Universidade Estadual de Campinas
Instituto de Computação

Guilherme Adriano Fôlego

ADNet: Computer-Aided Diagnosis for Alzheimer's Disease Using Whole-Brain 3D Convolutional Neural Network

ADNet: Diagnóstico Assistido por Computador para Doença de Alzheimer Usando Rede Neural Convolucional 3D com Cérebro Inteiro

CAMPINAS
2018

Guilherme Adriano Fôlego

ADNet: Computer-Aided Diagnosis for Alzheimer's Disease Using Whole-Brain 3D Convolutional Neural Network

ADNet: Diagnóstico Assistido por Computador para Doença de Alzheimer Usando Rede Neural Convolucional 3D com Cérebro Inteiro

Dissertação apresentada ao Instituto de Computação da Universidade Estadual de Campinas como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.

Thesis presented to the Institute of Computing of the University of Campinas in partial fulfillment of the requirements for the degree of Master in Computer Science.

Supervisor/Orientador: Prof. Dr. Anderson de Rezende Rocha
Co-supervisor/Coorientadora: Dra. Marina Weiler

Este exemplar corresponde à versão final da Dissertação defendida por Guilherme Adriano Fôlego e orientada pelo Prof. Dr. Anderson de Rezende Rocha.

CAMPINAS
2018

Agência(s) de fomento e nº(s) de processo(s): Não se aplica.
ORCID: https://orcid.org/0000-0002-6257-1058

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Matemática, Estatística e Computação Científica
Ana Regina Machado - CRB 8/5467

F698a

Fôlego, Guilherme Adriano, 1989-
ADNet : computer-aided diagnosis for Alzheimer's disease using whole-brain 3D convolutional neural network / Guilherme Adriano Fôlego. – Campinas, SP : [s.n.], 2018.

Orientador: Anderson de Rezende Rocha.
Coorientador: Marina Weiler.
Dissertação (mestrado) – Universidade Estadual de Campinas, Instituto de Computação.

1. Alzheimer, Doença de. 2. Diagnóstico auxiliado por computador. 3. Imagem por ressonância magnética. 4. Visão por computador. 5. Redes neurais convolucionais. I. Rocha, Anderson de Rezende, 1980-. II. Weiler, Marina, 1983-. III. Universidade Estadual de Campinas. Instituto de Computação. IV. Título.

Informações para Biblioteca Digital

Título em outro idioma: ADNet : diagnóstico assistido por computador para doença de Alzheimer usando rede neural convolucional 3D com cérebro inteiro
Palavras-chave em inglês:
Alzheimer's disease
Computer-aided diagnosis
Magnetic resonance imaging
Computer vision
Convolutional neural networks
Área de concentração: Ciência da Computação
Titulação: Mestre em Ciência da Computação
Banca examinadora:
Anderson de Rezende Rocha [Orientador]
Leticia Rittner
Sandra Eliza Fontes de Avila
Data de defesa: 18-12-2018
Programa de Pós-Graduação: Ciência da Computação

Universidade Estadual de Campinas
Instituto de Computação

Guilherme Adriano Fôlego

ADNet: Computer-Aided Diagnosis for Alzheimer's Disease Using Whole-Brain 3D Convolutional Neural Network

ADNet: Diagnóstico Assistido por Computador para Doença de Alzheimer Usando Rede Neural Convolucional 3D com Cérebro Inteiro

Banca Examinadora:

- Prof. Dr. Anderson de Rezende Rocha
  Instituto de Computação
- Profa. Dra. Leticia Rittner
  Faculdade de Engenharia Elétrica e de Computação
- Profa. Dra. Sandra Eliza Fontes de Avila
  Instituto de Computação

A ata da defesa, assinada pelos membros da Comissão Examinadora, consta no SIGA/Sistema de Fluxo de Dissertação/Tese e na Secretaria do Programa da Unidade.

Campinas, 18 de dezembro de 2018

Dicebat Bernardus Carnotensis
nos esse quasi nanos,
gigantium humeris insidentes,
ut possimus plura eis et remotiora videre,
non utique proprii visus acumine,
aut eminentia corporis,
sed quia in altum subvenimur et
extollimur magnitudine gigantea.

