre-trained networks as starting points to train on the MAS dataset, which is formally known as transfer learning. Therefore, the results for the ADNI dataset are presented first, where performance of 2D CNN, 3D CNN and and the proposed deep fusion network are compared.

The starting hypothesis was that 3D CNN should perform better than 2D CNN, as the former can capture the volumetric nature of MRI volumes better. The results for 2D CNN, 3D CNN as well as the novel deep fusion network trained on the ADNI training set and tested on the validation set are in Figure 5.7. This confirms the starting hypothesis, with the 3D CNN indeed performing better than the 2D CNN. Further to that, the deep fusion network performs better than the 2D and 3D CNNs in most of the experiments. This is because the deep fusion network attempts to utilize data from multiple modalities including NM based features and MRI images. The deep fusion network was designed to capture the intrinsic properties of both these modalities and co-optimize the network to discriminate between the classes considered. As deep fusion is an amalgamation between two networks, an ablation test was performed to evaluate the performance of each of the parts as well. The

performance comparison between the deep fusion network, a network trained with NM features alone (NM network) and 3D CNN trained with MRI images is presented in Figure 5.8. The NM network is comparable to deep fusion network after removing the imaging arm. It can be ascertained that NM features yield the best results, while the deep fusion network follows closely. This may be interpreted by examining two facets; (i) the deep fusion network is not overly biased towards one information mode; if that were the case, the performance of the deep fusion network would be on par with that of the network trained using NM features alone, for example, and (ii) while deep fusion improves performance when compared to 3D CNN, using NM features alone is still a better approach as literature reports that NM features are dominant. Overall, the deep fusion model constitutes a milestone in fusing different information information modalities in an end-to-end pipeline. As far as is known, this is the first proposed deep fusion pipeline for Alzheimer's disease recognition and it is also significantly different from comparable models due to the fact that the information modalities used are fundamentally different. Other approaches such as the one presented by Kumar et al. [KFFK18] focused on feature fusion, that too between similar modalities such as Positron Emission Tomography (PET) and Computed Tomography (CT) images. The intrinsic value of the deep fusion network proposed in this thesis is that it can be readily extended to fuse data from other modalities such as medical imaging (PET / CT) and supplementary data (hospital records, blood reports, gene records).

The 2D CNN, 3D CNN and deep fusion networks trained using the ADNI dataset were then used as the starting point for training on the MAS dataset. A typical issue when using CNNs for medical imaging problems is that there are no readily available pre-trained networks available for transfer learning. The available networks are usually trained on natural images, which are a fundamentally different modality to medical imaging. It has been demonstrated that transfer learning works best when pre-trained networks are from the same domain [TSK+18]. The per-

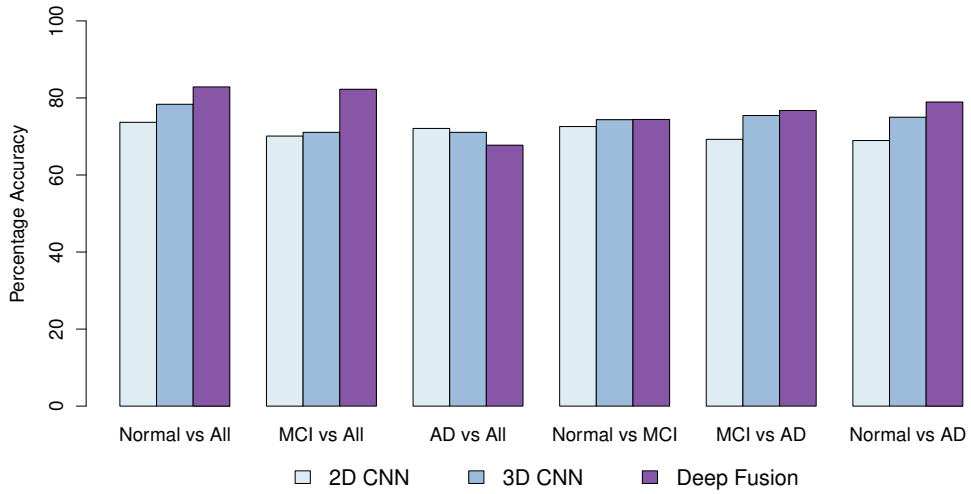


Figure 5.7: Comparison of performance between 2D CNN, 3D CNN and deep fusion network on ADNI dataset

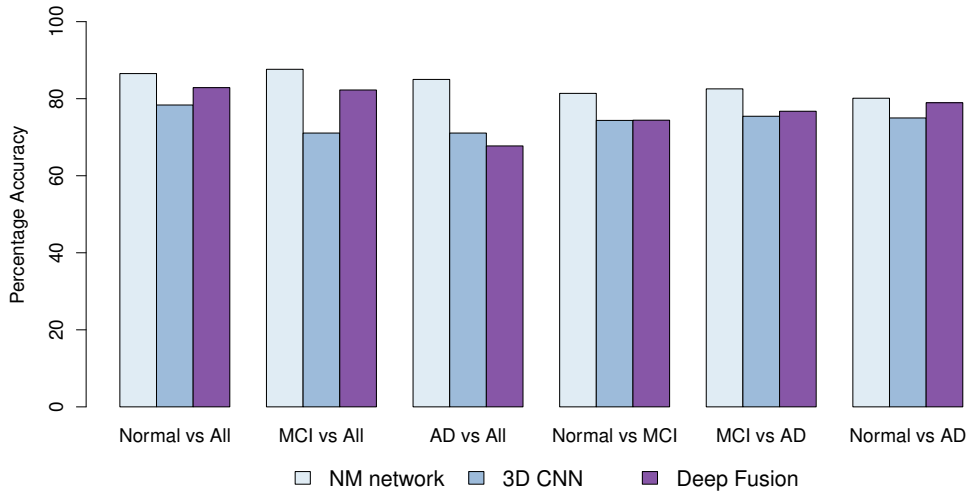


Figure 5.8: Ablative testing comparing the performance of the deep fusion network against its components

formance of the networks fine-tuned on the MAS dataset is presented in Figure 5.9. The networks fine-tuned on MAS dataset almost always perform equally well as those trained on the MAS dataset from scratch, if not better as evident by Figure 5.10. This demonstrates the generalizability and the transferability of the deep networks initially trained on the ADNI dataset, which can be considered as an extra level of validation.

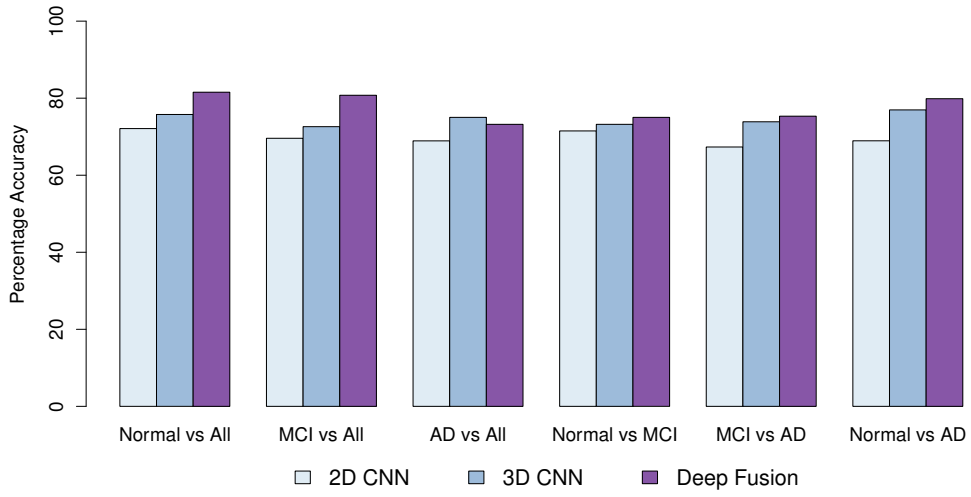


Figure 5.9: Performance of the fine-tuned network on MAS dataset

Additionally, the performance of the deep fusion networks directly trained on ADNI and the one subsequently fine-tuned on MAS are compared to the best results presented by Korolev et. al [KSBD17] in Figure 5.11; the latter can be considered as the closest experiments to the ones presented in this chapter as they report performance of two 3D CNN architectures based on VGG and ResNet [HZRS16] networks on the ADNI dataset. While they do not carry out all the experiments reported by this thesis, Figure 5.11 presents the results for similar experiments. It is evident that the deep fusion network exceeds the performance of the 3D volumetric CNN reported by Korolev et. al.

Finally, the ability of the networks to perform with partial information

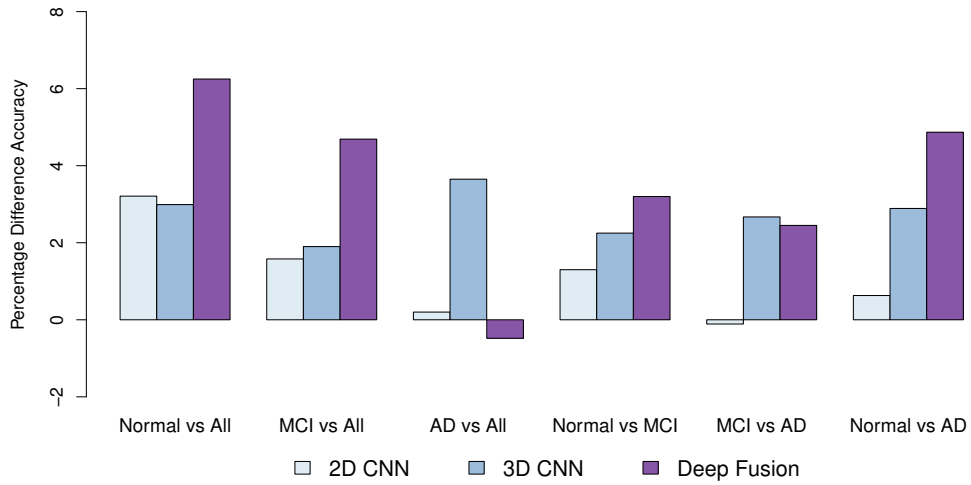


Figure 5.10: The percentage accuracy difference between fine-tuned network on MAS compared to networks trained from scratch on MAS

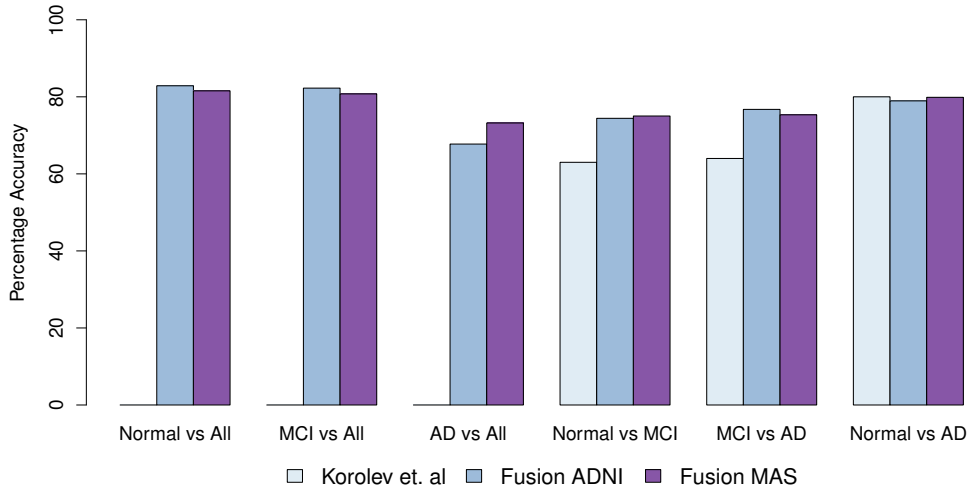


Figure 5.11: Comparison between the closest approach reported by Korolev et. al [KSBD17]. Results for the first three classification scenarios were not reported by them. Fusion ADNI refers to the deep fusion network directly trained on ADNI, while Fusion MAS refers to the deep fusion network trained on ADNI and fine-tuned on MAS. Each model was tested on its validation set.

was also tested using the model fine-tuned on MAS dataset. The performance on the validation set of the MAS dataset was evaluated while reducing the number of NM features input to the network. The results of this experiment are depicted in Figure 5.12, where the performance is plotted against the number of NM features clipped using the three clipping methods described. The network is robust to missing information upto a point and then the performance collapses significantly. However, adapting a systematic clipping method and only removing NM features with lower importance allows one to determine a minimal set of NM features required to effectively utilize the deep fusion network. Therefore, this can reduce the number of neuropsychological tests required for AD diagnosis, which will reduce the cost and time implications for patients.

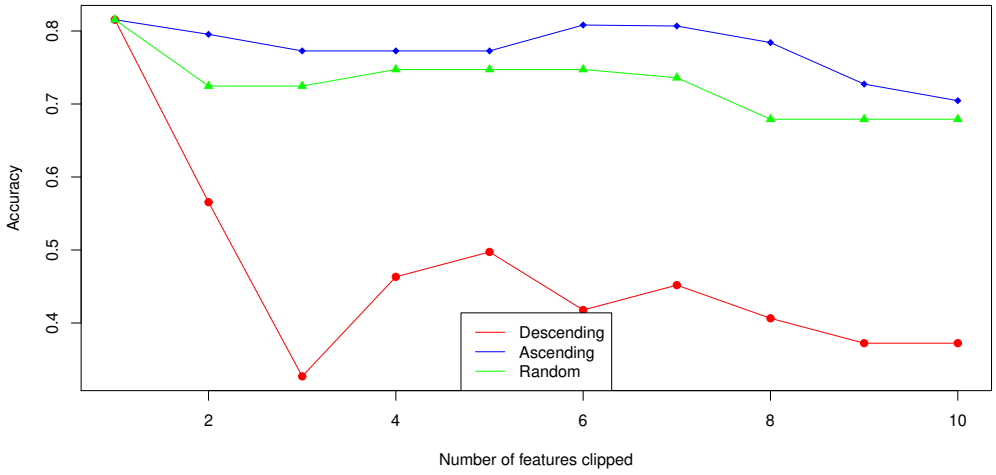


Figure 5.12: The effect of clipping neuropsychological measure based features

It can be ascertained that up to eight features (23% of the total number of features) can be removed without significantly affecting the performance of the network if the features are removed using the ascending importance method. The remaining features are presented in order of importance with respective feature importance in Table 5.4. While random clipping is not as robust, it still maintains upto 90% of the initial performance until 7 features are removed. However, if features

are removed in order of descending importance, the network quickly falls apart as expected, since the most important features are removed first.

It is gratifying to note that the most important features identified by the deep fusion network have also been suggested as discriminative biomarkers in the literature [MDP+, BWP+13, EGKB+03, KMA+15, HK14], thereby confirming the validity of the results.

Feature Name	Importance
FAS˙raw	0.1534091
RVLT˙intrusions˙raw	0.01818182
RVLTtotal˙raw	0.01534091
RVLT˙RPC˙raw	0.01534091
BNT˙raw	0.01136364
Animal˙raw	0.01079545
RVLT6˙raw	0.009090909
Age	0.008522727
TMTB˙raw	0.008522727
LM˙immed˙raw	0.007386364
GPegtest˙raw	0.00625
TMTA˙raw	0.004545455
Block˙raw	0.004545455
S˙raw	0.003409091
NART˙err˙raw	0.003409091
NART˙IQ	0.002272727
RVLT1˙raw	0.001704545
RVLT3˙raw	0.001136364
RVLT˙recTP˙raw	0.001136364
RVLT4˙raw	0.000568182

Table 5.4: Feature importance in descending order as identified by the deep fusion network

5.4 Summary

In this chapter, a number of deep learning pipelines are proposed to differentiate between cognitively normal individuals from individuals with mild cognitive impairment and Alzheimer's disease or dementia. The major motivation in designing these

pipelines is to make use of advanced medical imaging modalities such as magnetic resonance imaging. The main contribution of this chapter is a novel deep fusion pipeline that is capable of combining data from multiple modalities including volumetric MR and NM features. These two modalities exhibit the most disparity between their data sizes and complexities. In principle, this approach may be readily extended to incorporate other MR modalities such as diffusion tensor images or even other imaging modalities such as PET. The proposed network draws inspiration from a range of state-of-the-art deep learning architectures such as GoogLeNet, ResNet and DenseNet to effectively reduce the complexity of the datasets in order to enable fusion. As far as can be ascertained, this is the first work discussing the fusion of volumetric MR images with neuropsychological measure based features in an end-to-end deep learning pipeline.

