From the Division of Clinical Geriatrics
Department of Neurobiology, Care Sciences and Society
Karolinska Institutet, Stockholm, Sweden

QUANTIFYING NEURODEGENERATION FROM MEDICAL IMAGES WITH MACHINE LEARNING AND GRAPH THEORY

Gustav Mårtensson

Karolinska Institutet

Stockholm 2020

Cover illustration by Johan Klingstedt.
All previously published papers were reproduced with permission from the publisher.
Published by Karolinska Institutet.
Printed by Arkitektkopia AB, 2020

ISBN 978-91-7831-732-5

Quantifying neurodegeneration from medical images with machine learning and graph theory

THESIS FOR DOCTORAL DEGREE (Ph.D.)

The thesis will be defended in Erna Möller-salen in Neo, 5th floor, Karolinska Institutet, Campus Flemingsberg, Huddinge.

Friday May 15th 2020, 9.00 am.

By

Gustav Mårtensson

Principal Supervisor:
Professor Eric Westman
Karolinska Institutet
Department of Neurobiology,
Care Sciences and Society
Division of Clinical Geriatrics

Co-supervisors:
Assistant professor Joana B. Pereira
Karolinska Institutet
Department of Neurobiology,
Care Sciences and Society
Division of Clinical Geriatrics

Senior lecturer Giovanni Volpe
University of Gothenburg
Department of Physics
Soft Matter Lab

Opponent:
Senior Lecturer Jorge Cardoso
King's College
School of Biomedical Engineering
& Imaging Sciences

Examination Board:
Associate professor Pawel Herman
KTH Royal Institute of Technology
Dept. of Computational Biology
Division of Computational Science
and Technology

Professor Martin Lövdén
University of Gothenburg
Department of Psychology

Professor Katrine Riklund
Umeå University
Department of Radiation Sciences
Diagnostic Radiology

Abstract

Neurodegeneration (or *brain atrophy*) is part of the pathological cascade of Alzheimer's disease (AD) and is strongly associated with cognitive decline. In clinics, atrophy is measured through visual assessments of specific brain regions on medical images according to established rating scales.

In this thesis, we developed a model based on recurrent convolutional neural networks (*AVRA*: Automatic visual ratings of atrophy) that could predict scores from magnetic resonance images (MRI) according to commonly used clinical rating scales, namely: Scheltens' scale for medial temporal atrophy (MTA), Pasquier's frontal subscale of global cortical atrophy (GCA-F), and Koedam's posterior atrophy (PA) scale. AVRA was trained on over 2000 images rated by a single neuroradiologist and demonstrated similar inter-rater agreement levels on all three scales to what has reported between two "human raters" in previous studies.

We further applied different versions of AVRA, trained systematically on data with different levels of heterogeneity, in external data from multiple European memory clinics. We observed a general performance drop in the out-of-distribution (OOD) data compared to test sets sampled from the same cohort as the training data. By training AVRA on data from multiple sources, we show that the performance in external cohorts generally increased. AVRA demonstrated a notably low agreement in one memory clinic, despite good quality images, which suggests that it may be challenging to assess how well a machine learning model generalizes to OOD data.

For additional validation of our model, we compared AVRA's MTA ratings to two external radiologists' and the volumes of the hippocampi and inferior lateral ventricles. The images came from a longitudinal cohort that comprised individuals with subjective cognitive decline (SCD) and mild cognitive impairment (MCI) followed up over six years. AVRA showed substantial agreement to one of the radiologists, and lower rating agreement to the other. The two radiologists also showed low agreement

between each other. All sets of ratings were strongly associated with the subcortical volumes, suggesting that all three raters were reliable. We further observed that individuals with SCD and (probably) underlying AD pathology had a faster MTA progression than MCI patients with non-AD biomarker profile.

Finally, we evaluated a method to quantify patterns of atrophy through the use of graph theory. We compared structural gray matter networks between groups of healthy controls and AD patients, constructed from different subsamples and with different network construction methods. Our experiments suggested that structural gray matter networks may not be very stable. Our networks required more than 150 subjects/group to show convergence in the included network properties, which is a greater sample size than used in the majority of the studies applying these methods. The different graph construction methods did not yield consistent differences between the control and AD networks, which may explain why findings have been inconsistent across previous studies.

To conclude, we demonstrated that a machine learning model can successfully learn to mimic a radiologist's assessment of atrophy without intra-rater variability. The challenge going forward is to assert model consistency across clinics, scanners and image quality—nuisances that humans are better at ignoring than deep learning models.

Sammanfattning på svenska

Neurodegeneration (eller *hjärnatrofi*) drabbar patienter som lider av Alzheimer's sjukdom (AD), och är starkt förknippat med försämring av kognitiva förmågor. För att kvantifiera atrofi kliniskt så görs visuella bedömningar, där en tränad radiolog gör en skattning från 0 till 3 (eller 4) enligt etablerade skattningsskalor.

I den här avhandlingen så har vi utvecklat en automatisk metod för visuella skattningar som vi kallar *AVRA* (Automatic visual ratings of atrophy) och som bygger på neurala nätverk. De tre skalorna som automatiserades var för bedömning av medial temporallobsatrofi (MTA), frontal kortikalatrofi (GCA-F) och posterior atrofi (PA). AVRA:s skattningar överensstämde väl med radiologens bedömningar från samma kohort, och på samma nivå som tidigare studier har rapporterat.

För att förstå hur AVRA presterade på klinisk data, så undersökte vi hur väl skattningarna stämde överens i extern data från minneskliniker runt om i Europa. Vi noterade en generell minskning av överensstämningen bland skattningarna när testdatat inte kom från samma distribution som träningsdatat. Genom att inkludera bilder med större variation (d.v.s. från fler kohorter, kameror, och skanningsprotokoll) så ökade AVRA:s prestanda även i dessa dataset. Vi jämförde även AVRA med två andra radiologer, och där vi såg hög överenstämmelse mellan AVRA:s och ena radiologens MTA-bedömningar men lägre till den andre. (Samstämmigheten mellan radiologerna var också låg). Samtligas skattningar visade dock stark korrelation med de subkortikala strukturer som bedöms i MTA-skalan, vilket indikerar att båda radiologerna samt AVRA var pålitliga.

Med AVRA undersökte vi hur MTA förändras över tid hos individer med självupplevd försämring i kognition och patienter med mild kognitiv svikt (MCI). Målet var att karaktärisera hur kliniska MTA-skattningar ser ut i tidiga stadier av demens. Vi observerade att patienter med underliggande AD-patologi, men med mildare symptom, hade snabbare progression i MTA än MCI-patienterna utan abnormala AD-biomarkörer.

Våra resultat visade att MTA-skattningarna visade samma longitudinella trender som volymen av hippocampus, och att det främst är subjektiviteten i bedömningarna som begränsar visuella skattningar.

I den sista studien undersökte vi hur pålitliga resultat som erhålls när grafteori används för att kvantifiera kortikala atrofimönster i hjärnan. Vi jämförde nätverk konstruerade av data från AD-patienter mot friska kontroller – två grupper där vi antar att skillnaderna bör var stora. Våra resultat visade att det krävdes över 150 bilder från varje grupp för att få stabila resultat, vilket är betydligt fler än vad som vanligtvis använts i dessa typer av studier. Hur man definierade sin graf hade också stor påverkan, vilket kan förklara varför tidigare studier rapporterat olika resultat från studier på AD-nätverk.

Sammanfattningsvis så beskriver den här avhandlingen utvecklandet av en maskininlärgningsmodell som med god precision kan automatgenerera visuella skattningar som används kliniskt för att mäta atrofi. Utmaningen framåt är att se till att få dessa modeller att fungera i en klinisk verklighet där kameror, skanningsprotokoll och bildkvalitet kan variera kraftigt.

Contents

1 Introduction **1**
1.1 Aims and outline 2

2 Alzheimer's disease **3**
2.1 Diagnosing Alzheimer's disease 3
2.2 Neuropathology in Alzheimer's disease 7
2.2.1 What is causing AD? 8
2.3 Biomarkers in Alzheimer's disease 9

3 Artificial neural networks **15**
3.1 Multilayer perceptrons 16
3.1.1 Learning approaches 18
3.1.2 Training a network 20
3.2 Convolutional neural networks 23
3.2.1 CNN architectures and modules 25
3.3 Recurrent neural networks 28
3.4 Domain shift in medical imaging 29

4 Automatic visual ratings of atrophy **33**
4.1 Introduction 33
4.2 Visual rating scales 34
4.2.1 Reliability of human ratings 37
4.3 Automatic visual ratings 38
4.3.1 Network architecture 39
4.3.2 Preprocessing and data augmentation 42

4.3.3 Training procedure and hyperparameters 44
4.4 Assessing performance 44
4.4.1 Performance metric 44
4.4.2 Within-distribution data 46
4.4.3 Out-of-distribution data 48
4.4.4 Agreement to external radiologists and subcortical volumes 55
4.4.5 AVRA on longitudinal data 57
4.5 Conclusion 63

5 Quantifying atrophy through gray matter networks 65
5.1 Graph theory 65
5.2 Constructing graphs of neuroimaging data 67
5.2.1 Node definition 67
5.2.2 Edge definition 68
5.3 Graph theoretical measures 69
5.4 Reproducibility of gray matter networks 71
5.5 Conclusion 78

6 Concluding remarks 81

Nomenclature

Aβ Amyloid beta

AC-PC Anterior-posterior commissures

AChEI Acetylcholinesterase inhibitor

AD Alzheimer's disease

ADNI Alzheimer's disease neuroimaging initiative

AI Artificial intelligence

ApoE Apolipoprotein E

APP Amyloid precursor protein

ATN Amyloid; Tau; Neurodegeneration

AVRA Automatic visual ratings of atrophy

BioFINDER Biomarkers For Identifying Neurodegenerative Disorders Early and Reliably

C_1/C_2 Center 1 and Center 2 from E-DLB cohort

CNN Convolutional neural network

CSF Cerebrospinal fluid

CT Computed tomography

CTR Healthy control (also HC)

DL Deep learning

DLB Dementia with Lewy Bodies

DTI Diffusion tensor imaging

E-DLB European Dementia with Lewy Bodies consortium

EEG Electroencephalography

EOAD Early-onset Alzheimer's disease

FDG Fluoro-deoxyglucose

FL Federated learning

fMRI Functional MRI

FSL FMRIB Software Library

FTLD Frontotemporal lobe dementia

GAN Generative adversarial network

GCA Global cortical atrophy (visual rating scale)

GCA-F Global cortical atrophy, frontal subscale

GM Gray matter

GRU Gated recurrent unit

GT Graph theory

HC Hippocampus

ILV Inferior lateral ventricle

IWG International Working Group

LSTM Long short-term memory

MCI Mild cognitive impairment

ML Machine learning

MLP Multilayer perceptron

MMSE Mini mental state examination

MRI Magnetic resonance imaging

MS Multiple Sclerosis

MSE Mean squared error

MTA Medial temporal lobe atrophy (visual rating scale)

MTL Medial temporal lobe

NFT Neurofibrillary tangle

NIA-AA National Institute on Aging and Alzheimer's Association

NN Neural network

OOD Out-of-distribution

P-tau Phosphorylated tau

PA Posterior atrophy (visual rating scale)

PCS Posterior cingulate sulcus

PDD Parkinson's disease with dementia (PDD)

PET Positron emission tomography

POS Parieto-occipital sulcus

ReLU Rectified linear unit

RL	Reinforcement learning
RNN	Recurrent neural network
ROI	Region of interest
SCD	Subjective cognitive decline
SGD	Stochastic gradient descent
sMRI	Structural magnetic resonance imaging
SPM	Statistical Parametric Mapping
T-tau	Total tau
VBM	Voxel-based morphometry
WM	White matter

List of scientific papers

This thesis is based on the following original articles:

1. **Mårtensson G**, Ferreira D, Cavallin L, Muehlboeck J-S, Wahlund L-O, Wang C, Westman E. AVRA: Automatic Visual Ratings of Atrophy from MRI images using Recurrent Convolutional Neural Networks. *NeuroImage: Clinical*. 2019. 23(March), p. 101872. doi: 10.1016/j.nicl.2019.101872.

2. **Mårtensson G**, Ferreira D, Granberg T, Cavallin L, Oppedal K, Padovani A, Rektorova I, Bonanni L, Pardini M, Kramberger M, Taylor J-P, Hort J, Snædal J, Kulisevsky J, Blanc F, Antonini A, Mecocci P, Vellas B, Tsolaki M, Kłoszewska I, Soininen H, Lovestone S, Simmons A, Aarsland D, Westman E. The reliability of a deep learning model in clinical out-of-distribution MRI data: a multicohort study. *Manuscript under review*.

3. **Mårtensson G**, Håkansson C, Pereira JB, Palmqvist S, Hansson O, van Westen D[†], Westman E[†]. Medial temporal atrophy in preclinical dementia: visual and automated assessment during six year follow-up. *Manuscript under review*.
   [†]Shared last author

4. **Mårtensson G**, Pereira JB, Mecocci P, Vellas B, Tsolaki M, Kłoszewska I, Soininen H, Lovestone S, Simmons A, Volpe G, Westman E. Stability of graph theoretical measures in structural brain networks in Alzheimer's disease. *Scientific Reports*. 2018. 8(1), p. 11592. doi: 10.1038/s41598-018-29927-0.

Chapter 1

Introduction

Dementia is a category of neurodegenerative disorders that causes afflicted people's mind, memory and personality to change. A recent study has estimated that 47 million people in the world were suffering from dementia in 2015 (Wimo et al., 2017). As the life-expectancy increases, this number is predicted to reach 75 millions in 2030 (Prince et al., 2015), causing suffering and an enormous burden on caregivers and health systems. *Alzheimer's disease* (AD) is the most common type of dementia and constitutes about 2/3 of all dementia cases.

There is no cure for Alzheimer's disease available today. While there are drugs that are approved for clinical use, the long-term effects are small. Substantial efforts have gone into developing new drugs that can slow down the progression of Alzheimer's disease, but so far unfortunately without success. As neurodegeneration is irreversible, it is widely believed that treatment strategies need to be initiated in a very early stage of the disease—possibly already at an asymptomatic phase, which may last for 15 years prior to experiencing clinical symptoms. Therefore, it is important to have reliable, sensitive and early markers of the disease in order to detect these asymptomatic individuals and track the progression of the disorder. Senile plaques and neurofibrillary tangles in the brain are pathological hallmarks of AD, which leads to neurodegeneration (or *brain atrophy*) that is strongly associated with cognitive symptoms. The primary focus of this thesis is on different

measures of atrophy derived from medical images used in research and in clinics.

1.1 Aims and outline

The overall goal of this thesis is to investigate different methods used to quantify atrophy from structural magnetic resonance images (sMRI) during the continuum of dementia, and Alzheimer's disease in particular. This can be broken down into study-specific aims:

1. Develop a tool for automated predictions of radiologist ratings of atrophy according to established visual assessment scales used in the clinics.

2. Assess how reliable the tool is in external data from multiple European memory clinics.

3. Investigate the longitudinal progression of medial temporal atrophy in preclinical dementia using our tool compared to visual assessment and other neuroimaging softwares.

4. Study the utility of gray matter networks for quantifying atrophy in Alzheimer's disease.

The following two chapters of the thesis provide some background to Alzheimer's disease and artificial neural networks. In Chap. 4 studies 1 to 3 are discussed, where we propose and apply a deep learning model that predicts visual ratings of atrophy according to scales used in clinics. In the fourth study, we assess a different method, graph theory, for quantifying global atrophy patterns from sMRI images. This is covered in Chap. 5.

Chapter 2

Alzheimer's disease

The chapter aims to describe the pathological cascade in Alzheimer's disease. This includes clinical manifestations and diagnostic criteria, underlying pathological processes and biomarkers in AD.

2.1 Diagnosing Alzheimer's disease

A definitive diagnosis of AD can only be made postmortem through an autopsy, which shows the occurrence of senile plaques (Aβ pathology; Amyloid β) and neurofibrillary tangles (NFT; tau pathology) in the brain (McKhann et al., 1984). Researchers have tried to characterize the continuum of AD as a number of phases, ranging from pre-symptomatic to fully-developed dementia.

From a clinical and a patient perspective, the onset of the disease begins with cognitive abilities starting to deteriorate. Cognition is a rather abstract concept, which is difficult to quantify. In practice, it is done through a battery of neuropsychological tests which are designed to assess impairment in a number of specific cognitive domains, such as episodic memory, attention, language and visouspatial processing. However, even with these tests in place it can be tricky to measure cognitive decline. The word "decline" means that cognition has worsened from some previous baseline state. Since there is a degree of inter-subject

variability in cognitive baselines we compare the results of cognitive tests to normative values, which then really becomes a test of *impairment*.

An individual that has a self-perceived notion of worsening cognition, but where the cognition is still not considered abnormal based on neuropsychological tests, is said to have *subjective cognitive decline* (SCD). It can be the first symptomatic stage of dementia, such as AD, and "subtle cognitive deficits" is in fact part of the criteria of preclinical AD (Jack et al., 2011; Sperling et al., 2011). Jessen et al. (2014) have proposed (for the Subjective Cognitive Decline Initiative Working Group) research criteria for SCD subjects. These criteria are rather general and only include a self-perceived decline in cognitive ability (not caused by an acute event) while still performing "normal" (adjusted for age and education) on cognitive tests. The SCD population is a highly heterogeneous group, as the notion of cognitive decline varies among individuals. It has been shown that SCD individuals are more prevalent to progress to dementia compared to (self-perceived) cognitively unimpaired controls (Mitchell et al., 2014). Individuals with subjective memory complaints have, on a group level, been reported to have reduced volumes of the entorhinal cortex (Jessen et al., 2006), cortical thinning in brain regions affected in AD (Schultz et al., 2015), as well as increased amyloid burden (Perrotin et al., 2012) compared to controls.

Patients with worsening cognition that is detectable through neuropsychological tests, can be clinically diagnosed as having *mild cognitive impairment* (MCI). The severity is not sufficient for the patient to be diagnosed with dementia, so MCI is commonly viewed as an intermediate state between healthy aging and dementia. Diagnostic and research criteria for MCI have been suggested by Petersen (2004) who proposed two clinical subtypes of MCI: amnestic- and non-amnestic-MCI. The amnestic-MCI subtype (i.e. with impaired memory function) were suggested to have underlying AD pathology. Later, diagnostic guidelines of MCI *due to AD* were formalized by the National Institute on Aging and Alzheimer's Association (NIA-AA) (Albert et al., 2011) to characterize the pre-demented stage of AD:

- Cognition criteria:
  - Decline in cognition, reported by patient or informant.
  - Objective evidence of impairment in cognition (e.g. abnormal Mini Mental State Examination, MMSE, score).
  - Patient still have independence in functional abilities.
  - Not demented.
- Etiology:
  - Rule out traumatic, vascular or other non-AD causes where possible.
  - Evidence of longitudinal cognitive decline, if feasible.
  - History of AD genetic factors, when relevant.
- Biomarkers:
  - Aβ positive, detected in cerebrospinal fluid (CSF) or on a positron emission tomography (PET) scan.
  - Tau positive, detected in CSF or on PET.
  - Downstream neuronal injury (visible on MRI or PET).

Biomarkers were recommended to use in order to increases the likelihood of correctly diagnosing MCI due to AD, and not as a diagnostic requirement (Albert et al., 2011).

Diagnostic criteria for Alzheimer's disease dementia was proposed by McKhann et al. (1984) and is called the NINCDS-ADRDA [1] criteria. These were updated in 2011 by NIA-AA to be more in line with current research (McKhann et al., 2011). These guidelines describe AD as progression of three phases: preclinical AD, MCI due to AD (outlined above), and probable AD. The definition of preclinical AD was suggested for research purposes only and not for clinical diagnosis (Sperling et al., 2011). The three stages of preclinical AD start with abnormal levels of

[1] National Institute of Neurological and Communicative Disorders and Stroke (NINCDS) with Alzheimer's Disease and Related Disorders Association (ADRDA)

Aβ, followed by abnormal levels of tau and/or neurodegeneration, and finally also subtle cognitive deficits.

The proposed criteria for AD dementia include that the patient shows

- Functional decline from previous levels, not explained by delirium of psychiatric disorder.
- Clinically diagnosable cognitive decline.
- Cognitive or behavioral impairment in at least two of the following abilities:
  - Obtaining and remembering new information.
  - Reasoning of complex tasks s.a. poor decision-making ability.
  - Visuospatial, s.a. inability to recognize faces.
  - Language, s.a. experiencing problems with speaking or reading.
  - Personality, s.a. mood fluctuations or social withdrawal.

The use of biomarkers are not included in the core clinical NIA-AA's criteria from 2011 but it is stated that biomarkers increase the diagnostic certainty. They suggest using biomarkers in research, clinical trials and in cases when "deemed appropriate by the clinician". The main reasoning behind this was that the diagnostic accuracy of the core clinical criteria is good in most cases, but also due to lack of standardized biomarkers and procedures (McKhann et al., 2011). A European version of diagnostic guidelines was proposed by The International Working Group (IWG) by Dubois et al. (2007), and revised in 2014 (Dubois et al., 2014). They are called the IWG-2 criteria, and advocate the use of biomarkers in the diagnosis, but with the view that abnormality in these markers increases the probability of AD (and are not sufficient to diagnose AD pathology).

Recently, a fully biomarker-based diagnostic system was proposed by Jack et al. (2016) to be used in research: the *ATN system*. It is a binary system in which a patient that is e.g. $A^+T^-N^-$ has amyloid pathology ("amyloid positive"), normal tau levels ("tau negative") and

no abnormal neurodegeneration. The "N" has since been suggested to be put in parenthesis as "(N)", since markers for neurodegeneration are not necessarily AD-specific, whereas Aβ and tau are part of the pathological definition of AD (Jack et al., 2018a). The AT(N) system is thus completely disconnected from clinical symptoms, such as cognitive impairment, with the motivation that if an individual has the biomarker profile A^+T^+ he or she has Alzheimer's disease *by definition*—regardless of whether symptoms have developed or not. NIA-AA recently updated their criteria based on the AT(N) system, and are no longer based on clinical symptoms (Jack et al., 2018b). The framework thus has an opposite view on biomarkers compared to the IWG-2 criteria where the role of the AT biomarkers is to increase the probability of a correct AD diagnosis (Dubois et al., 2014).

It should be noted that both the NIA-AA and IWG are frameworks suggested mainly for research and not for clinical diagnosis of AD. In Sweden, the ICD-10 criteria (World Health Organization, 1990) are used for diagnosing AD.

2.2 Neuropathology in Alzheimer's disease

Despite enormous efforts in the scientific community, we still do not fully understand the cause and pathological process in AD. By definition, AD patients have senile plaques and neurofibrillary tangles in the brain. This is followed by neurodegeneration, affecting cognition and eventually leading to the death of the patient. The heterogeneity of the disease, the long time span, and that Aβ and tau pathology seem to precede clinical symptoms with many years makes it very challenging to study. The fact that other neuropathologies—such as Lewy bodies, cerebral vessel diseases and TDP-43—frequently appear alongside AD (Boyle et al., 2018) further adds to complexity of developing and assessing treatment strategies.

2.2.1 What is causing AD?

Among proposed disease mechanisms of AD, the *cholinergic hypothesis* was an early model of the disorder (Bartus et al., 1982). It was based on the observation that the levels of acetylcholine (ACh), a neurotransmitter, is lower in AD patients, and that acetylcholinesterase inhibitors (AChEIs) showed a positive effect on clinical symptoms. However, the leading hypothesis today in the research community is arguably the *amyloid hypothesis*.

The amyloid hypothesis was originally put forward by Hardy and Higgins (1992). The model proposes that an imbalance between production and clearance of Aβ is an initiating factor of AD. It was based on the discoveries that mutations in the gene coding for the amyloid precursor protein (APP) are associated with familial AD (Murrell et al., 1991), and that patients with Down syndrome (who are born with an extra chromosome carrying the APP gene) have a high risk of developing AD at a young age (Whalley, 1982). APP is cleaved by β- and γ-secretase which produces Aβ. If Aβ is not cleared at a proper rate, these Aβ monomers can aggregate to form fibrils, which subsequently can form Aβ plaques. In the hypothesis, increased levels of Aβ eventually causes aggregation of hyperphosphorylated tau (P-tau) protein and NFTs. Downstream effects of these events are neuronal synaptic dysfunction and neuronal loss, which eventually leads to a dementia diagnosis (Hardy and Higgins, 1992).

However, there are debates regarding the validity of the amyloid hypothesis (Selkoe and Hardy, 2016). This is mainly due to the lack of success in clinical trials targeting amyloid deposition together with weak associations between Aβ burden and neuronal loss (Ricciarelli and Fedele, 2017). It has been suggested that hyperphosphorylation of tau and Aβ depositions are independent or synergetic pathologies (Duyckaerts, 2011), or that they are independent but share an upstream cause (Small and Duff, 2008).

Spreading patterns of Aβ and tau aggregations in the brain were first described by Braak and Braak (1991). Their study showed that while the distribution of Aβ plaques varied widely across individuals,

the spreading of NFTs was more distinct. They differentiated six stages of the spreading of NFTs:

- Stages I-II: mild to moderate presence of NFTs in the transentorhinal region.
- Stages III-IV: progression of NFTs to the limbic region, including the hippocampus.
- Stages V-VI: progression into neocortex.

The distribution of Aβ plaques instead seems to follow the opposite direction; starting in neocortex and spreading to allocortical, basal ganglial and diencephalic structures (Brettschneider et al., 2015; Braak and Braak, 1991). The distribution of tau burden has been shown to correlate better with neuronal loss and memory impairment than the distribution of plaques (Gómez-Isla et al., 1997; Arriagada et al., 1992).

The cause of the disease may be debated, but there are a number of associated risk factors of sporadic AD, where age is the most important one. The most prominent genetic risk factor is the apolipoprotein E $\epsilon 4$ (ApoE $\epsilon 4$) allele, which has been shown to increase the risk of developing AD by a factor 3 (heterozygous) to 12 (homozygous) Corder et al. (1993); Farrer et al. (1997). Cardiovascular risk factors such as diabetes (Cheng et al., 2012), obesity (Kivipelto et al., 2005), smoking (Durazzo et al., 2014) and lack of physical activity (Norton et al., 2014) have been shown to also increase the risk of AD. Protective factors are ApoE $\epsilon 2$ homozygous (Farrer et al., 1997) as well as *cognitive reserve* that is built up through education and mental activity (Stern et al., 1994; Valenzuela and Sachdev, 2006).

2.3 Biomarkers in Alzheimer's disease

A definitive diagnosis of AD can, as previously stated, only be made postmortem. However, it is of great interest to be able to track the pathology *in vivo*—not only for a more reliable diagnosis but also in order to study the disease as well as developing treatments. That is, we

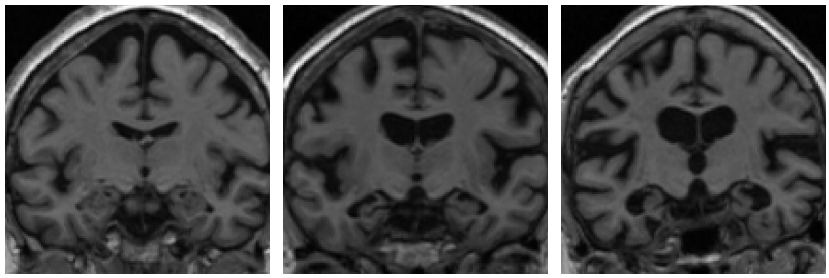
are interested in finding *biomarkers* of the disease, and to develop as precise measurement techniques as possible.

The two entities that have been the primary interest to quantify are Aβ and tau pathology, given that they are the pathophysiological hallmarks of Alzheimer's disease. They have typically been measured through CSF or PET, but recent advances have been made to measure them in blood plasma (Blennow and Hampel, 2003; Nordberg et al., 2010; Blennow and Zetterberg, 2018; Janelidze et al., 2020).

To track Aβ pathology in CSF, it is most common to measure the levels of Aβ_{42}, which is believed to be the most toxic Aβ peptide. In AD pathology, the CSF Aβ_{42} levels are reduced compared to normal aging due to less amyloid being cleared in the brain, and consequently less Aβ ending up in the CSF. This is an early marker that has been shown to precede clinical symptoms up to 15 years in autosomal dominant AD patients (Bateman et al., 2012). The level of P-tau in CSF reflects the degree of phosphorylated tau (and thus NFTs, i.e. tau pathology) in the brain and total tau (T-tau) has been considered to be a general, yet indirect, marker for neurodegeneration (Blennow and Hampel, 2003). Both P-tau and T-tau levels are increased in AD patients and can help in distinguishing AD from other neurological disorders (Vanmechelen et al., 2000).