(John of Salisbury)

Agradecimentos

Em primeiro lugar, eu gostaria de agradecer a meus pais, Filomena Maria da Silva Adriano Fôlego e Plínio Fôlego, por terem me dado a melhor educação possível ao longo de toda minha vida, além de sempre apoiarem e incentivarem meus estudos. Essa dissertação é resultado direto da dedicação e do esforço deles. Nesse sentido, também agradeço a toda minha família, pelo apoio e compreensão durantes esses anos de foco nos estudos.

Gostaria de agradecer ao meu orientador, Prof. Dr. Anderson de Rezende Rocha, por todo ensino e base de conhecimento fornecidos ao longo desses anos de parceria. Desde as disciplinas, passando pela orientação, até a conclusão desse trabalho, sempre houve bastante ajuda, cobrança e determinação em alcançar o melhor resultado possível. De forma similar, essa dissertação se tornou viável graças ao apoio de minha coorientadora, Dra. Marina Weiler, que é especialista na doença de Alzheimer e, conjuntamente ao Dr. Raphael Fernandes Casseb, auxiliou na interpretação e discussão dos resultados encontrados, além de nos fornecer melhor entendimento e compreensão sobre a doença.

Diversos colegas foram muito importantes para enriquecer meu conhecimento, principalmente os integrantes dos grupos Plataforma Tecnológica de Computação Cognitiva da Fundação CPqD, Oficina Tecnológica de Computação Visual do Instituto de Pesquisas Eldorado, Laboratório Reasoning for Complex Data (RECOD) do Instituto de Computação da Universidade Estadual de Campinas, e da turma ingressante em 2007 no curso de Ciência da Computação (CC07), oferecido pelo Instituto de Computação da Universidade Estadual de Campinas. Seria injusto citar apenas alguns, mas ressalto aqui aqueles que auxiliaram na revisão desse texto, Alan Godoy Souza Mello e Lucas Oliveira David.

Por fim, agradeço pela compreensão, incentivo e apoio financeiro das empresas em que trabalhei ao longo do desenvolvimento desse mestrado, Fundação CPqD e Instituto de Pesquisas Eldorado.

Resumo

Demência por doença de Alzheimer (DA) é uma síndrome clínica caracterizada por múltiplos problemas cognitivos, incluindo dificuldades na memória, funções executivas, linguagem e habilidades visuoespaciais. Sendo a forma mais comum de demência, essa doença mata mais do que câncer de mama e de próstata combinados, além de ser a sexta principal causa de morte nos Estados Unidos. A neuroimagem é uma das áreas de pesquisa mais promissoras para a detecção de biomarcadores estruturais da DA, onde uma técnica não invasiva é usada para capturar uma imagem digital do cérebro, a partir da qual especialistas extraem padrões e características da doença. Nesse contexto, os sistemas de diagnóstico assistido por computador (DAC) são abordagens que visam ajudar médicos e especialistas na interpretação de dados médicos, para fornecer diagnósticos aos pacientes. Em particular, redes neurais convolucionais (RNCs) são um tipo especial de rede neural artificial (RNA), que foram inspiradas em como o sistema visual funciona e, nesse sentido, têm sido cada vez mais utilizadas em tarefas de visão computacional, alcançando resultados impressionantes. Em nossa pesquisa, um dos principais objetivos foi utilizar o que há de mais avançado sobre aprendizagem profunda (por exemplo, RNC) para resolver o difícil problema de identificar biomarcadores estruturais da DA em imagem por ressonância magnética (IRM), considerando três grupos diferentes, ou seja, cognitivamente normal (CN), comprometimento cognitivo leve (CCL) e DA. Adaptamos redes convolucionais com dados fornecidos principalmente pela ADNI e avaliamos no desafio CADDementia, resultando em um cenário mais próximo das condições no mundo real, em que um sistema DAC é usado em um conjunto de dados diferente daquele usado no treinamento. Os principais desafios e contribuições da nossa pesquisa incluem a criação de um sistema de aprendizagem profunda que seja totalmente automático e comparativamente rápido, ao mesmo tempo em que apresenta resultados competitivos, sem usar qualquer conhecimento específico de domínio. Nomeamos nossa melhor arquitetura ADNet (Alzheimer's Disease Network) e nosso melhor método ADNet-DA (ADNet com adaptação de domínio), o qual superou a maioria das submissões no CADDementia, todas utilizando conhecimento prévio da doença, como regiões de interesse específicas do cérebro. A principal razão para não usar qualquer informação da doença em nosso sistema é fazer com que ele aprenda e extraia padrões relevantes de regiões importantes do cérebro automaticamente, que podem ser usados para apoiar os padrões atuais de diagnóstico e podem inclusive auxiliar em novas descobertas para diferentes ou novas doenças. Após explorar uma série de técnicas de visualização para interpretação de modelos, associada à inteligência artificial explicável (XAI), acreditamos que nosso método possa realmente ser empregado na prática médica. Ao diagnosticar pacientes, é possível que especialistas usem a ADNet para gerar uma diversidade de visualizações explicativas para uma determinada imagem, conforme ilustrado em nossa pesquisa, enquanto a ADNet-DA pode ajudar com o diagnóstico. Desta forma, os especialistas podem chegar a uma decisão mais informada e em menos tempo.