The efficacy of the proposed fusion pipeline has been demonstrated by comparing it against the best 3D CNN model and that of a recent work based on two 3D CNN models [KSBD17]. The major advantage of the proposed model is its ease of use which is important in a domain such as medical imaging, as domain experts may have neither the expertise nor the time to carefully design and train classifiers. It should also be noted that this approach may also be used in a scenario where MR imaging is available while NM based features are only partially available, which is typically the case since the full battery of neuropsychological testing is cost prohibitive compared to MR imaging. This has been verified by testing the resilience of the trained deep fusion network against an increasing set of missing NM features. It was demonstrated that the deep fusion network was able to maintain its accuracy upto a reduction of 23% of the commonly used NM features.

Finally, the efficiency of the pre-trained network as the starting point for another machine learning effort has been tested. The initial deep learning pipelines were trained on the ADNI dataset and the trained networks were then used to initial-

ize and fine-tune the deep learning pipelines using the MAS dataset. This approach allows one to seamlessly utilize a pre-trained network that has been standardized either as a classifier or as a starting point for a transfer learning experiment where it is fine-tuned using new data. Therefore, the expertise required to design and train a deep neural network from scratch is alleviated, which can assist neuroscientists and other medical professionals with limited machine learning knowledge.

Chapter 6

Survival Analysis for Dementia*

Survival analysis is used to evaluate the significance of prognostic / predictor variables (or covariates in statistical terms) in determining events such as death or recurrence of a disease such as cancer [YSK+16, CYA13, BT04, RA13]. The goals of survival analysis are three fold: (i) to estimate survival and hazard functions from survival data, (ii) to compare survival and hazard functions between groups and (iii) to assess the relationship between predictor variables and survival time [CAIH+16]. The focus of this chapter is on the first and the last goals. Three mandatory sets of information are necessary to perform survival analysis: (i) the covariates (either at baseline or time-varying), (ii) time to an event of interest and (iii) binary event outcome. Traditionally, survival analysis is a statistical technique and established methods are based on a statistics perspective. Generating a Kaplan-Meier (KM) curve from survival data to represent the probability of survival is the initial step

Portions of this chapter appear in:

1. Senanayake, U., Sowmya, A., Dawes, L., Kochan, N. A., Wen, W., Sachdev, P., "Deep Survival Analysis for Cognitive Impairment" (under preparation)

of any survival analysis task. The KM curve illustrates the estimate of the conditional probability of the time to failure calculated at each time point recording an event [YSK+16, CAIH+16]. Log rank test is used to compare the difference in survival between two or more groups, when evaluating treatment effectiveness against a control group [PCPA+77].

In this work, deep learning based survival analysis is applied to the domain of cognitive impairment and dementia for the first time. Traditional survival analysis has been explored for Alzheimer's disease and dementia using NM features [Hel09, aSWLM15] as well as MRI based features [EZELL+15]. However, an inherent issue with these studies is that a considerable amount of domain knowledge is needed to extract features, particularly for medical images. A study by Zeifman et al. [EZELL+15] used voxel level grey matter volume, which needs a significant level of expertise to extract. Using NM features on the other hand is comparatively straightforward, as the feature preprocessing can be trivial [aSWLM15]. The work in this chapter attempts to alleviate these difficulties by streamlining the feature extraction process, thereby allowing scientists without significant domain knowledge to carry out survival analysis experiments. Deep learning based survival analysis techniques already developed for other domains are then modified and applied to the MAS and ADNI datasets to demonstrate their effectiveness for survival analysis of dementia.

The rest of this chapter is structured as follows. In section 6.1, the dataset utilized is described, while section 6.2 discusses the methodology used for experiments. In section 6.3 the results of the study are presented and the chapter is concluded in the final section.

6.1 Dataset

The datasets used in this work are the same as those already described in Chapters 2, 4 and 5, namely the Sydney **M**emory and **A**geing **S**tudy (MAS) and the **A**lzheirmer's **D**isease **N**euroimaging **I**nitiative (ADNI) datasets. The dataset descriptions are not repeated. However, the preprocessing performed before carrying out survival analysis is described in this section. Both MAS and ADNI datasets are longitudinal studies that follow patients from a baseline. This allows one to derive two essential elements, namely the event indicator E and a failure event time T, to analyze survival on top of the covariates $\mathbf{x}$ that are available at each longitudinal time point. The event of interest can be a number of different things in survival analysis. This work focuses on the onset of dementia as the event of interest. Therefore, the preprocessing pipeline identifies individuals who become demented within the duration of the longitudinal study and these individuals are assigned a positive event indicator. The failure event time is derived as the time to event from baseline for individuals with a positive event indicator. Individuals without an event occurring in the duration of the study are known as censored individuals, in this case they are right-censored. Censoring happens when incomplete information is available about the failure time of certain individuals [LEA97]. What is known about these individuals is that they have survived up until the end of the study without undergoing an event, or until they were lost to the study. The transitions between different states that were observed for the ADNI and MAS studies are shown in Tables 6.1 and 6.2. Both ADNI and MAS datasets have a similar percentage of censored individuals, namely 70.74% and 66.07% of the cohort considered respectively.

Table 6.1: Transitions between states from baseline for patients in ADNI study

	End Stage			
Start Stage	NL	EMCI	LMCI	AD
NL	348	0	48	14
EMCI	26	17	225	36
LMCI	11	0	155	93
AD	0	0	1	232

Table 6.2: Transitions between states from baseline for patients in MAS study

	End Stage					
Start Stage	CN	sd-aMCI	md-aMCI	sd-naMCI	md-naMCI	Dementia
CN	410	32	11	39	1	11
sd-aMCI	18	63	7	5	3	7
md-aMCI	8	4	49	7	5	9
sd-naMCI	25	9	7	72	3	6
md-naMCI	0	0	5	3	16	2

6.2 Method

This section describes the method used to conduct survival analysis on the datasets discussed in section 6.1. The proposed techniques attempt to modify and utilize deep learning based survival analysis techniques inspired by other domains. These techniques are validated on the two datasets to understand their efficacy and generalizability. The main contribution of this work is the design, implementation and validation of a deep learning survival analysis pipeline that is capable of utilizing both NM and MRI based features without requiring significant domain knowledge to extract relevant features.

Survival and hazard functions are the fundamental functions of interest in survival analysis. The Survival function may be denoted as $S(t) = P_r(T > t)$, where $S(t)$ is the probability that an individual has survived beyond a certain time t. In this context, survival refers to remaining cognitively normal, ie: surviving dementia. The hazard function can be considered as a measure of risk at time t. A larger hazard value indicates a greater risk of an event occurring. The hazard function $\lambda(t)$

is defined as [KSC+18].

$$\lambda(t) = \lim_{\delta \to 0} \frac{P_r(t \leq T < t + \delta | T \geq t)}{\delta} \tag{6.1}$$

An individual's survival given their baseline data x is commonly modeled using the proportional hazards model. The latter assumes that the hazard function is composed of two functions, namely the baseline hazard function $\lambda_0(t)$ and a risk function, $h(x)$ that denotes the effects of an individual's covariates. It is assumed that the hazard function has the form $\lambda(t|x) = \lambda_0(t) \cdot e^{h(x)}$.

The rest of this section is organized as follows. In section 6.2.1, the statistical methods used as well as the deep learning methods adapted for survival analysis in this thesis are described, while section 6.2.2 discusses survival analysis using NM features. Survival analysis using MRI based features and fused features is described in section 6.2.3. Finally, section 6.2.4 details the methods used to validate the proposed methods and results.

6.2.1 Survival Analysis Models in Literature

A number of survival analysis models have been reported in the literature and the relevant models are discussed now. Cox Proportional Hazards Model is first described which is used to explore the survival characteristics of the datasets. This is followed by an overview of three different deep survival networks.

6.2.1.1 Cox Proportional Hazards Model

Cox Proportional Hazards model (CPH) was developed by Sir David Cox and is widely used in epidemiological and medical studies [Cox72]. It is a proportional

hazards model that uses a linear function $\hat{h}_\beta(x) = \beta^T x$ to estimate the risk function $h(x)$. The weights β are tuned to optimize Cox partial likelihood in order to perform Cox regression. The partial likelihood is the product of the probabilities at each event time T_i that the event has occurred to individual i, given the set of individuals still at risk at time T_i. It is defined as in Equation 6.2 and parameterized β [KSC+18] where T_i is the event time, E_i is the event indicator and x_i denotes the baseline data for the i^{th} observation. The term risk set $\mathbb{R}(\mathrm{t}) = \mathrm{i} : \mathrm{T_i} \geq \mathrm{t}$ denotes the patients still at risk of the event at time t. The covariates are assumed to have a multiplicative effect on the hazard. It is also assumed that this effect is constant over time.

$$L_c(\beta) = \prod_{i:E_i=1} \frac{exp(\hat{h}_\beta(x_i))}{\sum_{j \in \mathbb{R}(\mathrm{T_i})} exp(\hat{h}_\beta(x_j))} \tag{6.2}$$

6.2.1.2 DeepSurv

DeepSurv is a method proposed by Katzman et al. [KSC+18] that applies deep learning techniques to a nonlinear cox proportional hazards network. It is a multi-layer perceptron that predicts a patient's risk of an event. The output of the deep learning network is a single node that estimates the risk function $\hat{h}_\theta(x)$ parameterized by the weights of the network θ. The loss function used is the negative log partial likelihood as described in Equation 6.3.

$$l(\theta) = -\sum_{i:E_i=1} \left(\hat{h}_\theta(x_i) - log \sum_{j \in \mathbb{R}(\mathrm{T_i})} e^{\hat{h}_\theta(x_j)} \right) \tag{6.3}$$

The model uses multiple hidden layers and modern techniques such as weight decay regularization, Rectified Linear Units (ReLu) with batch normalization, Scaled Exponential Linear Units (SELU), dropouts, gradient descent optimization algorithms (Stochastic Gradient Descent and Adaptive Moment Estimation (ADAM)),

Nesterov momentum, gradient clipping and learning rate scheduling [KSC+18]. A random hyper-parameter optimization search is carried out to tune the network's hyper-parameters. The optimized network has 17 hidden layers, 3 dense layers, a dropout rate of 0.401 and ReLu as the activation function.

DeepSurv is used in this work in its original form in analysing NM features, which is discussed in section 6.2.2.

6.2.1.3 DeepHit

DeepHit is a deep learning based survival analysis approach proposed by Lee et al. [LZYvdS18]. Compared to a traditional approach such as the CPH, DeepHit has the ability to consider competing risks in survival analysis. These are settings where there is more than one possible event of interest. The formulation of the deep learning network and the input data is therefore different. In DeepHit, $K \geq 1$ possible events of interest are considered. It is assumed that exactly one event eventually occurs for each patient (eg: patient eventually dies, but can only die from one cause) [LZYvdS18]. Right censored data is also considered as an event of interest and is denoted by $\varnothing$ and therefore the set of possible events are $K = \varnothing, 1, ..., K$. The input data is mapped into a triple (x, s, k) where $x \in X$ is the feature vector or covariates of the patient, s is the time-to-event or censoring, and $k \in K$ is the unique event that occurred at time s. Experiments consider a single event of interest, namely the onset of dementia, therefore the set of possible events is limited to $K = \varnothing, 1$. True probability $P(s = s^*, k = k^*|x = x^*)$, ie: the true ex-ante probability that a patient with covariates x^* will experience the event k^* at time s^* is of interest given the tuple (x^*, s^*, k^*) with $k^* \neq \varnothing$. However, the true probability cannot be known considering a finite dataset and hence the network estimates $\hat{P}$ of true probabilities.

The original DeepHit neural network was proposed with cause-specific sub-

networks to cater to competing risks. In this study, the original DeepHit network is modified and trained to consider the scenario of a single risk, which is discussed in section 6.2.2.

6.2.1.4 Deep Recurrent Survival Analysis

Deep recurrent survival analysis (DRSA) is a technique proposed by Ren et al. [RQZ+19]. Traditionally, survival analysis has only been performed using baseline data [CAIH+16]. Although extensions exist that consider both baseline data and later data for CPH, the performance has been inconsistent [FL99]. Deep recurrent survival analysis has been proposed to take time varying covariates into consideration using recurrent neural networks. The problem formulation differs significantly from the models described so far. The true occurrence time for an event of interest is defined as z if the event has been tracked. This leads to the definition of the probabilistic density function of the true event time $p(z)$, which is the probability that the event truly occurs at time z [RQZ+19]. The survival rate at each time t is the cumulative distribution function defined as in Equation 6.4, which is the probability of the observed patient surviving (event not occurring), until the observed time t. Defining the event rate or the probability of the event occurring before the observation time t becomes trivial and is denoted by Equation 6.5.

$$S(t) = Pr(z > t) = \int_t^{\infty} p(z)dz \tag{6.4}$$

$$W(t) = Pr(z \leq t) = 1 - S(t) = \int_0^{t} p(z)dz \tag{6.5}$$

The input set for the model can be represented by a tuple (x, z, t) where $t > 0$ is the observation time for the given patient. Censored samples were marked

with null for variable z as the observation of true time was unavailable. Covariates are denoted by x at a given time t. Modelling the distribution of the true event time $p(z)$ over all the historical time-to-event while handling censored data, of which the true event time is unknown, can be considered as the goal of this model [RQZ+19]. Translating this to survival analysis, estimating the probability distribution $p(z|x)$ of event time, with regard to the sample feature vector x, is of interest. The model creates a mapping function T that learns the patterns within the data and predicts the event time distribution over the time space as $p(z|x) = T(x)$ [RQZ+19].

In order to build a recurrent neural network based model, a discrete time equivalent of the conditional hazard rate over time needs to be formulated. Over continuous time, the conditional hazard rate models the instant occurrence probability of the event at time t given that the event has not occurred before, and is represented by Equation 6.6.

$$h(t) = \lim_{\Delta \to 0} \frac{Pr\left(t < z \leq t + \Delta t | z > t\right)}{\Delta t} \tag{6.6}$$

Converting this to discrete space considering the grouping of continuous time as $l = 1, 2, ...L$ and uniformly dividing disjoint intervals $V_l = (t_{l-1}, t_l)$ where $t_0 = 0$ and t_l is the last observation interval boundary for the given sample, the event rate function and the survival rate function over discrete time space can be formulated as in Equation 6.7 and 6.8 respectively [RQZ+19].

$$W(t_l) = Pr(z \leq t_l) = \sum_{j \leq l} Pr(z \in V_j) \tag{6.7}$$

$$S(t_l) = Pr(z > t_l) = \sum_{j > t_l} Pr(z \in V_j) \tag{6.8}$$

Therefore, the discrete event time probability function at the $l-th$ time interval can be represented by Equation 6.9 and the discrete conditional hazard rate h_l, defined as the conditional probability approximating the continuous conditional hazard rate function $h(t_l)$ in Equation 6.6 as the intervals V_l become infinitesimal is represented by Equation 6.10.

$$p_l = Pr(z \in V_l) = W(t_l) - W(t_{l-1}) = S(t_{l-1}) - S(t_l) \tag{6.9}$$

$$h_l = Pr(z \in V_l | z > t_{l-1}) = \frac{Pr(z \in V_l)}{z > t_{l-1}} = \frac{p_l}{S(t_{l-1})} \tag{6.10}$$

So far, the discrete time model and the survival probability over the discrete time space have been described, which are needed for the definition of the recurrent neural network (RNN) model with the parameter θ [RQZ+19]. This enables the network to capture sequential patterns for conditional probability h_l^i at every time interval V_l for the i^{th} sample. At each time interval V_l, the $l-th$ RNN cell predicts the instant hazard rate h_l^i given the covariates x^i and the current time t_l conditioned upon the previous events as in Equation 6.11 where f_θ is the RNN function taking (x^i, t_l) as the input and h_l^i as the output. r_{l-1} is the hidden vector calculated from the RNN cell at the last time step which contains the information about the conditional. The RNN function is implemented as a standard LSTM unit.

$$h_l^i = Pr(z \in V_l | z > t_{l-1}, x^i; \theta) = f_\theta(x^i, t_l | r_{l-1}) \tag{6.11}$$

Using the probability chain rule, one can derive the survival rate function $S(t)$ and the corresponding event rate function $W(t)$ at time t for the i_{th} individual sample from Equations 6.7, 6.8, 6.10 and 6.11.

$$
\begin{aligned}
S(t|x^i;\theta) &= Pr(t < z|x^i;\theta) \\
&= Pr(z \notin V_1, z \notin V_2, ..., z \notin V_{l^i}|x^i;\theta) \\
&= Pr(z \notin V_1|x^i;\theta) \cdot Pr(z \notin V_2|z \notin V_1, x^i;\theta) \cdots Pr(z \notin V_{l^i}|z \notin V_1, ..., z \notin V_{l^i-1}, x^i;\theta) \\
&= \prod_{l:l\leq l^i} \left[1 - Pr(z \in V_l|z > t_{l-1}, x^i;\theta)\right] \\
&= \prod_{l:l\leq l^i} (1 - h_l^i)
\end{aligned}
\tag{6.12}
$$

$$
W(t|x^i;\theta) = Pr(t \geq z|x^i;\theta) = 1 - \prod_{l:l\leq l^i} (1 - h_l^i) \tag{6.13}
$$

$$
p_l^i = Pr(z \in V_{l^i}|x^i\theta) = h_{l^i}^i \prod_{l:l<l^i} (1 - h_l^i) \tag{6.14}
$$

In Equations 6.12 and 6.13, l_i denotes the time interval index for the i^{th} sample at t^i and equation 6.14 denotes the probability of time z lying in the interval of V_{l^i} for the i^{th} sample.