PET imaging has the advantage of providing (*in vivo*) spatial distribution of Aβ, tau and glucose metabolism; the latter being a marker of neuronal activity and implicitly for synaptic dysfunction and neurodegeneration. In AD, the retention of Aβ and tau is increased, whereas glucose metabolism is decreased (Nordberg et al., 2010; Schilling et al., 2016). PET can be a seen as a more direct marker of Aβ/tau pathology, whereas CSF provides indirect measures of pathology.

A third important and relevant biomarker in dementia and AD is *neurodegeneration*. The term refers to the degeneration of neuronal structure and function. This manifests in e.g. glucose hypometabolism, but also in *brain atrophy* visible on sMRI and computed tomography (CT) images. Quantifying brain atrophy (or gray matter loss) is the main focus of this thesis. Atrophy is common in many dementias, but different disorders may display distinct atrophy patterns (Harper et al., 2017).



(a) CTR (b) MCI (c) AD

Figure 2.1: Example of medial temporal atrophy and ventricle enlargement in a cognitively normal control (CTR), a patient with mild cognitive impairment (MCI), and an Alzheimer's disease (AD) patient.

The fact that regional atrophy correlates well with the distribution of NFTs (Whitwell et al., 2012) and cognitive deficits (Frisoni et al., 2010) suggests that atrophy is an important biomarker for AD and dementia.

Atrophy in the medial temporal lobe (MTL) is characteristic in AD and assessed in the diagnostic work-up (Jack et al., 1997; Barkhof et al., 2011), see Fig. 2.1. In the early stage of the disease atrophy is most pronounced in the medial temporal lobe, including hippocampus (HC) and entorhinal cortex, later affecting the basal temporal lobe and cortical regions in the parietal cortex such as the precuneus and posterior cingulate gyrus (Vemuri and Jack, 2010). However, it should be noted that not all AD patients display this atrophy pattern. Subtypes of AD has been identified based on distribution of NFTs by Murray et al. (2011) (typical AD, hippocampal-sparing, and limbic predominant), where each subtype displays a specific atrophy pattern (Whitwell et al., 2012). A minimal atrophy AD subtype has also been suggested (Byun et al., 2015). Early-onset AD (EOAD) patients tend to have more pronounced posterior atrophy Frisoni et al. (2007) that can help in distinguishing between frontotemporal lobe dementia (FTLD) and EOAD, which can have overlapping symptoms in early stages (Lehmann et al., 2012).

In research, atrophy has generally been studied through comparing volumes or thickness of specific brain regions, through vertex-based anal-

ysis, or using voxel-based morphometry (VBM). These methods provide sensitive measures that can be used to track atrophy progression. Studies that applied multivariate analysis methods or VBM-based machine learning (ML) methods to investigate patterns of atrophy have found that global cortical atrophy (i.e. not only in the medial temporal lobe) can provide additional information that is useful for differential diagnosis (Klöppel et al., 2008; Westman et al., 2011b). It is therefore of great interest to be able to quantify patterns of atrophy in a meaningful way, which has led to the studies on gray matter networks that is discussed in Chap. 5 of the thesis.

Focusing on hippocampal atrophy, several studies have demonstrated that HC volume alone can reliably distinguish AD patients from cognitively normal subjects Wahlund et al. (1999); Frisoni et al. (2010); Westman et al. (2011a) and that it is an early marker of the disease (Jack et al., 1997). It has been shown that the atrophy rate of HC is increased in MCI and AD (Jack et al., 2000; Henneman et al., 2009) (up to 8% loss/year Fox et al. (1996)), and correlates well with cognitive decline (Rusinek et al., 2003). Ridha et al. (2006) showed that brain atrophy rate was an even earlier marker of disease onset than cross-sectional volumes in familial AD cases.

In the clinical routine, atrophy is typically assessed through visual ratings, and not with automated measures. A trained radiologist gives a score of the degree of atrophy in a specific region according to an established rating scale, which is used in the diagnostic workup of dementia. Despite the simplicity[2] of visual ratings, they have been shown to provide similar diagnostic abilities as volumetric measures while being fast and reliable Wahlund et al. (1999); Westman et al. (2011a). Quantifying atrophy trough visual ratings is the main topic of this thesis and is discussed further in Chap. 4.

A temporal ordering of the aforementioned biomarkers has been proposed by Jack et al. (2013), see Fig. 2.2. This model follows the amyloid hypothesis discussed in the previous section, where Aβ abnormality in

[2] "Simple" in this context means in comparison to the advanced mathematical concepts underlying many computerized neuroimaging techniques—not that it is "simple" to become a reliable rater.

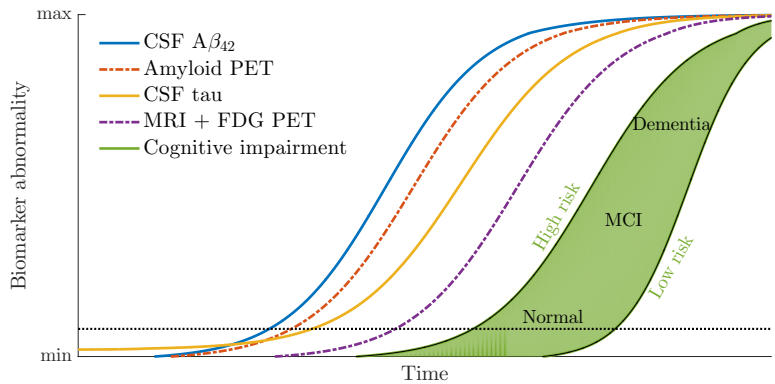


Figure 2.2: A model of the biomarkers in the Alzheimer's disease pathological cascade, where the dotted horizontal line represents the detection threshold. FDG (fluorodeoxyglucose) PET refers to a PET tracer developed to assess glucose metabolism. The figure is redrawn from the model presented in Jack et al. (2013).

the CSF represents the first (detectable) phase. This is followed by abnormal tau levels in the CSF (PET imaging of tau was still in a very early phase in 2013, but it has since been suggested that tau abnormality can be detected earlier in CSF than in PET (McDade and Bateman, 2018)), and brain atrophy before the onset of clinical symptoms. This model also highlights that some high-risk individuals may decline in cognition faster than others.

Chapter 3

Artificial neural networks

The enormous interest in *artificial intelligence* (AI) in the last decade is actually rooted in the success of *neural networks* (NNs). The increasing amount of data and computational resources has led to what we often call *deep learning*, which has achieved human-level performance in numerous fields and applications (Lecun et al., 2015). "Deep" in this context refers to that we now can train neural networks with many more layers ("deeper") than what was possible 30 years ago, although much of the fundamentals of NNs were established already in the 1950's (Haykin, 2008).

This chapter provides some necessary background to this branch of machine learning known as neural networks. My aim is to try to provide "non-ML-practitioners" an overview of how and why these methods work rather than a detailed mathematical description of modern neural networks. I start by describing the smallest computational unit, the *artificial neuron*, where multiple neurons together form an artificial neural network. I go on detailing *convolutional neural networks* (CNNs), which play a fundamental role in image recognition tasks today. They are also a key component in the works of this thesis together with *recurrent neural networks* (RNNs)—further detailed in Chap. 4.

Much of the information in this chapter regarding the basics of neural networks comes from the excellent resources Haykin (2008) and Goodfellow et al. (2016).

3.1 Multilayer perceptrons

The term *multilayer perceptron* (MLP) is commonly used do describe a feed-forward neural network where each neuron in a layers is connected to all neurons in the next layer [1]. In Fig. 3.1 we see a sketch of an MLP with one hidden layer, together with a more detailed illustration of a single neuron on the right.

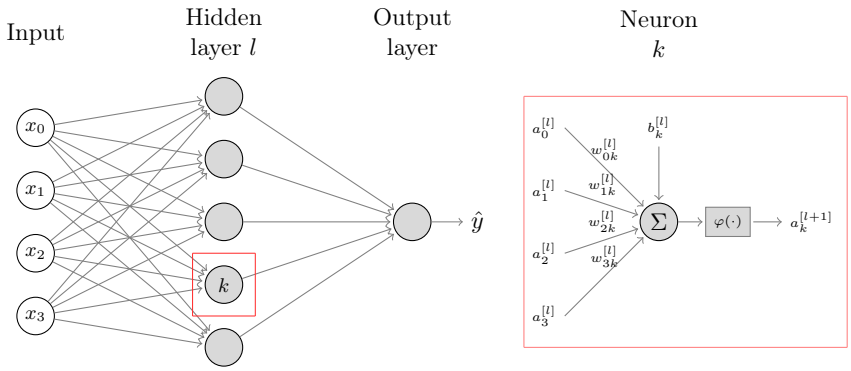
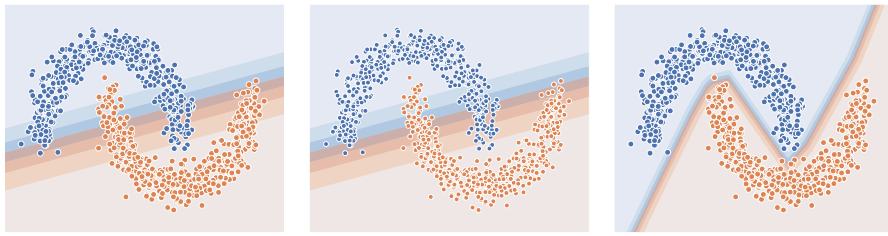


Figure 3.1: Schematics of a multilayer perceptron with input variables x_0, x_1, x_2 and x_3. These get propagated through the network and finally outputs $\hat{y}$. An illustration of a single neuron k (or perceptron) in hidden layer l is shown in the red box on the right, where the output $\mathbf{a}^{[l]}$ from a previous hidden layer is weighted and summed. This sum is passed through a non-linear activation function $\varphi(\cdot)$, e.g. a sigmoid function, and the output is forwarded as input to the next layer of neurons or to the output layer.

The neuron is the smallest computational unit of the network, inspired by the human neuron, and this model is called *Rosenblatt's perceptron* (Rosenblatt, 1958). Translating the model into how an actual neuron would work would go as follows: a neuron receives stimuli from other neurons in the form of chemical potentials to its dendrites, where x_i would be the signal received at the i:th dendrite. This input gets weighted by a factor w_i depending on the importance of the signal at the ith dendrite. The soma sums the weighted inputs $w_i x_i$ from all dendrites.

[1] Although a strict definition of MLPs includes having a step function as activation function.



(a) (b) (c)

Figure 3.2: Illustration of how well (a) a single neuron, (b) an MLP with one hidden layer but no activation function (i.e. $\varphi(z) = z$), and (c) which is the same architecture as in (b) but with $\varphi(z) = \max(0,z)$, can learn non-linear patterns. Each dot represent a data point with two input variables (x_1, x_2) belonging to one of two classes, and the background color show what class belonging the networks will predict for any (x_1, x_2) pair.

If this sum $\sum_i w_i x_i$ is greater than some threshold (bias) b, then the neuron "fires" (or "activates") and propagates (forwards) a signal in the neural network. What to propagate is computed through an *activation function* $\varphi(\cdot)$. These steps can be described mathematically as

$$a_k^{[l+1]} = \varphi\left(\sum_{j=0}^{m} w_{kj}^{[l]} a_j^{[l]}\right) \tag{3.1}$$

where $[l]$ is the specified hidden layer. In the first hidden layer ($l = 1$) the $a_i^{[0]}$ is the input data, i.e. $a_i^{[0]} = x_i$

A single neuron, as in the schematics on the right in Fig. 3.1, is the same as logistic regression if we use a sigmoid function as our activation function φ. These can only find linearly separable patterns (Minsky and Papert, 1988). By stacking multiple neurons in layers, we can create a neural network (i.e. an MLP) that can learn non-linear patterns as well. We illustrate this with an example in Fig. 3.2. In this figure, we also illustrate the importance of the activation function, which enables the network to learn non-linearities in the data.

Fig. 3.2c also demonstrates how we can use neural networks: if we input new data to a trained network we obtain a prediction of what

class this observation belongs to—just from having the network learn from previous observations.

3.1.1 Learning approaches

How do we train a neural network? There are three main concepts in machine learning.

Supervised learning

Supervised learning is perhaps the most intuitive approach to train a network, and it is the method used in this thesis. In a nutshell it means *learning from labeled data*, analogously to having a "supervisor" teach a "student" (i.e. the network) a task.

Assuming that we have some data $X = \{\mathbf{x}_i\}_{i=0}^{N}$ with associated labels $Y = \{y_i\}_{i=0}^{N}$, we want to learn a function f that maps $X \rightarrow Y$. For many problems, this function f is difficult or impossible to specify by hand. By training a neural network in a supervised way, we can approximate f through a network such as the one illustrated in Fig. 3.1. We do this by feeding input data $\mathbf{x}$ to the untrained network and compare the predicted output $\hat{y}$ with the label y. The weights of the network (θ, denoting all weights in the network) gets adjusted to minimize the difference between $\hat{y}$ and y. We elaborate on this in Sec. 3.1.2.

Unsupervised learning

Unsupervised learning is particularly useful when little or no annotated data is available. Thus, there is no paired label y_i associated to the observation $\mathbf{x}_i$ that we can use to optimize the network. Unsupervised learning aims to discover underlying patterns in the data.

Examples of common unsupervised learning methods in neural networks are *autoencoders* and *generative adversarial networks* (GANs).

An autoencoder is often used in conjunction with supervised learning, so called *semi-supervised learning*. It consist of two parts: an *encoder* network and a *decoder* network that can be trained end-to-end in a

supervised manner. The encoder maps the input data $\mathbf{x}$ to a lower-dimensional representation $\mathbf{z}$. The decoder tries to reconstruct the input $\mathbf{x}$ from $\mathbf{z}$. By minimizing the differences between the original input $\mathbf{x}$ and the reconstruction $\hat{\mathbf{x}}$, we force $\mathbf{z}$ to contain the most relevant and observation-specific information of $\mathbf{x}$ and removing redundancies—similar to a compression algorithm. This method was used e.g. by Payan and Montana (2015) for the purpose of pretraining, where they trained a neural network on the encoded $\mathbf{z}$ (discarding the decoding network) to diagnose AD from MRI images.

The branch of machine learning called GANs was originally proposed by Goodfellow et al. (2014). The idea is centered around game theory and the use of two neural networks competing against each other: one generative model G and one discriminative model D. We can use an example from medical imaging where the aim could be to generate synthetic CT images. The purpose of the G network is to generate as realistic synthetic CT images as possible, whereas the D network is trained to recognize whether an image is *real* (i.e. an actual CT image) or *fake* (generated by G). In the early training phases the images generated by G are poor and the D network would have no problem telling a real image from fake one. However, as training progresses the G network will produce more "realistic-looking images" and thus forcing D to be able to detect smaller and smaller differences between real and fake images.

Reinforcement learning

Reinforcement learning (RL) differs from the two previously discussed machine learning approaches. RL instead considers an *agent* acting in an *environment*, and is in a way the closest analogue to how humans learn: we explore the world by acting in it and where some actions reward us, which reinforces certain behaviors.

An example of an RL application can be letting a machine ("AI") learn how to play (video or board) games. Games are particularly suitable for developing RL methods, as it is 1) easy to quantify the state of the environment at any time, and 2) there is often a clear reward

signal to optimize towards. Using chess as an example, the *state* would be the positions of the pieces on the board (thus implicitly including the information of pieces already lost) and the reward would be capturing the opponent's pieces and winning the game. During training, the agent will start by playing (permissible) random moves, resulting in lost pieces and games. If the agent "accidentally" makes a good move and captures a piece, this will considered a reward and the agent will try to make more moves like this. By reinforcing this behavior (while occasionally exploring what happens if it makes an unexpected move) the agent will become a better and better chess player.

The most notable application of reinforcement learning is probably AlphaGo that beat the world champion of Go (Silver et al., 2016). This was considered to be a milestone in AI, as the game was believed to require "human intuition" [2]. A second version was later developed that was trained through self-play—completely without information from past (human) games (Silver et al., 2017). This model was called AlphaGo Zero, which could beat the previous "champion" AlphaGo. Similar advances have recently been demonstrated in team-based video games as well (Vinyals et al., 2019).

3.1.2 Training a network

This section is focused mainly on supervised training techniques, although many of the ideas discussed here pertains also to unsupervised and reinforcement learning. The aim of the supervised training procedure is to find the values of all adjustable parameters (such as weights and biases, collectively denoted θ) of f_θ that minimize a *cost function* J. An example of a cost function is the mean squared error (MSE) of all observations:

$$J = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - f_\theta(\mathbf{x}_i))^2 \tag{3.2}$$

[2] Due to the vast number of possible moves, it is impossible to calculate the optimal move by brute force.

where f_θ represents a network of neurons, described in Eq. (3.1). This cost function is thus small when predictions $\hat{y}$ are close to the labels y, and large if the predictions are far off.

During the training procedure we tune the parameters in θ to minimize this cost function J. To do this, we are interested in the property $\frac{\partial J}{\partial w_{ij}}$. That is: how does a change in a given weight w_{ij} affect the cost function J? This is equivalent to computing the gradient of the cost function with respect to the network parameters. If we know this entity, we can adjust the weights in the direction that decreases the cost function. By repeating this step multiple times we minimize the prediction error, ideally approaching a good approximation of the "true" function f. To calculate the partial derivative $\frac{\partial J}{\partial w_{ij}}$ can be computationally expensive. *Backpropagation* is a method to efficiently calculate the gradient with respect to the weights (Rumelhart et al., 1986), and is standard in modern ML frameworks.

There are a number of optimization methods that can be used to minimize the cost function

$$J(\theta) = \frac{1}{n} \sum_{i=1}^{n} J_i(\theta) \tag{3.3}$$

where the i subscript represents the ith out of n observations. That is, by summing the cost (prediction errors) of the n observations for all possible values of θ we would obtain a loss surface as a function of θ, and our aim is to find the values of θ that gives the smallest value of $J(\theta)$. Of course it would not be very efficient to assess all possible values of θ, particularly not as modern deep learning architectures often contain millions of parameters.

In *gradient descent* methods, we start with random weights and biases θ and calculate the gradient (slope) of the cost function $\nabla J(\theta)$ at this particular point on the loss surface. We then take a step of magnitude η (which is called the *learning rate*) in the direction in θ space that shows the steepest descent on the loss surface ($-\nabla J_i(\theta)$) and update the weights to this new position. Mathematically, each update

of a parameter $\theta_j \in \theta$ looks as follows:

$$\theta_j := \theta_j - \eta \nabla J(\theta_j) = \theta_j - \frac{\eta}{n} \sum_{i=1}^{n} \nabla J_i(\theta_j) \tag{3.4}$$

The optimization is thus done iteratively, and after some number of steps the solution will (hopefully) converge towards the global minima. The size of the learning rate η controls how large steps we take in each iteration. A large η can result in faster convergence, but may also cause us to "overshoot" the minima in our descent. In practice, one often gradually decreases η during training, as well as using a momentum term to dampen oscillations on the loss surface.

The method above is not very effective, as it involves calculating the gradient for *all* samples to do a single update. In stochastic gradient descent (SGD) we update the weights after a single (random) sample. This is faster—but also noisier—than performing gradient descent over the whole sample. The noise can however help to reduce the risk of the optimizer getting stuck in a local minima, which can occur due the smoother loss surface when averaging over the whole sample. Often one averages over mini-batches as it 1) is less noisy than SGD but more likely to not get stuck in local minima, 2) speeds up training because we can compute samples in parallel while still not having to run through all observations in each update.

There are other popular optimization algorithms used during training neural networks such as RMSprop and Adam (Kingma and Ba, 2014). Regardless of optimizer, it is customary to split the training data into three partitions: a training set, a development set, and a hold-out test set. The training set comprise the data on which we fit our network. However, at some point during training the network will likely start to *overfit* to the training data. That is, the network has learned features that is specific to the training data and does not generalize to unseen data. This is what the development set is for—to monitor the performance in data not used to fit the model on to be able to observe when overfitting occurs. However, as we are using the development set during the actual development of our model (such as tuning hyper-parameters) it does

not provide an unbiased estimate of the model's performance in unseen data. This is what the hold-out test set is for, which is to be assessed *once* at the end of the model development.

3.2 Convolutional neural networks

Convolutional neural networks (CNNs) are a class of neural networks particularly suitable for image related tasks. Instead of having a fully-connected neural network, such as in MLPs where all input variables are forwarded to all neurons, CNNs rely on *weight-sharing*. More specifically, they are built around *convolutions*.

The concept of convolutions is illustrated in Fig. 3.3, where we have an input image (in gray) that we wish to convolve with a *filter* (or *kernel*; in red). Images can be represented as a matrix containing pixel intensity values, and the filter as a matrix containing the values of the weights. We place the filter in the corner of the image, overlaying the pixel values. We multiply each pixel value with the filter weight on the overlaid position, and sum these values. We then move the filter to a new position, and repeat the multiplication and summation, until the whole input image has been filtered (convolved). The output of the convolution is a new 2D image called a *feature map*.

Describing a convolution as an equation (at a single filter position) looks as follows:

$$g[i,j] = W * x[i,j] = \sum_{s=-a}^{a} \sum_{t=-b}^{b} W[s,t]x[i-s, j-t] \tag{3.5}$$

where $*$ is the symbol for a convolution, W is the $(2a+1)$x$(2b+1)$ convolutional filter weight matrix, x the input image, and $g[i,j]$ is the pixel value of the feature map at position i, j. Note that this is very similar to the weighted summation of an artificial neuron in Eq. (3.1)! By adding an activation function $\varphi(\cdot)$ to Eq. (3.5) above we get the analogue of a neuron in CNNs. Having a small filter "slide" across the image is effectively weight-sharing of the network, and a filter can thus detect features regardless of their position in the image.

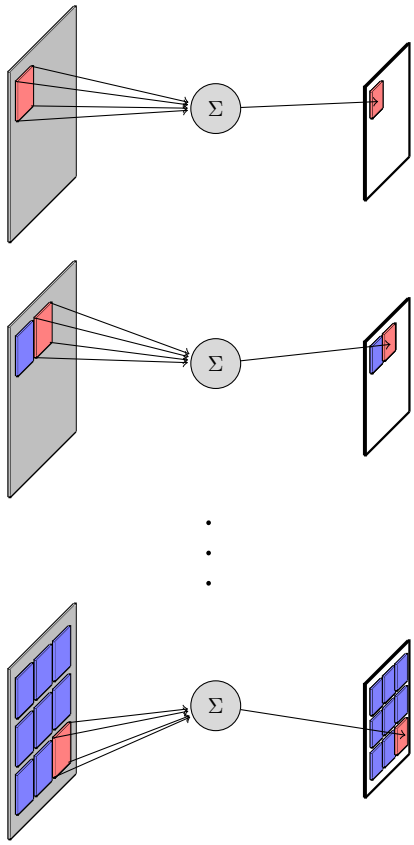
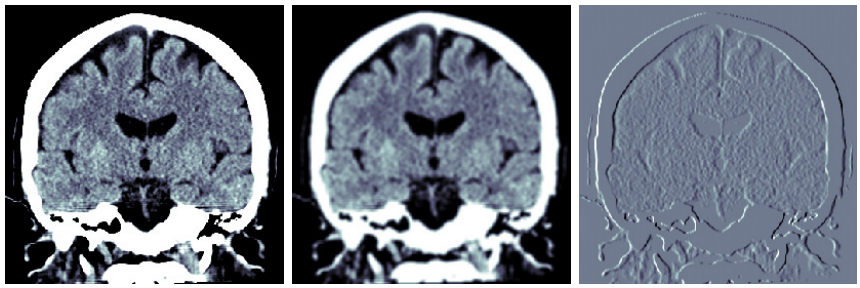


Figure 3.3: Sketch of a convolution, where a filter (in red) "slides" across the input image (in gray, on the left). The pixel values are multiplied element-wise, and the sum yields a feature map (right)

Convolutions are commonly used in image processing to filter images, where different filters yield different feature maps. In Fig. 3.4 we see examples of a CT image convolved with two different kernels. Fig. 3.4b shows a smoothing kernel (a low pass filter), that can be used to reduce noise in an image at the expense of blurring the image. In Fig. 3.4c, we see a Sobel kernel (a high pass filter) that effectively enhances horizontal edges in the image.



$$W = \frac{1}{9}\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \qquad W = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}$$

(a) Original image. (b) Mean filter. (c) Horizontal edge filter.

Figure 3.4: Examples of convolutional kernels W used in image processing, and their effects on the output. Connecting this to the illustration of convolutions in Fig. 3.3: (a) is the input image on the left; W is the red patch; (b,c) the feature map on the right.

Another way to to view the results in Fig. 3.4 is that different convolutional kernels can *detect* different features in the image. This is an intriguing property and gives us an intuition of why CNNs are so effective. It is the weights in convolutional filters that are learned during the training of the network. By stacking multiple convolutional filters to form a network it can learn to detect complex features in images which would be impossible to handcraft.

3.2.1 CNN architectures and modules

Here we will discuss some modules that are frequently used in CNNs, as well as some well-known network architectures. The aim is to provide a brief history of CNNs and to give a more detailed description of a CNN architecture that can be used for image recognition tasks.

One of the earliest successful implementation of a CNN trained using backpropagation was LeNet, which outperformed other models on recognizing hand-written digits (LeCun et al., 1989; Lecun et al.,

1998). It was the success of AlexNet (Krizhevsky et al., 2012), which demonstrated that deep CNNs trained on GPUs could yield superior image recognition performance, that reinvigorated the interest and funding of deep learning research.

We start with describing the VGG network (Simonyan and Zisserman, 2015) in more detail, as it is has a rather "clean" architecture (yet one that can still yield impressive performance). We use the VGG structure to explain how a basic feed-forward CNN can look and its components. An illustration of the architecture is shown in Fig. 3.5. It is based primarily on a module that is repeated multiple times in the network: a *3x3-filter convolution*, followed by a *rectified linear unit* (ReLU) activation function, and (sometimes) a *max pooling operation*.

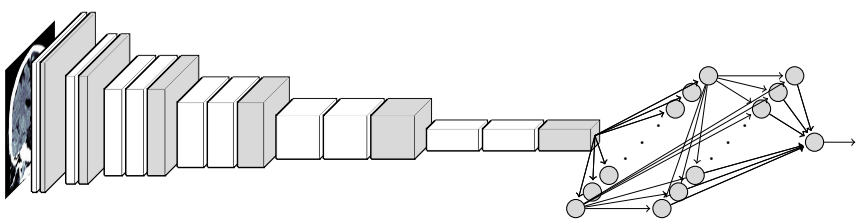


Figure 3.5: Sketch of the VGG16 network architecture. Each box represents a 3x3-filter convolution + ReLU activation. The gray boxes also includes a max pooling operation, reducing each dimension of the feature maps with a factor 2. The "length" of each box indicates how many filters is included in each layer, which increases deeper into the network. The output of the last block is flattened and used as input to a fully connected neural network that acts as a classifier.

We go through each of these steps in detail (omitting technical steps such as padding and striding). *3x3-filter convolutions* refers to convolutions with kernels of in-plane size 3x3 and a variable depth depending on the number of filters in the layer preceding it. Let us assume that we have an input image of size 3x224x224 (the "3" refers to the number of channels, e.g. RGB), and that we use 32 filters in the first two layers. Each of the filters in the first layer will have the dimensions 3x3x3. By convolving the input image with all filters we obtain 32 feature maps.

All these values are passed through a ReLU function—the activation function $\varphi(\cdot)$ used in VGG defined as

$$\text{ReLU} = \max(0, x) \tag{3.6}$$

which is a common choice of activation function in most neural networks today. The 32 feature maps from the first layer are used as input to the second layer with the dimensions 32x224x224 (i.e. 32 channels instead of 3). The filters of the second layer will thus comprise 32 kernels of size 32x3x3.

The output of the second layer (including ReLU activations) undergoes a *max pooling* operation. In the VGG network, this simply refers to dividing each feature map into 2x2 patches and forwarding only the maximum value in this patch to the next layer. This effectively reduces the input dimensions by a factor of 4, which speeds up training and lowers the memory consumption of the GPU. It also leads to substantially fewer neurons in the fully connected layers in the end, reducing the risk of overfitting.