Abstract

Dementia by Alzheimer's disease (AD) is a clinical syndrome characterized by multiple cognitive problems, including difficulties in memory, executive functions, language and visuospatial skills. Being the most common form of dementia, this disease kills more than breast cancer and prostate cancer combined, and it is the sixth leading cause of death in the United States. Neuroimaging is one of the most promising areas of research for early detection of AD structural biomarkers, where a non-invasive technique is used to capture a digital image of the brain, from which specialists extract patterns and features of the disease. In this context, computer-aided diagnosis (CAD) systems are approaches that aim at assisting doctors and specialists in interpretation of medical data to provide diagnoses for patients. In particular, convolutional neural networks (CNNs) are a special kind of artificial neural network (ANN), which were inspired by how the visual system works, and, in this sense, have been increasingly used in computer vision tasks, achieving impressive results. In our research, one of the main goals was bringing to bear what is most advanced in deep learning research (*e.g.*, CNN) to solve the difficult problem of identifying AD structural biomarkers in magnetic resonance imaging (MRI), considering three different groups, namely, cognitively normal (CN), mild cognitive impairment (MCI), and AD. We tailored convolutional networks with data primarily provided by ADNI, and evaluated them on the CADDementia challenge, thus resulting in a scenario very close to the real-world conditions, in which a CAD system is used on a dataset differently from the one used for training. The main challenges and contributions of our research include devising a deep learning system that is both completely automatic and comparatively fast, while also presenting competitive results, without using any domain specific knowledge. We named our best architecture ADNet (Alzheimer's Disease Network), and our best method ADNet-DA (ADNet with domain adaption), which outperformed most of the CADDementia submissions, all of them using prior knowledge from the disease, such as specific regions of interest of the brain. The main reason for not using any information from the disease in our system is to make it automatically learn and extract relevant patterns from important regions of the brain, which can be used to support current diagnosis standards, and may even assist in new discoveries for different or new diseases. After exploring a number of visualization techniques for model interpretability, associated with explainable artificial intelligence (XAI), we believe that our method can be actually employed in medical practice. While diagnosing patients, it is possible for specialists to use ADNet to generate a diversity of explanatory visualizations for a given image, as illustrated in our research, while ADNet-DA can assist with the diagnosis. This way, specialists can come up with a more informed decision and in less time.

List of Figures

List of Tables

Nomenclature

AD	Alzheimer's disease
ADNI	Alzheimer's disease neuroimaging initiative
AIBL	Australian imaging, biomarkers and lifestyle study of aging
aMCI	Amnestic mild cognitive impairment
ANN	Artificial neural network
AUC	Area under the receiver operating characteristic curve
Biomarker	Biological marker
CAD	Computer-aided diagnosis
CADDementia	Computer-aided diagnosis of dementia challenge
CN	Cognitively normal
CNN	Convolutional neural network
GPU	Graphics processing unit
ILSVRC	ImageNet large scale visual recognition challenge
MCI	Mild cognitive impairment
MCIc	Mild cognitive impairment converters
MCInc	Mild cognitive impairment non-converters
MMSE	Mini-mental state examination
MRI	Magnetic resonance imaging
PET	Positron-emission tomography
ROC	Receiver operating characteristic curve
sMRI	Structural magnetic resonance imaging
SVM	Support vector machine
TPF	True positive fraction
XAI	Explainable artificial intelligence