The loss function used to train DRSA is the maximum log-likelihood over the empirical data distribution. Three objectives are considered when deriving the loss function [RQZ+19]. The first loss minimizes the negative log-likelihood of the true event time $z = z^i$ over uncensored samples as in Equation 6.15 where l^i is the index of the interval of the true event time $z^i \in V_{l^i}$.

$$
\begin{aligned}
L_z &= -log \prod_{(x^i,z^i)\in\mathbb{D}_{uncensored}} Pr(z \in V_{l^i}|x^i;\theta) \\
&= -log \prod_{(x^i,z^i)\in\mathbb{D}_{uncensored}} h^i_{l^i} \prod_{l:l<l^i}(1-h^i_l) \\
&= -\sum_{(x^i,z^i)\in\mathbb{D}_{uncensored}} \left[log(h^i_{l^i}) + \sum_{l:l<l^i} log(1-h^i_l) \right]
\end{aligned}
\tag{6.15}
$$

Minimizing the negative partial log-likelihood of the event rate over the uncensored samples as in Equation 6.16 is the second loss. The third loss is derived considering the censored samples where partial log-likelihood embedded in the censored samples is used to correct the learning bias of the model as in Equation 6.17.

$$
\begin{aligned}
L_{uncensored} &= -log \prod_{(x^i,z^i)\in\mathbb{D}_{uncensored}} Pr(t^i \geq z|x^i;\theta) \\
&= -log \prod_{(x^i,z^i)\in\mathbb{D}_{uncensored}} W(t^i|x^i;\theta) \\
&= -\sum_{(x^i,z^i)\in\mathbb{D}_{uncensored}} log \left[1 - \prod_{l:l\leq l^i}(1-h^i_l) \right]
\end{aligned}
\tag{6.16}
$$

$$
\begin{aligned}
L_{censored} &= -log \prod_{(x^i,z^i)\in\mathbb{D}_{censored}} Pr(z > t^i|x^i\theta) \\
&= -log \prod_{(x^i,z^i)\in\mathbb{D}_{censored}} S(t^i|x^i;\theta) \\
&= -\sum_{(x^i,z^i)\in\mathbb{D}_{censored}} \sum_{l:l\leq l^i} log(1-h^i_l)
\end{aligned}
\tag{6.17}
$$

The combination of Equations 6.16 and 6.17 describes the classification of

survival status at time t^i of each sample as in Equation 6.17 which can be considered as the cross entropy loss for predicting the survival status at time t^i given x^i over all the data $\mathbb{D}_{full} = \mathbb{D}_{uncensored} \bigcup \mathbb{D}_{censored}$. Hence, the goal of the model is to minimize the negative log-likelihood over all the data samples as denoted by Equation 6.19 where an hyperparameter α controls the loss value balance between L_z and L_c. The deep recurrent survival analysis model is illustrated in Figure 6.1.

DRSA is used in this work in its original form in analysing NM features, which is discussed in section 6.2.2.

$$\begin{aligned} L_c &= L_{uncensored} + L_{censored} \\ &= -log \prod_{(x^i,z^i)\in\mathbb{D}_{full}} \left[S(t^i|x^i;\theta)\right]^{c^i} \cdot \left[1 - S(t^i|x^i;\theta)\right]^{1-c^i} \\ &= - \sum_{(x^i,z^i)\in\mathbb{D}_{full}} \left(c^i \cdot \log S(t^i|x^i\theta) + (1-c^i)\log[1 - S(t^i|x^i;\theta)]\right) \end{aligned} \tag{6.18}$$

$$arg\min_{\theta} \alpha L_z + (1-\alpha)L_c \tag{6.19}$$

6.2.2 Survival Analysis using NM Features

NM features that are used for survival analysis were already described in detail in Chapters 4 and 5. Both MAS and ADNI datasets are used in this work. The features were preprocessed to isolate the features at baseline and the subsequent diagnosis was used to derive the event indicator E and the time to failure T. An additional column was also derived, whose values indicate whether a subject has been censored or not. These collectively form the input dataset for this study. The event of interest is the onset of dementia or its prodromal condition, namely mild cognitive impairment

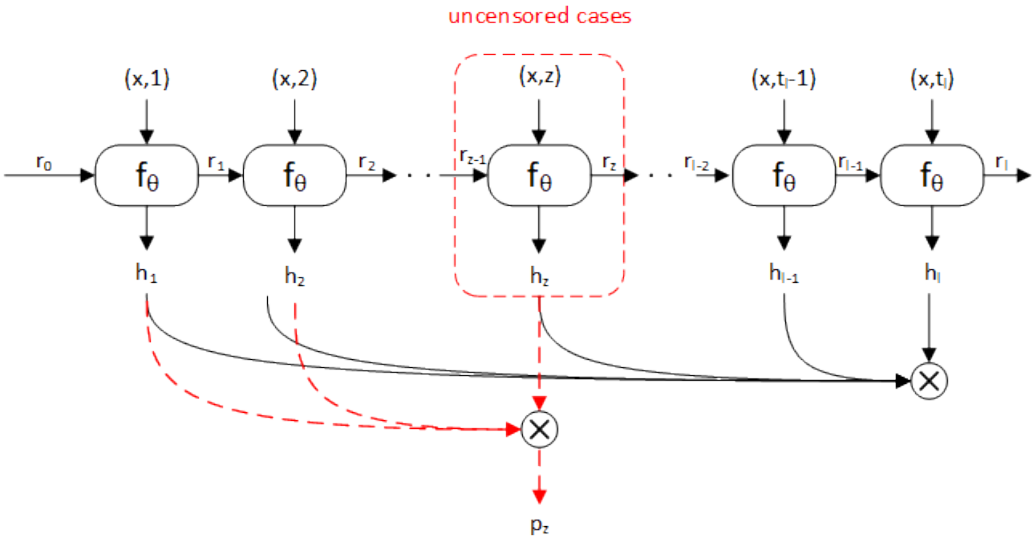


Figure 6.1: Detailed illustration of Deep Recurrent Survival Analysis model. Only uncensored samples have the true event time and can calculate p_z for the loss of L_z. The illustration was recreated from the original work [RQZ+19]

(MCI). Only patients with a cognitively normal diagnosis at baseline are considered for survival analysis. However, patients who are diagnosed with MCI at baseline are included as well when computing the Kaplan-Meier (KM) curve. This allows exploration of differential rates of survival between cognitively normal individuals and those with MCI.

The Kaplan-Meier (KM) curve may be considered as one of the best methods to measure the fraction of subjects surviving a certain amount of time after treatment [GKK10]. The effect of intervention is evaluated by measuring the number of patients who survived after that intervention over a period of time in clinical trials. KM curves are capable of handling censored data, as the latter is a practical reality in clinical trials where patients are either lost to follow-up or do not experience an event before the end of the study. In the current study, no control group exists. However, the same techniques can be applied to observe the differential effects on patients enrolling in the study as cognitively normal and patients who are cognitively impaired.

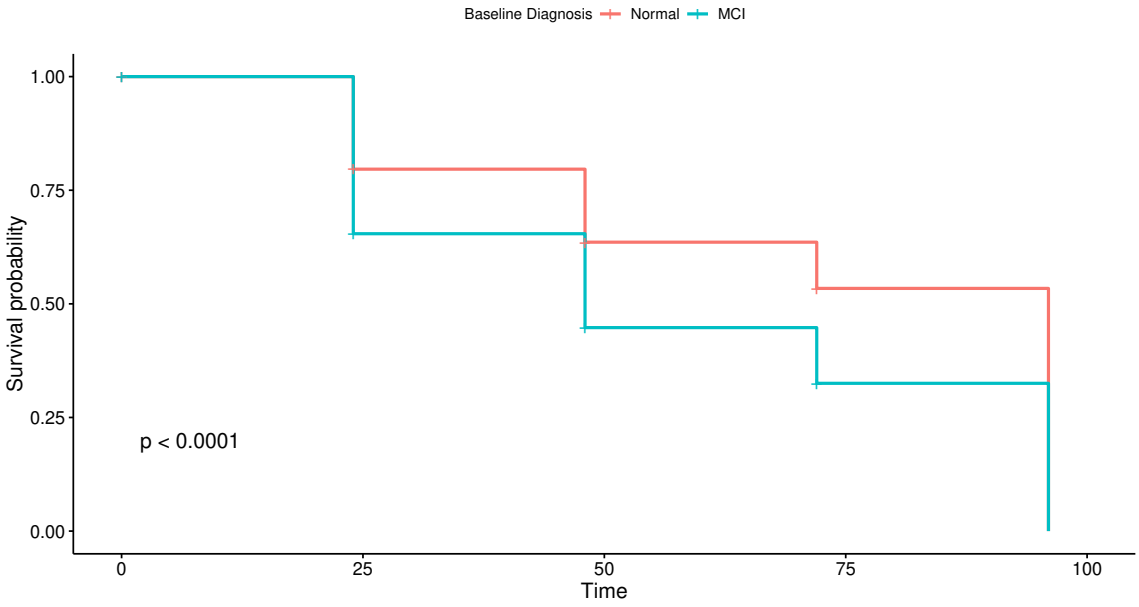


Figure 6.2: Kaplan-Meier curve for MAS dataset where time is in months

The KM curve for the MAS dataset is depicted in Figure 6.2. The vertical lines in this plot indicate censored data while their corresponding x-axis value indicates the time of censoring. The log-rank p-value of 0.0001 indicates a significant result considering $p < 0.05$ as the statistical significance level. For the MAS dataset, it appears that the baseline diagnosis significantly affects the survival probability. This affirms the observation that those with mild cognitive impairment have a higher risk of progressing to dementia.

The KM curve for the ADNI dataset is depicted in Figure 6.3. As the ADNI study has run for longer, it has a much lengthier x-axis. The log-rank p-value of 0.021 indicates a significant result considering $p < 0.05$ as the statistical significance level. Just as in the MAS dataset, it appears that the baseline diagnosis significantly affects the survival probability.

The DeepSurv network and DRSA are used in their original form with NM features. However, the original DeepHit neural network was proposed with cause-

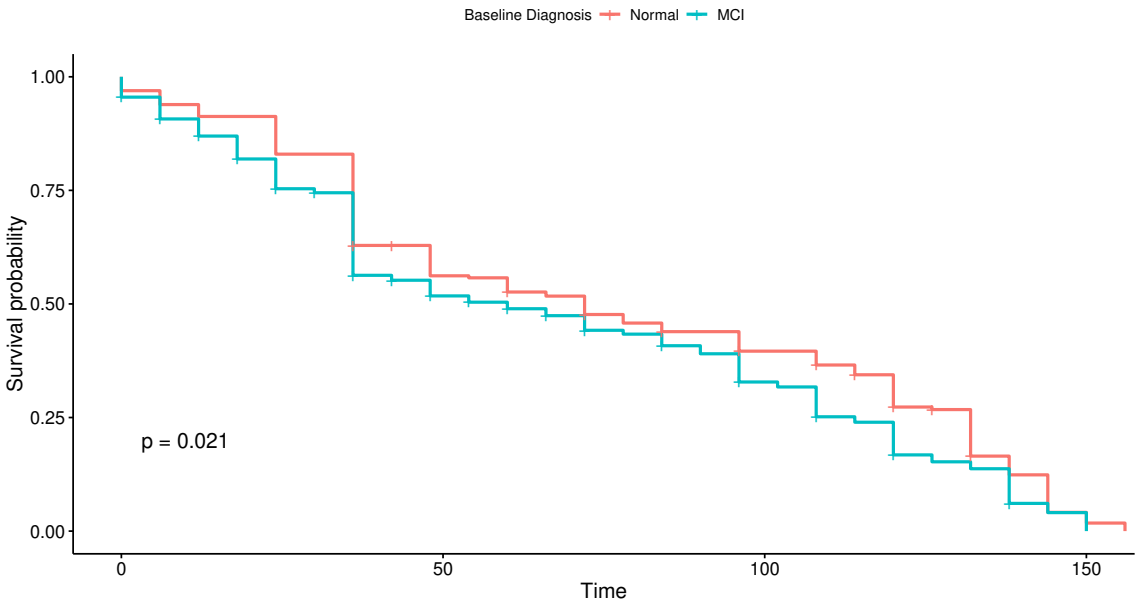
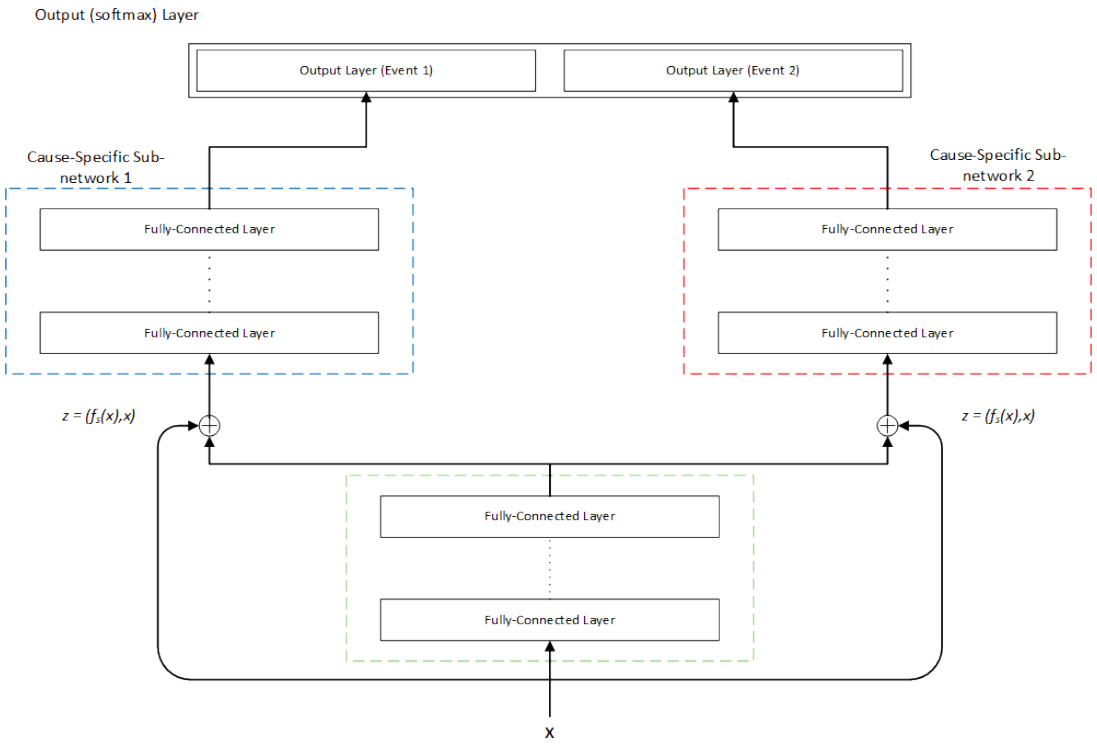


Figure 6.3: Kaplan-Meier curve for ADNI dataset where time is in months

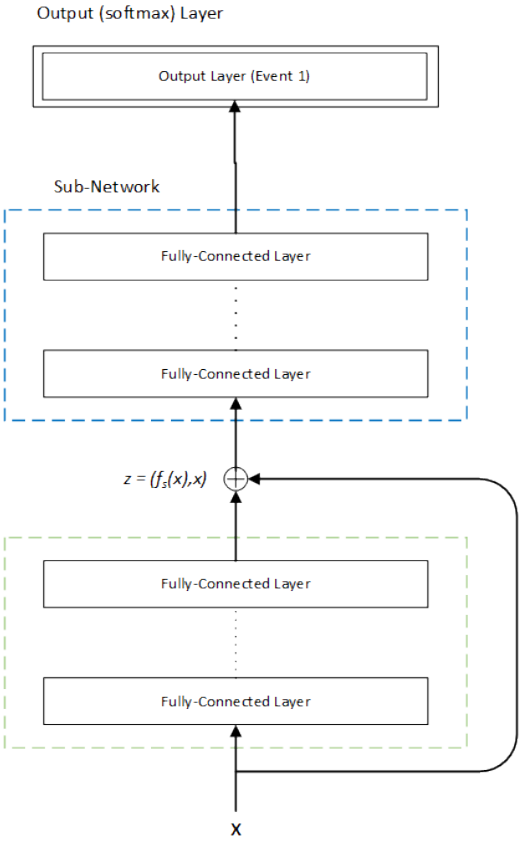
specific sub-networks to cater to competing risks. In this study, the original DeepHit network is modified and trained to consider the scenario of a single risk. The original network and the modified network are shown in Figure 6.4. The network architecture proposed in this thesis has a single sub-network node, compared to multiple sub-networks in DeepHit.