The modules in the shallow layers consist of two blocks of 3x3-filter convolutions+ReLU activations, followed by max pooling. The deeper layers contains three of these blocks, as well as more filters in each block. Common versions of the VGG architecture are called VGG16 and VGG19, containing 16 and 19 of these blocks (layers), respectively.

The "convolutional part" of the network is responsible for extracting features of the input images. These representations, encoded in the output of the convolutional layers, are flattened into a 1D vector and used as input to a "regular" fully-connected two-layer NN. This final part acts as a classifier based on the extracted features from the image.

There are numerous more concepts that have been developed since the VGG network was proposed, and are standard in most state-of-the-art architectures today. We go trough some of them here, with a focus on the ones used in our model described in Chap. 4.

Batch normalization was introduced by Ioffe and Szegedy (2015) and provided faster training and more robust models, as well as regularization of the network. It first transforms the values of each input to have a

zero mean and unit variance, and then scales and shifts the values with parameters γ and β, learned during training, according to

$$y = \frac{1 - \mathrm{E}[x]}{\sqrt{\mathrm{Var}[x] + \epsilon}}\gamma + \beta \tag{3.7}$$

where ϵ is a small constant added for numerical stability.

Residual networks were first proposed by He et al. (2016) with the ResNet architectures. They introduced *skip-connections*, which enabled deeper networks than was previously possible with improved performance. If the input to a layer is denoted $\mathbf{x}$, and the output of the convolutions is $F(\mathbf{x})$, the skip-connections introduces $F(\mathbf{x}) + \mathbf{x}$, which is what is forwarded to the next layer. That is, it adds the input to the output of the convolution layers. This helps alleviate the problem of vanishing gradients, which can make it difficult to train deep networks. Residual units are part of most modern network architectures today. As an example, U-Net, which is arguably the most popular network architecture for medical image segmentation, relies heavily on residual connections (Ronneberger et al., 2015).

3.3 Recurrent neural networks

So far, the networks we have described deals with single input and single output cases, and are not practical for sequential data with varying input length. Examples of this can be time-series data or text data used for natural language processing. A type of networks that is suitable for sequential data is called *recurrent neural networks* (RNNs).

A simple illustration of an RNN is shown in Fig. 3.6. The idea is that each cell is similar to a single layer in a vanilla NN—multiplying a weight matrix with an input vector and passing the sum through an activation function. The difference is that an RNN cell takes the i:th sequence of the data *and* the output from cell at the previous time step, which is called the hidden state (h_{i-1})[3]. This means that the cell can propagate relevant features from the previous time step to the next,

[3]The initial hidden state h_0 is typically an array of 0's.

together with the new input. It also means that we can use data of arbitrary sequence lengths. Just as for vanilla neural networks we can construct multi-layered RNNs. The second layer would then take the hidden state h_i from the previous layer as input instead of x_i.

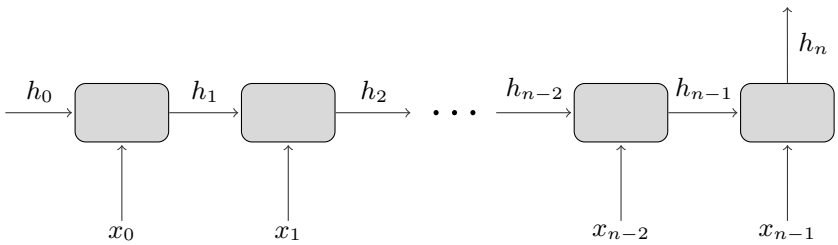


Figure 3.6: An example of a single-cell RNN, where the same cell (gray rectangle) at sequence i takes input x_i as input together with the hidden state h_i, which is the "internal memory" of the previous steps 0, 1, ..., $i-1$. The final state h_n can be used as input to a classifier.

Just as for convolutional neural networks there are different architectures for RNNs. Two popular architectures are Gated recurrent units (GRUs; Cho et al. (2014)) and Long short-term memory (LSTM; Hochreiter and Schmidhuber (1997)) cells. In this thesis we have used LSTM modules in our hybrid RNN-CNN network.

3.4 Domain shift in medical imaging

The phenomenon *domain shift* in the context of deep learning refers to when the training data comes from a different distribution than the data we wish to apply the model on (the "test data"). An example would be to train an image classifier on photos acquired with professional cameras where the aim is to apply the model on cell phone images. These would have visibly different quality that would not be an obstacle for a human, but the model will perform worse than on images from the training set distribution. We refer to to test data sampled from the same distribution as the training data as *within-distribution* data and external test data as *out-of-distribution* (OOD) data.

The domain shift topic has been studied surprisingly little in the context of medical imaging, where the implication of a failed prediction can have dire consequences. In neuroimaging, the domain shift can be attributed to a number of scenarios that could have a strong impact on the clinical applicability of machine learning models:

1. The clinical data is acquired with a different scanner, field strength and protocol than the images in the training set.
2. The patient belongs to a disease population not represented in the training data.
3. Image artifacts due to e.g. movement or metal implants, which are typically discarded in research settings.

Point 1) is particularly troublesome as it is impossible to include all relevant combinations of scanners and scanning protocols, where both hardware and software may be updated occasionally to improve image quality. Some studies have assessed the performance of DL models in data from external centers and have reported lower performance compared to their within-distribution test data (Kamnitsas et al., 2017; Perone et al., 2019; Yao et al., 2019). Albadawy et al. (2018) investigated the performance of a brain tumor segmentation tool on images from two institutions and found significant decreases in performance in OOD data compared to within-distribution test data. This study was conducted on a small set of 44 MRI images but nevertheless demonstrated the domain shift issue in neuroimaging data. Zech et al. (2018) investigated the performance of a CNN predicting pneumonia from chest x-ray images, finding lower overall performance in data from external centers. They further trained a classifier that managed to predict what center an image was acquired at with almost perfect accuracy. Since the disease prevalence was substantially different across sites, they performed additional experiments which led to the conclusion that their model leveraged the information of acquisition center in its predictions. This finding has also been reported for hip fracture radiographs (Badgeley et al., 2019). These studies give an example of why model performance can be lower in OOD data.

In Chap. 4 we investigate how a DL model is affected by the domain shift in multiple external memory clinic cohorts. We also address point 2) above to see if the predictive performance is degraded in disease populations not part of the training set.

Chapter 4

Automatic visual ratings of atrophy

4.1 Introduction

The way to measure neurodegeneration in clinics is typically not by software tools that can provide objective measures, such as volumes of specific brain regions, that allows us to track structural changes in the brain with high precision. In neuroimaging research we have been using software suites such as FreeSurfer, FSL, and SPM, which segment individual brains into anatomically distinct regions and calculate their volumes.

Instead, a trained neuroradiologist visually inspects each image and assigns an integer score of the degree of atrophy in a specific region according to established rating scales. Atrophy (or really its complement, as we technically measure what remains of a brain region, not the atrophy itself) is a property that is suitable to describe in terms of volume. So why is this not implemented widely in clinics? There are two main reasons for this:

1. Visual rating scales are well-established in clinical practice.
2. Software is (currently) not reliable enough.

To exemplify point 1): an MTA score (Scheltens' scale assessing medial temporal atrophy, detailed in the following section) of 3 conveys something to a clinician, but a hippocampal volume of 3100mm^3 does not (yet) give an intuitive picture of the degree of atrophy. There are also age related cut-offs in place for what is to be considered pathological for these scales. This is perhaps mainly a habitual issue, but as the mapping between e.g. HC volume and MTA score is not one-to-one (Wahlund et al., 1999; Cavallin et al., 2012a), clearly defined volumetric cut-offs would need to be established. Point 2) involves the concept of "reliability", which is expanded on in Sec. 4.2.1. Let us illustrate this with an example in which a patient undergoes an MRI scan, and the implemented software outputs a poor hippocampal segmentation and thus an inaccurate volume estimation. If this error goes undetected, the information may lead to a misdiagnosis of the patient at worst. If it is detected (either by the software or through manual quality control), it is very difficult to "save" that image in a clinical setting—at least in a time efficient manner[1]. One may thus need to fall back on visual ratings anyway in these cases. Further, since CT images have lower gray/white matter (GM/WM) contrast and are often acquired with a slice thickness that prohibits rendering them in 3D, GM volumes are not possible to estimate reliably. So implementing a software must outweigh the advantages of having two separate measures of atrophy for CT and MRI.

4.2 Visual rating scales

There are a number of proposed visual rating scales assessing different regions of the brain, providing fast and robust ways to quantify neurodegeneration (Wahlund et al., 1999). Some of the most commonly used scales (Vernooij et al., 2019) are listed in Table 4.1 (see Harper et al. (2015) for a more extensive review). All these scales, except for Fazekas', provide a framework to quantify gray matter atrophy through

[1] "Save" here means that if a segmentation fails, we would need to manually delineate the structure in order to calculate its volume.

Table 4.1: Description of commonly used rating scales in clinics.

Scale	Developer	Structure	Range	Modality
MTA	Scheltens et al. (1992)	Medial temporal lobe atrophy	0-4	T_1, CT
Fazekas'	Fazekas et al. (1987)	WM hyperintensities	0-3	T_2, FLAIR, CT
PA	Koedam et al. (2011)	Posterior atrophy	0-3	T_1, FLAIR, CT
GCA	Pasquier et al. (1996)	Global cortical atrophy	0-3	T_2, CT

a discrete scale ranging from 0-3 or 0-4, where the lowest score means "no atrophy" and the highest "end-stage atrophy". In this thesis we aimed to create a model that can predict scores of three rating scales: Scheltens' scale of medial temporal atrophy (MTA), Koedam's scale of posterior atrophy (PA), and Pasquier's frontal subscale of global cortical atrophy (GCA-F). Visual examples of these scales are shown in Fig. 4.1.

Scheltens' MTA scale is the most common scale to report in clinics today (Vernooij et al., 2019). A radiologist visually assesses the hippocampus, the choroid fissure and the inferior lateral ventricle (ILV) and provides a discrete score for each hemisphere according to Table 4.2. Several studies have reported on the diagnostic ability of the scale in distinguishing between healthy controls and AD patients (Scheltens et al., 1992; Wahlund et al., 1999; Westman et al., 2011a). It is rated in a single coronal slice from a T_1-weighted MRI or CT image. In the original paper proposing the scale by Scheltens et al. (1992), the assessed images were acquired with 5mm slice thickness parallel to the axis of the brainstem. With the emergence of 3D protocols in MRI, a radiologist would nowadays rotate the image to align the anterior and posterior commissures—so called AC-PC alignment—before locating the coronal slice just posterior to the amygdala and mammillary bodies from which the rating is performed. Despite its diagnostic value, the MTA scale has still been argued to be underreported in clinics (Torisson et al., 2015; Håkansson et al., 2019).

The PA scale for posterior atrophy was proposed by Koedam et al.

Table 4.2: Description of the MTA scale, where N denotes normal width or height. Adapted from Scheltens et al. (1992).

Score	Width of choroid fissure	Width of temporal horn	Height of hippocampal formation
0	N	N	N
1	↑	N	N
2	↑↑	↑	↓
3	↑↑↑	↑↑	↓↓
4	↑↑↑	↑↑↑	↓↓↓

(2011) to offer an addition to the MTA scale in characterizing a patient's atrophy pattern. The scale has been shown to be valuable in diagnosing AD in the absence of abnormal MTA scores, such as in early-onset AD (Lehmann et al., 2012; Möller et al., 2013). The image is rated in all three anatomical planes, ideally in both a T_1-weighted and a FLAIR sequence. It focuses on particular structures in the parietal lobe, namely the posterior cingulate sulcus (PCS), the precuneus, the parieto-occipatal sulcus (POS), and the parietal cortex. The ordinal scale reflects increased widening of PCS and POS, and increased atrophy in the precuneus and parietal cortex, illustrated in Fig. 4.1.

Pasquier et al. (1996) developed a scale for visual assessment of cerebral atrophy in 13 different brain regions. The scale has since been simplified into a global assessment of cortical atrophy rated from 0 (absent) to 3 (severe) called the GCA scale. The frontal subscale of GCA (GCA-F) rates the degree of frontal atrophy, which has been shown to be associated with executive dysfunction (Elliott, 2003). The GCA-F scale can aid in the diagnosis of FTLD (Ferreira et al., 2016). Pasquier et al. (1996) suggested to use T_2-weighted images for the assessment, but multiple studies have rated GCA in T_1-weighted images (Ferreira et al., 2016, 2017; Scheltens et al., 1997).

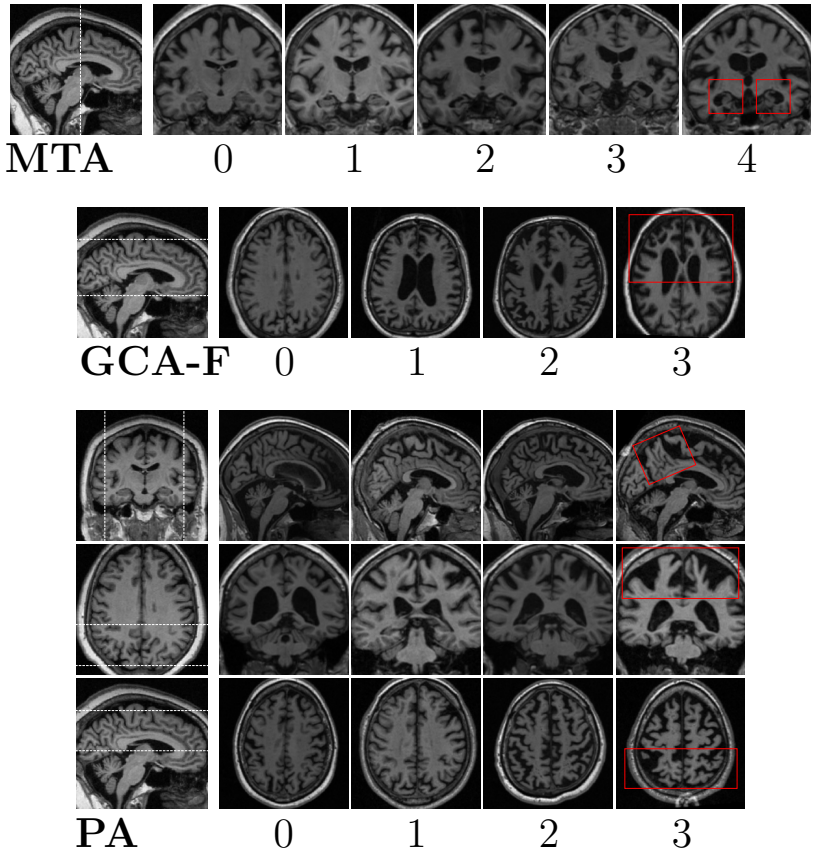


Figure 4.1: Example of visual rating scales used clinically today. The area between the white dotted lines in the left images show which slices are being assessed in the respective scales. MTA is rated in a single coronal slice. The red boxes indicate the regions that are being assessed in each scale.

4.2.1 Reliability of human ratings

Quantifying structural brain changes through visual assessment means that the ratings are subjective. That is, two radiologists assessing the same set of images may not assign the same ratings—commonly referred to as *inter-rater variability*. A less experienced radiologist practices

together with a more senior radiologist to "become" a reliable rater. It has been shown that the inter-rater agreement between radiologists not working together can be low, and that the *intra-rater agreement* drops for a rater if visual assessments are not performed on a regular basis (Cavallin et al., 2012b). This implies that if you pick two (experienced) radiologists at random, chances are that their rating agreement will be low. This may be due to them having different "rating styles", by which we mean that one of the radiologists is more conservative than the other. Due to the absence of ground truth ratings it is not easy to say which rater is "more reliable". One can compare ratings against volumetric measures of the involved structures, or planimetrics, but a stronger anticorrelation between e.g. MTA and hippocampal volume may not necessarily imply a "better" rater (although some level of correlation is of course required). We further assume a reliable rater to have a low intra-rater variability, which would indicate consistency.

An issue of most the common volumetric software tools used in research is that images acquired with different scanners will yield variations in the segmentation maps (Guo et al., 2019). This phenomena, i.e. performance drops in images from a different cohort, has been demonstrated in ML models in medical imaging as well (Klöppel et al., 2015; De Fauw et al., 2018; Zech et al., 2018). This domain shift problem is something that humans seem capable to handle, demonstrated by e.g. Wattjes et al. (2009) who showed excellent rating agreement between image modalities (CT and MRI).

Thus, "reliability" is a term that does not only entail great inter- and intra-rater agreement, but also the ability to assess images of low quality, with image artifacts, and from a wide range of scanners and protocols.

4.3 Automatic visual ratings

Visual ratings of atrophy currently have, as suggested in the previous section, many practical advantages over volumetric measures. The limitation of subjectivity, which gives rise to inter- and intra-rater variability,

is overcome if we can create a model to predict the ratings. Given the recent progress and capabilities demonstrated by deep learning applications—and the vast amount of rated images used in previous studies by our group—an approach based on convolutional neural networks was a suitable choice. As most deep learning models in medical imaging are assessed on within-distribution data, we were interested in how our model would perform in external memory clinics. That is, to investigate the domain shift in a systematic manner to learn in what ways a deep learning model may fail if naively implemented in new clinics.

The motivation behind developing the tool (which we will refer to as *AVRA* (Automatic Visual Ratings of Atrophy) from here on) was not to demonstrate the usefulness of deep learning in medical imaging but to provide new insights that may be useful to clinicians. We therefore applied AVRA on longitudinal data of individuals with SCD and MCI to investigate the progression rates of medial temporal atrophy expressed in MTA ratings and how they relate to the volumes of the subcortical structures assessed in the scale.

The aims of the studies in this thesis involving visual ratings can be summarized as follows:

- Develop an automated tool that can predict commonly used visual rating scales.
- Assess the model performance in out-of-distribution clinical data and compare rating agreement to external radiologists.
- Apply the model to research data to gain clinically relevant insights of the MTA scale in preclinical dementia.

In the following sections we will describe how we investigated each of these aims and what we learned from the process.

4.3.1 Network architecture

The choice of model architecture is connected to the training procedure as it is an iterative process, where different networks and hyperparameters

Table 4.3: Distribution of the ratings used for training during the development of AVRA. The MTA columns includes ratings of both left and right hemisphere.

Cohort	**Images**	**MTA**					**GCA-F**				**PA**			
		0	1	2	3	4	0	1	2	3	0	1	2	3
ADNI	1966	425	1581	1147	555	224	1449	468	49	0	1188	611	157	10
MemClin	384	23	265	296	139	45	279	89	14	2	210	127	43	4
Total	2350	448	1846	1443	694	269	1728	557	63	2	1398	738	200	14
Ratio (%)		9.5	39.3	30.7	14.8	5.7	73.5	23.7	2.7	0.09	59.5	31.4	8.5	0.6

are assessed to find the optimal performance on the development set. During the development phase of AVRA we trained and evaluated the model in data from the memory clinic at Karolinska Sjukhuset (*MemClin*) and the Alzheimer's disease neuroimaging initiative (ADNI), making up a set of 2350 images in total. The characteristics of these two cohorts are described in Table 4.5, together with the external memory clinic data used to assess the domain shift (we elaborate further on these in Sec. 4.4.3).

We pooled the ADNI and the MemClin data and split the set into a training (80%) and a test set (20%). The training set was subsequently split into five partitions for cross-validation, where we enforced similar distribution of ratings in each subset. This was important mainly for the GCA-F and PA models where cases with high ratings were scarce, see Table 4.3. All images were rated by a single expert neuroradiologist (Lena Cavallin), who has previously demonstrated excellent inter- and intra-rater agreement in research and taught inexperienced radiologist in how to perform these ratings (Cavallin et al., 2012b,a).

Apart from being able to obtain good performance, we also wanted to use the same (except for input dimensions) network architecture for all three rating scales. Ideally, we also wanted to use an architecture that would we suitable for CT images, a modality far more common than MRI in the diagnostic workup (Falahati et al., 2015) and hence where a tool such as AVRA would have the greatest clinical impact. This meant that we could not assume that all images were isometric 3D volumes.

We settled on a hybrid network architecture where we use a CNN to extract features that are propagated to an LSTM network (Hochreiter and Schmidhuber, 1997; Gers et al., 2000). A sketch of the network architecture is shown in Fig. 4.2. This solution is inspired by how a radiologist would visually assess an image: scrolling through the image slice-by-slice and remembering relevant information from each slice, and in the end give a composite score based on this information. The architecture can handle input of different slice thickness and anatomical planes while remaining relatively "light" ($\sim$1.5 million weights). It would thus work on all three rating scales that we aimed to automate, with the potential to be applied also to 2.5D CT images. Using recurrent convolutional neural networks for MRI images is not very common, but not novel either (Ypsilantis and Montana, 2016; Poudel et al., 2017; Grewal et al., 2018), and has previously been applied to video predictions (Karpathy et al., 2014; Donahue et al., 2015). In our initial experiments during the network development phase this architecture yielded consistent results across all three rating scales, whereas more common 2D and 3D CNN architectures performed poorly on the PA and GCA-F scales.

For the feature extraction we used a slimmed version of the Residual attention network developed by Wang et al. (2017), see Fig. 4.3. It effectively combines the properties of residual modules—which allow for deeper networks—and spatial focusing attributed to the attention modules (Bahdanau et al., 2015; Xu et al., 2015). Initial experiments showed that a slimmed version (i.e. fewer filters in each layer) did not yield a drop in performance compared to wider nets but reduced GPU memory consumption substantially, facilitating the hyperparameter tuning. The PyTorch (Paszke et al., 2017, 2019) implementation of AVRA's architecture can be found in `github.com/gsmartensson/avra_public`.

In the MTA scale each hemisphere is rated individually. (This can be done for the other scales as well, but we did not have enough bilateral ratings for this). In AVRA, only the left hemisphere is rated in each forward pass. To predict MTA of the right hemisphere we mirror the input image.

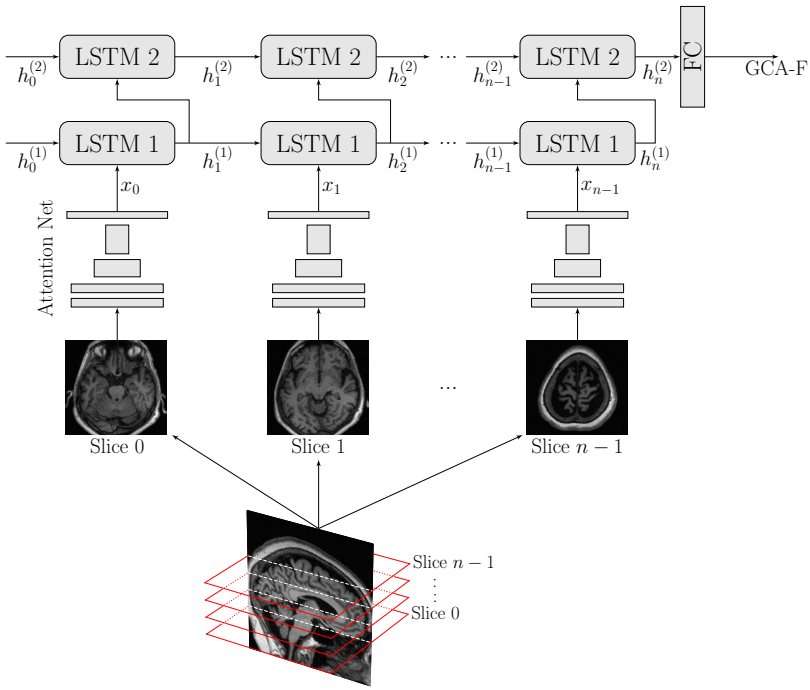


Figure 4.2: Overview of the network architecture of AVRA, used for all three rating scales. Features from each slice are extracted by a residual attention network and used as input to a two-layer LSTM network. Once all slices have been processed, a fully connected (FC) neural network is used to predict the score.

4.3.2 Preprocessing and data augmentation

Prior to training, all images underwent a registration procedure to the MNI brain using FSL FLIRT 6.0 (FMRIB's Linear Image Registration Tool) (Jenkinson et al., 2002; Jenkinson and Smith, 2001; Greve and Fischl, 2009). This was a rigid registration, i.e. rotation and translation, which is similar to an AC-PC alignment that radiologist performs (or accounts for) when rating. This was very convenient from an engineering perspective as the registered images are conformed to the same isometric resolution (1x1x1mm^3), and the brains are centered. The procedure makes it easier to locate the rating slices automatically. For example,

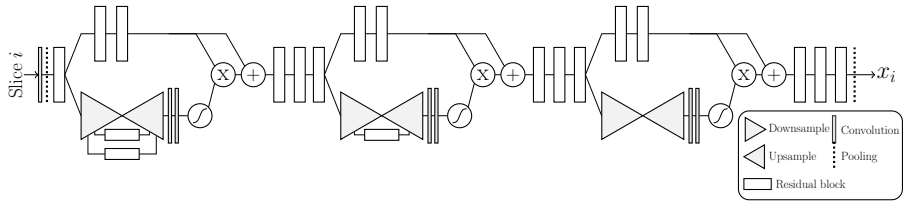


Figure 4.3: Schematics of the residual attention network, which constitutes the CNN part of the AVRA architecture. The downsampling block comprised repeated maxpooling operations and residual modules, whereas the upsampling was done through bilinear interpolation. Flow chart is redrawn based on Wang et al. (2017).

the MTA scale is rated in a single slice but the issue of (automatically) locating this particular slice is still a practical challenge. Since we want the model to learn from a single rating on each image (as opposed to segmentation tasks were the annotations contain much more information) we want to remove as many redundant slices as possible while still being certain that the correct rating slice is fed to the network. (When rating new images the registration is done within the pipeline.) During the training and inference we normalize the intensity of each individual image to have a zero mean and unit variance.

For the MTA scale the MRI volumes are cropped to comprise 22 coronal slices of size 128x128mm^2 with 1x1mm^2 resolution. (An example of the cropping in the coronal plane is shown in Fig. 4.4.) To the GCA-F model we feed the network 40 axial slices of size 160x192mm^2 of the frontal lobe, with 2mm slice thickness during training and 1mm during evaluation. As the PA scale is assessed in all three planes, we stack 37 axial, 28 coronal and 34 sagittal slices from the parietal lobe of size 128x128mm^2 and 2mm slice thickness as input the network. The approximate cropping areas are illustrated in 4.1.

During training data augmentation was performed by random shifting of center crop voxel within $\pm$10mm, scaling and mirroring. Some subjects in the ADNI cohort had multiple images from the same timepoint (typically from the same scanner session but some had both 1.5T and 3T acquisitions). For those cases a random image was chosen during training.

4.3.3 Training procedure and hyperparameters

We trained five (one per cross-validation set) separate models for each rating scale. Since the scales are ordinal (and discrete) we chose to treat the task as a regression problem, as opposed to a classification problem. That is, instead of trying to predict the most likely rating (class) the model predicts a single continuous score. During training, we calculate the MSE between the prediction and the (discrete) radiologist rating as our cost function. Other hyperparameters used to train the models were:

- Number of epochs: 200
- Optimization method: Stochastic gradient descent (SGD[2]) with momentum=0.9 and no weight decay.
- Learning rate: Varying cyclically between 0.01 and 0.0005 governed by cosine annealing schedule restarted after 100 epochs (Loshchilov and Hutter, 2016; Huang et al., 2017).
- Minibatch size: 20

The images were randomly shuffled during training, but where less frequent ratings were sampled more often to alleviate the class imbalance issue. Random oversampling has previously been shown to improve performance of convolutional neural networks (Buda et al., 2017).

4.4 Assessing performance

4.4.1 Performance metric

There are multiple aspects that can be considered when evaluating the *performance* of a model. First of all, the choice of metric to assess performance is important, where a single metric may not be sufficient. We have mainly used Cohen's (linearly) weighted kappa κ_w in this thesis, in combination with mean squared error.

[2]As implemented in PyTorch.

Kappa statistics is commonly used in the literature to quantify inter- and intra-rater agreement, and visual ratings studies are no exceptions. It takes into account the distribution of ratings and how likely a rater is to get the "correct" rating due to chance. (E.g. in a sample consisting of 99% healthy individuals and 1% infected a model would achieve a 99% accuracy by just predicting "healthy", but it would not be a very useful model). The kappa metric was introduced to account for this, where the weighted variant considers the cases where the different categories are ordinal (Cohen, 1960, 1968). The κ_w measure is a suitable metric to use for our model as it is well-established in the literature and can give us an idea of the "human-level performance". It is also a metric that the radiological community is familiar with.