Contents

Chapter 1

Introduction

Dementia by Alzheimer's disease (AD) is a clinical syndrome characterized by multiple cognitive problems, including difficulties in memory, executive functions, language and visuospatial skills. Inexorably eroding the lifetime of memories and cognitive capacities that define us as individuals, AD robs patients of their unique identity, leading them to complete dependency for basic functions of daily life and ultimately to death.

Being the most common form of dementia, this disease kills more than breast cancer and prostate cancer combined, and it is the sixth leading cause of death in the United States [4]. Over a decade ago, nearly 25 million people lived with dementia worldwide, and 4.6 million new cases arise every year [23]. In Brazil specifically, there is a lack of general information regarding incidence and prevalence of AD dementia, but previous studies suggested that around 7% of the aged population is affected, and the incidence is estimated at 55 000 new cases each year [33]. By far, the single greatest risk for AD is aging, as there is almost a 15-fold increase in the prevalence of dementia between the ages of 60 and 85 years [21]. Markedly, the projected burden of the disease represents a looming healthcare crisis as the population of most industrialized countries continues to grow older.

The classic neuropathology of AD prominently includes intracellular aggregates of hyperphosphorylated tau protein that disrupt microtubule organization, and diffuse extracellular amyloid β-protein (Aβ) deposition [65]. These pathological events are accompanied by reactive microgliosis, oxidative stress and brain inflammation [31]. The loss of neurons and synapses result in a slow and progressive degeneration of brain structures, which can be seen as a dramatic cerebral shrinkage in structural magnetic resonance imaging (sMRI). Atrophy is especially severe in the hippocampus and temporal structures, which are areas that play a key role in the formation of new memories, and other cortical regions are also affected, such as parietal and frontal cortices.

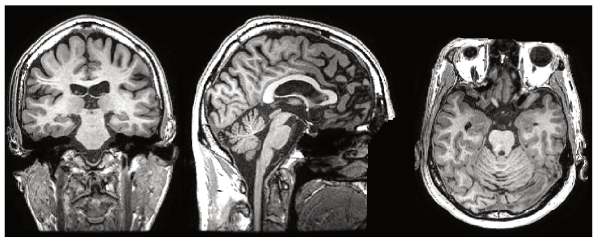
Although there is still not a cure, it is possible to treat both cognitive and behavioral symptoms of AD. The early diagnosis of the disease is paramount, not to say the most important hope, since it benefits patients' treatment and gives them more time to plan for the future. Moreover, clinical trials in AD tend to enroll subjects at earlier time-points, before neuronal degeneration has achieved a certain stage and treatment might be more effective. Even though detecting AD at early stages is difficult, a few biological markers (biomarkers) have been studied and defined. In short, a biomarker is an objectively

measurable characteristic, which supports the accuracy of diagnosis, being particularly important for AD, given that about 10% to 15% of AD cases are misdiagnosed [76]. Furthermore, biomarkers also serve as indirect measures of disease severity, and changes in biomarkers following intervention can be important indicators of alterations in the severity or stage of a disease. One of the main biomarkers for AD is the level of Aβ and tau proteins from the cerebrospinal fluid [53], which can be obtained through a considerably invasive procedure. As such, it is important to consider and evaluate different potential AD biomarkers, especially less invasive ones, for instance, using magnetic resonance imaging (MRI) or positron-emission tomography (PET).