6.2.2.1 Evaluation

Concordance index (C-index) is the standard measure used to evaluate a survival analysis model. The ordering of predicted survival times are evaluated by this measure. A score of 0.5 is the expected outcome from random predictions while 1.0 is the perfect concordance [RSK+07]. Let T be the event time, Z be a $px1$ covariate vector and $g(Z)$ be the theoretical counterpart of the estimated risk score for the subject with Z. Considering two independent copies $(T_1, Z_1, g(Z_1))', (T_2, Z_2, g(Z_2))'$ of $(T, Z, g(Z))'$, commonly used concordance measure is [UCP+11]



(a) DeepHit Network by Lee et al. [LZYvdS18]



(b) Modified DeepHit

Figure 6.4: Original and modified DeepHit Networks. The original DeepHit network illustration was recreated from the original work [LZYvdS18]

$$C = pr(g(Z_1) > g(Z_2)|T_2 > T_1) \tag{6.20}$$

When there's right censoring, a modified version of C-index is considered with a fixed, pre-specified follow-up period $(0, \tau)$ where [UCP+11]

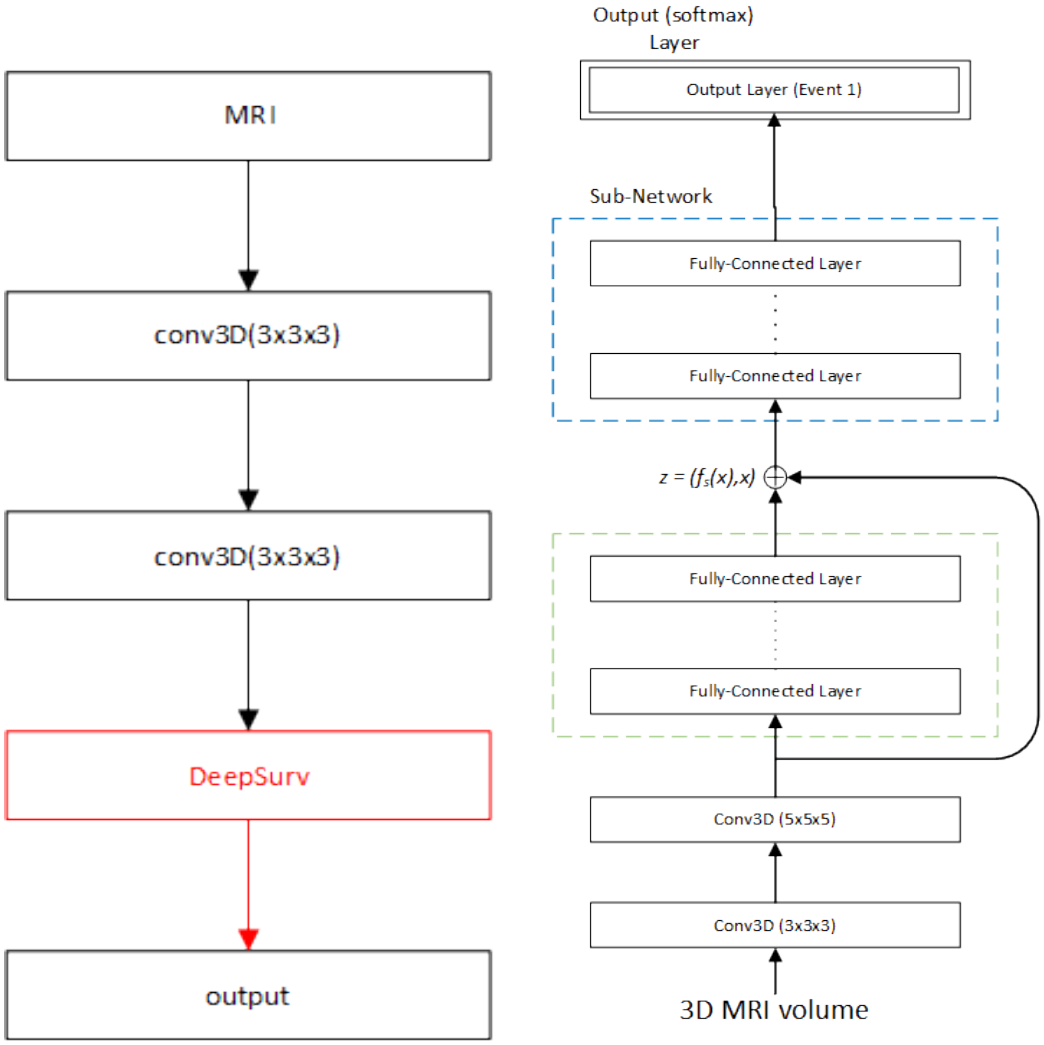
$$C_\tau = pr(g(Z_1) > g(Z_2)|T_2 > T_1, T_1 < \tau) \tag{6.21}$$

6.2.3 Survival Analysis using MRI based Features

The main contribution of this work is the use of MRI based features for survival analysis. This is particularly important in the context of dementia and Alzheimer's disease, as imaging is considered to be the cheaper and quicker alternative to conducting a battery of neuropsychological tests which takes longer and costs more. Two approaches are attempted with MRI based features: (i) a convolutional layer is directly used to capture the MRI based features and (ii) MRI based features are extracted from the best models trained in section 5.2.3 and used as input to survival analysis. DeepSurv is used as the baseline survival analysis method followed by DeepHit. The architecture diagram for the convolutional layer based DeepSurv network and DeepHit network are depicted in Figure 6.5, and the feature extraction methods used for the experiments are discussed below.

6.2.3.1 Feature Extraction

When 3D MRI images are directly used in the end-to-end survival analysis network, it could be argued that the number of image samples is insufficient to learn the risk function due to the complexity of the MRI images and the shallow network



(a) The Architecture of the CNN based DeepSurv network

(b) The Architecture of CNN based DeepHit network

Figure 6.5: End-to-end 3D MRI based survival analysis using DeepSurv and DeepHit. The parameters were optimized against a validation set and the displayed parameters are those with the best performance. Conv3D stands for a 3D convolution kernel and Conv3D(3x3x3) stand for a field-of-view of 3x3x3

used. Therefore, extracted features were also used separately to train DeepSurv in order to evaluate the performance against the end-to-end model. In doing this, two methods of feature extraction were attempted. From using the best deep learning network trained in section 5.2.2, features were extracted from the last layer before the classification layer. However, as the DeepFusion network yielded better results overall, it is also possible to extract fused features, with both MRI based features and NM features fused. The corresponding architectural diagrams are in Figure 6.6.

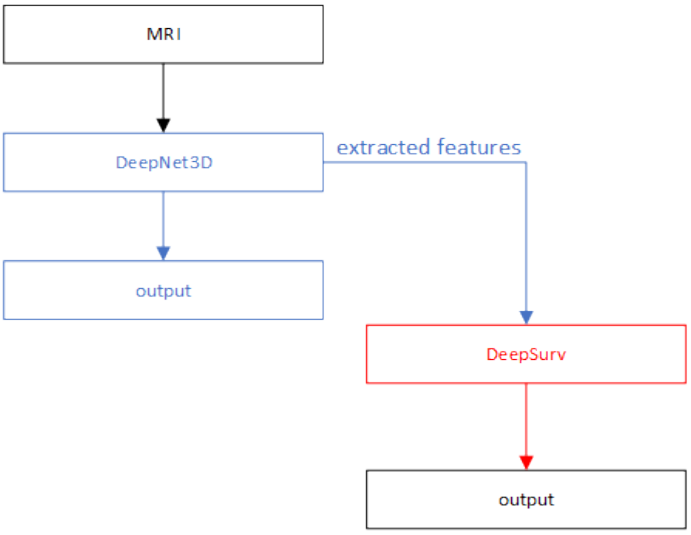
6.2.4 Validation of Results

The comparison experiments were carried out with the same subset across the different techniques attempted. Each experiment was ten-fold cross validated to generalize the results and reduce variance. The folds were randomly picked. The results presented in section 6.3 show the averaged results.

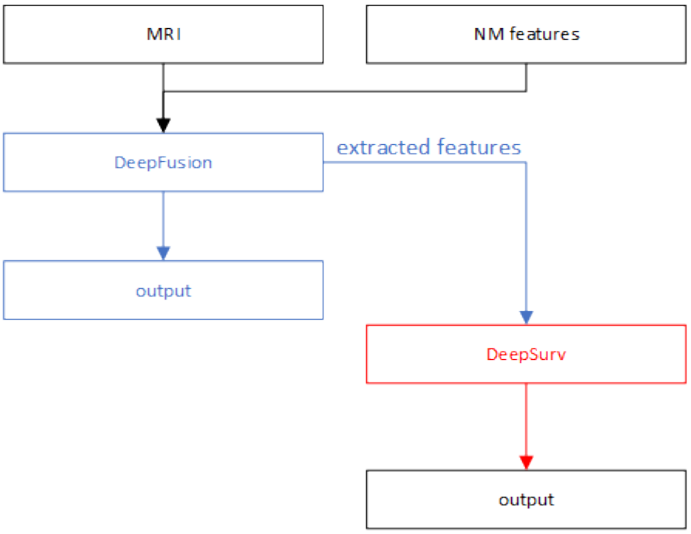
6.3 Results

The intention of this chapter was to take inspiration from statistics based survival analysis practice that is prevalent in clinical studies, and explore its application to predict the survival of a cognitively normal individual with respect to progressing to dementia. There were two datasets that were available for experimentation, namely, the ADNI and MAS datasets. Each of these datasets contains multiple modalities, namely, NM scores and structural MR images. Four techniques were used to compare and contrast the performance between them in the task of survival analysis.

First, a comparison between the different techniques applied to the ADNI dataset is presented in Figure 6.7. Although only two types of data modalities were considered, the results are for three post-processed modalities, namely, the NM fea-



(a) Extracting 3D MRI based features



(b) Extracting fused features using 3D MRI and NM features

Figure 6.6: The feature extraction and training pipeline for MRI based survival analysis

tures alone, extracted MRI image based features alone and fusion based features. An end-to-end deep learning network was designed for structural MRI, where the 3D MRI volume was directly fed as input, and the results for this approach are presented in Figure 6.8. As is evident, this did not result in good performance, as the shallow network architecture prevalent in survival analysis coupled with 3D convolution was not complex enough to learn the feature representation and the survival function concurrently. This is the reason why extracted features from a pre-trained 3D convolutional neural network, discussed in Chapter 5 was used next. The third set of features was derived by extracting features from the deep fusion network where the latent representation from both NM features and MRI volumes were considered. The pre-trained network used for this purpose was presented in sections 5.2.2 and 5.2.3.

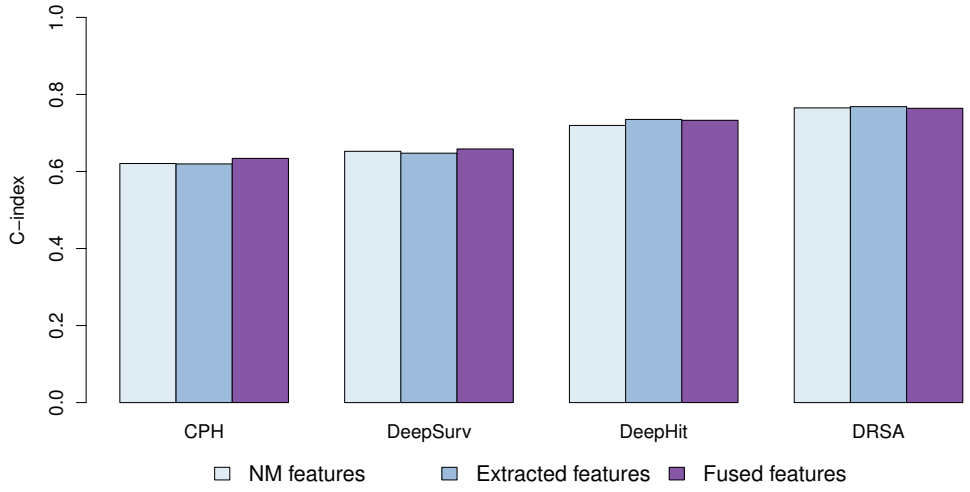


Figure 6.7: Comparison of different techniques used for survival analysis using multiple modalities of data for ADNI dataset

CPH was used as the baseline and as expected, it has the lowest performance across all feature modalities as evident from Figure 6.7. DeepSurv, a deep neural network technique inspired by CPH has performed slightly better but on par with CPH. Both DeepHit and DRSA are deep neural network based methods and they have performed very well on this task. It should be noted that while CPH, DeepSurv

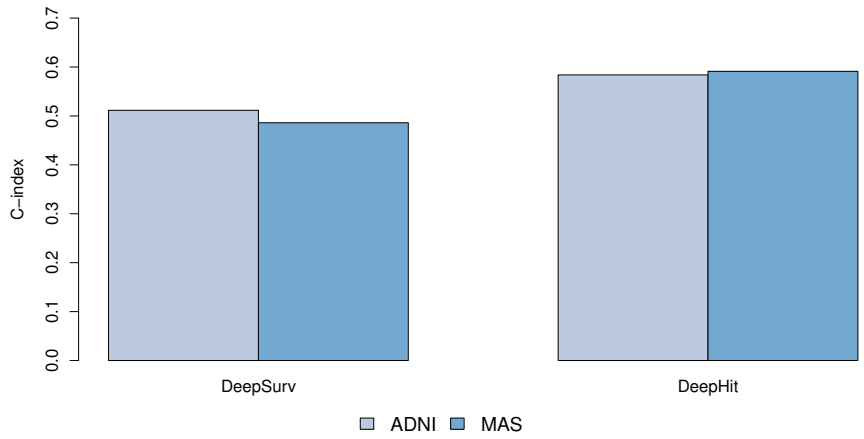


Figure 6.8: The performance of modified DeepSurv and DeepHit techniques using 3D MRI volumes as input

and DeepHit only use the baseline covariates of the patients, DRSA is capable of using the subsequently available covariates as well. The recurrent cells used in DRSA thus enable the capture of time-varying features, which traditional survival analysis lacks. The best performing survival analysis method is DRSA, although DeepHit is close.

The same experiments were repeated on the MAS dataset in order to verify the performance of the different techniques. As presented in Figure 6.9, DRSA is indeed the best performing survival analysis method for MAS as well. The gap between the performance of DeepHit and DRSA has also widened compared to performance on ADNI. DeepSurv closely follows the performance of CPH as expected. In order to isolate these variations across datasets, the results based on extracted features and fused features are presented in Figure 6.10. These results shed light on the effect of input modalities on the techniques applied and the value of determining the complexity of the techniques and matching them to the data complexity and size. With both extracted features and fused features, ADNI dataset provides better results than MAS in Figure 6.10. However, when it comes to DeepHit and DRSA, MAS dataset has better performance than ADNI. The MRI volumes used in the MAS study was of 3T resolution, while the ones in ADNI study were of 1.5T resolution.

Given the difference in resolution, it can be clearly deduced that the amount of information input to the model is richer for the MAS dataset than the ADNI dataset. This in turn means that the extracted features could be a richer latent representation in MAS compared to ADNI. It is more than likely that this is why MAS dataset has better performance on both DeepHit and DRSA, which are models capable of consuming complex input information. On the other hand, both CPH and DeepSurv are comparatively shallow techniques that are unable to fully utilize the richer data. Therefore, it is important that the complexity of the technique is matched to the dataset complexity.