This raises the question whether it is the inter- or intra-rater agreement levels that we should aim for as the performance target, as AVRA is trained to mimic the rater of our test set. As a reference, 244 images were rated more than once, yielding the radiologist intra-rater agreement of $\kappa_w = 0.83$ (MTA left); $\kappa_w = 0.79$ (MTA right); $\kappa_w = 0.46$ (GCA-F); $\kappa_w = 0.65$ (PA). These are slightly lower than previously reported intra-rater agreements (notably so for the GCA-F scale) and could be due to that the time between the rating sessions were long: up to 16 months.

The drawback of using κ_w for our application is that it requires integer ratings, forcing us to round AVRA's prediction to the nearest integer. That means that a continuous rating of 1.49 and 0.51 is considered "just as wrong" if the radiologist rating is 2. The MSE metric considers AVRA's ratings as continuous, and is thus more sensitive, but possibly at the expense of being less intuitive.

There are other ways to assess the performance of a model such as AVRA though:

- How does the model perform in external (clinical) data?
- How does AVRA's ratings compare to other radiologists?
- Are AVRA's ratings sufficiently sensitive and robust to be applied in longitudinal data?

These points are further discussed in Secs. 4.4.3, 4.4.4 and 4.4.5, respectively.

4.4.2 Within-distribution data

When evaluating on test data drawn from the same distribution (cohorts) as the training data, we used the five models (from the cross-validation training) as an ensemble model considering their average prediction as AVRA's rating. The model predicted rating agreements similar to inter-rater agreements between two radiologists previously published, see Table 4.4 [3]. Compared to the human intra-rater agreement levels on the training set ("Rad. in study" entries), we see that AVRA achieved slightly higher scores for the GCA-F and PA scales. This was a bit strange, as those κ_w values should suggest an upper limit of AVRA's performance on that rating scale. We believe that this can partly be explained by the skewed rating distributions of the GCA-F and PA scale, as highly uneven rating prevalence can hurt the kappa value (Byrt et al., 1993). We still argue that kappa, despite this caveat, is the preferred metric to use due to its popularity in the field, even though the GCA-F results may be difficult to interpret.

To understand why the MTA model did not reach the radiologist's intra-rater agreement level we looked deeper into some images with MTA ratings of 2's and 3's (Fig. 4.4). These would be the most critical ratings to predict correctly, as the threshold for what is considered pathological is in this interval[4] (Cavallin et al., 2012a). We let the radiologist re-assess the images in Fig. 4.4 (without knowledge of her previous ratings or AVRA's predictions) and describe the reasoning process behind her ratings. The cases that AVRA predicted correctly were given the same ratings on re-assessment, with the ratings of MTA (Rad: 2, AVRA: 2.4) and (2, 2.6) being described as "between MTA 2 and 3". The ratings (2, 2.8) and (2, 3.0) were re-rated as 3's, following the automated ratings. However, AVRA got the cases (3, 2.0), (3, 2.2)

[3] The selection was based on studies reporting Cohen's weighted kappa values.

[4] Slight variations of age-dependent cut-offs have been suggested in different studies, but an average MTA between 2-3 seems to be the most common.

Table 4.4: Previously reported weighted kappa agreements and the results of the trained models in this study. "Rad. in study" refers to the calculated intra-rater agreement on the images rated more than once in the ANDI cohort. The "VGG16" entries show the results when training a VGG16 network on the same data for comparison.

Study	Scale	N	Intra-rater agreement (κ_w)	Inter-rater agreement (κ_w)
Cavallin et al. (2012b)		100	0.83-0.94	0.72 - 0.84
Cavallin et al. (2012a)		100	0.84-0.85	—
Westman et al. (2011a)		100	0.93	—
Velickaite et al. (2017)		20/50	0.79-0.84	0.6-0.65
Ferreira et al. (2017)	**MTA**	120	0.89-0.94	0.70-0.71
Koedam et al. (2011)		29/118	0.91-0.95	0.82-0.90
Rad. in study		244	0.79-0.83	—
VGG16		464	1	0.58 - 0.59
AVRA		**464**	**1**	**0.72 - 0.74**
Ferreira et al. (2016)		100	0.70	0.59
Ferreira et al. (2017)		120	0.83	0.79
Rad. in study	**GCA-F**	244	0.46	—
VGG16		464	1	0.56
AVRA		**464**	**1**	**0.62**
Koedam et al. (2011)		29/118	0.93-0.95	0.65-0.84
Ferreira et al. (2017)		120	0.88	0.88
Rad. in study	**PA**	244	0.65	—
VGG16		464	1	0.63
AVRA		**464**	**1**	**0.74**

and (3, 2.4) wrong. On closer inspection we see that there are so called "hippocampal adhesions" in both (3, 2.0) and (3, 2.2) that is likely the cause of the wrongful predictions. A radiologist would "mentally picture" how the image would look without this adhesion between the hippocampus and the cerebral white matter. These cases are not very prevalent, yet not that uncommon either, and from Fig. 4.4 it seems that the model has not learned to properly account for this.

To have a reference of the added value of using the fairly complicated network architecture of AVRA, we trained additional networks with the VGG16 architecture (Simonyan and Zisserman, 2015), only modifying

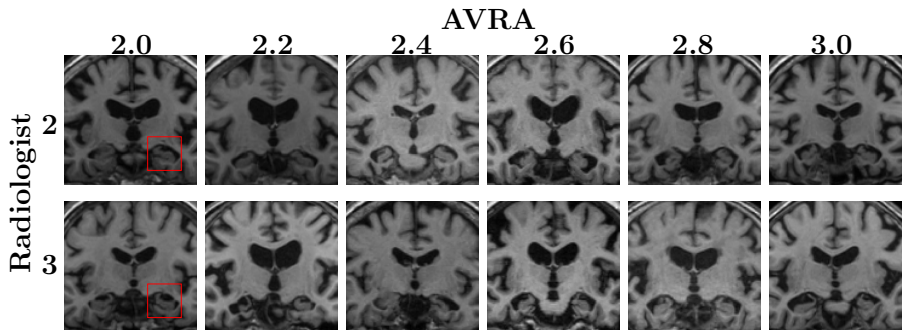


Figure 4.4: Post hoc qualitative assessment of predictions. Top (bottom) row: are all images from the test set with right-hand side, indicated by the red box, rated as MTA 2 (3) by the radiologist. Each column represents the score AVRA predicted.

the input dimension. The rating agreements for all three scales were lower than for the AVRA architecture. These results, together with our overall impression when exploring other architectures, suggest that the hybrid model provided added value over more common CNN architectures. However, we wish to emphasize that these results are merely suggestive, as we spent substantially less time optimizing the performance of the VGG model compared to AVRA.

4.4.3 Out-of-distribution data

In the previous section we showed that AVRA generalized well to new data coming from the same cohorts. For a model to have practical value (besides as proof-of-concept, which of course is valuable in other ways) it needs to show good performance in data from external cohorts as well. This includes images acquired from different scanners and protocols, but also from different sample populations than what comprised the training set. As clinical data can be difficult to acquire in the large quantities often necessary to train deep learning models, many models are trained primarily on public research cohorts. In the field of neuroimaging and Alzheimer's disease, the ADNI data set is arguably the most extensively used for machine learning due to its size, amount of clinical information (annotations) and image homogeneity. Multiple studies have developed

deep learning models trained and evaluated on ADNI data (see Jo et al. (2019) for a review), but few have investigated their model's performance in external cohorts.

We were interested in seeing how AVRA performed in images from external memory clinics. This data is more reflective of the images collected as part of the clinical routine, both in terms of image variability and disease population. To make the study more general, and of more interest to the medical machine learning community, we widened the scope to investigate what effect the level of heterogeneity of the training set has on performance in OOD data. To make the study more comprehensible, and to reduce the number of time and energy consuming training procedures, we focused only on the MTA scale since it is the most common scale used in the clinical routine and we have two ratings per images (one for each hemisphere).

The data sets used are described in Table 4.5. This included ADNI, AddNeuroMed (research cohort similar to ADNI in regards to disease population and scanning protocols), MemClin (used for training in the original study), and data from the European DLB consortium (*E-DLB*). This last cohort consisted of data from 12 memory clinics across Europe comprising healthy, AD, Parkinson's disease with dementia (PDD), and dementia with Lewy Bodies (DLB). From the E-DLB cohort we created subsets of data to isolate specific characteristics (described further in Table 4.5).

We were interesting in studying the performance in out-of-distribution data, specifically in:

- Clinical data from multiple institutions.
- Research data, similar to the training data.
- Different disease populations (DLB and PDD).

Table 4.5: Overview of data sources used for training and/or evaluation, and why these specific subsets were of interest. N_{train}/N_{test} refers to the number of labeled images used during training/evaluation, where some cohorts were split into training and test sets. Abbreviations: Deep Learning (DL); Out-of-distribution (OOD) data; Alzheimer's disease (AD); Healthy controls (CTR); Frontotemporal lobe dementia (FTLD); Dementia with Lewy Bodies (DLB); Parkinson's disease with dementia (PDD).

Cohort	**Scanners/Protocols**	**Disease population**	**Purpose of inclusion**
ADNI N_{train}=1568 N_{test}=398	Multiple scanners and sites, but strictly harmonized with phantom. Both 1.5T and 3T.	AD spectrum and CTR.	Common cohort to train and evaluate DL models in, which we hypothesize should not generalize well.
AddNeuroMed N=122	Harmonized, designed to be compatible with ADNI.	AD patients only.	Assess AVRA in an external research cohort similar to ADNI.
MemClin N_{train}=318 N_{test}=66	Unharmonized, part of clinical routine from a single memory clinic.	Mainly AD spectrum and CTR, with 37 FTLD patients.	Large clinical cohort with similar disease population as ADNI and AddNeuroMed.
E-DLB_{all} N=645	Retrospective unharmonized data of varying quality from 12 European sites as part of their clinical routine.	Mainly DLB spectrum, but also CTR, AD and PDD.	To assess performance of AVRA in a large, realistic clinical cohort.
E-DLB_{AD} N=193	Same as E-DLB_{all}	Only individuals with AD pathology from E-DLB_{all}.	To isolate effects of scanners/protocols not seen during training from disease population.
E-$DLB_{\{DLB,PDD\}}$ N={266,97}	Same as E-DLB_{all}	Only individuals with DLB or PDD pathology from E-DLB_{all}, respectively.	To assess the impact scanners/protocols *and* disease populations not seen during training have on AVRA performance.
E-$DLB_{\{25\%,50\%\}}$ N_{train}={173,312} N_{test}= 333	Same as E-DLB_{all}	Randomly selected images with a probability of 25% (or 50%) from all centers in E-DLB_{all}.	To assess effect of including training data from test set distribution has on AVRA performance.
E-$DLB_{\{C_1,C_2\}}$ N={101,165}	Both centers have used a single scanner (3T) and protocol.	Only images from center C_1 and C_2 from E-DLB_{all}, respectively.	"External validation sets": how would AVRA perform if deployed in two external memory clinics?
E-$DLB_{C_{3-12}}$ N=379	Same as E-DLB_{all}	All images in E-DLB_{all} *except* from center C_1 and C_2.	Large clinical cohort with a more heterogeneous disease population than MemClin.
Test-Retest N=72	Three Siemens scanners (two 1.5T, one 3T) with similar protocols but unharmonized.	Young (38 $\pm$ 13 years old) MS patients and healthy controls.	Systematic evaluation of the impact scanner variability has on AVRA predictions.

We constructed multiple training sets by combining subsets of these

cohorts in various ways. This allowed us to control the level of heterogeneity in the training set from low (only $\text{ADNI}^{\text{train}}$), to medium-low ($\text{ADNI}^{\text{train}}$ + AddNeuroMed), medium-high ($\text{ADNI}^{\text{train}}$ + $\text{MemClin}^{\text{train}}$), to high ($\text{ADNI}^{\text{train}}$ + $\text{MemClin}^{\text{train}}$ + data from some memory clinics in the E-DLB set: $\text{E-DLB}_{\text{C}_{3-12}}$).

In an effort to reduce confounding sources we kept the training set sizes fixed to $N = 1568$. This was the number of rated images in $\text{ADNI}^{\text{train}}$, which needed to be part of all training sets in order to reach adequate training set sizes. Further, all images were annotated by the same expert neuroradiologist, which removes the confounding factor of inter-observer variability[5]. When adding a second cohort to the training set, we removed subjects from ADNI with the same ratings. By doing this, the rating distribution was also kept fixed. To avoid the process of hyper-parameter tuning, we replicated the training procedure described in Sec. 4.3.3. This included training five individual models on 4/5 of the training set data. To provide some estimates of model variability we present mean and standard deviation of both κ_w and MSE from these five trained models—trained from scratch with different weight initialization and different partitions of that data—instead of a single metric from the ensemble model[6]. The MSE metric is a more sensitive measure to use for assessing how much the model performance degrades, whereas κ_w is useful in order to relate our results to human levels of rating agreement. Presenting the results in mean±std makes it easier to interpret trends and patterns in the results.

To study the generalization across disease populations we stratified the E-DLB cohort into subsets of patients with AD, PDD, and DLB pathology. Further, we selected two memory clinics, C_1 and C_2, where data was collected with a single scanner and protocol, as "external centers where we wish to implement AVRA". That is, assuming that we have developed a model showing good performance in within-distribution (and possibly even OOD) data, what would be the out-of-the-box performance

[5]Comparing AVRA's prediction to an second rater would be difficult to say whether rating disagreement is due to domain-shift or differences in rating styles. We look further into this in Sec. 4.4.4

[6]These results are included as supplementary data in Mårtensson et al. (2019)

if we implemented this into new clinics? (C_1 and C_2 do not fully reflect clinical data due to the single scanner acquisitions, but the consistency makes it easier to interpret the findings.)

In Table 4.6 we see the agreement (κ_w and mean squared error) between AVRA and the same radiologist for all training/test set combinations. By looking at trends in the agreements, we interpreted the results as follows:

- Within-distribution performance was higher than OOD performance (although still on an "acceptable" level in most test sets).
- The model generalized well to a similar research cohort (AddNeuroMed) when trained only on ADNI.
- OOD performance varied across memory clinics, notably in:
  - E-DLB$_{C_1}$, where the performance of AVRA was very low.
  - E-DLB$_{C_2}$, where the results were close to within-distribution test set performances when only trained on ADNI.
- Increasing heterogeneity of the training data improved the model's overall performance.
- The model generalized across disease populations (e.g. when training on AD cohorts and applying the model on a DLB cohort).

The first two findings were expected, as it has been demonstrated by multiple previous studies investigating the domain shift (Kamnitsas et al., 2017; Perone et al., 2019; Albadawy et al., 2018; Zech et al., 2018; Yao et al., 2019). The later points were more interesting. The fact that the results were so different in C_1 and C_2 was very difficult to predict *a priori*; both centers used a single 3T scanner and protocol, and from visual inspection the quality of all images were high. Additional experiments revealed that AVRA systematically predicted too low ratings in the E-DLB$_{C_1}$ data when only trained on ADNI. This discrepancy between rating agreements in C_1 and C_2 illustrates that it may be challenging to assess the OOD performance of DL models, as doing it in a single

Table 4.6: The agreement between AVRA's ratings and the radiologist's on different test sets (rows) when training models on different subsets of data (columns). The ✓ symbol indicates that the cohort on that row was included in the training set in that column. E.g. the first column shows κ_w and MSE for different test sets when trained only on ADNI, the second when trained on ADNI+AddNeuroMed, etc. There was no overlap of images or subjects in training and test sets for the reported results. The greatest agreement values for each test set are in bold.

Cohort	Cohorts incl. in training									
$\text{ADNI}^{\text{train}}$	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
AddNeuroMed		✓				✓	✓			✓
$\text{MemClin}^{\text{train}}$			✓		✓		✓			✓
$\text{E-DLB}_{\text{C}_{3-12}}$				✓	✓	✓	✓			
$\text{E-DLB}^{\text{train}}_{25\%}$								✓		
$\text{E-DLB}^{\text{train}}_{50\%}$									✓	✓
	Cohen's κ_w									
$\text{ADNI}^{\text{test}}$	0.67±.02	**0.69±.01**	0.67±.02	0.69±.02	0.66±.01	0.67±.02	0.67±.02	0.66±.01	0.67±.01	0.67±.02
AddNeuroMed	**0.66±.01**	—	0.64±.02	0.65±.01	0.61±.05	—	—	0.63±.03	0.63±.03	—
MemClin	0.62±.02	0.62±.02	—	0.63±.02	—	**0.64±.03**	—	0.62±.05	0.61±.03	—
$\text{MemClin}^{\text{test}}$	0.64±.04	0.65±.03	0.72±.03	0.67±.03	**0.74±.04**	0.66±.05	0.69±.02	0.65±.07	0.59±.05	0.71±.02
$\text{E-DLB}_{\text{all}}$	0.58±.02	0.58±.02	**0.61±.01**	—	—	—	—	—	—	—
$\text{E-DLB}^{\text{test}}_{50\%}$	0.59±.02	0.58±.01	0.60±.02	—	—	—	—	0.62±.02	0.63±.02	**0.65±.02**
E-DLB_{AD}	0.52±.03	0.52±.01	**0.57±.03**	—	—	—	—	—	—	—
$\text{E-DLB}_{\text{DLB}}$	0.59±.03	0.58±.03	**0.61±.01**	—	—	—	—	—	—	—
$\text{E-DLB}_{\text{PDD}}$	0.58±.04	0.58±.06	**0.60±.05**	—	—	—	—	—	—	—
$\text{E-DLB}_{\text{C}_1}$	0.30±.04	0.31±.04	0.49±.07	0.42±.07	0.51±.05	0.52±.05	**0.52±.03**	—	—	—
$\text{E-DLB}_{\text{C}_2}$	0.64±.04	0.61±.02	0.64±.01	0.64±.04	**0.65±.02**	0.63±.03	0.64±.02	—	—	—
	Mean squared error									
$\text{ADNI}^{\text{test}}$	0.31±.02	**0.29±.01**	**0.29±.01**	**0.29±.01**	0.32±.01	0.30±.02	0.30±.02	0.31±.01	0.31±.01	0.31±.02
AddNeuroMed	**0.27±.01**	—	0.28±.01	0.30±.01	0.32±.05	—	—	**0.27±.01**	0.29±.03	—
MemClin	0.34±.02	0.31±.02	—	0.31±.02	—	**0.28±.02**	—	0.31±.04	0.32±.02	—
$\text{MemClin}^{\text{test}}$	0.33±.02	0.29±.04	0.23±.02	0.27±.03	**0.22±.03**	0.26±.03	0.24±.01	0.29±.03	0.31±.04	0.25±.02
$\text{E-DLB}_{\text{all}}$	0.41±.02	0.41±.03	**0.36±.02**	—	—	—	—	—	—	—
$\text{E-DLB}^{\text{test}}_{50\%}$	0.41±.02	0.40±.03	0.36±.03	—	—	—	—	0.35±.02	0.34±.02	**0.33±.01**
E-DLB_{AD}	0.50±.05	0.48±.02	**0.39±.05**	—	—	—	—	—	—	—
$\text{E-DLB}_{\text{DLB}}$	0.41±.04	0.42±.03	**0.38±.01**	—	—	—	—	—	—	—
$\text{E-DLB}_{\text{PDD}}$	0.30±.03	0.30±.05	**0.27±.02**	—	—	—	—	—	—	—
$\text{E-DLB}_{\text{C}_1}$	0.83±.11	0.79±.13	0.49±.12	0.53±.09	0.46±.08	0.45±.04	**0.44±.05**	—	—	—
$\text{E-DLB}_{\text{C}_2}$	**0.28±.03**	0.32±.02	0.30±.01	0.30±.04	0.29±.03	0.30±.02	0.30±.03	—	—	—

external center is likely not enough. This may further be an obstacle for large-scale deployment of DL models in clinics, where models may need to be fine-tuned on data from the center it is implemented in, or at least carefully validated.

A more positive finding was that by using images from multiple sites in the training data, the overall OOD performance increased. We also did not notice any performance drops due to disease population. These results may be useful when deciding on what images to annotate when curating training data for DL applications in medical imaging.

A reliable model should provide the same output regardless of MRI scanner and protocol used to image the subject's brain. To assess this, we ran AVRA on images from nine subjects: six with Multiple Sclerosis (MS) and three healthy controls. These individuals, included in an earlier test-retest study (Guo et al., 2019), were scanned twice, with repositioning, in three different scanners. In Fig. 4.5 we see AVRA's (ensemble) predictions when trained on ADNI data only (top row) and when widening the training set to also include MemClin, AddNeuroMed and E-DLB$_{C_{3-12}}$ data (bottom row). We see that the differences are indeed quite small for most subjects, particularly when considering that it is a discrete scale. The predictions based on the images from the 3T scanner seems to be the greatest "outliers", which has been reported in a previous machine learning study as well (Abdulkadir et al., 2011). The within-scanner variability was very small, which suggests that AVRA could be reliable in longitudinal studies where follow-up data is generally acquired with the same scanner (or with a harmonized protocol).

Based on the experience from the OOD results, we trained a new version of AVRA on all available data described in Table 4.5. We denote this version v0.8, and is the trained model we made publicly available at `github.com/gsmartensson/avra_public`.

Figure 4.5: Boxplot of AVRA's ensemble ratings of left MTA (left column) and right MTA (right column) for all participants in the test-retest dataset. Top row: model trained only on ADNI. Bottom row: model trained on ADNI+AddNeuroMed+MemClin+E-DLB$_{C_{3-12}}$. Each subject was scanned twice with repositioning in three different scanners, and each image's AVRA rating is plotted in different colors depending on scanner. Individuals denoted with the prefix "HC" were healthy controls and "MS" were patients with Multiple Sclerosis.

4.4.4 Agreement to external radiologists and subcortical volumes

High rating agreement to a single radiologist is not sufficient to conclude that a model, such as AVRA, is reliable—even across multiple cohorts. Rating scales have been compared to volumetric measures and VBM analysis in previous literature as a mean to establish validity (Wahlund et al., 1999; Cavallin et al., 2012a; Ferreira et al., 2016; Möller et al., 2014), and we can assume that a "good" rater should show strong correlations to atrophy in the brain structures the scale assesses. A

reliable rater should also show good agreement to other experienced radiologists to be trustworthy. (These two statements also implicitly test that the radiologist who rated the training set is reliable.)

We ran AVRA v0.8 on data from the BioFINDER (Biomarkers For Identifying Neurodegenerative Disorders Early and Reliably) study (`www.biofinder.se`). In total, the analyzed data comprised 372 images from 93 subjects, acquired at 4 different time points over 6 years for each subject. These individuals were classified as SCD or MCI at baseline, i.e. in preclinical stages of dementia. Two neuroradiologists who performs ratings in clinics on a regular basis, rated all images according to Scheltens' MTA scale (blinded to age, sex, diagnosis and ID). They had not trained together prior to rating the images.

Additionally, we ran FreeSurfer's longitudinal pipeline to segment all images (Dale et al., 1999; Fischl et al., 2004; Reuter et al., 2012). We were mainly interested in hippocampal volumes and the volumes of the inferior lateral ventricles, as these are the structures which are part of the rating scale. (The choroid fissure is very small and often non-existent in the FreeSurfer segmentations). It is also the two structures which another study predicted MTA scores from using a piece-wise linear model with normalized volumes as input (Koikkalainen et al., 2019). Thus, the volumes of these structures should show a strong correlation to both AVRA's and the radiologists' ratings.

The rating agreement between all raters ("Rad. 1", "Rad. 2", and AVRA's predictions rounded to nearest integer), together with their respective Spearman correlations with the subcortical volumes, are shown in Table 4.7. Interestingly, the agreement with Rad. 2 was low for both Rad. 1 and AVRA, while their mutual agreement was around $\kappa_w \sim 0.6$. Does this mean that Rad. 2 is a "worse" rater than AVRA and Rad. 1? Not necessarily, and at least not in this case. Rad. 2's ratings showed similar correlations to the subcortical measures as the other raters. By looking at the violinplots in Fig. 4.6, plotting MTA against HC and ILV volumes respectively, it seemed that Rad. 2 had a tendency to give lower scores than AVRA and Rad. 1.
Spearman correlation was selected in favor of tau correlation since previous studies comparing HC volume to MTA ratings used this metric,

Table 4.7: Inter-rater agreements (κ_w) and Spearman correlations (r_s) between ratings and hippocampal (HC) volumes, inferior lateral ventricle (ILV) volumes, MMSE and ADAS delayed word recall. In the Spearman correlation tests we only used one timepoint per subjects to not violate i.i.d. assumption.

		Rad. 1		Rad. 2		AVRA	
Measure	Metric	Left	Right	Left	Right	Left	Right
Rad. 1	κ_w			0.30	0.36	0.58	0.61
Rad. 2	κ_w	0.30	0.36			0.30	0.35
AVRA	κ_w	0.58	0.61	0.30	0.35		
HC vol.	r_s	-0.58	-0.51	-0.58	-0.50	-0.58	-0.61
ILV vol.	r_s	0.82	0.82	0.85	0.87	0.89	0.89
MMSE	r_s	-0.54	-0.54	-0.49	-0.43	-0.45	-0.44
ADAS-DWR	r_s	0.51	0.51	0.55	0.52	0.56	0.56

allowing for comparisons. These studies reported modest associations of r_s between -0.26 and -0.37 (Wahlund et al., 1999; Cavallin et al., 2012a), which is substantially weaker than the values in Table 4.7 and adds to the evidence that the raters in this study were in fact reliable.

To conclude, by studying AVRA's ratings in relation to the radiologists' and subcortical volumes it seems that the model outputs reliable MTA predictions. At least they display the same characteristics as the human ratings, which implies that AVRA can be used to rate large-scale data sets instead of having a radiologist spend days on doing it manually. Further, the continuous predictions may allow for performing more sensitive analyses with the MTA scale. An example could be to establish more sensitive cut-offs, but also to study longitudinal changes in different disease populations or in healthy aging.

4.4.5 AVRA on longitudinal data

In the previous section we have only analyzed the data in a cross-sectional manner with the aim to assess how reliable AVRA's ratings are in comparison to neuroradiologists. Here we investigate the longitudinal aspects of the data with AVRA's continuous ratings, in an effort to derive clinically useful knowledge of the progression of MTL atrophy in SCD and MCI patients. More specifically we wanted to answer the

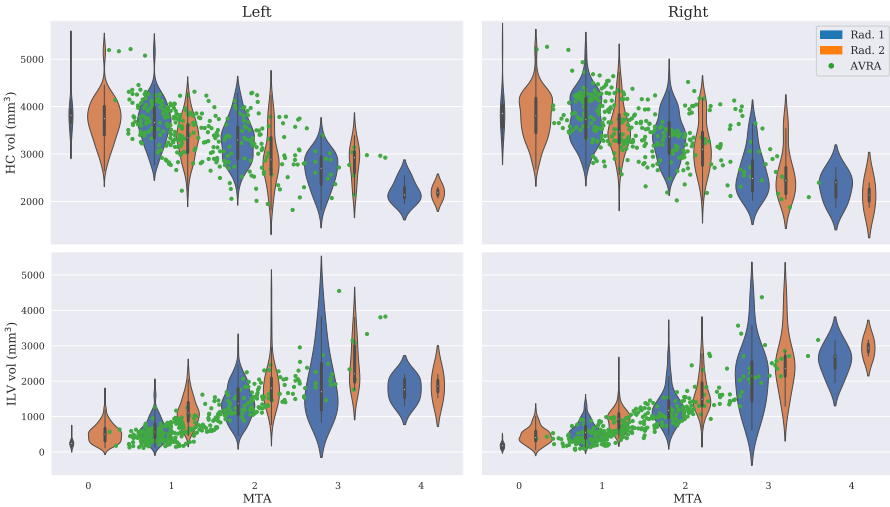


Figure 4.6: Violinplots of the radiologists' MTA ratings and corresponding hippocampal volume (HC; *top*) and inferior lateral ventricle volume (ILV; *bottom*). The width of the violins shows the distribution over volumes for each rating and rater, and the area indicates the number of images given a specific rating. The green dots show AVRA's MTA rating for each image.

question:

- What are the expected atrophy progression rates—expressed in Scheltens' MTA scale—in different preclinical stages of dementia?

To this end, we stratified the individuals into subgroups based on their CSF biomarker (A^-T^-, A^+T^-, and A^+T^+) and cognitive (SCD and MCI) profile. AVRA's MTA ratings, HC volumes, and ILV volumes of the left hemisphere are plotted for each individual with respect to age in Fig. 4.7.