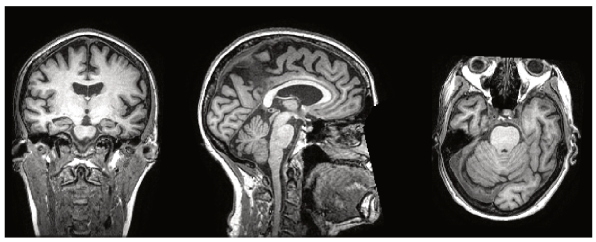
Neuroimaging is one of the most promising areas of research for early detection of AD, where a non-invasive technique is used to capture a digital image of the brain, from which specialists extract patterns and features of the disease. In 2011, anatomical MRI was included as evidence for neurodegeneration even in the prodromal stage of AD [70], *i.e.*, mild cognitive impairment (MCI), helping with the diagnosis by identifying specific patterns of atrophy that are characteristic of the disease. Amnestic MCI (aMCI), specifically, is a well-recognized risk factor for AD development, which is characterized by a cognitive decline in memory and possibly other domains, but without significant impairment in social and functional performance, *i.e.*, absence of dementia [2].

Computer-aided diagnosis (CAD) systems are approaches that aim at assisting doctors and specialists in interpretation of medical data to provide diagnoses for patients. This is an interdisciplinary field, joining forces from medicine and computer science, that gained traction in 1980s, mostly due to performance improvements at the time and acceptance by radiologists [17].

Some samples from the training set of the computer-aided diagnosis of dementia (CADDementia) challenge [9], based on structural MRI data, are illustrated in Figure 1.1, where,



(a) Cognitively normal (*train_ vumc_ 007*).



(b) Alzheimer's disease (*train_ emc_ 009*).

Figure 1.1: Anatomical planes of cognitively normal and Alzheimer's disease individuals from CADDementia training set. In (b), we can observe an atrophy mainly in temporal structures, such as the hippocampus, and posterior parts of the parietal cortex.

in (b), we can observe an atrophy mainly in temporal structures, such as the hippocampus, and posterior parts of the parietal cortex. In particular, the Alzheimer's Disease Neuroimaging Initiative (ADNI) [54] certainly is spearheading most of current efforts for data collection and research goals, and many researchers have used data provided by ADNI in order to evaluate CAD systems for AD, including techniques from machine learning and computer vision areas.

In short, machine learning is an area of computer science, which explores algorithms that can learn patterns from data, and then make further predictions. Considering a CAD system based on image data, there will typically be a medical specialist explaining the patterns and characteristics within the images to a computer vision specialist, which will then translate this knowledge into image processing techniques and machine learning models.

In general, these systems learn from data by adapting a set of internal parameters, which are internal configuration variables that define and control the behavior of the algorithm. Usually, models that have a larger number of parameters are capable of learning more complex patterns and arrangements from the provided data. On the other hand, having too many parameters will also imply in longer times for training (optimization) and execution, as well as in an increased chance of overfitting, where a model memorizes training data, instead of actually learning meaningful patterns. The ability to accurately predict previously unseen data is referred to as generalization.

Deep learning, an alias for artificial neural networks (ANNs), is a machine-learning technique inspired by how the brain works. Historically, traditional machine-learning approaches involved specialists for manually designing hand-crafted features for each task, which were then fed to a classifier or regressor. On the other hand, ANNs are capable of taking raw data as input, and automatically learn discriminative representations in a hierarchical way. This is an interesting approach to both corroborate previous findings by specialists, and to eventually assist in new discoveries. In particular, convolutional neural networks (CNNs) are a special kind of ANNs, which were inspired by how the visual system works, and, in this sense, have been increasingly used in computer vision tasks, achieving important results.

The primary unit in ANNs is a *neuron*, which basically receives a value as input, processes it by applying a mathematical function, and then outputs the result. A traditional ANN architecture is composed of a number of layers, where a layer is simply a stack of one or more neurons. These layers are organized hierarchically, where a given layer receives input from the previous layer, processes this information, and then passes these values to the next layer. In general, an ANN has one input layer, which is the provided data, one output layer, which is the prediction made by the network, and an arbitrary number of hidden layers between them, where more layers represent higher complexity. For further information on deep learning, we refer the reader to Goodfellow et al. [29].

1.1 Related Work

In terms of AD, prior research analyzed a number of classification tasks, including distinguishing between cognitively normal (CN), MCI, and AD. Additionally, considering conversions from diagnosed MCI to AD range only from 10.2% to 33.6% in a year [83], there is also the challenging task of differentiating between MCI patients who will convert to AD (MCI converters, MCIc) and MCI patients who will not convert to AD (MCI non-converters, MCInc). Even though we do not intend to review CAD systems for AD, we present here prior research somewhat extensively, in order to emphasize our challenges and contributions.