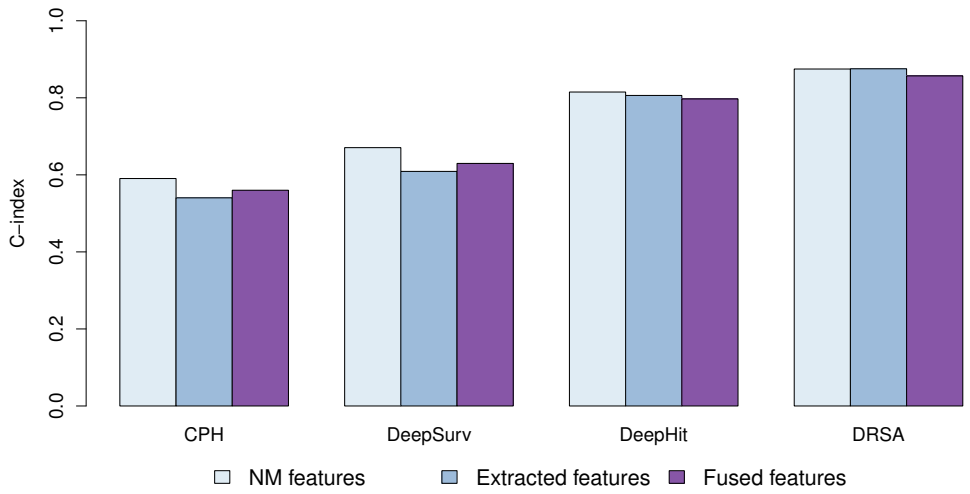
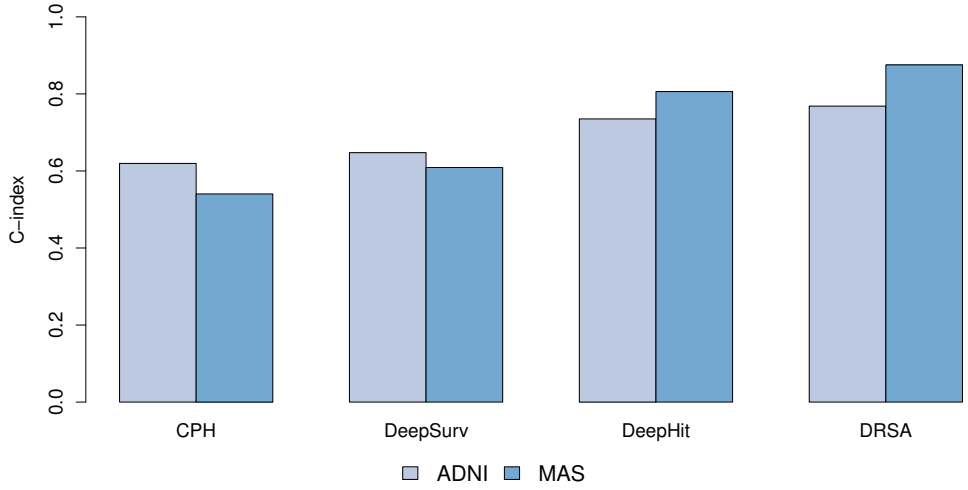
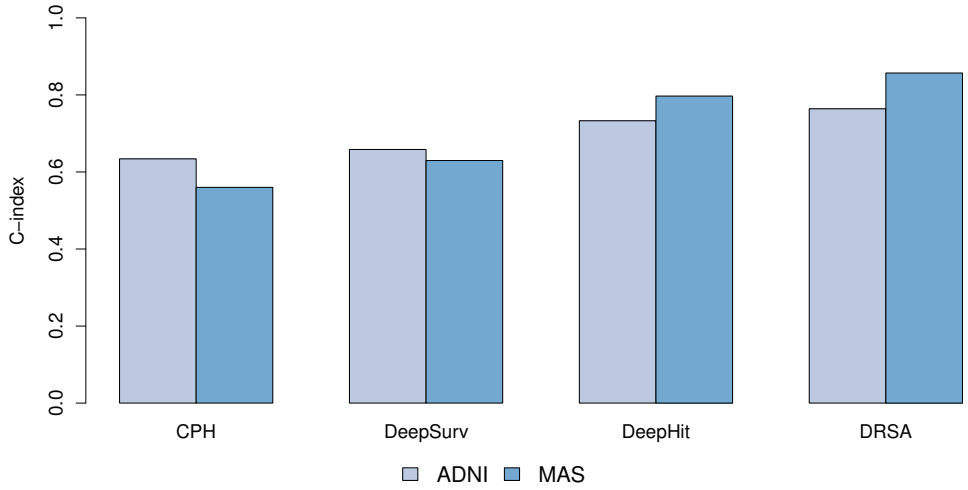


Figure 6.9: Comparison of different techniques used for survival analysis using multiple modalities of data for MAS dataset

Finally, the same pattern is confirmed by comparing the results of the best performing technique, namely DRSA, for both MAS and ADNI. As depicted in Figure 6.11, it is evident that the models trained on the MAS dataset outperforms those trained on the ADNI dataset.



(a) Comparison of performance when using extracted features from 3D MRI based CNN



(b) Comparison of performance when using extracted features from DeepFusion network

Figure 6.10: Comparison between the performance of extracted features from 3D MRI based CNN and DeepFusion network

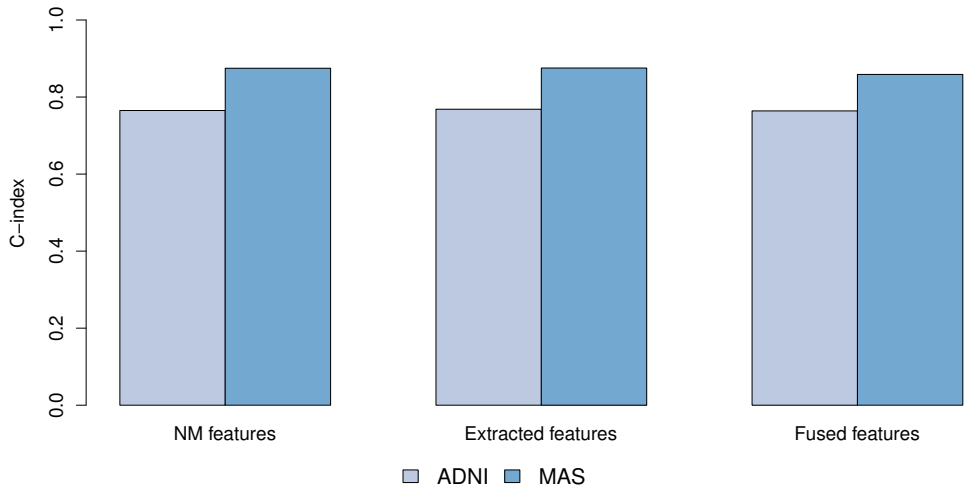


Figure 6.11: The performance of DRSA model for MAS and ADNI across different feature modalities

6.4 Summary

In this chapter, the survival of a cognitively normal patient with respect to dementia and MCI was studied. The event of interest considered is the onset of dementia or its prodromal conditions such as MCI. Survival analysis models estimate the survival and hazard functions for patients. These can then be utilized to determine the risk of progression to dementia for a cognitively normal patient. Traditionally, this is done with shallow data modalities such as NM scores and methods such as proportional hazard models. While such models are optimized for shallow data modalities, they are neither optimized for more complex data modalities such as medical images nor are they free of underlying assumptions such as proportionality or absence of competing causes.

This chapter proposes a collection of novel techniques that can be used to overcome the inherent assumptions and utilize complex data modalities at the same time. This is particularly important for dementia progression prediction, because a

battery of neuropsychological tests can be both cost prohibitive and time consuming to conduct, which does not bode well for elderly individuals. As far as is known, this is the first effort to incorporate a medical imaging modality such as MRI into deep survival analysis. It was demonstrated that deep learning techniques such as DeepHit and DRSA can effectively consume more complex data modalities. It was established that the ability to use time-varying covariates can significantly improve the performance, as demonstrated by the performance difference between DeepHit and DRSA. In this regard, it is recommended to use DRSA as the technique of choice for survival analysis if there are time-varying covariates. However, DeepHit can be considered as a comparable technique if the study does not have curated time-varying covariates and only contains baseline data.

Finally, the efficacy of using extracted features from structural MRI images and fused features from the deep fusion network which fused both structural MRI and NM features was demonstrated. This cements the main effort of this chapter, by proposing a technique to handle complex input data models with survival analysis. This also adds further credibility to the experiments and the work carried out in Chapter 5, as the trained networks therein were used to extract features that are also useful in survival analysis, confirming that the networks have indeed learned the latent representation of the input data space.

Chapter 7

Conclusion

The goal of this thesis was to design, build and test a set of computer aided diagnostic techniques for differentiating cognitively normal individuals from individuals with mild cognitive impairment and dementia, and track their progression to dementia using survival analysis techniques. Two datasets were used in this work: an in-house dataset and a publicly available dataset, which allows additional validation of the proposed techniques. As far as is known, neuropsychological measure based features have not been used to differentiate between MCI subtypes and dementia in a machine learning or deep learning setting, even though they play a large part in diagnosis. Further, fusing of information from disparate data modalities such as NM data and MR images has not been reported before. This was a strong motivation for the thesis, as it allows researchers to leverage all information modalities in a computer aided diagnostic system rather than relying on a single data modality at a time. This thesis presents an end-to-end deep fusion pipeline that is capable of fusing information from multiple modalities and validates the model using two datasets. Finally, this thesis also explores deep survival analysis techniques to better understand the progression to dementia. As far as is known, this is the first attempt to use deep survival analysis for tracking cognitive impairment and dementia diagnosis. This is also the first time

that survival analysis is proposed with MR images for the same diagnoses, as far as can be ascertained.

The remaining chapter is organized as follows. A summary of the thesis is presented in section 7.1 while thesis contributions are described in section 7.2. Limitations and possible future work are reported in section 7.3 and section 7.4 concludes the thesis.

7.1 Thesis Summary

A set of techniques that can be used for diagnosis of cognitive impairment and dementia and their progression tracking has been presented in this thesis. This includes novel methods to train automated classification systems using multi-modal data and survival analysis techniques to better understand the progression to cognitive impairment and dementia.

First, NM based features were used to train conventional machine learning classifiers to differentiate between individuals with cognitive impairment and those who are cognitively normal. As far as is known, NM based features have not been used for cognitive impairment diagnosis using machine learning. This was used as the baseline and extended to a deep learning pipeline based on architectures such as stacked auto-encoders and convolutional neural networks. This improved on the baseline results, establishing a new benchmark for the in-house dataset. It was shown that while CNNs are often recommended when adjacent features are correlated, they do not have to be in a strict order to leverage the architectures.

A typical problem when using NM based features is that the tests used to acquire the data are both lengthy and expensive. As cognitive impairment is prevalent in the older population, it is also important to make the testing as non-

intrusive as possible. To that end, a non-invasive medical imaging technique like structural MR images can be a huge improvement. Traditionally, neuropsychologists use MR images to diagnose cognitive impairment only when the NM based features values are ambiguous. This thesis proposes an end-to-end deep learning pipeline to train a computer aided diagnostics system with MR images using 2D and 3D convolutional neural networks. The added advantage of this approach is that the need for feature extraction or engineering is eliminated. The implementation of the deep fusion pipeline enabled the seamless fusion of data from multiple disparate modalities of vastly different dimensions such as NM based features and MR images. Although fusion has been attempted before for different problems, it was between similar data modalities such as MR images and PET images. The challenge in fusing data from disparate modalities is in ensuring that neither modality over-saturates the learning process. This was achieved through the architecture proposed by this thesis.

The deep fusion pipeline was also demonstrated to be robust against missing information. This enables patients with incomplete NM based features to benefit from the computer aided diagnostics system, which is otherwise impossible. Additionally, the deep fusion pipeline demonstrated the efficacy of transfer learning by model training on the ADNI dataset and finetuning of the pre-trained model on a smaller dataset such as the MAS dataset. This effectively addresses the data paucity problem that researchers typically face in medical imaging studies.

Another important motivation for this thesis was to better understand the progression to dementia from MCI and normal cognition. Survival analysis techniques were employed for this purpose. First, conventional survival analysis techniques were used to determine the time to survival (or time till progression to cognitive impairment or dementia in this case). These initial results were then used as the baseline and the method extended using deep learning techniques to achieve better

and more robust results. As far as is known, this is the first time that survival analysis techniques have been used to understand the progression of cognitive decline. The techniques were designed so that the structural MR images and fused features can be used for survival analysis, which also demonstrates another use case for the deep fusion pipeline as well as making feature engineering redundant.

7.2 Thesis Contributions

While accomplishing the set goals, the thesis also contributes to two different fields; computer vision, machine learning and deep learning, and cognitive impairment diagnosis and prediction.

7.2.1 Contributions to Computer Vision, Machine Learning and Deep Learning

The main contributions of this thesis to computer vision, machine learning and deep learning are summarized below.

i Creating a baseline for classification systems using NM based features for cognitive impairment diagnosis.

ii Design of 1D, 2D and 3D convolutional neural networks that uniquely exploit the characteristics of the data modalities they are trained on for cognitive impairment diagnosis.

iii Establishing the limitations of 2D convolutional neural networks trained on low resolution 2D structural MR images.

iv Establishing the utility and applicability of machine learning and deep learning algorithms for cognitive impairment diagnosis.

v Design of an end-to-end deep fusion pipeline that fuses information from disparate multiple modalities of vastly different dimensions such as NM based features and MR images.

vi Establishing that the proposed deep fusion pipeline is robust to missing information.

vii Establishing a baseline for transfer learning based model fine-tuning for deep fusion pipeline across two datasets.

viii Design of deep learning based survival analysis techniques to better understand the progression to cognitive impairment and dementia that uses MR images and fused features without having to extract or engineer features by itself.

7.2.2 Contributions to Cognitive Impairment Diagnosis and Prediction

This thesis also makes important contributions to cognitive impairment diagnosis and prediction as summarised below.

i Development of a computer aided classification system to diagnose cognitive impairment using NM based features.

ii Development of a computer aided classification system to diagnose cognitive impairment using multi-modal data such as NM based features and MR images.

iii Demonstrating a computer aided classification system that is robust to missing information.

iv Development of a computer aided survival analysis system to better understand the progression to cognitive impairment and dementia.

v Establishing and facilitating a range of computer aided diagnostic techniques for the in-house MAS dataset, with ADNI as a benchmark.

7.3 Limitations and Future Work

The set of techniques proposed in this thesis effectively diagnoses stages of cognitive decline and elicits better understanding of progression to dementia. As two datasets were used for this thesis, cross comparison between models was possible. However, the datasets used also have some limitations, with the two datasets having some differences in features such as different NM features or differing resolutions of MR images. This is a potential limitation, and the performance of the proposed system could be further improved if the two datasets were similar. Another limitation of the survival analysis technique stems from the fact that only four time points were used for the MAS dataset, which limits the use of advanced deep learning techniques such as recurrent neural networks to better model the longitudinal nature of the dataset. The scope for future work includes the following:

i The survival analysis techniques should be validated on a lengthier longitudinal dataset to improve the performance.

ii The deep fusion pipeline can utilize higher resolution MR images if available, which can be used to improve the performance.

iii Techniques such as Generative Adversarial Networks (GANs) can be used to improve the quality of the dataset, effectively improving the performance of the whole system.

iv Techniques such as GANs can also increase the size of the dataset by generating new data which can address the limited dataset size.

v Active learning techniques can be developed to incorporate a feedback loop from the clinicians, in order to continuously improve the system performance.

vi Personalised modelling techniques may be adapted to better understand progression to dementia as more longitudinal data becomes available.

7.4 Concluding Remarks

Automated diagnosis of cognitive impairment and dementia continues to be important. The techniques and frameworks presented in this thesis build on earlier attempts while also addressing the shortcomings identified. Novel methods have been proposed, implemented and validated, including NM based cognitive impairment diagnosis, a deep fusion pipeline to fuse data from multiple disparate modalities and deep survival analysis based techniques to better understand the progression to dementia. The comprehensive battery of experimental results show that the proposed methods are effective solutions and could further benefit from more data and experiments.

Appendix A

AUC values for Figures 4.4, 4.5, 4.6 and 4.7 are tabulated in Table A1, A2, A3 and A4.