Assuming that MTL atrophy can not reverse with age, we should expect the MTA score to be monotonically increasing with age for all subjects. From Fig. 4.7, it looks like this is true for most subjects but not all. Our impression upon visual inspection is that some of this noise is attributed to irregularities in the FSL registrations, which is the first step

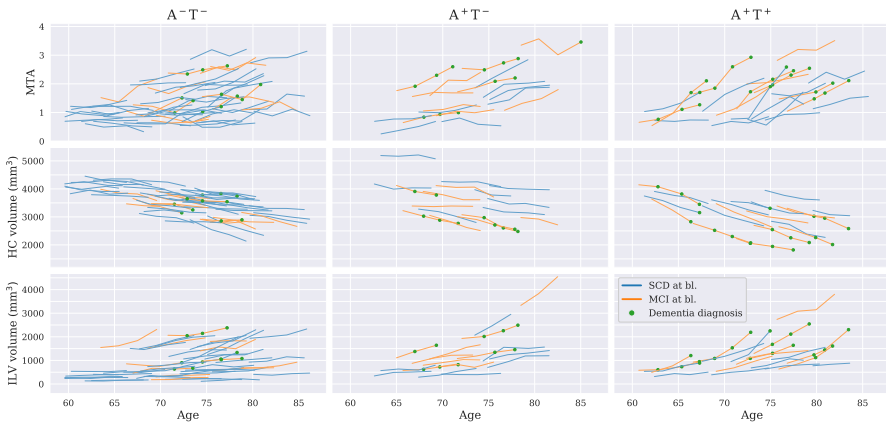


Figure 4.7: AVRA's MTA ratings (top), hippocampal (HC) volumes (middle), and inferior lateral ventricle volumes (ILV) plotted for each individual and biomarker profile. Blue lines represent individuals with subjective cognitive decline (SCD) at baseline, and orange mild cognitive impairment (MCI). Green dots show if a subject has a dementia diagnosis at that timepoint.

of AVRA's pipeline. In Fig. 4.8 we see some examples of four individuals with consistent registrations (chosen for visualization purposes), and that AVRA's ratings are monotonically increasing. The volumetric measures look less noisy in Fig. 4.7, but it should be noted that we did discard roughly 10% of the images based on poor segmentation quality upon visual inspection, whereas all AVRA predictions were included.

We fitted linear slopes, through least-square error, for each individual's MTA, HC volume, and ILV volume progression. The average baseline values and annual rates ("Δ") for all MTL measures of each biomarker profile are shown in Table 4.8. From these results we made the following observations:

- For baseline measures:
  - MCI patients had more severe atrophy than those with SCD.
  - No clear trends in the biomarker abnormalities.

Table 4.8: Baseline (bl) ratings and volumes, and the annual rates ("Δ"), for the three biomarker profiles. Abbreviations: hippocampus (HC); inferior lateral ventricle (ILV); subjective cognitive decline (SCD); mild cognitive impairment (MCI); amyloid (A); tau (T).

	A^-T^-		A^+T^-		A^+T^+	
Measure	Left	Right	Left	Right	Left	Right
AVRA: MTA at bl.	**1.26 ± 0.58**	**1.26 ± 0.56**	**1.39 ± 0.71**	**1.40 ± 0.64**	**1.20 ± 0.58**	**1.28 ± 0.64**
SCD only	1.18 ± 0.55	1.24 ± 0.55	1.10 ± 0.50	1.33 ± 0.69	1.02 ± 0.43	1.01 ± 0.50
MCI only	1.54 ± 0.60	1.34 ± 0.60	1.62 ± 0.77	1.46 ± 0.58	1.39 ± 0.65	1.57 ± 0.65
AVRA: ΔMTA/year	**0.04 ± 0.04**	**0.04 ± 0.04**	**0.07 ± 0.05**	**0.08 ± 0.05**	**0.13 ± 0.08**	**0.11 ± 0.08**
SCD only	0.04 ± 0.05	0.04 ± 0.04	0.07 ± 0.05	0.07 ± 0.05	0.11 ± 0.09	0.09 ± 0.08
MCI only	0.04 ± 0.04	0.06 ± 0.04	0.07 ± 0.05	0.09 ± 0.05	0.15 ± 0.07	0.14 ± 0.07
Rad. 1: MTA at bl.	**1.17 ± 0.66**	**1.17 ± 0.63**	**1.56 ± 0.68**	**1.28 ± 0.45**	**1.67 ± 0.78**	**1.43 ± 0.58**
SCD only	1.07 ± 0.63	1.10 ± 0.61	1.38 ± 0.48	1.38 ± 0.48	1.27 ± 0.45	1.18 ± 0.39
MCI only	1.50 ± 0.65	1.42 ± 0.64	1.70 ± 0.78	1.20 ± 0.40	2.10 ± 0.83	1.70 ± 0.64
Rad. 1: ΔMTA/year	**0.05 ± 0.09**	**0.05 ± 0.08**	**0.04 ± 0.07**	**0.07 ± 0.09**	**0.09 ± 0.11**	**0.12 ± 0.10**
SCD only	0.05 ± 0.09	0.04 ± 0.08	0.04 ± 0.06	0.03 ± 0.05	0.07 ± 0.10	0.08 ± 0.11
MCI only	0.05 ± 0.09	0.06 ± 0.08	0.04 ± 0.07	0.11 ± 0.10	0.11 ± 0.11	0.16 ± 0.07
Rad. 2: MTA at bl.	**0.50 ± 0.71**	**0.56 ± 0.79**	**0.61 ± 0.76**	**0.50 ± 0.69**	**0.62 ± 0.79**	**0.81 ± 0.85**
SCD only	0.36 ± 0.65	0.52 ± 0.79	0.50 ± 0.71	0.62 ± 0.70	0.27 ± 0.45	0.45 ± 0.66
MCI only	1.00 ± 0.71	0.67 ± 0.75	0.70 ± 0.78	0.40 ± 0.66	1.00 ± 0.89	1.20 ± 0.87
Rad. 2: ΔMTA/year	**0.05 ± 0.08**	**0.06 ± 0.09**	**0.11 ± 0.10**	**0.13 ± 0.07**	**0.15 ± 0.12**	**0.13 ± 0.12**
SCD only	0.06 ± 0.09	0.06 ± 0.09	0.04 ± 0.07	0.10 ± 0.06	0.11 ± 0.12	0.10 ± 0.10
MCI only	0.03 ± 0.07	0.07 ± 0.09	0.16 ± 0.10	0.14 ± 0.07	0.19 ± 0.10	0.17 ± 0.13
HC vol at bl. (mm^3)	**3629 ± 432**	**3753 ± 506**	**3698 ± 586**	**3834 ± 567**	**3331 ± 487**	**3433 ± 494**
SCD only	3697 ± 414	3773 ± 479	3999 ± 571	4023 ± 547	3530 ± 325	3659 ± 309
MCI only	3409 ± 415	3686 ± 579	3431 ± 456	3666 ± 531	3151 ± 536	3229 ± 538
ΔHC/year (mm^3/year)	**-36.3 ± 26.9**	**-39.3 ± 25.5**	**-53.4 ± 29.7**	**-55.4 ± 31.3**	**-93.4 ± 33.2**	**-99.3 ± 42.0**
SCD only	-34.7 ± 27.4	-35.7 ± 25.1	-36.8 ± 20.2	-46.0 ± 23.8	-79.4 ± 21.3	-87.8 ± 27.1
MCI only	-41.4 ± 24.5	-50.9 ± 23.2	-68.1 ± 29.0	-63.8 ± 34.6	-106.0 ± 36.7	-109.7 ± 49.7
ΔHC/year (%/year)	**-1.0 ± 0.9**	**-1.1 ± 0.8**	**-1.6 ± 1.0**	**-1.5 ± 0.9**	**-2.9 ± 1.0**	**-2.9 ± 1.2**
ILV vol at bl. (mm^3)	**777 ± 529**	**724 ± 523**	**1053 ± 739**	**860 ± 629**	**858 ± 507**	**817 ± 412**
SCD only	700 ± 481	683 ± 472	804 ± 545	839 ± 772	698 ± 241	652 ± 329
MCI only	1029 ± 596	856 ± 644	1274 ± 813	879 ± 465	1001 ± 627	966 ± 423
ΔILV/year (mm^3/year)	**38.1 ± 41.3**	**38.9 ± 44.6**	**83.1 ± 74.2**	**84.1 ± 98.5**	**117.1 ± 76.5**	**107.8 ± 94.9**
SCD only	37.7 ± 43.5	36.3 ± 45.5	69.8 ± 63.6	98.8 ± 123.3	86.7 ± 83.8	49.3 ± 44.9
MCI only	39.6 ± 33.3	47.1 ± 40.6	95.0 ± 80.7	71.0 ± 66.6	144.4± 56.8	160.4 ± 97.2
ΔILV/year (%/year)	**4.8 ± 4.0**	**4.5 ± 3.9**	**7.9 ± 4.3**	**9.9 ± 7.0**	**14.7 ± 9.0**	**13.9 ± 10.0**

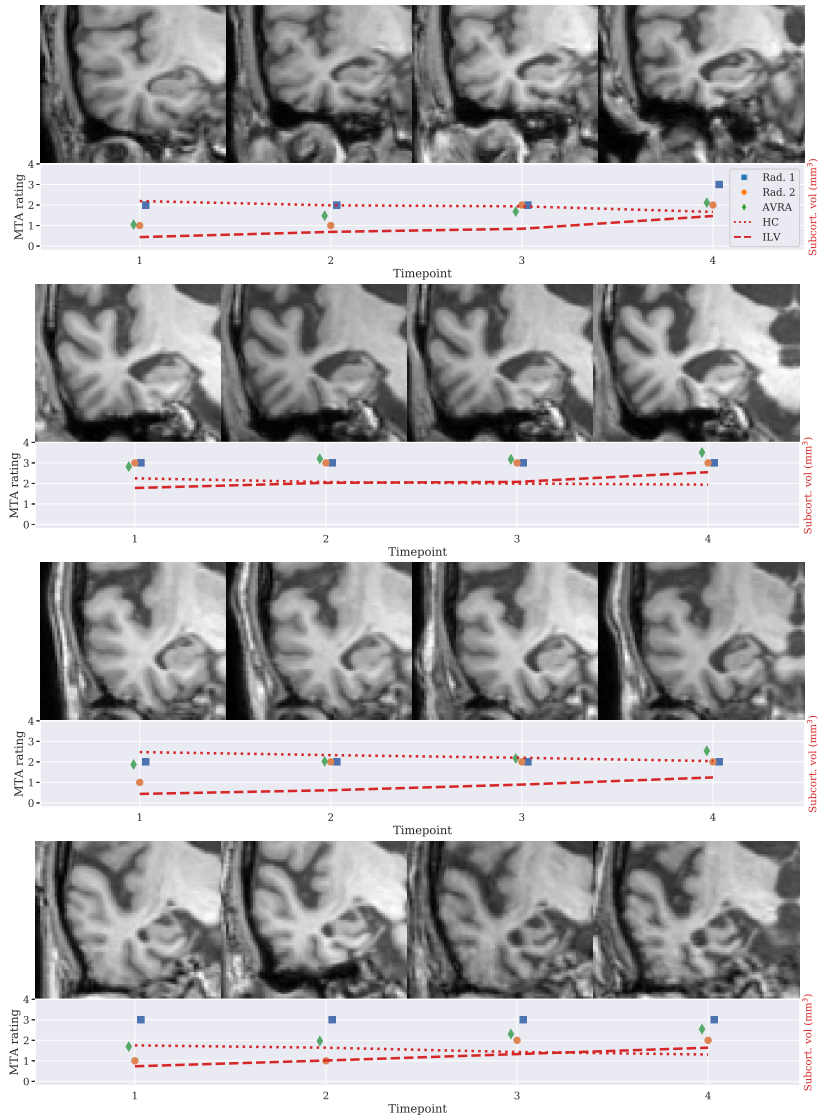


Figure 4.8: Rating slice at each timepoint for four study participants with corresponding MTA ratings and MTL volumes.

- For progression rates:
  - Increased progression rates with more biomarker abnormalities, i.e. (rate of A^-T^-) < (rate of A^+T^-) < (rate of A^+T^+).
  - Atrophy rates were faster in (A^+T^+, SCD) subjects than in (A^-T^-, MCI) patients.
- The trends for the MTA ratings were similar to the subcortical volumes.

The finding that MCI patients has greater medial temporal atrophy than SCD has been demonstrated earlier (Yue et al., 2018). Pettigrew et al. (2017) investigated longitudinal medial temporal lobe atrophy with the same biomarker profile (but cognitively unimpaired) and found the same trends. The data showed that the SCD subjects with biomarker profile A^+T^+ had greater progression rates than the MCI group with biomarker profile A^-T^-. This may be due to that these MCI patients do not have AD pathology and are suffering from another disorder in which MTL atrophy is less prominent.

An interesting aspect of our results was that the MTA ratings (both from AVRA and the radiologists) seemed to capture these trends described above as well, and not just the volume measures. This suggests that the MTA scale is rather sensitive. Of course, these results are based on averaging the human ratings of multiple subjects, and does not translate directly onto individual cases since the scale is discrete. In Fig. 4.8 a few cases are shown, and we can see that AVRA's continuous ratings follow the atrophy progression quite well.

In conclusion, the MTA ratings display similar longitudinal trends as subcortical volumes do. The main drawback of "human" ratings are that they are inherently subjective, and that the difference between two discrete rating steps is too large to capture MTL changes if time between scans is short.

4.5 Conclusion

Through the studies in this chapter we have proposed a DL model for automated visual ratings, tested its performance in external cohorts, compared its ratings to external radiologists, and used it for characterizing a preclinical dementia population with the MTA scale. In this final section, I will discuss what we learned from our studies, and elaborate on how to develop a DL model that can be implemented in clinics.

We demonstrated that a degradation in performance can arise when applying a trained model on data from external clinics. There are several things that would likely mitigate these degradations in OOD data, such as more image preprocessing and data augmentation. We had an initial idea that AVRA should be able to rate "as raw images as possible", demonstrating the same capabilities as a human reader. We also wanted to keep preprocessing to a minimum in order to not have to rely too heavily on external software (which may propagate errors that the model can leverage). We did experiment with more aggressive image preprocessing and data augmentation during the development phase of AVRA but it did not boost the performance. However, this was assessed in within-distribution data and the upside of preprocessing and data normalization may be seen primarily in OOD data.

More training data from a wider range of protocols and scanners could, as we reported, have a positive effect. Medical imaging data can be difficult to get access to—particularly if you want to train on real clinical data. *Federated learning* (FL) is an approach where one does not need to have all images gathered centrally on a single server or computer, and all stakeholders (such as clinics) are able keep their data, that we need for training, locally. The key idea is that instead of transferring data between centers, or to a central server, the model's parameters are being transferred. Say we want to train a model on data from two clinics where regulations and privacy concerns prevent us from having data uploaded to a central server, where we train our models. In FL, we let each center train the model locally for a few iterations, and upload the updated model parameters to our server. The trained models (one from each center) get aggregated on the server, and this updated model is

sent back to the centers who resumes their local training. By repeating this process during training we make the model benefit from data from all participating centers without actually transferring any sensitive data. There are many technical challenges with these techniques that remains to be investigated, but the advantage of being able to train on clinical data makes FL a promising approach for future research (Rieke et al., 2020).

In the works of this thesis we were fortunate to have access to a large number of images that was rated by a single expert neuroradiologist. Ground truth annotations do not exist for many medical imaging problems, and most annotations can only be provided by medical experts which introduces some degree of subjectivity. Inter-observer variability is a big practical issue for many DL projects in medical imaging, as this confounding factor can make it difficult to evaluate the performance of your model. If we were to use ratings from Rad. 2 for an external validation of AVRA we would have a hard time distinguishing if it was rating style differences or domain shift that caused the low agreement[7]. However, this was a relatively simple confounder to identify *a priori*—even for someone without substantial domain knowledge. A recent study investigated how accurate labels (generated automatically using natural language processing) of a large public chest X-ray dataset were, and found that many of the images showed no signs of the pathology suggested by their label (Oakden-Rayner, 2020). This, on the other hand, would be difficult for someone without domain expertise to detect (just as I would have missed that AVRA's MTA model had problems accounting for hippocampal adhesions in Fig. 4.4). Radiologists and clinicians should ideally be involved in all stages of the development of a DL model that aims to become practically useful in clinics. This includes formulating the problem that needs solving, curating the data, validating the model and—maybe most importantly—detecting when it fails.

[7] Actually, we still cannot be certain that the domain shift did not have a substantial impact on AVRA's ratings since the BioFINDER data has not been rated by our training set rater.

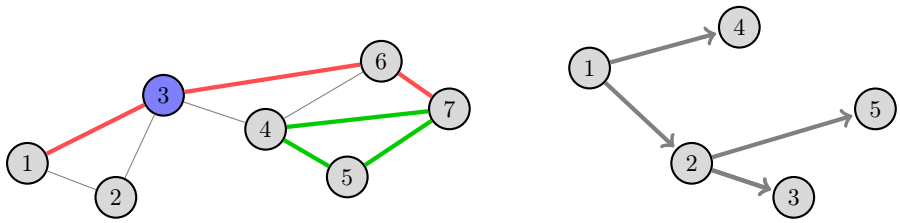
Chapter 5

Quantifying atrophy through gray matter networks

5.1 Graph theory

Graph theory (GT) has been applied in several domains to model relational systems, such as in physics, biology and computer science. It is based on the concept of representing a state or a problem as a graph: a framework to model the relation or interaction between two objects. A graph comprises two key components: *nodes* (or vertices) and *edges* (or links). In Fig. 5.1a we see a sketch of an undirected graph, where nodes are represented by circles and the edges are the lines between some of the nodes. Directed graphs (Fig. 5.1b) are useful when studying systems where there is a directionality between nodes, such as in structural causal modeling (Pearl, 1982). Directed graphs have not been explored further in this thesis, as defining meaningful connections with directionality between nodes from sMRI data is not readily done.

We can illustrate the concept of graphs with the example of a train network. To formulate this as a graph, the train stations would be nodes. If station i and station j have a railway connecting them (with no other station on that section of the rail) they would have a non-zero edge. In the example in Fig. 5.1a station 1 is connected to station 2 but not to



(a) Undirected graph. (b) Directed graph.

Figure 5.1: Sketch to illustrate basic graph theoretical concepts. In fig (a), node 3 in blue represents a hub as well as a node with high clustering. The thick red line shows shortest path between node 1 and node 7, and the green lines show an example of a triangle.

station 4. It is also clear from the figure that station 3 is an important part of the train network, since it connects trains from stations 1-2 with stations 4-7. The train example could be expanded to contain additional information in the graph, such as distance or number of travelers. This is called a *weighted network*, as opposed to a *binary network* (i.e. rail or no rail).

Representing the train network as a graph makes it possible to use the numerous mathematical tools developed for graph theory to derive more abstract features of the network, such as its efficiency, and identify critical nodes in a network. Given that the brain can be thought of as a network in many aspects—the connectome being perhaps the most natural association—there has been a large interest in applying graph theoretical methods to neuroimaging data. Only in Alzheimer's disease, GT has been applied extensively to electroencephalography (EEG) (Stam et al., 2007), PET (Duan et al., 2017), diffusion tensor imaging (DTI) (Lo et al., 2010), functional MRI (fMRI) (Sanz-Arigita et al., 2010), and sMRI (He et al., 2008; Yao et al., 2010; Pereira et al., 2016; Voevodskaya et al., 2018; Tijms et al., 2017; Dicks et al., 2018; Ferreira et al., 2019) data. Gray matter networks are common to study—facilitated by the easy access to T_1-weighted MRI images—and have been analyzed in several other neurological disorders as well, such

as schizophrenia (Bassett et al., 2008; Zhang et al., 2012), epilepsy (Bernhardt et al., 2011), Parkinson's disease (Pereira et al., 2015), but also in healthy aging (Fan et al., 2011; Khundrakpam et al., 2013).

5.2 Constructing graphs of neuroimaging data

Performing network analysis on neuroimaging data starts with defining a graph. This process involves making numerous methodological choices, some dependent on modality, which may all affect the analysis results. A description of many of these methods were put together by Rubinov and Sporns (2010) in their seminal paper, which has been instrumental in the use of graph analysis in brain imaging. Next, we elaborate on some of these methodological choices—and their potential issues—mainly from a perspective of sMRI data, as this was the modality we used to explore gray matter networks in this thesis.

5.2.1 Node definition

As opposed to train networks, the nodes of a brain are not as readily defined (at least not on the resolution *in vivo* imaging can provide) (Fornito et al., 2013). The most common method is to use a neuroanatomical atlas, and define a node as a ROI from this atlas. There are, however, a number of different neuroanatomical atlases to chose from, such as the Desikan (Desikan et al., 2006) and Destrieux (Destrieux et al., 2010) atlases from FreeSurfer, and the AAL atlas in SPM (Tzourio-Mazoyer et al., 2002). These atlases all have different number of ROIs leading to networks with different number of nodes N, which has been shown to impact GT measures (Zalesky et al., 2010; van Wijk et al., 2010). (Further, if you are analyzing fMRI data then these atlases, which are based on structural landmarks (sulci and gyri), may not be the best option. Instead an atlas such as the one proposed by Thomas Yeo et al. (2011), with ROIs defined from functional activity, is a more sensible choice). There are also examples of studies using atlas-free models through cubical GM volumes (Tijms et al., 2012). These methods will end up with a different N for each subject depending mainly on

their intracranial volume but also the overall atrophy (which correlates with disease progression). This can make it difficult to disentangle the added value of applying graph theory as opposed to just comparing GM volumes, which is also easier to interpret.

5.2.2 Edge definition

What can be considered a "connection" between two nodes in MRI data? For DTI data (imaging WM tracts) it is possible to define it as "physical" connection; i.e. is there a WM tract connecting node i and j? This would constitute a structural and "biologically meaningful" connection, one similar to the train network example earlier. In functional data, such as fMRI or EEG, the connectivity strength is typically determined through some statistical measure of similarity, such as correlation ρ. Here it is less clear what constitutes a connection: is it correlations $\rho > 0.5$, or maybe $\rho > 0.3$? Or should *all* nodes be considered connected, but with different strengths? There is not a straightforward answer to this, nor is there a general consensus within the neuroscience field on how to do this. Most studies apply some form thresholding to "remove spurious connections", as Drakesmith et al. (2015) showed that even one spurious connection can influence network properties greatly (even tractography studies typically involve some level of thresholding). This is done by setting all connection strengths below a specified threshold to 0, thus "disconnecting the nodes". But what threshold should be applied? And should network density (# connected nodes / # possible connections) be kept constant between compared networks, at the expense of allowing weaker connections in some networks?

In gray matter group networks created from sMRI data, examples of choices involved in the definition of edges are:

- Choice of cortical measure:
  - E.g. thickness or volume?
- Preprocessing/corrections:
  - E.g. regress effect of total intracranial volume, age and sex?

- Statistical measure defining connectivity strength:
  - E.g. Pearson correlation, rank correlation, or partial correlation?
- Thresholding:
  - Compare between absolute threshold values or against network densities?
- Binary or weighted graphs?

Thus, the number of possible ways to create gray matter networks are many, and rarely the same across studies. This has been offered as one explanation to why reported findings across AD studies often show little agreement (Guye et al., 2010; Fornito et al., 2013; Dai and He, 2014; Dimitriadis et al., 2017; Muldoon et al., 2016). Apart from Phillips et al. (2015), who showed that depending on what correlation measure they applied they could obtain both significantly shorter and longer path length in AD networks, few studies have investigated how some of these choice impact the subsequent results. This is important in order to interpret results and compare findings across studies.

5.3 Graph theoretical measures

There are a number of different network measures that have been used in graph theoretical studies on neuroimaging data. The most commonly used are described in (Rubinov and Sporns, 2010), with many of them having both directed, undirected, and weighted analogues. Deciding which measures to analyze and report is thus another methodological choice the investigator has to make, as there are no "standard measures" that all studies report.

There are both *nodal* and *global* network measures, where global metrics refer to properties of the whole graph. Here we discuss some binary, undirected network measures relevant to the work in this thesis, mainly coming from (Rubinov and Sporns, 2010). We denote nodal measures with a subscript i, referring to node i.

The most basic measure is the *degree* of a node, which shows how many nodes are connected to node i:

$$k_i = \sum_{j \in N} a_{ij} \tag{5.1}$$

where N is the set of all nodes, and a_{ij} is 1 if nodes i, j are connected, otherwise 0.

The *shortest path length*, d_{ij}, relates to the distance between two nodes i and j (see Fig. 5.1a). In the train network example, this the number of stations we need to travel between to get from i to j. The global measure of this is called the *characteristic path length*, L, and describes the average path length for each node to every other node:

$$L = \frac{1}{n} \sum_{i \in N} \frac{\sum_{j \in N, j \neq i} d_{ij}}{n-1}, \tag{5.2}$$

with n being the number of nodes in the network. In unconnected networks, where one or more nodes are disconnected from the rest, the path length to those nodes are infinite. The characteristic path length measure then becomes irrelevant. An alternative measure, that can handle disconnected graphs, is *global efficiency* (Latora and Marchiori, 2001):

$$E = \frac{1}{n} \sum_{i \in N} \frac{\sum_{j \in N, j \neq i} (d_{ij})^{-1}}{n-1}, \tag{5.3}$$

which can be considered an inverse measure to L. Thus, as evident by Eq. (5.3), infinite path lengths do not contribute to the global efficiency.

While L and E are considered measures of integration, there is another family of equations describing network segregation. Two of these measure are *clustering* and *transitivity*. They describe the fraction of closed loops (or *triangles*, i.e. three nodes all connected to each other, see Fig. 5.1a) compared to the degree k_i. In clustering, this is compared on a nodal level, and averaged for the global measure C:

$$C = \frac{1}{n} \sum_{i \in N} \frac{2t_i}{k_i(k_i - 1)}, \tag{5.4}$$

where t_i is the number of triangles of node i. A potential issue with C is that nodes with low degree can greatly influence the global metric. To remedy this, the transitivity measure T was proposed which normalizes on the degree on a global level instead:

$$T = \frac{\sum_{i \in N} 2t_i}{\sum_{i \in N} k_i(k_i - 1)}, \tag{5.5}$$

The clustering metric have so far been more commonly applied than transitivity in gray matter network studies, possibly due to that C is part of the *small-world* measure S

$$S = \frac{C/C_{\text{rand}}}{L/L_{\text{rand}}} \tag{5.6}$$

where $_{\text{rand}}$ denotes metrics from random networks, sometimes used to normalize network measures. Many biological networks seems to be somewhere between highly regular and completely random. These networks, which are clustered as regular networks yet have short path length as random networks, are called "small-world" (Watts and Strogatz, 1998), as illustrated in Fig 5.2. The small-world measure is possibly the most commonly reported metric in brain network studies, alongside path length (Tijms et al., 2013).

As the gray matter networks we generated in this thesis were not fully connected (i.e. the path length between some nodes were infinite) we were not able to report the reliability of the characteristic path length and small-world measures.

5.4 Reproducibility of gray matter networks

As the number of neuroscience studies using GT increases it becomes important to investigate the robustness of these methods when applied to neuroimaging data. These studies have largely been lacking so far

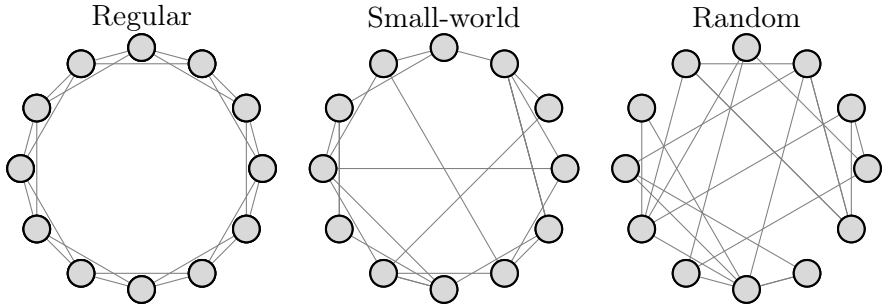


Figure 5.2: Illustration of networks with increased randomness. *Left*: a highly regular network; *Middle*: a small-world network; *Right*: a random network. Image adapted from `jhnet.co.uk/projects/figures/watts_strogatz`.

in the field, possibly due to that there are so many ways to construct a brain network that it is difficult to say how well these findings generalize to other methods. (For instance, are findings robust to preprocessing procedure also when changing neuroanatomical atlas to define nodes?) There are a few exceptions however; Phillips et al. (2015) investigated different graph creation methods for gray matter networks and found that the choice of statistical measure (to quantify connectivity strength) as well as using weighted graphs had large impact on the network measures.