Works in this area have recurrently considered only a small number of subjects and images, often with curated data (*i.e.*, reviewed, prepared and organized by experts), such as ADNI's Standardized MRI Data Sets [88]. Additionally, with the lack of a standard evaluation protocol, each study employed its own criteria, with its own random data split. This not only hinders comparison between different methods, but it also usually overestimates their performance in a real-world scenario, where data will not be readily preprocessed, and will most likely come from different sources. In this sense, a few works reviewed multiple techniques for the Alzheimer's biomarker identification task, and, more recently, a few challenges with standard protocols and hidden test labels were launched, such as the CADDementia challenge [9]. As a side note, the ineffectiveness in comparing results was our main motivation to focus in describing techniques, rather than their performance metrics in this work.

Training deep learning systems usually requires large amounts of data, and most datasets are in the range of a few hundred samples. In order to overcome this limitation, studies that make use of deep-learning methods usually extract multiple small regions of the brain, thus generating thousands of input data. Differently from this approach, our method considers the whole brain when optimizing the networks, in an exploratory fashion, making it particularly interesting for automatically determining the most effective regions. This was also a reason for our emphasis on dataset sizes and their respective image dimensions in the description of related works.

1.1.1 Alzheimer's Disease

Many CAD systems for AD using sMRI have been proposed in prior art, and they usually achieve promising results. However, most of these works evaluate systems on their own non-disclosed data, or use their own split of available data, which makes comparisons between different methods very difficult.

Falahati et al. [22] reviewed several AD classification and MCI conversion prediction studies, focusing on sMRI. The main idea was to train a system for classification between CN and AD, and then evaluate them on CN *vs.* AD, and MCIc *vs.* MCInc. The authors indicated that performances for methods using small sample sizes are usually superior than the ones using larger datasets or external validation sets, which is probably due to overfitting or very small homogeneous samples. Besides data, different approaches for feature extraction, feature selection, classification, and, more importantly, validation

hamper comparisons across different works.

Cuingnet et al. [14] evaluated ten methods, namely, five based on voxel, three based on cortical thickness, and two based on hippocampus. Using 509 subjects from the ADNI database, they compared these methods on three different scenarios, which were classification between CN and AD, CN and MCIc, and MCIc and MCInc. They found that methods based on voxel or cortical thickness achieved high accuracies on CN *vs.* AD. However, CN *vs.* MCIc had considerably inferior results, while MCIc *vs.* MCInc did not perform statistically different from chance. Additionally, the authors noted that most classification errors were oldest controls and youngest patients, which was partially explained by the brain atrophy associated with normal aging.

Similarly, Sabuncu and Konukoglu [62] collected 810 samples from ADNI and 415 from OASIS (cross-sectional study [49]), and analyzed a combination of four feature sets with three algorithms, to distinguish CN *vs.* AD, and CN *vs.* MCI. Features included volumes of anatomical structures, average thickness within specific cortical parcellations, and cortical thickness, while algorithms included support vector machine (SVM), neighborhood approximation forest, and relevance voxel machine. Results indicated that accuracy and relevance of image-derived measurements are more important than the prediction algorithm for the overall performance, but data quality and sample size play even bigger roles. Moreover, even though cross-validation is an interesting technique to measure accuracy, it is generally optimistic in terms of generalization, especially when the trained system is applied to an independent dataset, due to data acquisition protocol, composition of populations, and application of diagnostic criteria or clinical tests.

Advancing research for comparison of different methods, the Alzheimer's Disease Big Data DREAM Challenge #1 [3], in subchallenge 3, proposed a standard evaluation protocol, including defined data for training (628 entries from ADNI), and testing (182 entries from AddNeuroMed [46]). Available data contained sMRI, and other variables, such as years of education, and genotypes. The objective was to predict Mini-Mental State Examination (MMSE) scores and diagnostic classes, and the winning method achieved a correct diagnosis percentage of 60.2%, considering CN, MCI, and AD classes. This method combined both clinical and image features, with volume of hippocampus being the most important one.