	CN vs MCI	CN vs aMCI	CN vs naMCI	MCI Subtypes	aMCI Subtypes	naMCI Subtypes
AB	0.93	0.97	0.85	0.95	0.87	0.74
RF	0.94	0.98	0.95	0.94	0.89	0.82
ES	0.92	0.98	0.94	0.95	0.88	0.8

Table A1: AUC for the three best conventional machine learning methods considered in one-vs-one classification on first wave

	naMCI	aMCI	md-aMCI	md-naMCI	sd-aMCI	sd-naMCI
AB	0.77	0.96	0.73	0.55	0.73	0.66
RF	0.91	0.97	0.92	0.94	0.92	0.86
ES	0.9	0.98	0.9	0.93	0.91	0.84

Table A2: AUC for the three best conventional machine learning methods considered in one-vs-all classification on first wave

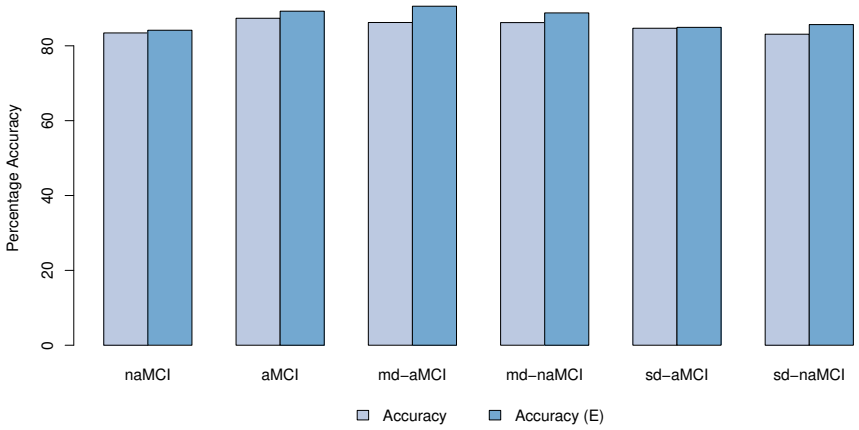
	CN vs MCI	CN vs aMCI	CN vs naMCI	MCI Subtypes	aMCI Subtypes	naMCI Subtypes
W1	0.93	0.97	0.85	0.95	0.87	0.74
W2	.95	.96	.84	.92	.88	.79
W3	.91	.93	.86	.89	.87	.75
W4	0.88	.91	0.87	0.90	0.86	0.75

Table A3: AUC for the best conventional machine learning method (AB)in one-vs-one classification across four waves

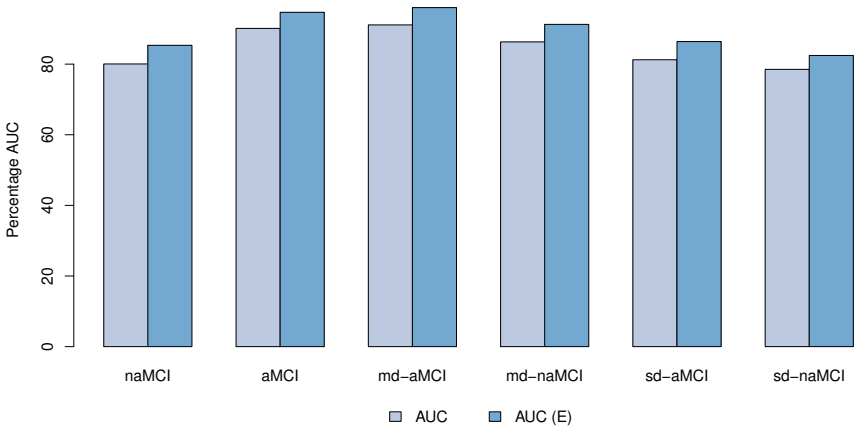
	naMCI	aMCI	md-aMCI	md-naMCI	sd-aMCI	sd-naMCI
W1	0.77	0.96	0.73	0.55	0.73	0.66
W2	0.81	.92	.78	.69	.78	.72
W3	.8	.93	.82	.68	.81	.75
W4	.83	.91	.79	.70	.81	.71

Table A4: AUC for the best conventional machine learning method (AB)in one-vs-all classification across four waves

An ensemble of SAEs at the model level was built to improve performance in Chapter 4 where Figure 4.12 depicted the results on wave 3. The results on wave 1, 2 and 4 are presented in Figures A1, A2 and A3.

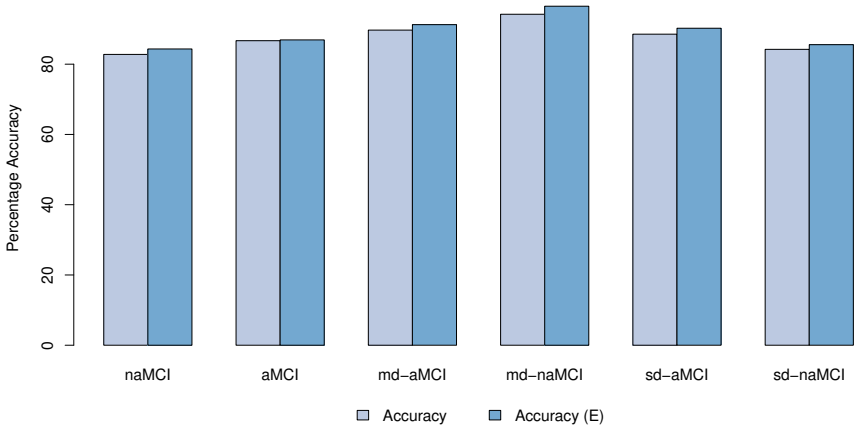


(a) Percentage Accuracy of SAE classifier and SAE Ensemble classifier

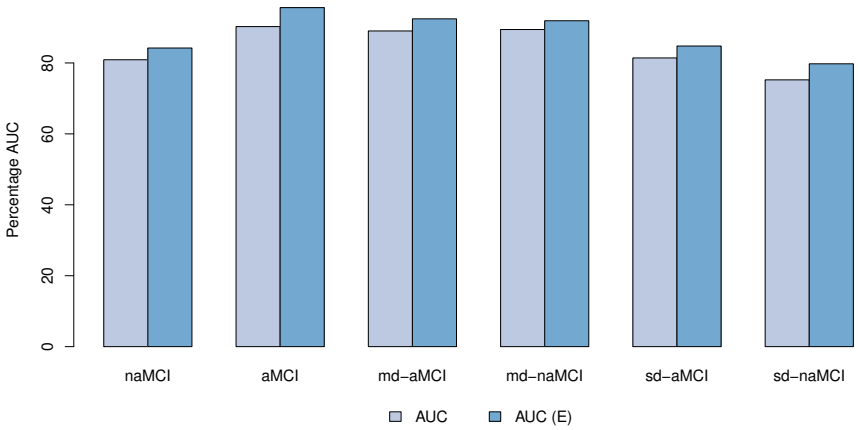


(b) Percentage AUC of SAE classifier and SAE Ensemble classifier

Figure A1: Comparison of best SAE classifier results against SAE Ensemble classifier results for Wave 1. Accuracy (E) and AUC (E) stands for the accuracy and AUC of the SAE ensemble classifier

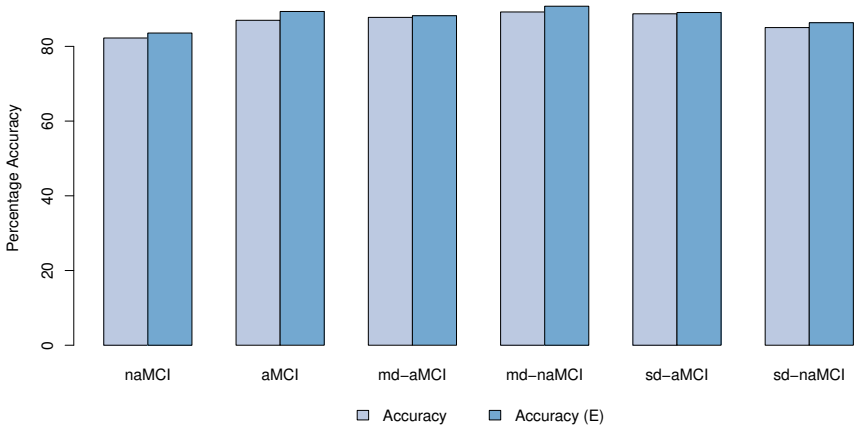


(a) Percentage Accuracy of SAE classifier and SAE Ensemble classifier

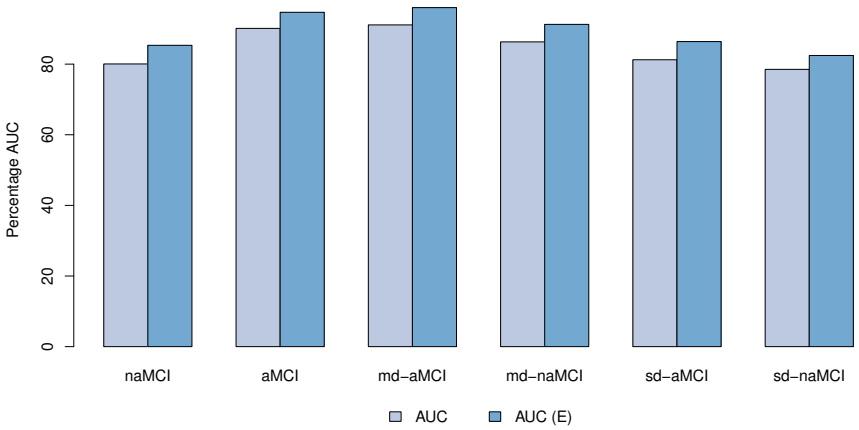


(b) Percentage AUC of SAE classifier and SAE Ensemble classifier

Figure A2: Comparison of best SAE classifier results against SAE Ensemble classifier results for Wave 2. Accuracy (E) and AUC (E) stands for the accuracy and AUC of the SAE ensemble classifier



(a) Percentage Accuracy of SAE classifier and SAE Ensemble classifier



(b) Percentage AUC of SAE classifier and SAE Ensemble classifier

Figure A3: Comparison of best SAE classifier results against SAE Ensemble classifier results for Wave 4. Accuracy (E) and AUC (E) stands for the accuracy and AUC of the SAE ensemble classifier

Measures other than accuracy used to evaluate the model performance in Chapter 4 are reported in Table A5.

		CN vs MCI		CN vs aMCI		CN vs naMCI		MCI Subtypes		aMCI Subtypes		naMCI Subtypes	
		TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR
2D CNN	Wave 1	93.8492	14.7147	92.2619	13.5135	94.246	18.9189	81.3953	18.9189	85.4369	12.1951	94.2623	11.5385
	Wave 2	86.4198	15.3846	91.8519	8.609	94.5545	7.3394	89.404	13.7615	92.4731	13.7931	96.5517	9.09
	Wave 3	89.94	14.705	91.7989	16.66	93.915	21.25	84.44	7.5	97.72	8.69	100	0
	Wave 4	90.102	17.948	92.832	13.33	91.808	12.22	83.80	8.889	96.82	7.142	94.73	7.142
1D CNN	Wave 1	86.904	14.41	96.428	9.189	89.881	12.837	88.648	8.108	82.524	12.195	97.541	7.692
	Wave 2	91.11	11.15	90.37	10.596	86.67	15.59	92.05	10.09	86.021	6.89	96.551	9.09
	Wave 3	94.17	5.29	91.798	10.0	89.68	15.0	88.89	6.25	93.18	10.86	95.31	12.5
	Wave 4	91.467	11.28	93.51	10.47	85.66	15.56	89.52	4.44	88.89	9.52	94.73	7.14
AdaBoost	Wave 1	89.68	10.81	96.03	7.02	91.07	6.08	89.18	4.05	78.64	10.97	86.88	15.38
	Wave 2	89.38	10.76	96.79	5.96	90.37	9.17	93.37	11.0	80.64	12.06	82.75	18.18
	Wave 3	89.94	11.17	97.35	4.44	91.26	7.5	91.11	8.75	77.27	17.39	75.0	25.0
	Wave 4	88.72	11.79	94.53	6.67	89.07	7.78	83.80	11.11	76.19	21.42	84.21	14.28

Table A5: Comparison of Results for 1D CNN, 2D CNN and AdaBoost across multiple waves using True Positive Rate (TPR) and False Positive Rate (FPR)

Bibliography

[ada84] A new rating scale for alzheimer's disease. *American Journal of Psychiatry*, 141(11):1356–1364, 1984. PMID: 6496779.

[ADD+11] Marilyn S Albert, Steven T DeKosky, Dennis Dickson, Bruno Dubois, Howard H Feldman, Nick C Fox, Anthony Gamst, David M Holtzman, William J Jagust, Ronald C Petersen, Peter J Snyder, Maria C Carrillo, Bill Thies, and Creighton H Phelps. The diagnosis of mild cognitive impairment due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, 7(3):270–279, December 2011.

[AHK+17] Liesbeth Aerts, Megan Heffernan, Nicole A. Kochan, John D. Crawford, Brian Draper, Julian N. Trollor, Perminder S. Sachdev, and Henry Brodaty. Effects of mci subtype and reversion on progression to dementia in a community sample. *Neurology*, 88(23):2225–2232, 2017.

[ALLF07] Andrew L. Alexander, Jee Eun Lee, Mariana Lazar, and

Aaron S. Field. Diffusion tensor imaging of the brain. *Neurotherapeutics*, 4(3):316–329, Jul 2007. 17599699[pmid].

[APATFoNaAPACoNa00] Statistics American Psychiatric Association. Task Force on Nomenclature and and Statistics American Psychiatric Association. Committee on Nomenclature and. *Diagnostic and statistical manual of mental disorders*. DSM-IV. Washington, D.C. : American Psychiatric Association, Washington, D.C., 4th ed. edition, 2000. Includes bibliographical references and index.

[ARS+14] Hossein Azizpour, Ali Sharif Razavian, Josephine Sullivan, Atsuto Maki, and Stefan Carlsson. Factors of Transferability for a Generic ConvNet Representation. *arXiv:1406.5774 [cs]*, June 2014. arXiv: 1406.5774.

[aSWLM15] Carina a. S. Wattmo, Elisabet Y Londos, and Lennart Minthon. Longitudinal associations between survival in alzheimer's disease and cholinesterase inhibitor use, progression, and community-based services. *Dementia and geriatric cognitive disorders*, 40 5-6:297–310, 2015.

[Axe01] Bradley N Axelrod. Administration duration for the wechsler adult intelligence scale-iii and wechsler memory scale-iii. *Archives of Clinical Neuropsychology*, 16(3):293 – 301, 2001.

[Bal12] Pierre Baldi. Autoencoders, Unsupervised Learning, and Deep Architectures. *ICML Unsupervised and Transfer Learning*, pages 37–50, 2012.

[Ber84] Leonard Berg. Clinical dementia rating. *The British journal of psychiatry : the journal of mental science*, 145:339, 10 1984.

[BGS+87] Nelson Butters, Eric Granholm, David Salmon, Igor Grant, and Jessica Wolfe. Episodic and semantic memory: A comparison of amnesic and demented patients. *Journal of clinical and experimental neuropsychology*, 9:479–97, 11 1987.

[Bis07] Christopher M. Bishop. Pattern Recognition and Machine Learning. *Journal of Electronic Imaging*, 16(4):049901, 2007.

[Bow14] F. DuBois Bowman. Brain Imaging Analysis. *Annual Review of Statistics and Its Application*, 1:61–85, 2014.

[BT04] Eric Bair and Robert Tibshirani. Semi-supervised methods to predict patient survival from gene expression data. *PLOS Biology*, 2(4), 04 2004.

[BWP+13] Roberta Biundo, Luca Weis, Manuela Pilleri, Silvia Facchini, Patrizia Formento-Dojot, Annamaria Vallelunga, and Angelo Antonini. Diagnostic and screening power of neuropsychological testing in detecting mild cognitive impairment in parkinson's disease. *Journal of Neural Transmission*, 03 2013.

[Bü12] Peter Bühlmann. Bagging, boosting and ensemble methods. *Handbook of Computational Statistics*, 01 2012.

[CAIH+16] Chatree Chai-Adisaksopha, Alfonso Iorio, Christopher Hillis, Wendy Lim, and Mark Crowther. A systematic

review of using and reporting survival analyses in acute lymphoblastic leukemia literature. *BMC Hematology*, 16, 12 2016.

[Cha19] Niel Chah. Down the deep rabbit hole: Untangling deep learning from machine learning and artificial intelligence. *First Monday*, 24(2), 2019.

[CLE+05] G Chételat, B Landeau, F Eustache, F Mézenge, F Viader, V de la Sayette, B Desgranges, and J-C Baron. Using voxel-based morphometry to map the structural changes associated with rapid conversion in MCI: a longitudinal MRI study. *NeuroImage*, 27(4):934–46, October 2005.

[CMMK05] John C. Marshall and Gillian Morriss-Kay. Functional anatomy of the brain. *Journal of Anatomy*, 207:1 – 2, 07 2005.

[CNMCK04] Rich Caruana, Alexandru Niculescu-Mizil, Geoff Crew, and Alex Ksikes. Ensemble selection from libraries of models. In *Proceedings of the Twenty-first International Conference on Machine Learning*, ICML '04, pages 18–, New York, NY, USA, 2004. ACM.

[Cox72] D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

[CSL+12] Yue Cui, Perminder S. Sachdev, Darren M. Lipnicki, Jesse S. Jin, Suhuai Luo, Wanlin Zhu, Nicole a. Kochan, Simone Reppermund, Tao Liu, Julian N. Trollor, Henry Brodaty, and Wei Wen. Predicting the development of

mild cognitive impairment: A new use of pattern recognition. *NeuroImage*, 60(2):894–901, 2012.