We aimed to shed light on the effect of some of the choices discussed in Sec. 5.2, namely:

1. Do we obtain similar results when basing nodes on ROIs defined by two different neuronatatomical atlases?
2. Do we see similar patterns when basing edges on cortical thickness compared to volumes?
3. How many participants are necessary to include when constructing group networks to obtain reliable results?

To understand to what degree any of these choices impact GT measures, it is necessary to have a frame of reference. We compared networks based on AD patients to ones based on healthy controls (CTR). These diagnostic groups have been compared in multiple studies, and allows

us to relate our findings to previous results. Further, we expect the differences in GT measures between CTR and AD groups to be greater than between e.g. CTR and MCI or even between A^- and A^+ cognitively unimpaired individuals.

To motivate point 3), we consider the following example: A study investigates e.g. clustering in two networks constructed from 100 AD patients and 100 CTR subjects, respectively, and finds significantly lower clustering in the AD network. If this finding is generalizable to the disease population at large—i.e. to be able to claim that clustering is decreased in AD gray matter networks and not just in these specific group compositions—then it is reasonable to believe that if we remove five random subjects from each group the results would remain the same. If we do not get the same results, then it is highly unlikely that the finding is reliable. At the same time, however, there must be a minimal number of subjects required to create a stable network.

We investigated the points mentioned above in data of 293 AD patients and 293 healthy controls from the research cohorts ADNI-1 and AddNeuroMed. They were segmented and parcellated with FreeSurfer 5.3 according to two different atlases: the Desikan atlas, comprising 68 ROIs, and the Destrieux atlas with 148 ROIs (Desikan et al., 2006; Destrieux et al., 2010).

An overview of the procedure is described in pseudocode in Alg. 1. To describe some individual steps in more detail: Each connectivity matrix was defined through the Pearson correlation matrix, where an entry at position $[i, j]$ was the Pearson correlation between the cortical measures of ROI i and j across all subjects. The matrix was binarized according to a threshold, so that all correlations below t were set to 0 and all above to 1. The threshold was chosen to yield the specified network density d, and were thus different for the CTR and AD networks for the same density. From these binarized networks we computed clustering, transitivity and global efficiency for the AD and CTR networks. We calculated p-values through permutation tests to test whether a specific group composition showed significant ($\alpha < .05$) differences, simulating how a "regular" study might perform hypothesis testing. Thus, for each density d and group size N, we get p-values for 100 random group

```
Data: Cortical measure C from atlas A
i ← 0;
// Loop 100 different group compositions
while i < 100 do
    N ← 50 // Number of subjects in each group,
       randomly selected
    // Loop over number of subjects
    while N <= 290 do
        Construct connectivity matrices ← N, C, A;
        d ← 0.05 // Network density
        // Loop over network densities
        while d <= 0.35 do
            Construct networks with density d;
            Calculate GT measures;
            Compute p-value of difference;
            d ← d + 0.01;
        end
        N ← N + 5// Add 5 random subjects
    end
    i++;
end
```

Algorithm 1: Overview of procedure to investigate stability in graph theoretical measures in structural gray matter networks for a single cortical measure C (e.g. volumes) and atlas A (e.g. Desikan).

compositions. We use these to calculate a *significance ratio*, referring to how frequently a significant difference is observed. This entity can be related to the risk of making a type I (false positive) or type II error (false negative). GT measures and statistics were calculated using BRAPH [1] (Mijalkov et al., 2017).

In Fig. 5.3 – 5.5 we see the results for clustering, transitivity and global efficiency, respectively. Our overall interpretation of these results can be summarized as follows:

[1] Freely available at `braph.org`

- The direction of differences between AD and CTR networks seems to be consistent between the two atlases, at least if the differences are "large" (well-separable).
- GM networks based on cortical thickness correlations do not yield similar results as cortical volumes do.
- GT metrics from networks generated from a small number of subjects (<150) were highly dependent on the specific group composition, raising the question of how well GM network findings generalize to the disease population that one wishes to characterize.

Out of the three GT measures, clustering is most commonly used in GM network studies, e.g. (Hosseini et al., 2012; Pereira et al., 2015; Zou et al., 2018; Voevodskaya et al., 2018; Liu et al., 2019). With our choice of network creation methods, the clustering metric seems to converge with relatively few subjects, but does a poor job in discriminating between AD and CTR networks (Fig. 5.3). However, multiple studies have reported significant differences in in this metric between these two groups (Li et al., 2012; Yao et al., 2010; Phillips et al., 2015; Pereira et al., 2016). These studies have used slightly different edge or node definitions, which may explain their findings. We note from our results that there was a roughly 20% risk of obtaining a false positive—particularly at small group sizes. Whether these reported differences are spurious findings (caused by the specific group compositions) or that other graph creation methods yields distinct different clustering patterns is difficult to say from study. The results do, however, suggest future studies to be cautious when interpreting clustering results.

The closely related transitivity measure, normalized at a global level, showed substantially better discriminative abilities between the groups. Transitivity has been reported to show significant differences over a large range of network densities in previous gray matter AD network studies (Mijalkov et al., 2017; Pereira et al., 2016). The metric did require roughly 150 subjects in each group in order for the results to be significant in 95% of the group compositions. This number is very large in relation to the sample sizes in most gray matter network studies—sometimes less than 30 patients for diseases with low prevalence—raising

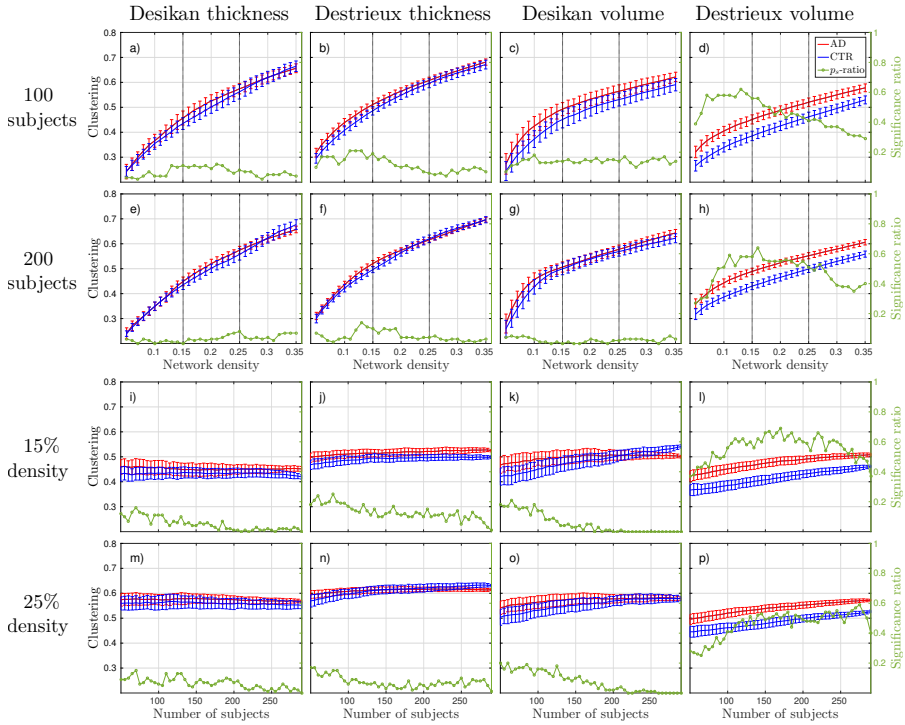


Figure 5.3: The top (second) row shows clustering results as a function of density for networks created with 100 (200) subjects using cortical thickness or volume measures defined with the Desikan or the Destrieux atlas. The two bottom rows show how the GT measure changes when creating networks with different number of subjects at the fixed network densities of 15% and 25%. The error bars show the standard deviation from 100 random group compositions. The green lines illustrate the significance ratio, i.e. how many of these 100 random group compositions yielded significant differences at $p<0.05$.

concerns of the usefulness of studying gray matter networks based on covariance matrices.

Similar results as for transitivity were found for the global efficiency measure in Fig. 5.5. Again, we found reliable differences between controls and AD patients but mainly at large sample sizes.

All investigated measures showed that only cortical thickness mea-

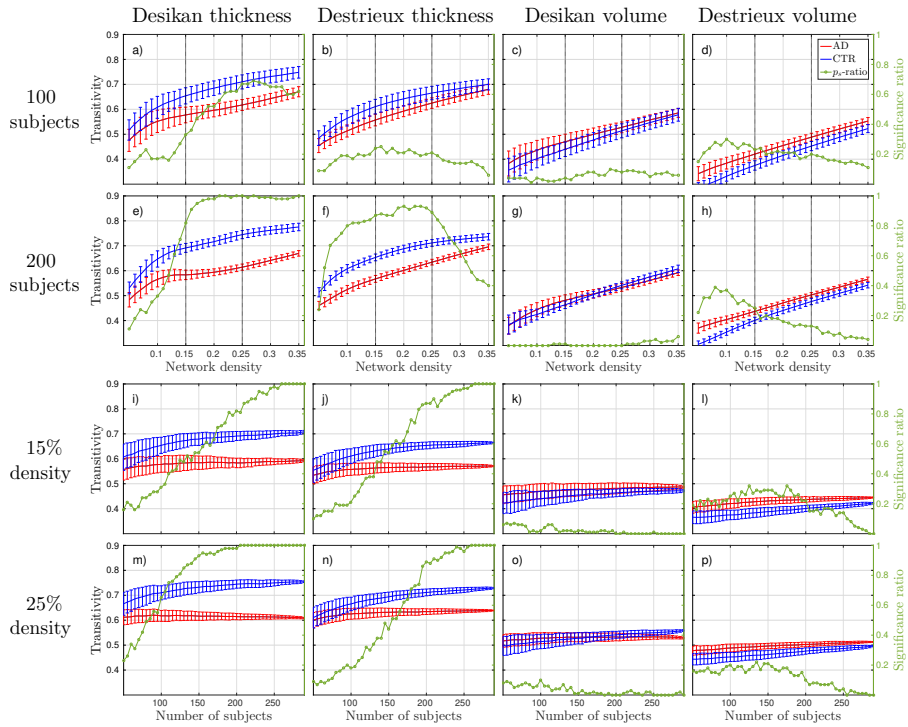


Figure 5.4: The top (second) row shows transitivity results as a function of density for networks created with 100 (200) subjects using cortical thickness or volume measures defined with the Desikan or the Destrieux atlas. The two bottom rows show how the GT measure changes when creating networks with different number of subjects at the fixed network densities of 15% and 25%. The error bars show the standard deviation from 100 random group compositions. The green lines illustrate the significance ratio, i.e. how many of these 100 random group compositions yielded significant differences at $p<0.05$.

sures displayed stable differences between the group networks. Cortical volumes may display larger differences in other disorders affecting the brain, and from this study we cannot conclude that they should not be used in other disease populations or in conjunction with other measures defining edges.

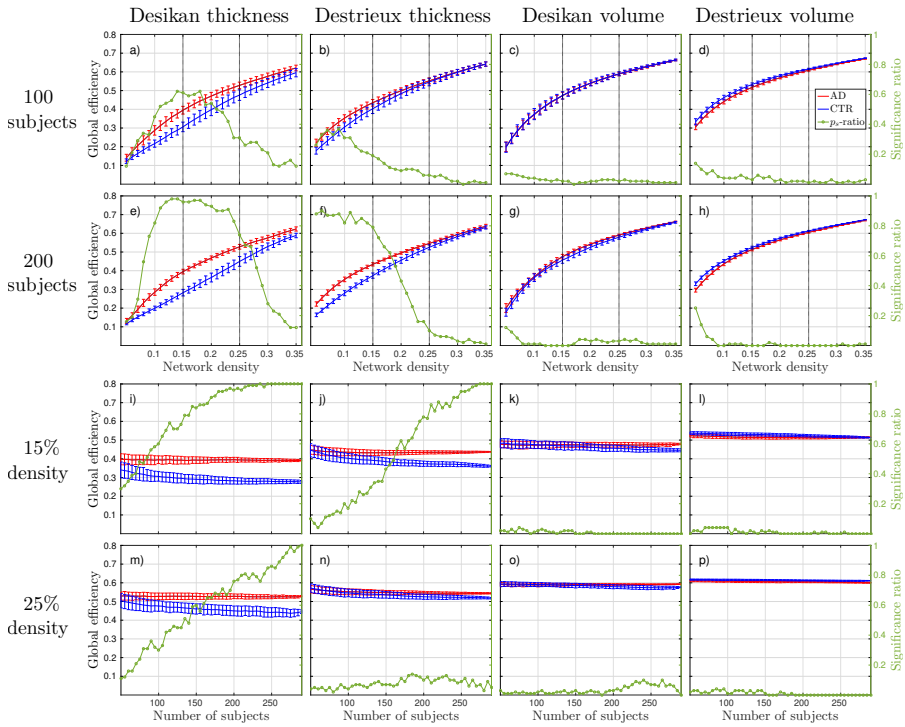


Figure 5.5: The top (second) row shows global efficiency results as a function of density for networks created with 100 (200) subjects using cortical thickness or volume measures defined with the Desikan or the Destrieux atlas. The two bottom rows show how the GT measure changes when creating networks with different number of subjects at the fixed network densities of 15% and 25%. The error bars show the standard deviation from 100 random group compositions. The green lines illustrate the significance ratio, i.e. how many of these 100 random group compositions yielded significant differences at $p<0.05$.

5.5 Conclusion

So, how reliable are structural gray matter networks? To give a complete or definitive answer to this is of course impossible. This is largely attributed to that the number of possible combinations of graph creation parameters—that may have large impact on the analyses results—is

practically infinite. Thus, we cannot say to what degree our results generalizes to other methods and disease populations. Our study does however show that gray matter networks results *can* be very fragile when constructed from small sample sizes. And this when comparing a disease in which severe atrophy is a pathological hallmark. It is difficult to imagine *a priori* two populations that should display larger differences in network properties than CTR and AD. It would thus be reasonable to hypothesize that the differences in network properties are even smaller in e.g. preclinical stages of Alzheimer's disease compared to healthy controls—not to mention disorders not typically associated with pronounced atrophy.

Our results suggest that many studies on gray matter networks may run the risk of reporting findings that do not generalize to larger sample sizes. We propose that future studies should repeat their network analyses on random subsamples of their data set. Comparing the results of these repeated measurements can provide an indication to whether findings are robust or not.

Chapter 6

Concluding remarks

The works in this thesis followed two themes:

- The development and usage of our proposed deep learning model, AVRA.
- Assessing the reliability of two techniques commonly applied to sMRI data: deep learning and graph theory.

Quantifying neurodegeneration through visual assessment does admittedly not sound like the best tool for measuring atrophy. Robust computerized methods that can provide tissue maps, volumetric information, or quantify disease specific atrophy patterns, are more likely to be used in the clinical routine in the future. So what was the purpose of creating a tool such as AVRA and making a trained version publicly available?

I see two main areas where I believe AVRA can be useful. First, AVRA's ratings can provide a benchmark for future tools that are proposed. To give an example: let us say you are developing an automated tool for distinguishing between stable and progressive MCI patients from MRI images. The performance of the proposed tool is really only relevant in contrast to how well this can be done with current methods used in the clinics. That is: what is the added clinical value of your tool? By using AVRA as a "proxy" for a radiologist, this can facilitate this comparison. The second use case is similar to how we applied AVRA

in the longitudinal study. That is, compare visual ratings in large data sets to other clinical markers—possibly leveraging continuous ratings for added sensitivity. This can help bridge the gap between neuroscientific research and the clinics.

The second theme—assessing how robust two separate techniques applied to neuroimaging are—may have a negative ring to it. By assessing *one* deep learning model and *one* graph construction method in larger data sets than most studies have access to, we showed that these can yield overly optimistic results. Through the design of the studies we cannot confidently say that these findings are representative of other deep learning models and graph construction methods, nor do we offer a solution to these issues. I do believe, however, that these types of studies are important for both fields. They demonstrate why experimental results should be interpreted with caution and hopefully contribute to reproducible science.

Acknowledgments

Many people have made my PhD studies a joyful experience and whom I wish to thank.

First and foremost, my supervisor **Eric**, for all discussions, support and guidance these years. For encouraging me to explore my own research interests while always being able to guide me on the right path when running into problems. I could not have asked for a better supervisor.

My co-supervisors **Joana** and **Giovanni** for your feedback, hard work, scientific guidance and support during my PhD studies.

Dani, for your infectious enthusiasm, great collaborations, and for always finding time to share your knowledge when I needed someone to discuss ideas with.

Lena Cavallin, for introducing me to clinical radiology, and lending me your visual ratings for me to try to mimic. I hope I did them justice!

Anette, for taking care of me and patiently helping with all my questions.

Sebastian, for fun TCs and helping me out with everything computer-related during these years. The works in this thesis would have taken twice as long without theHive and your help.

Moa, for finding me this PhD position and then encouraging me to apply. For inviting me to be part of her work, and giving me a glimpse of the insane efforts behind gathering data that I was previously ignorant of.

To my collaborators over the years: **Danielle and the people in Lund**, for a fun and efficient collaboration. **Chunliang, and your**

colleagues at KTH, for re-introducing me to neural networks and for the incredibly helpful discussions and advice. **Tobias**, for your feedback and swift support. The **AddNeuroMed** and **E-DLB** consortia for allowing me to use your excellent and important cohorts.

To my friends from the office: **Olga**, **Konstantinos**, **Una**, **Ale**, **Nira**, **Patri**, **Lissett**, **Rosaleena**, **Atef**. You truly are an amazing bunch of people! Incredibly talented, funny, and always with something interesting to say. Academia is lucky to have, or have had, you.

To the extended neuroimaging group: **Farshad**, **Soheil**, **Urban**, **Anna**, **Divya**, **Anna**, **Abbe**, **Milan**, **Kris**, **Lucia**, **Irene**, **Love**. Thank for all the help and discussions along the way. It's been a pleasure working with you. **Olof**, for inviting me early in my PhD to collaborate with you on your paper. **Mite**, for taking care of me and showing me around on my first research trip to Bilkent. **Lars-Olof**, for starting the division and creating the spirit of community that still surrounds it.

To the people in, and around, Neo. **Emilia**, **Juraj**, **Kostas**, **Elena**, **Antoine**, **Mona-Lisa**, **Amit**, **Laetitia**, **Marco**, **Médoune**, **Lorena**, **Axel**, **Vesna**, **Agneta**, **Mia**. For all the lunch room chats, coffee breaks, and interesting discussions these years. It's always a joy coming to the office when there are so many fun and bright people around. **Nenad**, for teaching us about neuroanatomy and for answering my many ignorant questions. **Tales**, for your enthusiasm and keeping me and everyone else in shape.

To **Fabian**, for being my mentor during this PhD project.

To family and friends for their support. Special shoutout to **Johan**, my first scientific partner-in-crime who also made the cover illustration, and **Uffe** (PhD) for his helpful feedback during the writing of this thesis.

To **Johanna** and our furry family. Tack för att du uppmuntrade mig till att säga upp mig från mitt jobb för att börja doktorera, och att tålmodigt lyssnat på mina långa utläggningar om forskning, icke-forskning och allt mittemellan. Att få komma hem till er är alltid höjdpunkten på dagen.

Bibliography

A. Abdulkadir, B. Mortamet, P. Vemuri, C. R. Jack, G. Krueger, and S. Klöppel. Effects of hardware heterogeneity on the performance of SVM Alzheimer's disease classifier. *NeuroImage*, 58(3):785–792, oct 2011. ISSN 10538119. doi: 10.1016/j.neuroimage.2011.06.029. URL https://linkinghub.elsevier.com/retrieve/pii/S1053811911006471.

E. A. Albadawy, A. Saha, and M. A. Mazurowski. Deep learning for segmentation of brain tumors: Impact of cross-institutional training and testing: Impact. *Medical Physics*, 45 (3):1150–1158, 2018. ISSN 00942405. doi: 10.1002/mp.12752.

M. S. Albert, S. T. DeKosky, D. Dickson, B. Dubois, H. H. Feldman, N. C. Fox, A. Gamst, D. M. Holtzman, W. J. Jagust, R. C. Petersen, P. J. Snyder, M. C. Carrillo, B. Thies, and C. 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 and Dementia*, 7(3):270–279, 2011. ISSN 15525260. doi: 10.1016/j.jalz.2011.03.008. URL http://dx.doi.org/10.1016/j.jalz.2011.03.008.

P. V. Arriagada, J. H. Growdon, E. T. Hedley-Whyte, and B. T. Hyman. Neurofibrillary tangles but not senile plaques parallel duration and severity of Alzheimer's disease. *Neurology*, 42(3):631–631, 1992. ISSN 0028-3878. doi: 10.1212/WNL.42.3.631. URL http://www.neurology.org/cgi/doi/10.1212/WNL.42.3.631.

M. A. Badgeley, J. R. Zech, L. Oakden-Rayner, B. S. Glicksberg, M. Liu, W. Gale, M. V. McConnell, B. Percha, T. M. Snyder, and J. T. Dudley. Deep learning predicts hip fracture using confounding patient and healthcare variables. *npj Digital Medicine*, 2(1), 2019. ISSN 2398-6352. doi: 10.1038/s41746-019-0105-1. URL http://dx.doi.org/10.1038/s41746-019-0105-1.

D. Bahdanau, K. H. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, pages 1–15, 2015.

F. Barkhof, N. C. Fox, A. J. Bastos-Leite, and P. Scheltens. *Neuroimaging in Dementia*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011. ISBN 978-3-642-00817-7. doi: 10.1007/978-3-642-00818-4. URL http://link.springer.com/10.1007/978-3-642-00818-4.

R. Bartus, R. Dean, B. Beer, and A. Lippa. The cholinergic hypothesis of geriatric memory dysfunction. *Science*, 217(4558):408–414, 1982. ISSN 0036-8075. doi: 10.1126/science.7046051.

D. S. Bassett, E. Bullmore, B. A. Verchinski, V. S. Mattay, D. R. Weinberger, and A. Meyer-Lindenberg. Hierarchical Organization of Human Cortical Networks in Health and Schizophrenia. *Journal of Neuroscience*, 28(37):9239–9248, 2008. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.1929-08.2008. URL `http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.1929-08.2008`.

R. J. Bateman, C. Xiong, T. L. Benzinger, A. M. Fagan, A. Goate, N. C. Fox, D. S. Marcus, N. J. Cairns, X. Xie, T. M. Blazey, D. M. Holtzman, A. Santacruz, V. Buckles, A. Oliver, K. Moulder, P. S. Aisen, B. Ghetti, W. E. Klunk, E. McDade, R. N. Martins, C. L. Masters, R. Mayeux, J. M. Ringman, M. N. Rossor, P. R. Schofield, R. A. Sperling, S. Salloway, and J. C. Morris. Clinical and biomarker changes in dominantly inherited Alzheimer's disease. *New England Journal of Medicine*, 367(9):795–804, 2012. ISSN 15334406. doi: 10.1056/NEJMoa1202753.

B. C. Bernhardt, Z. Chen, Y. He, A. C. Evans, and N. Bernasconi. Graph-theoretical analysis reveals disrupted small-world organization of cortical thickness correlation networks in temporal lobe epilepsy. *Cerebral Cortex*, 21(9):2147–2157, 2011. ISSN 10473211. doi: 10.1093/cercor/bhq291.

K. Blennow and H. Hampel. Review CSF markers for incipient Alzheimer ' s disease CSF markers for incipient AD. *The Lancet*, 2(October):605–613, 2003. URL `http://www.ncbi.nlm.nih.gov/pubmed/14505582`.

K. Blennow and H. Zetterberg. The Past and the Future of Alzheimer's Disease Fluid Biomarkers. *Journal of Alzheimer's Disease*, 62(3):1125–1140, 2018. ISSN 13872877. doi: 10.3233/JAD-170773. URL `http://www.medra.org/servlet/aliasResolver?alias=iospress{&}doi=10.3233/JAD-170773`.

P. A. Boyle, L. Yu, R. S. Wilson, S. E. Leurgans, J. A. Schneider, and D. A. Bennett. Person-specific contribution of neuropathologies to cognitive loss in old age. *Annals of Neurology*, 83(1):74–83, 2018. ISSN 15318249. doi: 10.1002/ana.25123.

H. Braak and E. Braak. Neuropathological stageing of Alzheimer-related changes. *Acta Neuropathologica*, 82(4):239–259, 1991. ISSN 0001-6322. doi: 10.1007/BF00308809. URL `http://link.springer.com/10.1007/BF00308809`.

J. Brettschneider, K. Del Tredici, V. M. Y. Lee, and J. Q. Trojanowski. Spreading of pathology in neurodegenerative diseases: A focus on human studies. *Nature Reviews Neuroscience*, 16(2):109–120, 2015. ISSN 14710048. doi: 10.1038/nrn3887. URL `http://dx.doi.org/10.1038/nrn3887`.

M. Buda, A. Maki, and M. A. Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *arXiv preprint*, pages 1–23, 2017. URL `http://arxiv.org/abs/1710.05381`.

T. Byrt, J. Bishop, and J. B. Carlin. Bias, prevalence and kappa. *Journal of Clinical Epidemiology*, 46(5):423–429, 1993. ISSN 08954356. doi: 10.1016/0895-4356(93)90018-V.

M. S. Byun, S. E. Kim, J. Park, D. Yi, Y. M. Choe, B. K. Sohn, H. J. Choi, H. Baek, J. Y. Han, J. I. Woo, and D. Y. Lee. Heterogeneity of regional brain atrophy patterns associated with distinct progression rates in Alzheimer's disease. *PLoS ONE*, 10(11): 1–16, 2015. ISSN 19326203. doi: 10.1371/journal.pone.0142756.

L. Cavallin, L. Bronge, Y. Zhang, A. R. Øksengard, L. O. Wahlund, L. Fratiglioni, and R. Axelsson. Comparison between visual assessment of MTA and hippocampal volumes in an elderly, non-demented population. *Acta Radiologica*, 53(5):573–579, 2012a. ISSN 02841851. doi: 10.1258/ar.2012.110664.

L. Cavallin, K. Løken, K. Engedal, A. R. Øksengård, L. O. Wahlund, L. Bronge, and R. Axelsson. Overtime reliability of medial temporal lobe atrophy rating in a clinical setting. *Acta Radiologica*, 53(3):318–323, 2012b. ISSN 02841851. doi: 10.1258/ar.2012.110552.

G. Cheng, C. Huang, H. Deng, and H. Wang. Diabetes as a risk factor for dementia and mild cognitive impairment: A meta-analysis of longitudinal studies. *Internal Medicine Journal*, 42(5):484–491, 2012. ISSN 14440903. doi: 10.1111/j.1445-5994.2012.02758.x.

K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pages 1724–1734, 2014. doi: 10.3115/v1/d14-1179.

J. Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960. ISSN 15523888. doi: 10.1177/001316446002000104.

J. Cohen. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213–220, 1968. ISSN 00332909. doi: 10.1037/h0026256.

E. Corder, A. Saunders, W. Strittmatter, D. Schmechel, P. Gaskell, G. Small, A. Roses, J. Haines, and M. Pericak-Vance. Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science*, 261(5123):921–923, 1993. ISSN 0036-8075. doi: 10.1126/science.8346443. URL `http://www.sciencemag.org/cgi/doi/10.1126/science.8346443`.

Z. Dai and Y. He. Disrupted structural and functional brain connectomes in mild cognitive impairment and Alzheimer's disease. *Neuroscience Bulletin*, 30(2):217–232, 2014. ISSN 19958218. doi: 10.1007/s12264-013-1421-0.

A. M. Dale, B. Fischl, and M. I. Sereno. Cortical surface-based analysis: I. Segmentation and surface reconstruction. *NeuroImage*, 9(2):179–194, 1999. ISSN 10538119. doi: 10.1006/nimg.1998.0395.

J. De Fauw, J. R. Ledsam, B. Romera-Paredes, S. Nikolov, N. Tomasev, S. Blackwell, H. Askham, X. Glorot, B. O'Donoghue, D. Visentin, G. van den Driessche, B. Lakshminarayanan, C. Meyer, F. Mackinder, S. Bouton, K. Ayoub, R. Chopra, D. King, A. Karthikesalingam, C. O. Hughes, R. Raine, J. Hughes, D. A. Sim, C. Egan, A. Tufail,

H. Montgomery, D. Hassabis, G. Rees, T. Back, P. T. Khaw, M. Suleyman, J. Cornebise, P. A. Keane, and O. Ronneberger. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature medicine*, 24(9):1342–1350, 2018. ISSN 1546-170X. doi: 10.1038/s41591-018-0107-6. URL `http://www.ncbi.nlm.nih.gov/pubmed/30104768`.