Equivalently, Bron et al. [9] launched the CADDementia challenge, a standard comparison between different methods for classification within CN, MCI, and AD classes, with the same defined data and evaluation protocol. Teams had only 30 sMRI scans available for training, and 354 scans with hidden diagnoses (including group priors) for testing. The main idea was to leverage existing public data to optimize the classification system, which could then be optimized again or just fine-tuned on a different small dataset, leaving a larger number of unseen images to evaluate performance. This is a very interesting protocol, as it more closely relates to a real-world scenario, where an algorithm would need to be adapted for a practical clinical setting. In this challenge, most approaches used volume, thickness, intensity, and shape features from specific regions of interest, along with regression, support vector machine, and random forest for classification.

The winning method [69] achieved an accuracy of 63.0%, using a number of individual MRI imaging biomarkers, with hippocampal volume, ventricular volume, hippocampal

texture, and parietal lobe thickness being the most important characteristics. Data from both ADNI and AIBL [19] were used for training the system, which used a linear discriminant analysis classifier. It is important to note that hippocampal shape scores used transductive inference, thus needing the CADDementia test data to be calculated. Even though this is a valid approach, it deviates from the original proposal of a practical clinical setting. Additionally, their pipeline failed to process three scans from the CADDementia test set, requiring manual intervention. The analysis of each subject took 19 hours of computation time.

The second best team [82] employed a domain-adaptation approach. Their idea was to weight samples from a source dataset according to a target dataset distribution, and five different weighting techniques were evaluated on 751 subjects from ADNI, and 215 from AIBL. However, the system submitted to the challenge corresponds to the fourth best accuracy in their experiments, indicating that there is still room for improvement. More specifically for the challenge, optimization was done on the union of ADNI and CADDementia training sets, with equal weights for each sample. Classification was done by a generalized linear model, using volume, cortical thickness, and shape features. The analysis of each subject took 17.4 hours of computation time.

1.1.2 Deep Learning

One of the first successful real-world applications of convolutional neural networks (CNNs) was a system to read checks [41, 42]. More recently, a variety of computer vision approaches have been evaluated at the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) [61], which is held yearly and involves different tasks, such as general object recognition and localization. Since the first entry using deep CNNs [40], this technique has dominated the top results in this contest. Thus, it is very interesting to have a better understanding on how it would behave in the AD task.

Suk and Shen [72] presented one of the first works using ANN for the Alzheimer's biomarker identification task, extracting some low-level features, which were then fed to a stacked auto-encoder (SAE) with three hidden layers. To optimize the parameters, a greedy layer-wise unsupervised learning approach was used. Then, learned layers were stacked together with a new output layer, and the whole network was fine-tuned with labeled data. Next, original low-level features were concatenated with SAE latent feature representation, *i.e.*, outputs from the last hidden layer, generating an augmented feature vector. Finally, a multi-task and multi-kernel support vector machine was trained to predict class labels, MMSE, and AD assessment scale-cognitive subscale (ADAS-Cog) scores.

Li et al. [44] presented one of the first methods using 3D CNNs for the Alzheimer's biomarker identification task, although indirectly. The main idea was to adopt both MRI and PET scans to perform the diagnosis; however, a number of subjects only had the MRI modality. Given an MRI, the proposed method used a CNN to predict the corresponding PET scan. Images were first resized to $64 \times 64 \times 64$, and 50 000 random patches of size $15 \times 15 \times 15$ were extracted from each image. Finally, with this input patch, a CNN with two hidden convolutional layers and 37 761 parameters outputs a $3 \times 3 \times 3$ PET patch.

Suk et al. [73] introduced an approach using deep Boltzmann machine (DBM), which comprises restricted Boltzmann machine (RBM) as building blocks. Similarly to previous methods, both MRI and PET scans were used, being resized to $64 \times 64 \times 64$ voxels, with a final voxel size of $4 \times 4 \times 4\ mm^3$. Discriminative patches of size $11 \times 11 \times 11$ were extracted, and then fed to a Gaussian DBM. This architecture comprised two networks of two hidden layers each for independent MRI and PET patches processing, which were concatenated and then followed by three hidden layers for a multimodal DBM (MM-DBM). Features extracted by this DBM were used in an image-level hierarchical SVM classifier.