[CSM+96] Deborah A. Cahn, David P. Salmon, Andreas U. Monsch, Nelson Butters, W.C. Wiederholt, Jody Corey-Bloom, and Elizabeth Barrett-Connor. Screening for dementia of the alzheimer type in the community: The utility of the clock drawing test. *Archives of Clinical Neuropsychology*, 11(6):529 – 539, 1996.

[CV95] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Mach. Learn.*, 20(3):273–297, September 1995.

[CWC+09] Terence C Chua, Wei Wen, Xiaohua Chen, Nicole Kochan, Melissa J Slavin, Julian N Trollor, Henry Brodaty, and Perminder S Sachdev. Diffusion tensor imaging of the posterior cingulate is a useful biomarker of mild cognitive impairment. *The American journal of geriatric psychiatry : official journal of the American Association for Geriatric Psychiatry*, 17(July):602–613, 2009.

[CWL+12] Yue Cui, Wei Wen, Darren M. Lipnicki, Mirza Faisal Beg, Jesse S. Jin, Suhuai Luo, Wanlin Zhu, Nicole a. Kochan, Simone Reppermund, Lin Zhuang, Reddy Raamana, Tao Liu, Julian N. Trollor, Lei Wang, Henry Brodaty, and Perminder S. Sachdev. Automated detection of amnestic mild cognitive impairment in community-dwelling elderly adults: A combined spatial atrophy and white matter alteration approach. *NeuroImage*, 59(2):1209–1217, 2012.

[CWSS08] Terence C Chua, Wei Wen, Melissa J Slavin, and Perminder S Sachdev. Diffusion tensor imaging in mild cognitive

impairment and Alzheimer's disease : a review. *Current Opinions in Neurology*, 2008.

[CYA13] Wei-Yi Cheng, Tai-Hsien Ou Yang, and Dimitris Anastassiou. Development of a prognostic model for breast cancer survival in an open challenge environment. *Science Translational Medicine*, 5(181):181ra50–181ra50, 2013.

[DBS+11] Christos Davatzikos, Priyanka Bhatt, Leslie M. Shaw, Kayhan N. Batmanghelich, and John Q. Trojanowski. Prediction of MCI to AD conversion, via MRI, CSF biomarkers, and pattern classification. *Neurobiology of Aging*, 32(12):2322.e19–2322.e27, 2011.

[dMSB+XX] Leyla deToledo Morrell, T. R. Stoub, M. Bulgakova, R. S. Wilson, D. A. Bennett, S. Leurgans, J. Wuu, and D. A. Turner. Mri-derived entorhinal volume is a good predictor of conversion from mci to ad. *Neurobiology of Aging*, 25(9):1197–1203, 2015/06/14 XXXX.

[EGKB+03] Armando Estévez-González, Jaime Kulisevsky, Anunciación Boltes, Pilar Otermín, and Carmen García-Sánchez. Rey verbal learning test is a useful tool for differential diagnosis in the preclinical phase of alzheimer's disease: comparison with mild cognitive impairment and normal aging. *International Journal of Geriatric Psychiatry*, 18(11):1021–1028, 2003.

[Elw91] Richard W. Elwood. The wechsler memory scale—revised: Psychometric characteristics and clinical application. *Neuropsychology Review*, 2(2):179–201, Jun 1991.

[EWT+12] Michael Ewers, Cathal Walsh, John Q. Trojanowski, Leslie M. Shaw, Ronald C. Petersen, Clifford R. Jack, Howard H. Feldman, Arun L W Bokde, Gene E. Alexander, Philip Scheltens, Bruno Vellas, Bruno Dubois, Michael Weiner, and Harald Hampel. Prediction of conversion from mild cognitive impairment to Alzheimer's disease dementia based upon biomarkers and neuropsychological test performance. *Neurobiology of Aging*, 33(7):1203–1214.e2, 2012.

[EZELL+15] Lubov E Zeifman, William Eddy, Oscar L Lopez, Lewis H Kuller, Cyrus Raji, Paul Thompson, and James Becker. Voxel level survival analysis of grey matter volume and incident mild cognitive impairment or alzheimer's disease. *Journal of Alzheimer's disease : JAD*, 46, 02 2015.

[fDCP19] Centers for Disease Control and Prevention. Cognitive impairment: A call for action, now!, 08 2019.

[FDM98] Philip S. Fastenau, Natalie L. Denburg, and Beth A. Mauer. Parallel short forms for the boston naming test: Psychometric properties and norms for older adults. *Journal of Clinical and Experimental Neuropsychology*, 20(6):828–834, 1998. PMID: 10484693.

[FFM75] Marshal F. Folstein, Susan E. Folstein, and Paul R. McHugh. "mini-mental state": A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, 12(3):189 – 198, 1975.

[FL99] Lloyd D. Fisher and D. Y. Lin. Time-dependent covariates in the cox proportional-hazards regression model. *An-*

nual Review of Public Health, 20(1):145–157, 1999. PMID: 10352854.

[FS99] Yoav Freund and Robert E. Schapire. A short introduction to boosting, 1999.

[Fym18] Alain Fymat. Dementia: A review. *Journal of Clinical Psychiatry and Neuroscience*, pages 27–34, 05 2018.

[GKK10] Manish Goel, Pardeep Khanna, and Jugal Kishore. Understanding survival analysis: Kaplan-meier estimate. *International journal of Ayurveda research*, 1:274–8, 10 2010.

[GMH13] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech Recognition With Deep Recurrent Neural Networks. *Icassp*, (3):6645–6649, 2013.

[GRZ+06] Serge Gauthier, Barry Reisberg, Michael Zaudig, Ronald C Petersen, Karen Ritchie, Karl Broich, Sylvie Belleville, Henry Brodaty, David Bennett, Howard Chertkow, Jeffrey L Cummings, Mony de Leon, Howard Feldman, Mary Ganguli, Harald Hampel, Philip Scheltens, Mary C Tierney, Peter Whitehouse, and Bengt Winblad. Mild cognitive impairment. *The Lancet*, 367(9518):1262 – 1270, 2006.

[GSRK08] Ellen Grober, Martin Sliwinsk, and Saul R. Korey. Development and validation of a model for estimating premorbid verbal intelligence in the elderly. *Journal of Clinical and Experimental Neuropsychology*, 13:933–949, 01 2008.

[GSS+11] M Ganguli, B E Snitz, J A Saxton, C-C H Chang, C-W Lee, J Vander Bilt, T F Hughes, D A Loewenstein,

F W Unverzagt, and R C Petersen. Outcomes of mild cognitive impairment depend on definition: a population study. *Archives of neurology*, 68(6):761–767, June 2011.

[Gus96] L Gustafson. What is dementia? *Acta Neurologica Scandinavica*, 94:22–24, 1996.

[Hal00] Mark Hall. Correlation-based feature selection for machine learning. *Department of Computer Science*, 19, 06 2000.

[Hel09] Elizabeth Helzner. Survival in alzheimer disease: A multiethnic, population-based study of incident cases (vol 71, pg 1489, 2008). *Neurology*, 72:861–861, 03 2009.

[HG04] Trey Hedden and John D E Gabrieli. Insights into the ageing mind: a view from cognitive neuroscience. *Nat Rev Neurosci*, 5(2):87–96, February 2004.

[Hin98] Ian Hindmarch. The bayer activities of daily living scale (b-adl). *Dementia and Geriatric Cognitive Disorders*, 1998.

[HK14] Yun Hwang and Hyanghee Kim. Utility of the boston naming test in differentiating between mild cognitive impairment and normal elderly: A meta-analysis. *Communication Sciences Disorders*, 19:501–512, 12 2014.

[HLW16] Gao Huang, Zhuang Liu, and Kilian Q. Weinberger. Densely connected convolutional networks. *CoRR*, abs/1608.06993, 2016.

[HMH+13] S Haller, P Missonnier, F R Herrmann, C Rodriguez, M-P Deiber, D Nguyen, G Gold, K-O Lovblad, and P Giannakopoulos. Individual classification of mild cognitive

impairment subtypes by support vector machine analysis of white matter DTI. *AJNR. American journal of neuroradiology*, 34(2):283–91, February 2013.

[HOT06] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. A Fast Learning Algorithm for Deep Belief Nets. *Neural Computation*, 18(7):1527–1554, 2006.

[HS13] Michiel Hermans and Benjamin Schrauwen. Training and Analyzing Deep Recurrent Neural Networks. *Nips*, pages 190–198, 2013.

[HSXJ11] Chris Hinrichs, Vikas Singh, Guofan Xu, and Sterling C. Johnson. Predictive markers for AD in a multi-modality framework: An analysis of MCI progression in the ADNI population. *NeuroImage*, 55(2):574–589, 2011.

[HZRS16] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, June 2016.

[IMS+92a] Robert J. Ivnik, James F. Malec, Glenn E. Smith, Eric G. Tangalos, Ronald C. Petersen, Emre Kokmen, and Leonard T. Kurland. Mayo's older americans normative studies: Updated avlt norms for ages 56 to 97. *Clinical Neuropsychologist*, 6(sup001):83–104, 1992.

[IMS+92b] Robert J. Ivnik, James F. Malec, Glenn E. Smith, Eric G. Tangalos, Ronald C. Petersen, Emre Kokmen, and Leonard T. Kurland. Mayo's older americans normative

studies: Wais-r norms for ages 56 to 97. *Clinical Neuropsychologist*, 6(sup001):1–30, 1992.

[Ini05] Alzheimer's Disease Neuroimaging Initiative. Alzheimer's disease neuroimaging initiative protocol. 2005.

[Ins19] New Zealand Brain Research Institute. Magnetic resonance imaging at 3 tesla - the first in new zealand, 2019.

[JNS19] Taeho Jo, Kwangsik Nho, and Andrew Saykin. Deep learning in alzheimer's disease: Diagnostic classification using neuroimaging data. 05 2019.

[JPX+99] Clifford R. Jack, Ronald C. Petersen, Yue Cheng Xu, Peter C. O'Brien, Glenn E. Smith, Robert J. Ivnik, Bradley F. Boeve, Stephen C. Waring, Eric G. Tangalos, and Emre Kokmen. Prediction of ad with mri-based hippocampal volume in mild cognitive impairment. *Neurology*, 52(7):1397–1403, Apr 1999. 10227624[pmid].

[JSL+14] Jiyang Jiang, Perminder Sachdev, Darren M. Lipnicki, Haobo Zhang, Tao Liu, Wanlin Zhu, Chao Suo, Lin Zhuang, John Crawford, Simone Reppermund, Julian Trollor, Henry Brodaty, and Wei Wen. A longitudinal study of brain atrophy over two years in community-dwelling older individuals. *NeuroImage*, 86:203–211, 2014.

[KCK+00] Daniel I. Kaufer, Jeffrey L. Cummings, Patrick Ketchel, Vanessa Smith, Audrey MacMillan, Timothy Shelley, Oscar L. Lopez, and Steven T. DeKosky. Validation of the npi-q, a brief clinical form of the neuropsychiatric inven-

tory. *The Journal of Neuropsychiatry and Clinical Neurosciences*, 12(2):233–239, 2000. PMID: 11001602.

[KEbS15] Hany Kasban, Mohsen El-bendary, and Dina Salama. A comparative study of medical imaging techniques. *International Journal of Information Science and Intelligent System*, 4:37–58, 04 2015.

[KFFK18] Ashnil Kumar, Michael Fulham, David Dagan Feng Feng, and Jinman Kim. Co-learning feature fusion maps from pet-ct images of lung cancer, 10 2018.

[KFH+14] Nikola Kasabov, Valery Feigin, Zeng-Guang Hou, Yixiong Chen, Linda Liang, Rita Krishnamurthi, Muhaini Othman, and Priya Parmar. Evolving spiking neural networks for personalised modelling, classification and prediction of spatio-temporal patterns with a case study on stroke. *Neurocomputing*, 134:269 – 279, 2014. Special issue on the 2011 Sino-foreign-interchange Workshop on Intelligence Science and Intelligent Data Engineering (IScIDE 2011) Learning Algorithms and Applications.

[KGW83] Edith Kaplan, Harold Goodglass, and Sandra Weintraub. *Boston naming test*. Lea & Febiger, Philadelphia, 1983.

[KH10] Nikola Kasabov and Yingjie Hu. Integrated optimisation method for personalised modelling and case studies for medical decision support. *I. J. Functional Informatics and Personalised Medicine*, 3:236–256, 01 2010.

[KJS18] Oxana Korzh, Mikel Joaristi, and Edoardo Serra. Convolutional neural network ensemble fine-tuning for extended

transfer learning. In Francis Y. L. Chin, C. L. Philip Chen, Latifur Khan, Kisung Lee, and Liang-Jie Zhang, editors, *Big Data – BigData 2018*, pages 110–123, Cham, 2018. Springer International Publishing.

[KLB+18] Qiuhong Ke, Jun Liu, Mohammed Bennamoun, Senjian An, Ferdous Sohel, and Farid Boussaid. Chapter 5 - computer vision for human–machine interaction. In Marco Leo and Giovanni Maria Farinella, editors, *Computer Vision for Assistive Healthcare*, Computer Vision and Pattern Recognition, pages 127 – 145. Academic Press, 2018.

[KMA+15] Yuriko Katsumata, Melissa Mathews, Erin L Abner, Gregory A Jicha, Allison Caban-Holt, Charles D Smith, Peter T Nelson, Richard J Kryscio, Frederick A Schmitt, and David W Fardo. Assessing the discriminant ability, reliability, and comparability of multiple short forms of the boston naming test in an alzheimer's disease center cohort. *Dementia and geriatric cognitive disorders*, 39(3-4):215—227, 2015.

[Kon94] Igor Kononenko. Estimating attributes: Analysis and extensions of relief. In Francesco Bergadano and Luc De Raedt, editors, *Machine Learning: ECML-94*, volume 784 of *Lecture Notes in Computer Science*, pages 171–182. Springer Berlin Heidelberg, 1994.

[Kot07] Sotiris B. Kotsiantis. Supervised machine learning: A review of classification techniques. pages 3–24, 2007.

[KSB+10] Nicole A Kochan, Melissa J Slavin, Henry Brodaty, John D Crawford, Julian N Trollor, Brian Draper, and

Perminder S Sachdev. Effect of Different Impairment Criteria on Prevalence of “Objective” Mild Cognitive Impairment in a Community Sample. *The American Journal of Geriatric Psychiatry*, 18(8):711–722, December 2010.

[KSBD17] S. Korolev, A. Safiullin, M. Belyaev, and Y. Dodonova. Residual and plain convolutional neural networks for 3d brain mri classification. In *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pages 835–838, April 2017.

[KSC+18] Jared L. Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18(1):24, Feb 2018.

[KSH12a] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[KSH12b] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. *Advances In Neural Information Processing Systems*, pages 1–9, 2012.

[Lab16] Stanford Vision Lab. Imagenet dataset, 2016.

[LBBH98] Y LeCun, L Bottou, Yoshua Bengio, and P Haffner. Gradient Based Learning Applied to Document Recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[LBH15a] Yann LeCun, Y Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521:436–44, 05 2015.

[LBH15b] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, may 2015.

[LD14] Dong Yu Li Deng. Deep learning: Methods and applications. Technical report, May 2014.

[Le15] Quoc V Le. A Tutorial on Deep Learning Part 2: Autoencoders, Convolutional Neural Networks and Recurrent Neural Networks. *Tutorial*, pages 1–20, 2015.

[LEA97] Kwan-Moon Leung, Robert M. Elashoff, and Abdelmonem A. Afifi. Censoring issues in survival analysis. *Annual Review of Public Health*, 18(1):83–104, 1997. PMID: 9143713.

[Lea19] Lumen Learning. The four major regions of the brain, 08 2019.

[Lip15] Zachary Chase Lipton. A Critical Review of Recurrent Neural Networks for Sequence Learning. *CoRR*, abs/1506.0:1–38, 2015.

[LL14] Kenneth M. Langa and Deborah A. Levine. The diagnosis and management of mild cognitive impairment: a clinical review. *JAMA*, 312(23):2551–2561, Dec 2014. 25514304[pmid].

[LSG+12] Luís Lemos, Dina Silva, Manuela Guerreiro, Isabel Santana, Alexandre de Mendonça, Pedro Tomás, and Sara C. Madeira. Discriminating alzheimer's disease from mild cognitive impairment using neuropsychological data. *KDD 2012*, 2012.

[LTT+14] Feng Li, Loc Tran, Kim-Han Thung, Shuiwang Ji, Dinggang Shen, and Jiang Li. Robust Deep Learning for Improved Classification of AD/MCI Patients. *Machine Learning in Medical Imaging*, 8679:240–247, 2014.

[LW02] Andy Liaw and Matthew Wiener. Classification and Regression by randomForest. *R News*, 2(3):18–22, 2002.

[LZYvdS18] Changhee Lee, William R. Zame, Jinsung Yoon, and Mihaela van der Schaar. Deephit: A deep learning approach to survival analysis with competing risks. In *AAAI*, 2018.