R. S. Desikan, F. Ségonne, B. Fischl, B. T. Quinn, B. C. Dickerson, D. Blacker, R. L. Buckner, A. M. Dale, R. P. Maguire, B. T. Hyman, M. S. Albert, and R. J. Killiany. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage*, 31(3):968–980, 2006. ISSN 10538119. doi: 10.1016/j.neuroimage.2006.01.021.

C. Destrieux, B. Fischl, A. Dale, and E. Halgren. Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *NeuroImage*, 53(1):1–15, 2010. ISSN 10538119. doi: 10.1016/j.neuroimage.2010.06.010.

E. Dicks, B. M. Tijms, M. ten Kate, A. A. Gouw, M. R. Benedictus, C. E. Teunissen, F. Barkhof, P. Scheltens, and W. M. van der Flier. Gray matter network measures are associated with cognitive decline in mild cognitive impairment. *Neurobiology of Aging*, 61:198–206, 2018. ISSN 15581497. doi: 10.1016/j.neurobiolaging.2017.09.029. URL `https://doi.org/10.1016/j.neurobiolaging.2017.09.029`.

S. I. Dimitriadis, M. Drakesmith, S. Bells, G. D. Parker, D. E. Linden, and D. K. Jones. Improving the reliability of network metrics in structural brain networks by integrating different network weighting strategies into a single graph. *Frontiers in Neuroscience*, 11 (DEC):1–17, 2017. ISSN 1662453X. doi: 10.3389/fnins.2017.00694.

J. Donahue, L. A. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, and T. Darrell. Long-term Recurrent Convolutional Networks for Visual Recognition and Description. *arXiv*, pages 1–14, 2015.

M. Drakesmith, K. Caeyenberghs, A. Dutt, G. Lewis, A. S. David, and D. K. Jones. Overcoming the effects of false positives and threshold bias in graph theoretical analyses of neuroimaging data. *NeuroImage*, 118:313–333, 2015. ISSN 10959572. doi: 10.1016/j.neuroimage.2015.05.011.

H. Duan, J. Jiang, J. Xu, H. Zhou, Z. Huang, Z. Yu, and Z. Yan. Differences in Aβ brain networks in Alzheimer's disease and healthy controls. *Brain Research*, 1655(October 2016):77–89, 2017. ISSN 18726240. doi: 10.1016/j.brainres.2016.11.019. URL `http://dx.doi.org/10.1016/j.brainres.2016.11.019`.

B. Dubois, H. H. Feldman, C. Jacova, S. T. DeKosky, P. Barberger-Gateau, J. Cummings, A. Delacourte, D. Galasko, S. Gauthier, G. Jicha, K. Meguro, J. O'Brien, F. Pasquier, P. Robert, M. Rossor, S. Salloway, Y. Stern, P. J. Visser, and P. Scheltens. Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS-ADRDA criteria. *Lancet Neurology*, 6(8):734–746, 2007. ISSN 14744422. doi: 10.1016/S1474-4422(07)70178-3.

B. Dubois, H. H. Feldman, C. Jacova, H. Hampel, J. L. Molinuevo, K. Blennow, S. T. Dekosky, S. Gauthier, D. Selkoe, R. Bateman, S. Cappa, S. Crutch, S. Engelborghs, G. B. Frisoni, N. C. Fox, D. Galasko, M. O. Habert, G. A. Jicha, A. Nordberg, F. Pasquier,

G. Rabinovici, P. Robert, C. Rowe, S. Salloway, M. Sarazin, S. Epelbaum, L. C. de Souza, B. Vellas, P. J. Visser, L. Schneider, Y. Stern, P. Scheltens, and J. L. Cummings. Advancing research diagnostic criteria for Alzheimer's disease: The IWG-2 criteria. *The Lancet Neurology*, 13(6):614–629, 2014. ISSN 14744465. doi: 10.1016/S1474-4422(14)70090-0.

T. C. Durazzo, N. Mattsson, and M. W. Weiner. Smoking and increased Alzheimer's disease risk: A review of potential mechanisms. *Alzheimer's & Dementia*, 10:S122–S145, jun 2014. ISSN 15525260. doi: 10.1016/j.jalz.2014.04.009. URL `http://doi.wiley.com/10.1016/j.jalz.2014.04.009`.

C. Duyckaerts. Tau pathology in children and young adults: Can you still be unconditionally baptist? *Acta Neuropathologica*, 121(2):145–147, 2011. ISSN 00016322. doi: 10.1007/s00401-010-0794-7.

R. Elliott. Executive functions and their disorders. *Imaging neuroscience: clinical frontiers for diagnosis and management*, 65(March):49–59, 2003. doi: 10.1093/bmb/ldg65.049.

F. Falahati, S. M. Fereshtehnejad, D. Religa, L. O. Wahlund, E. Westman, and M. Eriksdotter. The use of MRI, CT and lumbar puncture in dementia diagnostics: Data from the svedem registry. *Dementia and Geriatric Cognitive Disorders*, 39:81–91, 2015. ISSN 14219824. doi: 10.1159/000366194.

Y. Fan, F. Shi, J. K. Smith, W. Lin, J. H. Gilmore, and D. Shen. Brain anatomical networks in early human brain development. *NeuroImage*, 54(3):1862–1871, 2011. ISSN 10538119. doi: 10.1016/j.neuroimage.2010.07.025. URL `http://dx.doi.org/10.1016/j.neuroimage.2010.07.025`.

L. A. Farrer, L. A. Cupples, J. L. Haines, B. Hyman, W. A. Kukull, R. Mayeux, R. H. Myers, M. A. Pericak-Vance, N. Risch, and C. M. Van Duijn. Effects of age, sex, and ethnicity on the association between apolipoprotein E genotype and Alzheimer disease: A meta-analysis. *Journal of the American Medical Association*, 278(16):1349–1356, 1997. ISSN 00987484. doi: 10.1001/jama.278.16.1349.

F. Fazekas, J. B. Chawluk, A. Alavi, H. I. Hurtig, and R. A. Zimmerman. MR Signal Abnormalities At 1.5-T in Alzheimer Dementia and Normal Aging. *American Journal of Roentgenology*, 149(2):351–356, 1987. ISSN 0361-803X. doi: 10.2214/ajr.149.2.351.

D. Ferreira, L. Cavallin, T. Granberg, O. Lindberg, C. Aguilar, P. Mecocci, B. Vellas, M. Tsolaki, I. Kłoszewska, H. Soininen, S. Lovestone, A. Simmons, L. O. Wahlund, and E. Westman. Quantitative validation of a visual rating scale for frontal atrophy: associations with clinical status, APOE e4, CSF biomarkers and cognition. *European Radiology*, 26(8):2597–2610, 2016. ISSN 14321084. doi: 10.1007/s00330-015-4101-9.

D. Ferreira, C. Verhagen, J. A. Hernández-Cabrera, L. Cavallin, C. J. Guo, U. Ekman, J. S. Muehlboeck, A. Simmons, J. Barroso, L. O. Wahlund, and E. Westman. Distinct subtypes of Alzheimer's disease based on patterns of brain atrophy: Longitudinal trajectories and clinical applications. *Scientific Reports*, 7(April):1–13, 2017. ISSN 20452322. doi: 10.1038/srep46263. URL `http://dx.doi.org/10.1038/srep46263`.

D. Ferreira, J. B. Pereira, G. Volpe, and E. Westman. Subtypes of Alzheimer's disease display distinct network abnormalities extending beyond their pattern of brain atrophy. *Frontiers in Neurology*, 10(MAY), 2019. ISSN 16642295. doi: 10.3389/fneur.2019.00524.

B. Fischl, D. H. Salat, A. J. W. Van Der Kouwe, N. Makris, F. Ségonne, B. T. Quinn, and A. M. Dale. Sequence-independent segmentation of magnetic resonance images. *NeuroImage*, 23(SUPPL. 1):69–84, 2004. ISSN 10538119. doi: 10.1016/j.neuroimage.2004.07.016.

A. Fornito, A. Zalesky, and M. Breakspear. Graph analysis of the human connectome: Promise, progress, and pitfalls. *NeuroImage*, 80:426–444, 2013. ISSN 10538119. doi: 10.1016/j.neuroimage.2013.04.087.

N. C. Fox, E. K. Warrington, P. A. Freeborough, P. Hartikainen, A. M. Kennedy, J. M. Stevens, and M. N. Rossor. Presymptomatic hippocampal atrophy in Alzheimer's disease. A longitudinal MRI study. *Brain : a journal of neurology*, 119 (Pt 6(1996):2001–7, 1996. ISSN 0006-8950. doi: 10.1093/brain/119.6.2001. URL `http://www.ncbi.nlm.nih.gov/pubmed/9010004`.

G. B. Frisoni, M. Pievani, C. Testa, F. Sabattoli, L. Bresciani, M. Bonetti, A. Beltramello, K. M. Hayashi, A. W. Toga, and P. M. Thompson. The topography of grey matter involvement in early and late onset Alzheimer's disease. *Brain*, 130(3):720–730, 2007. ISSN 00068950. doi: 10.1093/brain/awl377.

G. B. Frisoni, N. C. Fox, C. R. Jack, P. Scheltens, and P. M. Thompson. The clinical use of structural MRI in Alzheimer disease. *Nature Reviews Neurology*, 6(2):67–77, 2010. ISSN 17594758. doi: 10.1038/nrneurol.2009.215.

F. A. Gers, J. Schmidhuber, and F. Cummins. Learning to Forget: Continual Prediction with LSTM. *Neural Computation*, 12(10):2451–2471, oct 2000. ISSN 0899-7667. doi: 10.1162/089976600300015015. URL `http://www.mitpressjournals.org/doi/10.1162/089976600300015015`.

T. Gómez-Isla, R. Hollister, H. West, S. Mui, J. H. Growdon, R. C. Petersen, J. E. Parisi, and B. T. Hyman. Neuronal loss correlates with but exceeds neurofibrillary tangles in Alzheimer's disease. *Annals of Neurology*, 41(1):17–24, 1997. ISSN 03645134. doi: 10.1002/ana.410410106.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative Adversarial Nets. *Advances in Neural Information Processing Systems 27*, pages 2672–2680, 2014. ISSN 10495258. doi: 10.1017/CBO9781139058452. URL `http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf`.

I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. MIT Press, 2016.

D. N. Greve and B. Fischl. Accurate and robust brain image alignment using boundary-based registration. *NeuroImage*, 48(1):63–72, 2009. ISSN 10538119. doi: 10.1016/j.neuroimage.2009.06.060. URL `http://dx.doi.org/10.1016/j.neuroimage.2009.06.060`.

M. Grewal, M. M. Srivastava, P. Kumar, and S. Varadarajan. RADnet: Radiologist level accuracy using deep learning for hemorrhage detection in CT scans. *Proceedings - International Symposium on Biomedical Imaging*, 2018-April:281–284, 2018. ISSN 19458452. doi: 10.1109/ISBI.2018.8363574.

C. Guo, D. Ferreira, K. Fink, E. Westman, and T. Granberg. Repeatability and reproducibility of FreeSurfer, FSL-SIENAX and SPM brain volumetric measurements and the effect of lesion filling in multiple sclerosis. *European Radiology*, 29(3):1355–1364, 2019. ISSN 14321084. doi: 10.1007/s00330-018-5710-x.

M. Guye, G. Bettus, F. Bartolomei, and P. J. Cozzone. Graph theoretical analysis of structural and functional connectivity MRI in normal and pathological brain networks, 2010. ISSN 09685243.

C. Håkansson, G. Torisson, E. Londos, O. Hansson, and D. van Westen. Structural imaging findings on non-enhanced computed tomography are severely underreported in the primary care diagnostic work-up of subjective cognitive decline. *Neuroradiology*, 61(4): 397–404, 2019. ISSN 14321920. doi: 10.1007/s00234-019-02156-6.

J. Hardy and G. Higgins. Alzheimer's disease: the amyloid cascade hypothesis. *Science*, 256(5054):184–185, 1992. ISSN 0036-8075. doi: 10.1126/science.1566067. URL `http://www.sciencemag.org/cgi/doi/10.1126/science.1566067`.

L. Harper, F. Barkhof, N. C. Fox, and J. M. Schott. Using visual rating to diagnose dementia: A critical evaluation of MRI atrophy scales. *Journal of Neurology, Neurosurgery and Psychiatry*, 86(11):1225–1233, 2015. ISSN 1468330X. doi: 10.1136/jnnp-2014-310090.

L. Harper, F. Bouwman, E. J. Burton, F. Barkhof, P. Scheltens, J. T. O'Brien, N. C. Fox, G. R. Ridgway, and J. M. Schott. Patterns of atrophy in pathologically confirmed dementias: A voxelwise analysis. *Journal of Neurology, Neurosurgery and Psychiatry*, 88(11):908–916, 2017. ISSN 1468330X. doi: 10.1136/jnnp-2016-314978.

S. Haykin. *Neural Networks and Learning Machines*, volume 3. 2008. ISBN 9780131471399. doi: 978-0131471399.

K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (1):770–778, 2016. ISSN 1664-1078. doi: 10.1109/CVPR.2016.90. URL `http://ieeexplore.ieee.org/document/7780459/`.

Y. He, Z. Chen, and A. Evans. Structural Insights into Aberrant Topological Patterns of Large-Scale Cortical Networks in Alzheimer's Disease. *Journal of Neuroscience*, 28 (18):4756–4766, 2008. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.0141-08.2008. URL `http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.0141-08.2008`.

W. J. Henneman, J. D. Sluimer, J. Barnes, W. M. Van Der Flier, I. C. Sluimer, N. C. Fox, P. Scheltens, H. Vrenken, and F. Barkhof. Hippocampal atrophy rates in Alzheimer disease: Added value over whole brain volume measures. *Neurology*, 72(11):999–1007, 2009. ISSN 1526632X. doi: 10.1212/01.wnl.0000344568.09360.31.

S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9: 1735–1780, 1997.

S. M. Hosseini, D. Koovakkattu, and S. R. Kesler. Altered small-world properties of gray matter networks in breast cancer. *BMC Neurology*, 12, 2012. ISSN 14712377. doi: 10.1186/1471-2377-12-28.

G. Huang, Y. Li, G. Pleiss, Z. Liu, J. E. Hopcroft, and K. Q. Weinberger. Snapshot Ensembles: Train 1, get M for free. *arXiv*, pages 1–14, 2017. URL `http://arxiv.org/abs/1704.00109`.

S. Ioffe and C. Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *arXiv preprint*, 2015. ISSN 0717-6163. doi: 10.1007/s13398-014-0173-7.2. URL `http://arxiv.org/abs/1502.03167`.

C. R. Jack, R. C. Petersen, Y. C. Xu, S. C. Waring, P. C. O'Brien, E. G. Tangalos, G. E. Smith, R. J. Ivnik, and E. Kokmen. Medial temporal atrophy on MRI in normal aging and very mild Alzheimer's disease. *Neurology*, 49(3):786–794, 1997. ISSN 00283878. doi: 10.1212/WNL.49.3.786.

C. R. Jack, R. C. Petersen, Y. Xu, P. C. O'Brien, G. E. Smith, R. J. Ivnik, B. F. Boeve, E. G. Tangalos, and E. Kokmen. Rates of hippocampal atrophy correlate with change in clinical status in aging and AD. *Neurology*, 55(4):484–489, 2000. ISSN 00283878. doi: 10.1212/wnl.55.4.484.

C. R. Jack, M. S. Albert, D. S. Knopman, G. M. McKhann, R. A. Sperling, M. C. Carrillo, B. Thies, and C. H. Phelps. Introduction to the recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's and Dementia*, 7(3):257–262, 2011. ISSN 15525260. doi: 10.1016/j.jalz.2011.03.004. URL `http://dx.doi.org/10.1016/j.jalz.2011.03.004`.

C. R. Jack, D. S. Knopman, W. J. Jagust, R. C. Petersen, M. W. Weiner, P. S. Aisen, L. M. Shaw, P. Vemuri, H. J. Wiste, S. D. Weigand, T. G. Lesnick, V. S. Pankratz, M. C. Donohue, and J. Q. Trojanowski. Tracking pathophysiological processes in Alzheimer's disease: An updated hypothetical model of dynamic biomarkers. *The Lancet Neurology*, 12(2):207–216, 2013. ISSN 14744422.

C. R. Jack, H. J. Hampel, S. Universities, M. Cu, and R. C. Petersen. A new classification system for AD , independent of cognition A / T / N : An unbiased descriptive classification scheme for Alzheimer disease biomarkers. *Neurology*, 0(July):1–10, 2016.

C. R. Jack, D. A. Bennett, K. Blennow, M. C. Carrillo, B. Dunn, S. B. Haeberlein, D. M. Holtzman, W. Jagust, F. Jessen, J. Karlawish, E. Liu, J. L. Molinuevo, T. Montine, C. Phelps, K. P. Rankin, C. C. Rowe, P. Scheltens, E. Siemers, H. M. Snyder, R. Sperling, C. Elliott, E. Masliah, L. Ryan, and N. Silverberg. NIA-AA Research Framework: Toward a biological definition of Alzheimer's disease. *Alzheimer's and Dementia*, 14(4): 535–562, 2018a. ISSN 15525279. doi: 10.1016/j.jalz.2018.02.018.

C. R. Jack, D. A. Bennett, K. Blennow, M. C. Carrillo, B. Dunn, S. B. Haeberlein, D. M. Holtzman, W. Jagust, F. Jessen, J. Karlawish, E. Liu, J. L. Molinuevo, T. Montine,

C. Phelps, K. P. Rankin, C. C. Rowe, P. Scheltens, E. Siemers, H. M. Snyder, R. Sperling, C. Elliott, E. Masliah, L. Ryan, and N. Silverberg. NIA-AA Research Framework: Toward a biological definition of Alzheimer's disease. *Alzheimer's & Dementia*, 14(4):535–562, apr 2018b. ISSN 15525260. doi: 10.1016/j.jalz.2018.02.018. URL `https://linkinghub.elsevier.com/retrieve/pii/S1552526018300724`.

S. Janelidze, N. Mattsson, S. Palmqvist, R. Smith, h. G. Beach, G. E. Serrano, X. Chai, N. K. Proctor, U. Eichenlaub, H. Zetterberg, K. Blennow, E. M. Reiman, E. Stomrud, J. L. Dage, and O. Hansson. Plasma P-tau181 in Alzheimer's disease: relationship to other biomarkers, differential diagnosis, and longitudinal progression to Alzheimer's dementia. *Nature Medicine*, 26(March):1–8, 2020. ISSN 1546-170X. doi: 10.1038/s41591-020-0755-1. URL `http://dx.doi.org/10.1038/s41591-020-0755-1`.

M. Jenkinson and S. Smith. A global optimisation method for robust affine registration of brain images. *Medical Image Analysis*, 5(2):143–156, 2001. ISSN 13618415. doi: 10.1016/S1361-8415(01)00036-6.

M. Jenkinson, P. Bannister, M. Brady, and S. Smith. Improved optimization for the robust and accurate linear registration and motion correction of brain images. *NeuroImage*, 17 (2):825–841, 2002. ISSN 10538119. doi: 10.1016/S1053-8119(02)91132-8.

F. Jessen, L. Feyen, K. Freymann, R. Tepest, W. Maier, R. Heun, H. H. Schild, and L. Scheef. Volume reduction of the entorhinal cortex in subjective memory impairment. *Neurobiology of Aging*, 27(12):1751–1756, 2006. ISSN 01974580. doi: 10.1016/j.neurobiolaging.2005.10.010.

F. Jessen, R. E. Amariglio, M. Van Boxtel, M. Breteler, M. Ceccaldi, G. Chételat, B. Dubois, C. Dufouil, K. A. Ellis, W. M. Van Der Flier, L. Glodzik, A. C. Van Harten, M. J. De Leon, P. McHugh, M. M. Mielke, J. L. Molinuevo, L. Mosconi, R. S. Osorio, A. Perrotin, R. C. Petersen, L. A. Rabin, L. Rami, B. Reisberg, D. M. Rentz, P. S. Sachdev, V. De La Sayette, A. J. Saykin, P. Scheltens, M. B. Shulman, M. J. Slavin, R. A. Sperling, R. Stewart, O. Uspenskaya, B. Vellas, P. J. Visser, and M. Wagner. A conceptual framework for research on subjective cognitive decline in preclinical Alzheimer's disease. *Alzheimer's and Dementia*, 10(6):844–852, 2014. ISSN 15525279. doi: 10.1016/j.jalz.2014.01.001.

T. Jo, K. Nho, and A. J. Saykin. Deep Learning in Alzheimer's Disease: Diagnostic Classification and Prognostic Prediction Using Neuroimaging Data. *Frontiers in Aging Neuroscience*, 11(August), 2019. ISSN 16634365. doi: 10.3389/fnagi.2019.00220.

K. Kamnitsas, C. Baumgartner, C. Ledig, V. Newcombe, J. Simpson, A. Kane, D. Menon, A. Nori, A. Criminisi, D. Rueckert, and B. Glocker. Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 10265 LNCS:597–609, 2017. ISSN 16113349. doi: 10.1007/978-3-319-59050-9_47.

A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-fei. Large-scale Video Classification with Convolutional Neural Networks. *arXiv preprint*, 2014.

B. S. Khundrakpam, A. Reid, J. Brauer, F. Carbonell, J. Lewis, S. Ameis, S. Karama, J. Lee, Z. Chen, S. Das, and A. C. Evans. Developmental changes in organization of structural brain networks. *Cerebral Cortex*, 23(9):2072–2085, 2013. ISSN 10473211. doi: 10.1093/cercor/bhs187.

D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations*, pages 1–15, 2014. ISSN 09252312. doi: http://doi.acm.org.ezproxy.lib.ucf.edu/10.1145/1830483.1830503. URL `http://arxiv.org/abs/1412.6980`.

M. Kivipelto, T. Ngandu, L. Fratiglioni, M. Viitanen, I. Kåreholt, B. Winblad, E. L. Helkala, J. Tuomilehto, H. Soininen, and A. Nissinen. Obesity and vascular risk factors at midlife and the risk of dementia and Alzheimer disease. *Archives of Neurology*, 62(10):1556–1560, 2005. ISSN 00039942. doi: 10.1001/archneur.62.10.1556.

S. Klöppel, C. M. Stonnington, C. Chu, B. Draganski, R. I. Scahill, J. D. Rohrer, N. C. Fox, C. R. Jack, J. Ashburner, and R. S. Frackowiak. Automatic classification of MR scans in Alzheimer's disease. *Brain*, 131(3):681–689, 2008. ISSN 00068950. doi: 10.1093/brain/awm319.

S. Klöppel, J. Peter, A. Ludl, A. Pilatus, S. Maier, I. Mader, B. Heimbach, L. Frings, K. Egger, J. Dukart, M. L. Schroeter, R. Perneczky, P. Häussermann, W. Vach, H. Urbach, S. Teipel, M. Hüll, and A. Abdulkadir. Applying Automated MR-Based Diagnostic Methods to the Memory Clinic: A Prospective Study. *Journal of Alzheimer's Disease*, 47(4):939–954, aug 2015. ISSN 13872877. doi: 10.3233/JAD-150334. URL `http://www.medra.org/servlet/aliasResolver?alias=iospress{&}doi=10.3233/JAD-150334`.

E. L. Koedam, M. Lehmann, W. M. Van Der Flier, P. Scheltens, Y. A. Pijnenburg, N. Fox, F. Barkhof, and M. P. Wattjes. Visual assessment of posterior atrophy development of a MRI rating scale. *European Radiology*, 21(12):2618–2625, 2011. ISSN 09387994. doi: 10.1007/s00330-011-2205-4.

J. R. Koikkalainen, H. F. M. Rhodius-Meester, K. S. Frederiksen, M. Bruun, S. G. Hasselbalch, M. Baroni, P. Mecocci, R. Vanninen, A. Remes, H. Soininen, M. van Gils, W. M. van der Flier, P. Scheltens, F. Barkhof, T. Erkinjuntti, and J. M. P. Lötjönen. Automatically computed rating scales from MRI for patients with cognitive disorders. *European Radiology*, 13(7): P1108, feb 2019. ISSN 0938-7994. doi: 10.1007/s00330-019-06067-1. URL `http://ovidsp.ovid.com/ovidweb.cgi?T=JS{&}PAGE=reference{&}D=emexa{&}NEWS=N{&}AN=620612139http://link.springer.com/10.1007/s00330-019-06067-1`.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. *Advances In Neural Information Processing Systems*, pages 1–9, 2012. ISSN 10495258. doi: http://dx.doi.org/10.1016/j.protcy.2014.09.007.

V. Latora and M. Marchiori. Efficient behavior of small-world networks. *Physical Review Letters*, 87(19):198701–1–198701–4, 2001. ISSN 10797114. doi: 10.1103/PhysRevLett.87.198701.

Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to digit recognition, 1989.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. ISSN 00189219. doi: 10.1109/5.726791. URL http://ieeexplore.ieee.org/document/726791/.

Y. Lecun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015. ISSN 14764687. doi: 10.1038/nature14539.

M. Lehmann, E. L. Koedam, J. Barnes, J. W. Bartlett, N. S. Ryan, Y. A. Pijnenburg, F. Barkhof, M. P. Wattjes, P. Scheltens, and N. C. Fox. Posterior cerebral atrophy in the absence of medial temporal lobe atrophy in pathologically-confirmed Alzheimer's disease. *Neurobiology of Aging*, 33(3):627.e1–627.e12, 2012. ISSN 15581497. doi: 10.1016/j.neurobiolaging.2011.04.003. URL http://dx.doi.org/10.1016/j.neurobiolaging.2011.04.003.

Y. Li, Y. Wang, G. Wu, F. Shi, L. Zhou, W. Lin, and D. Shen. Discriminant analysis of longitudinal cortical thickness changes in Alzheimer's disease using dynamic and network features. *Neurobiology of Aging*, 33(2):427.e15–427.e30, 2012. ISSN 01974580. doi: 10.1016/j.neurobiolaging.2010.11.008.

H. Liu, H. Jiang, W. Bi, B. Huang, X. Li, M. Wang, X. Wang, H. Zhao, Y. Cheng, X. Tao, C. Liu, T. Huang, C. Jin, T. Zhang, and J. Yang. Abnormal Gray Matter Structural Covariance Networks in Children With Bilateral Cerebral Palsy. *Frontiers in Human Neuroscience*, 13(October):1–13, 2019. ISSN 16625161. doi: 10.3389/fnhum.2019.00343.

C.-Y. Lo, P.-N. Wang, K.-H. Chou, J. Wang, Y. He, and C.-P. Lin. Diffusion Tensor Tractography Reveals Abnormal Topological Organization in Structural Cortical Networks in Alzheimer's Disease. *Journal of Neuroscience*, 30(50):16876–16885, 2010. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.4136-10.2010. URL http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.4136-10.2010.

I. Loshchilov and F. Hutter. SGDR: Stochastic Gradient Descent with Warm Restarts. *arXiv*, pages 1–16, 2016. doi: 10.1002/fut. URL http://arxiv.org/abs/1608.03983.

G. Mårtensson, D. Ferreira, T. Granberg, L. Cavallin, K. Oppedal, A. Padovani, I. Rektorova, L. Bonanni, M. Pardini, M. Kramberger, J.-P. Taylor, J. Hort, J. Snædal, J. Kulisevsky, F. Blanc, A. Antonini, P. Mecocci, B. Vellas, M. Tsolaki, I. Kłoszewska, H. Soininen, S. Lovestone, A. Simmons, D. Aarsland, and E. Westman. The reliability of a deep learning model in clinical out-of-distribution MRI data: a multicohort study. *arXiv preprint*, pages 1–18, 2019. URL http://arxiv.org/abs/1911.00515.

E. McDade and R. J. Bateman. Tau Positron Emission Tomography in Autosomal Dominant Alzheimer Disease. *JAMA Neurology*, 75(5):536, may 2018. ISSN 2168-6149. doi: 10.1001/jamaneurol.2017.4026. URL http://archneur.jamanetwork.com/article.aspx?doi=10.1001/jamaneurol.2017.4026.