[LZZW17] D. Li, J. Zhang, Q. Zhang, and X. Wei. Classification of ecg signals based on 1d convolution neural network. In *2017 IEEE 19th International Conference on e-Health Networking, Applications and Services (Healthcom)*, pages 1–6, Oct 2017.

[MDP+] Michelle McDonnell, Lauren Dill, Stella Panos, Stacy Amano, Warren Brown, Shadee Giurgius, Gary Small, and Karen Miller. Verbal fluency as a screening tool for mild cognitive impairment. *International Psychogeriatrics*, page 1–8.

[MFD09] Chandan Misra, Yong Fan, and Christos Davatzikos. Baseline and longitudinal patterns of brain atrophy in

MCI patients, and their use in prediction of short-term conversion to AD: Results from ADNI. *NeuroImage*, 44(4):1415–1422, 2009.

[Mit97] Thomas M. Mitchell. *Machine Learning*. McGraw-Hill, Inc., New York, NY, USA, 1 edition, 1997.

[MJ19] Matt A. Morgan and Jeremy Jones. T1 weighted image, 2019.

[MR58] Ralph M. Reitan. Validity of the trail making test as an indicator of organic brain damage. *Perceptual and Motor Skills*, 8:271–276, 12 1958.

[MRSea04] Grundman M, Petersen RC, Ferris SH, and et al. Mild cognitive impairment can be distinguished from alzheimer disease and normal aging for clinical trials. *Archives of Neurology*, 61(1):59–66, 2004.

[MSF09] A. J. Mitchell and M. Shiri-Feshki. Rate of progression of mild cognitive impairment to dementia – meta-analysis of 41 robust inception cohort studies. *Acta Psychiatrica Scandinavica*, 119(4):252–265, 2009.

[Mur98] SreeramaK. Murthy. Automatic construction of decision trees from data: A multi-disciplinary survey. *Data Mining and Knowledge Discovery*, 2(4):345–389, 1998.

[Nas84] Ruth Nass. The assessment of aphasia and related disorders by harold goodglass and edith kaplan philadelphia, lea & febiger, 1983 illustrated, $27.50 (package). *Annals of Neurology*, 16(5):625–625, 1984.

[NH10] Vinod Nair and Geoffrey E Hinton. Rectified Linear Units Improve Restricted Boltzmann Machines. *Proceedings of the 27th International Conference on Machine Learning*, (3):807–814, 2010.

[NM01] Tormod Næs and Bjørn-Helge Mevik. Understanding the collinearity problem in regression and discriminant analysis. *Journal of Chemometrics*, 15(4):413–426, 2001.

[NO78] Hazel E. Nelson and Anne O'Connell. Dementia: The estimation of premorbid intelligence levels using the new adult reading test. *Cortex*, 14(2):234 – 244, 1978.

[OGS+09] Emilio Soria Olivas, Jose David Martin Guerrero, Marcelino Martinez Sober, Jose Rafael Magdalena Benedito, and Antonio Jose Serrano Lopez. *Handbook Of Research On Machine Learning Applications and Trends: Algorithms, Methods and Techniques - 2 Volumes*. Information Science Reference - Imprint of: IGI Publishing, Hershey, PA, 2009.

[Org19a] World Health Organization. Global health estimate 2016, 08 2019.

[Org19b] World Health Organization. New guide for carers of people with dementia, 08 2019.

[PCPA+77] R Peto, M C Pike, P.E. Armitage, N E Breslow, David Cox, S V Howard, N Mantel, K McPherson, J Peto, and Peter Smith. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. i.

introduction and design. *British journal of cancer*, 34:585–612, 01 1977.

[Pet04] R. C. Petersen. Mild cognitive impairment as a diagnostic entity. *Journal of Internal Medicine*, 256(3):183–194, 2004.

[PKB+09] Ronald C Petersen, David S Knopman, Bradley F Boeve, Yonas E Geda, Robert J Ivnik, Glenn E Smith, Rosebud O Roberts, and Clifford R Jack. Mild Cognitive Impairment: Ten Years Later. *Archives of neurology*, 66(12):1447–1455, December 2009.

[PKH+82] R. I. Pfeffer, T. T. Kurosaki, Jr. Harrah, C. H., J. M. Chance, and S. Filos. Measurement of Functional Activities in Older Adults in the Community1. *Journal of Gerontology*, 37(3):323–329, 05 1982.

[PLF+07] Brenda Plassman, Kenneth Langa, Gwenith Fisher, S.G. Heeringa, David Weir, Mary Beth Ofstedal, James Burke, M D Hurd, G.G. Potter, Willard Rodgers, D.C. Steffens, Robert Willis, and R.B. Wallace. Prevalence of dementia in the united states: The aging, demographics, and memory study. *Neuroepidemiology*, 29:125–32, 02 2007.

[RA13] Patrick Royston and Douglas G. Altman. External validation of a cox prognostic model: principles and methods. *BMC Medical Research Methodology*, 13(1):33, Mar 2013.

[Rey64] Andre Rey. *L'examen clinique en psychologie.* Presses universitaires de France, Paris, 1964.

[RKB+13] Pradeep Reddy, Nicole Kochan, Henry Brodaty, Perminder Sachdev, Lei Wang, Mirza Faisal Beg, and Wei Wen. Novel ThickNet features for the discrimination of amnestic MCI subtypes. *NeuroImage Clinical*, 6:284–295, 2013.

[Rot11] Carole Roth. Boston naming test. In JeffreyS. Kreutzer, John DeLuca, and Bruce Caplan, editors, *Encyclopedia of Clinical Neuropsychology*, pages 430–433. Springer New York, 2011.

[RQZ+19] Kan Ren, Jiarui Qin, Lei Zheng, Zhengyu Yang, Weinan Zhang, Lin Qiu, and Yong Yu. Deep recurrent survival analysis. 2019.

[RSK+07] Vikas Raykar, Harald Steck, Balaji Krishnapuram, Cary Oberije, and Philippe Lambin. On ranking in survival analysis: Bounds on the concordance index. volume 20, 01 2007.

[RSRF12] Martin Reuter, Nicholas J. Schmansky, Herminia Diana Rosas, and Bruce Fischl. Within-subject template estimation for unbiased longitudinal image analysis. *NeuroImage*, 61(4):1402–1418, 2012.

[RW17] Waseem Rawat and Zenghui Wang. Deep convolutional neural networks for image classification: A comprehensive review. *Neural Computation*, 29(9):2352–2449, 2017. PMID: 28599112.

[RWK+14] Pradeep Reddy Raamana, Wei Wen, Nicole a. Kochan, Henry Brodaty, Perminder S. Sachdev, Lei Wang, and Mirza Faisal Beg. The sub-classification of amnestic mild

cognitive impairment using MRI-based cortical thickness measures. *Frontiers in Neurology*, pages 1–10, 2014.

[RZW+14] S. Reppermund, L. Zhuang, W. Wen, M. J. Slavin, J. N. Trollor, H. Brodaty, and P. S. Sachdev. White matter integrity and late-life depression in community-dwelling individuals: diffusion tensor imaging study using tract-based spatial statistics. *The British Journal of Psychiatry*, 205:315–320, 2014.

[SBR+10] Perminder S. Sachdev, Henry Brodaty, Simone Reppermund, Nicole A. Kochan, Julian N. Trollor, Brian Draper, Melissa J. Slavin, John Crawford, Kristan Kang, G. Anthony Broe, Karen A. Mather, and Ora Lux. The sydney memory and ageing study (mas): methodology and baseline medical and neuropsychiatric characteristics of an elderly epidemiological non-demented cohort of australians aged 70–90 years. *International Psychogeriatrics*, 22:1248–1264, 12 2010.

[Sim18] Osvaldo Simeone. A very brief introduction to machine learning with applications to communication systems. *CoRR*, abs/1808.02342, 2018.

[SJSY04] M Symms, H R Jäger, K Schmierer, and T A Yousry. A review of structural magnetic resonance neuroimaging. *Journal of Neurology, Neurosurgery & Psychiatry*, 75(9):1235–1244, 2004.

[SKF+11] Claire E. Sexton, Ukwuori G. Kalu, Nicola Filippini, Clare E. Mackay, and Klaus P. Ebmeier. A meta-

analysis of diffusion tensor imaging in mild cognitive impairment and Alzheimer's disease. *Neurobiology of Aging*, 32(12):2322.e5–2322.e18, 2011.

[SLC+13] Perminder S. Sachdev, Darren M. Lipnicki, John Crawford, Simone Reppermund, Nicole a. Kochan, Julian N. Trollor, Wei Wen, Brian Draper, Melissa J. Slavin, Kristan Kang, Ora Lux, Karen a. Mather, Henry Brodaty, and Ageing Study Team. Factors Predicting Reversion from Mild Cognitive Impairment to Normal Cognitive Functioning: A Population-Based Study. *PLoS ONE*, 8(3):1–10, 2013.

[SLJ+14] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. *CoRR*, abs/1409.4842, 2014.

[SLJS14] Christian Szegedy, Wei Liu, Yangqing Jia, and Pierre Sermanet. Going deeper with convolutions. *arXiv preprint arXiv: 1409.4842*, 2014.

[SM10] Gt Stebbins and Cm Murphy. Diffusion tensor imaging in Alzheimer's disease and mild cognitive impairment. *Behav Neurol*, 21(1):39–49, 2010.

[SS13] Heung Il Suk and Dinggang Shen. Deep learning-based feature representation for AD/MCI classification. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8150 LNCS(0 2):583–590, 2013.

[SSH19] Jill Seladi-Schulman and Seunggu Han. Brain overview, 08 2019.

[SSM+16] Suraj Srinivas, Ravi Kiran Sarvadevabhatla, Konda Reddy Mopuri, Nikita Prabhu, Srinivas S S Kruthiventi, and R. Venkatesh Babu. A Taxonomy of Deep Convolutional Neural Nets for Computer Vision. jan 2016.

[SSS06] E. Strauss, E.M.S. Sherman, and O. Spreen. *A Compendium of Neuropsychological Tests: Administration, Norms, and Commentary*. Oxford University Press, 2006.

[Ste05] Peter H. Steck. A revision of a. l. benton's visual retention test (bvrt) in two parallel forms. *Archives of Clinical Neuropsychology*, 20(3):409 – 416, 2005.

[SY86] Javaid I Sheikh and Jerome A Yesavage. Geriatric Depression Scale (GDS): Recent evidence and development of a shorter version., 1986.

[SZBW13] P S Sachdev, L Zhuang, N Braidy, and W Wen. Is Alzheimer's a disease of the white matter? *Curr Opin Psychiatry*, 26(3):244–251, 2013.

[TKR99] Tom N Tombaugh, Jean Kozak, and Laura Rees. Normative data stratified by age and education for two measures of verbal fluency: {FAS} and animal naming. *Archives of Clinical Neuropsychology*, 14(2):167 – 177, 1999.

[Tom04] Tom N Tombaugh. Trail making test a and b: Normative data stratified by age and education. *Archives of Clinical Neuropsychology*, 19(2):203 – 214, 2004.

[TSG+16] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang. Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning? *IEEE Transactions on Medical Imaging*, 35(5):1299–1312, May 2016.

[TSK+18] Chuanqi Tan, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu. A survey on deep transfer learning. *CoRR*, abs/1808.01974, 2018.

[TWZ+12] Senthil Thillainadesan, Wei Wen, Lin Zhuang, John Crawford, Nicole Kochan, Simone Reppermund, Melissa Slavin, Julian Trollor, Henry Brodaty, and Perminder Sachdev. Changes in mild cognitive impairment and its subtypes as seen on diffusion tensor imaging. *International Psychogeriatrics*, 24:1483–1493, 2012.

[UCP+11] Hajime Uno, Tianxi Cai, Michael J. Pencina, Ralph B. D'Agostino, and L. J. Wei. On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine*, 30(10):1105–1117, 2011.

[VK15] Juan Jose Vaquero and Paul Kinahan. Positron emission tomography: Current challenges and opportunities for technology advances in clinical and pre-clinical imaging systems. *Annual Reviews of Biomedical Engineering*, 17, 01 2015.

[VLBMR16] Kristina Vatcheva, MinJae Lee, Joseph B. McCormick, and Mohammad Rahbar. Multicollinearity in regression

analyses conducted in epidemiologic studies. *Epidemiology open access*, 06:227, 03 2016.

[VPM17] Sandra Vieira, Walter H.L. Pinaya, and Andrea Mechelli. Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications. *Neuroscience and Biobehavioral Reviews*, 74(Part A):58 – 75, 2017.

[Wec81] David Wechsler. The psychometric tradition: Developing the Wechsler Adult Intelligence Scale., 1981.

[WHC+13] Bradley T. Wyman, Danielle J. Harvey, Karen Crawford, Matt A. Bernstein, Owen Carmichael, Patricia E. Cole, Paul K. Crane, Charles DeCarli, Nick C. Fox, Jeffrey L. Gunter, Derek Hill, Ronald J. Killiany, Chahin Pachai, Adam J. Schwarz, Norbert Schuff, Matthew L. Senjem, Joyce Suhy, Paul M. Thompson, Michael Weiner, and Clifford R. Jack. Standardization of analysis sets for reporting results from adni mri data. *Alzheimer's Dementia*, 9(3):332 – 337, 2013.

[WJK+11] Robin Wolz, Valtteri Julkunen, Juha Koikkalainen, Eini Niskanen, Dong Ping Zhang, Daniel Rueckert, Hilkka Soininen, and Jyrki Lötjönen. Multi-method analysis of MRI images in early diagnostics of Alzheimer's disease. *PLoS ONE*, 6(10):1–9, 2011.

[WPK+04] B. Winblad, K. Palmer, M. Kivipelto, V. Jelic, L. Fratiglioni, L.-O. Wahlund, A. Nordberg, L. Bäckman, M. Albert, O. Almkvist, H. Arai, H. Basun, K. Blennow,

M. De Leon, C. DeCarli, T. Erkinjuntti, E. Giacobini, C. Graff, J. Hardy, C. Jack, A. Jorm, K. Ritchie, C. Van Duijn, P. Visser, and R.C. Petersen. Mild cognitive impairment – beyond controversies, towards a consensus: report of the international working group on mild cognitive impairment. *Journal of Internal Medicine*, 256(3):240–246, 2004.

[WPN+07] Jennifer L. Whitwell, Ronald C. Petersen, Selamawit Negash, Stephen D. Weigand, Kejal Kantarci, Robert J. Ivnik, David S. Knopman, Bradley F. Boeve, Glenn E. Smith, and Clifford R. Jack. Patterns of atrophy differ among specific subtypes of mild cognitive impairment. *Arch Neurol*, 64(8):1130–1138, Aug 2007. 17698703[pmid].

[Yil15] Pinar Yildirim. Filter based feature selection methods for prediction of risks in hepatitis disease. *International Journal of Machine Learning and Computing*, 5:258–263, 08 2015.

[YK15] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *CoRR*, abs/1511.07122, 2015.

[YNDT18] Rikiya Yamashita, Mizuho Nishio, Richard Kinh Gian Do, and Kaori Togashi. Convolutional neural networks: an overview and application in radiology. *Insights into Imaging*, 9(4):611–629, Aug 2018.

[YSK+16] Robert W. Yeh, Eric A. Secemsky, Dean J. Kereiakes, Sharon-Lise T. Normand, Anthony H. Gershlick, David J.

Cohen, John A. Spertus, Philippe Gabriel Steg, Donald E. Cutlip, Michael J. Rinaldi, Edoardo Camenzind, William Wijns, Patricia K. Apruzzese, Yang Song, Joseph M. Massaro, Laura Mauri, and for the DAPT Study Investigators. Development and Validation of a Prediction Rule for Benefit and Harm of Dual Antiplatelet Therapy Beyond 1 Year After Percutaneous Coronary Intervention-Prediction Rule for Long-term Dual Antiplatelet Therapy After PCIPrediction Rule for Long-term Dual Antiplatelet Therapy After PCI. *JAMA*, 315(16):1735–1749, 04 2016.

[ZST+12] L. Zhuang, P. S. Sachdev, J. N. Trollor, N. a. Kochan, S. Reppermund, H. Brodaty, and W. Wen. Microstructural white matter changes in cognitively normal individuals at risk of amnestic MCI. *Neurology*, 79:748–754, 2012.

[ZST+13] Lin Zhuang, Perminder S. Sachdev, Julian N. Trollor, Simone Reppermund, Nicole a. Kochan, Henry Brodaty, and Wei Wen. Microstructural White Matter Changes, Not Hippocampal Atrophy, Detect Early Amnestic Mild Cognitive Impairment. *PLoS ONE*, 8(3):1–10, 2013.