G. McKhann, D. Drachman, M. Folstein, R. Katzman, D. Price, and E. M. Stadlan. Clinical diagnosis of Alzheimer's disease: Report of the NINCDS-ADRDA Work Group* under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. *Neurology*, 34(7):939–939, 1984. ISSN 0028-3878. doi: 10.1212/WNL.34.7.939. URL http://www.neurology.org/cgi/doi/10.1212/WNL.34.7.939.

G. M. McKhann, D. S. Knopman, H. Chertkow, B. T. Hyman, C. R. Jack, C. H. Kawas, W. E. Klunk, W. J. Koroshetz, J. J. Manly, R. Mayeux, R. C. Mohs, J. C. Morris, M. N. Rossor, P. Scheltens, M. C. Carrillo, B. Thies, S. Weintraub, and C. H. Phelps. The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's and Dementia*, 7(3):263–269, 2011. ISSN 15525260. doi: 10.1016/j.jalz.2011.03.005. URL `http://dx.doi.org/10.1016/j.jalz.2011.03.005`.

M. Mijalkov, E. Kakaei, J. B. Pereira, E. Westman, and G. Volpe. BRAPH: A graph theory software for the analysis of brain connectivity. *PLoS ONE*, 12(8):e0178798, aug 2017. ISSN 19326203. doi: 10.1371/journal.pone.0178798. URL `http://dx.plos.org/10.1371/journal.pone.0178798`.

M. Minsky and S. Papert. *Perceptrons (An introduction to computational geometry): Epilogue*. 1988. ISBN 0262631113. URL `http://cdsweb.cern.ch/record/114106`.

A. J. Mitchell, H. Beaumont, D. Ferguson, M. Yadegarfar, and B. Stubbs. Risk of dementia and mild cognitive impairment in older people with subjective memory complaints: Meta-analysis. *Acta Psychiatrica Scandinavica*, 130(6):439–451, 2014. ISSN 16000447. doi: 10.1111/acps.12336.

C. Möller, H. Vrenken, L. Jiskoot, A. Versteeg, F. Barkhof, P. Scheltens, and W. M. van der Flier. Different patterns of gray matter atrophy in early- and late-onset Alzheimer's disease. *Neurobiology of Aging*, 34(8):2014–2022, 2013. ISSN 01974580. doi: 10.1016/j.neurobiolaging.2013.02.013. URL `http://dx.doi.org/10.1016/j.neurobiolaging.2013.02.013`.

C. Möller, W. M. Van Der Flier, A. Versteeg, M. R. Benedictus, M. P. Wattjes, E. L. G. M. Koedam, P. Scheltens, F. Barkhof, and H. Vrenken. Quantitative regional validation of the visual rating scale for posterior cortical atrophy. *European Radiology*, 24(2):397–404, 2014. ISSN 09387994. doi: 10.1007/s00330-013-3025-5.

S. F. Muldoon, E. W. Bridgeford, and D. S. Bassett. Small-world propensity and weighted brain networks. *Scientific Reports*, 6(February):1–13, 2016. ISSN 20452322. doi: 10.1038/srep22057. URL `http://dx.doi.org/10.1038/srep22057`.

M. E. Murray, N. R. Graff-Radford, O. A. Ross, R. C. Petersen, R. Duara, and D. W. Dickson. Neuropathologically defined subtypes of Alzheimer's disease with distinct clinical characteristics: A retrospective study. *The Lancet Neurology*, 10(9):785–796, 2011. ISSN 14744422. doi: 10.1016/S1474-4422(11)70156-9. URL `http://dx.doi.org/10.1016/S1474-4422(11)70156-9`.

J. Murrell, M. Farlow, B. Ghetti, and M. D. Benson. A mutation in the amyloid precursor protein associated with hereditary Alzheimer's disease. *Science*, 254(5028):97–99, 1991. ISSN 00368075. doi: 10.1126/science.1925564.

A. Nordberg, J. O. Rinne, A. Kadir, and B. Lngström. The use of PET in Alzheimer disease. *Nature Reviews Neurology*, 6(2):78–87, 2010. ISSN 17594758. doi: 10.1038/nrneurol.2009.217. URL `http://dx.doi.org/10.1038/nrneurol.2009.217`.

S. Norton, F. E. Matthews, D. E. Barnes, K. Yaffe, and C. Brayne. Potential for primary prevention of Alzheimer's disease: An analysis of population-based data. *The Lancet Neurology*, 13(8):788–794, 2014. ISSN 14744465. doi: 10.1016/S1474-4422(14)70136-X.

L. Oakden-Rayner. Exploring Large-scale Public Medical Image Datasets. *Academic Radiology*, 27(1):106–112, 2020. ISSN 18784046. doi: 10.1016/j.acra.2019.10.006. URL `https://doi.org/10.1016/j.acra.2019.10.006`.

F. Pasquier, D. Leys, J. G. Weerts, F. Mounier-Vehier, F. Barkhof, and P. Scheltens. Inter-and intraobserver reproducibility of cerebral atrophy assessment on mri scans with hemispheric infarcts. *European Neurology*, 36(5):268–272, 1996. ISSN 14219913. doi: 10.1159/000117270.

A. Paszke, G. Chanan, Z. Lin, S. Gross, E. Yang, L. Antiga, and Z. Devito. Automatic differentiation in PyTorch. In *31st Conference on Neural Information Processing Systems*, pages 1–4, 2017. ISBN 9788578110796. doi: 10.1017/CBO9781107707221.009.

A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Advances in Neural Information Processing Systems*, pages 8024–8035, dec 2019. URL `http://arxiv.org/abs/1912.01703`.

A. Payan and G. Montana. Predicting Alzheimer's disease: a neuroimaging study with 3D convolutional neural networks. *arXiv preprint*, pages 1–9, 2015. ISSN 10769757. doi: 10.1613/jair.301. URL `http://arxiv.org/abs/1502.02506`.

J. Pearl. Reverend Bayes on Inference Engines: a Distributed Hierarchical Approach. pages 133–136, 1982.

J. B. Pereira, D. Aarsland, C. E. Ginestet, A. V. Lebedev, L. O. Wahlund, A. Simmons, G. Volpe, and E. Westman. Aberrant cerebral network topology and mild cognitive impairment in early Parkinson's disease. *Human Brain Mapping*, 36(8):2980–2995, 2015. ISSN 10970193. doi: 10.1002/hbm.22822.

J. B. Pereira, M. Mijalkov, E. Kakaei, P. Mecocci, B. Vellas, M. Tsolaki, I. Kłoszewska, H. Soininen, C. Spenger, S. Lovestone, A. Simmons, L. O. Wahlund, G. Volpe, and E. Westman. Disrupted Network Topology in Patients with Stable and Progressive Mild Cognitive Impairment and Alzheimer's Disease. *Cerebral Cortex*, 26(8):3476–3493, 2016. ISSN 14602199. doi: 10.1093/cercor/bhw128.

C. S. Perone, P. Ballester, R. C. Barros, and J. Cohen-Adad. Unsupervised domain adaptation for medical imaging segmentation with self-ensembling. *NeuroImage*, 194:1–11, 2019. ISSN 10959572. doi: 10.1016/j.neuroimage.2019.03.026.

A. Perrotin, E. C. Mormino, C. M. Madison, A. O. Hayenga, and W. J. Jagust. Subjective cognition and amyloid deposition imaging: A Pittsburgh compound B positron emission tomography study in normal elderly individuals. *Archives of Neurology*, 69(2):223–229, 2012. ISSN 00039942. doi: 10.1001/archneurol.2011.666.

R. C. Petersen. Mild cognitive impairment as a diagnostic entity. *Journal of Internal Medicine*, 256(3):183–194, 2004. ISSN 09546820. doi: 10.1111/j.1365-2796.2004.01388.x.

C. Pettigrew, A. Soldan, K. Sloane, Q. Cai, J. Wang, M. C. Wang, A. Moghekar, M. I. Miller, and M. Albert. Progressive medial temporal lobe atrophy during preclinical Alzheimer's disease. *NeuroImage: Clinical*, 16(August):439–446, 2017. ISSN 22131582. doi: 10.1016/j.nicl.2017.08.022. URL `https://doi.org/10.1016/j.nicl.2017.08.022`.

D. J. Phillips, A. McGlaughlin, D. Ruth, L. R. Jager, and A. Soldan. Graph theoretic analysis of structural connectivity across the spectrum of Alzheimer's disease: The importance of graph creation methods. *NeuroImage: Clinical*, 7:377–390, 2015. ISSN 22131582. doi: 10.1016/j.nicl.2015.01.007. URL `http://dx.doi.org/10.1016/j.nicl.2015.01.007`.

R. P. Poudel, P. Lamata, and G. Montana. Recurrent fully convolutional neural networks for multi-slice MRI cardiac segmentation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 10129 LNCS:83–94, 2017. ISSN 16113349. doi: 10.1007/978-3-319-52280-7_8.

M. Prince, A. Wimo, M. Guerchet, A. Gemma-Claire, Y.-T. Wu, and M. Prina. World Alzheimer Report 2015: The Global Impact of Dementia - An analysis of prevalence, incidence, cost and trends. *Alzheimer's Disease International*, page 84, 2015. doi: 10.1111/j.0963-7214.2004.00293.x. URL `www.alz.co.uk`.

M. Reuter, N. J. Schmansky, H. D. Rosas, and B. Fischl. Within-subject template estimation for unbiased longitudinal image analysis. *NeuroImage*, 61(4):1402–1418, jul 2012. ISSN 10538119. doi: 10.1016/j.neuroimage.2012.02.084. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3624763/pdf/nihms412728.pdfhttps://linkinghub.elsevier.com/retrieve/pii/S1053811912002765`.

R. Ricciarelli and E. Fedele. The Amyloid Cascade Hypothesis in Alzheimer's Disease: It's Time to Change Our Mind. *Current Neuropharmacology*, 15(6):926–935, 2017. ISSN 1570159X. doi: 10.2174/1570159x15666170116143743.

B. H. Ridha, J. Barnes, J. W. Bartlett, A. Godbolt, T. Pepple, M. N. Rossor, and N. C. Fox. Tracking atrophy progression in familial Alzheimer's disease: a serial MRI study. *Lancet Neurology*, 5(10):828–834, 2006. ISSN 14744422. doi: 10.1016/S1474-4422(06)70550-6.

N. Rieke, J. Hancox, W. Li, F. Milletari, H. Roth, S. Albarqouni, S. Bakas, M. N. Galtier, B. Landman, K. Maier-Hein, S. Ourselin, M. Sheller, R. M. Summers, A. Trask, D. Xu, M. Baust, and M. J. Cardoso. The Future of Digital Health with Federated Learning. *arXiv*, 2020. URL `http://arxiv.org/abs/2003.08119`.

O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9351:234–241, may 2015. ISSN 16113349. doi: 10.1007/978-3-319-24574-4_28. URL `http://link.springer.com/10.1007/978-3-319-24574-4{_}28http://arxiv.org/abs/1505.04597`.

F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408, 1958. ISSN 0033295X. doi: 10.1037/h0042519.

M. Rubinov and O. Sporns. Complex network measures of brain connectivity: Uses and interpretations. *NeuroImage*, 52(3):1059–1069, 2010. ISSN 10538119. doi: 10.1016/j.neuroimage.2009.10.003. URL `http://dx.doi.org/10.1016/j.neuroimage.2009.10.003`.

D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986. ISSN 00280836. doi: 10.1038/323533a0.

H. Rusinek, S. De Santi, D. Frid, W. H. Tsui, C. Y. Tarshish, A. Convit, and M. J. De Leon. Regional Brain Atrophy Rate Predicts Future Cognitive Decline: 6-Year Longitudinal MR Imaging Study of Normal Aging. *Radiology*, 229(3):691–696, 2003. ISSN 00338419. doi: 10.1148/radiol.2293021299.

E. J. Sanz-Arigita, M. M. Schoonheim, J. S. Damoiseaux, S. A. R. B. Rombouts, E. Maris, F. Barkhof, P. Scheltens, and C. J. Stam. Loss of 'Small-World' Networks in Alzheimer's Disease: Graph Analysis of fMRI Resting-State Functional Connectivity. *PLoS ONE*, 5 (11):1–14, 2010. ISSN 19326203. doi: 10.1371/journal.pone.0013788.

P. Scheltens, D. Leys, F. Barkhof, D. Huglo, H. C. Weinstein, P. Vermersch, M. Kuiper, M. Steinling, E. C. Wolters, and J. Valk. Atrophy of medial temporal lobes on MRI in "probable" Alzheimer's disease and normal ageing: diagnostic value and neuropsychological correlates. *Journal of Neurology Neurosurgery, and Psychiatry*, 55:967–972, 1992. ISSN 0022-3050. doi: 10.1136/jnnp.55.10.967.

P. Scheltens, F. Pasquier, J. G. Weerts, F. Barkhof, and D. Leys. Qualitative assessment of cerebral atrophy on MRI: inter- and intra- observer reproducibility in dementia and normal aging. *European Neurology*, 37(2):95–99, 1997. ISSN 0014-3022.

L. P. Schilling, E. R. Zimmer, M. Shin, A. Leuzy, T. A. Pascoal, A. L. Benedet, W. V. Borelli, A. Palmini, S. Gauthier, and P. Rosa-Neto. Imaging Alzheimer's disease pathophysiology with PET. *Dementia e Neuropsychologia*, 10(2):79–90, 2016. ISSN 19805764. doi: 10.1590/S1980-5764-2016DN1002003.

S. A. Schultz, J. M. Oh, R. L. Koscik, N. M. Dowling, C. L. Gallagher, C. M. Carlsson, B. B. Bendlin, A. LaRue, B. P. Hermann, H. A. Rowley, S. Asthana, M. A. Sager, S. C. Johnson, and O. C. Okonkwo. Subjective memory complaints, cortical thinning, and cognitive dysfunction in middle-age adults at risk of AD. *Alzheimer's and Dementia: Diagnosis, Assessment and Disease Monitoring*, 1(1):33–40, 2015. ISSN 23528729. doi: 10.1016/j.dadm.2014.11.010. URL `http://dx.doi.org/10.1016/j.dadm.2014.11.010`.

D. J. Selkoe and J. Hardy. The amyloid hypothesis of Alzheimer's disease at 25 years. *EMBO Molecular Medicine*, 8(6):595–608, 2016. ISSN 1757-4676. doi: 10.15252/emmm.201606210. URL `http://embomolmed.embopress.org/lookup/doi/10.15252/emmm.201606210`.

D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis. Mastering the game of Go with deep neural networks and tree search.

Nature, 529(7587):484–489, 2016. ISSN 14764687. doi: 10.1038/nature16961. URL http://dx.doi.org/10.1038/nature16961.

D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, T. Lillicrap, F. Hui, L. Sifre, G. Van Den Driessche, T. Graepel, and D. Hassabis. Mastering the game of Go without human knowledge. *Nature*, 550(7676):354–359, 2017. ISSN 14764687. doi: 10.1038/nature24270. URL http://dx.doi.org/10.1038/nature24270.

K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv*, sep 2015. ISSN 14732262. doi: 10.1016/j.infsof.2008.09.005. URL http://arxiv.org/abs/1409.1556.

S. A. Small and K. Duff. Linking Aβ and Tau in Late-Onset Alzheimer's Disease: A Dual Pathway Hypothesis. *Neuron*, 60(4):534–542, 2008. ISSN 08966273. doi: 10.1016/j.neuron.2008.11.007. URL http://dx.doi.org/10.1016/j.neuron.2008.11.007.

R. A. Sperling, P. S. Aisen, L. A. Beckett, D. A. Bennett, S. Craft, A. M. Fagan, T. Iwatsubo, C. R. Jack, J. Kaye, T. J. Montine, D. C. Park, E. M. Reiman, C. C. Rowe, E. Siemers, Y. Stern, K. Yaffe, M. C. Carrillo, B. Thies, M. Morrison-Bogorad, M. V. Wagster, and C. H. Phelps. Toward defining the preclinical stages of Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's and Dementia*, 7(3):280–292, 2011. ISSN 15525260. doi: 10.1016/j.jalz.2011.03.003. URL http://dx.doi.org/10.1016/j.jalz.2011.03.003.

C. J. Stam, B. F. Jones, G. Nolte, M. Breakspear, and P. Scheltens. Small-world networks and functional connectivity in Alzheimer's disease. *Cerebral Cortex*, 17(1):92–99, 2007. ISSN 10473211. doi: 10.1093/cercor/bhj127.

Y. Stern, B. Gurland, T. K. Tatemichi, M. X. Tang, D. Wilder, and R. Mayeux. Influence of Education and Occupation on the Incidence of Alzheimers-Disease. *Jama-Journal of the American Medical Association*, 271(13):1004–1010, 1994. ISSN 0098-7484. doi: 10.1001/jama.271.13.1004.

B. T. Thomas Yeo, F. M. Krienen, J. Sepulcre, M. R. Sabuncu, D. Lashkari, M. Hollinshead, J. L. Roffman, J. W. Smoller, L. Zöllei, J. R. Polimeni, B. Fisch, H. Liu, and R. L. Buckner. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of Neurophysiology*, 106(3):1125–1165, 2011. ISSN 00223077. doi: 10.1152/jn.00338.2011.

B. M. Tijms, P. Seris, D. J. Willshaw, and S. M. Lawrie. Similarity-based extraction of individual networks from gray matter MRI scans. *Cerebral Cortex*, 22(7):1530–1541, 2012. ISSN 10473211. doi: 10.1093/cercor/bhr221.

B. M. Tijms, A. M. Wink, W. de Haan, W. M. van der Flier, C. J. Stam, P. Scheltens, and F. Barkhof. Alzheimer's disease: connecting findings from graph theoretical studies of brain networks. *Neurobiology of Aging*, 34(8):2023–2036, 2013. ISSN 01974580. doi: 10.1016/j.neurobiolaging.2013.02.020. URL http://dx.doi.org/10.1016/j.neurobiolaging.2013.02.020.

B. M. Tijms, M. ten Kate, A. A. Gouw, A. Borta, S. Verfaillie, C. E. Teunissen, P. Scheltens, F. Barkhof, and W. M. van der Flier. Grey matter networks and clinical progression in subjects with pre-dementia Alzheimer's disease. *Neurobiology of Aging*, 61:75–81, 2017. ISSN 01974580. doi: 10.1016/j.neurobiolaging.2017.09.011. URL `http://linkinghub.elsevier.com/retrieve/pii/S0197458017303044`.

G. Torisson, D. Van Westen, L. Stavenow, L. Minthon, and E. Londos. Medial temporal lobe atrophy is underreported and may have important clinical correlates in medical inpatients. *BMC Geriatrics*, 15(1):1–8, 2015. ISSN 14712318. doi: 10.1186/s12877-015-0066-4.

N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, and M. Joliot. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage*, 15(1):273–289, 2002. ISSN 10538119. doi: 10.1006/nimg.2001.0978.

M. J. Valenzuela and P. Sachdev. Brain reserve and dementia: A systematic review. *Psychological Medicine*, 36(4):441–454, 2006. ISSN 00332917. doi: 10.1017/S0033291705006264.

B. C. M. van Wijk, C. J. Stam, and A. Daffertshofer. Comparing brain networks of different size and connectivity density using graph theory. *PLoS ONE*, 5(10):1–13, 2010. ISSN 19326203. doi: 10.1371/journal.pone.0013701.

E. Vanmechelen, H. Vanderstichele, P. Davidsson, E. Van Kerschaver, B. Van Der Perre, M. Sjögren, N. Andreasen, and K. Blennow. Quantification of tau phosphorylated at threonine 181 in human cerebrospinal fluid: A sandwich ELISA with a synthetic phosphopeptide for standardization. *Neuroscience Letters*, 285(1):49–52, 2000. ISSN 03043940. doi: 10.1016/S0304-3940(00)01036-3.

V. Velickaite, D. Ferreira, L. Cavallin, L. Lind, H. Ahlström, L. Kilander, E. Westman, and E. M. Larsson. Medial temporal lobe atrophy ratings in a large 75-year-old population-based cohort: gender-corrected and education-corrected normative data. *European Radiology*, pages 1–9, 2017. ISSN 14321084. doi: 10.1007/s00330-017-5103-6.

P. Vemuri and C. R. Jack. Role of structural MRI in Alzheimer's disease. *Alzheimer's Research and Therapy*, 2(4), 2010. ISSN 17589193. doi: 10.1186/alzrt47.

M. W. Vernooij, F. B. Pizzini, R. Schmidt, M. Smits, T. A. Yousry, N. Bargallo, G. B. Frisoni, S. Haller, and F. Barkhof. Dementia imaging in clinical practice: a European-wide survey of 193 centres and conclusions by the ESNR working group. *Neuroradiology*, 61(6):633–642, 2019. ISSN 14321920. doi: 10.1007/s00234-019-02188-y.

O. Vinyals, I. Babuschkin, W. M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D. H. Choi, R. Powell, T. Ewalds, P. Georgiev, J. Oh, D. Horgan, M. Kroiss, I. Danihelka, A. Huang, L. Sifre, T. Cai, J. P. Agapiou, M. Jaderberg, A. S. Vezhnevets, R. Leblond, T. Pohlen, V. Dalibard, D. Budden, Y. Sulsky, J. Molloy, T. L. Paine, C. Gulcehre, Z. Wang, T. Pfaff, Y. Wu, R. Ring, D. Yogatama, D. Wünsch, K. McKinney, O. Smith, T. Schaul, T. Lillicrap, K. Kavukcuoglu, D. Hassabis, C. Apps, and D. Silver. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, 2019. ISSN 14764687. doi: 10.1038/s41586-019-1724-z. URL `http://dx.doi.org/10.1038/s41586-019-1724-z`.

O. Voevodskaya, J. B. Pereira, G. Volpe, O. Lindberg, E. Stomrud, D. van Westen, E. Westman, and O. Hansson. Altered structural network organization in cognitively normal individuals with amyloid pathology. *Neurobiology of Aging*, 64:15–24, 2018. ISSN 15581497. doi: 10.1016/j.neurobiolaging.2017.11.014. URL `https://doi.org/10.1016/j.neurobiolaging.2017.11.014`.

L.-O. Wahlund, P. Julin, J. Lindqvist, and P. Scheltens. Visual assessment of medial temporal lobe atrophy in demented and healthy control subjects: correlation with volumetry. *Psychiatry Research: Neuroimaging*, 90(3):193–199, 1999. ISSN 09254927. doi: 10.1016/S0925-4927(99)00016-5. URL `https://ac.els-cdn.com/S0925492799000165/1-s2.0-S0925492799000165-main.pdf?{_}tid=33ac94fc-d111-4137-88d9-2f6ca702583e{&}acdnat=1526123326{_}dfd495990f9ade84488074d8bdf84427{%}0Ahttp://linkinghub.elsevier.com/retrieve/pii/S0925492799000165`.

F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang. Residual Attention Network for Image Classification. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, number 1, pages 6450–6458. IEEE, jul 2017. ISBN 978-1-5386-0457-1. doi: 10.1109/CVPR.2017.683. URL `http://ieeexplore.ieee.org/document/8100166/`.

M. P. Wattjes, W. J. P. Henneman, W. M. van der Flier, O. de Vries, F. Träber, J. J. G. Geurts, P. Scheltens, H. Vrenken, and F. Barkhof. Diagnostic Imaging of Patients in a Memory Clinic: Comparison of MR Imaging and 64–Detector Row CT. *Radiology*, 253(1):174–183, 2009. ISSN 0033-8419. doi: 10.1148/radiol.2531082262. URL `http://pubs.rsna.org/doi/10.1148/radiol.2531082262`.

D. J. Watts and S. H. Strogatz. Collective dynamics of 'small world' networks. *Nature*, 393 (June):440–442, 1998.

E. Westman, L. Cavallin, J. S. Muehlboeck, Y. Zhang, P. Mecocci, B. Vellas, M. Tsolaki, I. Kłoszewska, H. Soininen, C. Spenger, S. Lovestone, A. Simmons, and L. O. Wahlund. Sensitivity and specificity of medial temporal lobe visual ratings and multivariate regional MRI classification in Alzheimer's disease. *PLoS ONE*, 6(7), 2011a. ISSN 19326203. doi: 10.1371/journal.pone.0022506.

E. Westman, A. Simmons, Y. Zhang, J. S. Muehlboeck, C. Tunnard, Y. Liu, L. Collins, A. Evans, P. Mecocci, B. Vellas, M. Tsolaki, I. Kloszewska, H. Soininen, S. Lovestone, C. Spenger, and L. O. Wahlund. Multivariate analysis of MRI data for Alzheimer's disease, mild cognitive impairment and healthy controls. *NeuroImage*, 54(2):1178–1187, 2011b. ISSN 10538119. doi: 10.1016/j.neuroimage.2010.08.044.

L. J. Whalley. The Dementia of Down's Syndrome and its Relevance to Aetiological Studies of Alzheimer's Disease. *Annals of the New York Academy of Sciences*, 396(1):39–53, 1982. ISSN 17496632. doi: 10.1111/j.1749-6632.1982.tb26842.x.

J. L. Whitwell, D. W. Dickson, M. E. Murray, S. D. Weigand, N. Tosakulwong, M. L. Senjem, D. S. Knopman, B. F. Boeve, J. E. Parisi, R. C. Petersen, C. R. Jack, and K. A. Josephs. Neuroimaging correlates of pathologically defined subtypes of Alzheimer's disease: A case-control study. *The Lancet Neurology*, 11(10):868–877, 2012. ISSN 14744422. doi: 10.1016/S1474-4422(12)70200-4.

A. Wimo, M. Guerchet, G. C. Ali, Y. T. Wu, A. M. Prina, B. Winblad, L. Jönsson, Z. Liu, and M. Prince. The worldwide costs of dementia 2015 and comparisons with 2010. *Alzheimer's and Dementia*, 13(1):1–7, 2017. ISSN 15525279. doi: 10.1016/j.jalz.2016.07.150.

W. World Health Organization. *The ICD-10 classification of mental and behavioural disorders: Diagnostic criteria for research*. World Health Organization, 1990. ISBN 9241544554.

K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *International Conference on Machine Learning*, 2015.

L. Yao, J. Prosky, B. Covington, and K. Lyman. A Strong Baseline for Domain Adaptation and Generalization in Medical Imaging. *arXiv preprint*, pages 1–5, apr 2019. URL `http://arxiv.org/abs/1904.01638`.

Z. Yao, Y. Zhang, L. Lin, Y. Zhou, C. Xu, and T. Jiang. Abnormal cortical networks in mild cognitive impairment and alzheimer's disease. *PLoS Computational Biology*, 6(11):1–11, 2010. ISSN 1553734X. doi: 10.1371/journal.pcbi.1001006.

P.-P. Ypsilantis and G. Montana. Recurrent Convolutional Networks for Pulmonary Nodule Detection in CT Imaging. *arXiv preprint*, pages 1–36, 2016.

L. Yue, T. Wang, J. Wang, G. Li, J. Wang, X. Li, W. Li, M. Hu, and S. Xiao. Asymmetry of hippocampus and amygdala defect in subjective cognitive decline among the community dwelling Chinese. *Frontiers in Psychiatry*, 9(JUN):1–11, 2018. ISSN 16640640. doi: 10.3389/fpsyt.2018.00226.

A. Zalesky, A. Fornito, I. H. Harding, L. Cocchi, M. Yücel, C. Pantelis, and E. T. Bullmore. Whole-brain anatomical networks: Does the choice of nodes matter? *NeuroImage*, 50(3):970–983, 2010. ISSN 10538119. doi: 10.1016/j.neuroimage.2009.12.027. URL `http://dx.doi.org/10.1016/j.neuroimage.2009.12.027`.

J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLoS Medicine*, 15(11):1–17, 2018. ISSN 15491676. doi: 10.1371/journal.pmed.1002683.

Y. Zhang, L. Lin, C. P. Lin, Y. Zhou, K. H. Chou, C. Y. Lo, T. P. Su, and T. Jiang. Abnormal topological organization of structural brain networks in schizophrenia. *Schizophrenia Research*, 141(2-3):109–118, 2012. ISSN 09209964. doi: 10.1016/j.schres.2012.08.021. URL `http://dx.doi.org/10.1016/j.schres.2012.08.021`.

T. X. Zou, L. She, C. Zhan, Y. Q. Gao, and H. J. Chen. Altered topological properties of gray matter structural covariance networks in minimal hepatic encephalopathy. *Frontiers in Neuroanatomy*, 12(November):1–10, 2018. ISSN 16625129. doi: 10.3389/fnana.2018.00101.