Payan and Montana [56] randomly extracted $5 \times 5 \times 5$ patches from $68 \times 95 \times 79$ images, which were flattened into an 125 dimensional input array, and fed to an overcomplete sparse autoencoder with 150 units. Each unit was rearranged into a $5 \times 5 \times 5$ filter, which was convolutionally applied to the complete original image. Next, outputs went through max-pooling, and a fully-connected layer, followed by the output layer. Therefore, even though the proposed is conceptually similar to a 3D CNN, in practice, it was a reinterpretation of the autoencoder units, followed by a traditional neural network.

Hosseini-Asl et al. [34] employed a stack of unsupervised 3D convolutional autoencoder (3D-CAE), and used the whole brain. All 30 images from the CADDementia training set were used as source domain, being preprocessed and normalized to $200 \times 150 \times 150$. These images were used to greedily train three stacked 3D-CAEs in a layer-wise fashion. The output of the last layer was flattened and used as features in a traditional fully-connected layer with two hidden layers, which was trained on a target domain. Even though this method could be directly applied to the CADDementia challenge, it was only cross-validated on ADNI.

Sarraf et al. [64] decomposed 4D resting-state functional MRI (rs-fMRI) and 3D structural MRI into 2D images, with $2 \times 2 \times 2\ mm^3$ resolution in standard space. Then, two 2D CNN architectures were evaluated for the CN *vs.* AD task, namely, LeNet-5 [42] and GoogLeNet [74]. Despite being a reasonable approach, it is unclear which axis shall provide best results, and, even if the decomposition is performed along all axes, this method will still not be able to find discriminative 3D patterns within the data.

Korolev et al. [39] is one of the works that most closely relates to ours. They designed 3D CNNs based on smaller versions of VGG [68] and ResNet [32] architectures, which were trained on whole-brain images of size $110 \times 110 \times 110$. The 231 images used were a subset of previously processed MRIs from ADNI, and included CN, early MCI (eMCI), late MCI (lMCI), and AD subjects. However, they only considered binary classification tasks, which were evaluated using cross-validation on ADNI, hindering better comparisons with our method. On the other hand, we proposed a multiclass approach that included very deep CNN architectures, with large input images, and which was evaluated on the CADDementia challenge, making our results more reliable.

Dolph et al. [18] were the first group to successfully propose a deep-learning approach to the CADDementia challenge, with a technique similar to Suk and Shen [72]. Basically, they extracted sub-cortical features, such as cortical thickness, surface area, and volumetric measurements, along with texture features from gray-level co-occurrence matrix in fractal dimension. These values were used to greedily layer-wise train a stacked auto-encoder with three hidden layers, achieving competitive results in the challenge.

Finally, Brosch and Tam [10] proposed and evaluated a fast training method for CNNs using fast Fourier transforms (FFTs). The authors stated that this approach made it practical to train a 3D CNN with two hidden layers, considering input images as large as $128 \times 128 \times 128$ voxels. This work illustrates the difficulty in training very deep 3D CNNs, especially with high-resolution images.

The main drawbacks of the aforementioned deep-learning approaches are the small depth of proposed networks, and the small dimensions of input images. While a network with only a few hidden layers is not able to identify complex patterns within the data, having a small resolution image only makes this task even more difficult. These characteristics are present across different methods largely due to hardware constraints. In order to reduce computational costs, we custom-tailor specific network architectures herein, and adapt traditional optimization approaches, making such training practical.

1.2 Contributions

In our research, one of the main goals was bringing to bear what is most advanced in deep learning research (*e.g.*, CNN) to solve the difficult problem of identifying AD in MRI, considering three different groups, namely, CN, MCI, and AD. From LeNet-5 [42] to Residual Nets [32], we explored a number of state-of-the-art CNN architectures, which are better described in Chapter 2, along with details of our methodology. We tailored convolutional networks with data primarily provided by ADNI [54], and evaluated them on the CADDementia challenge [9], thus resulting in a scenario very close to real-world conditions, in which a CAD system is used on a dataset differently from the one used for training. Our experimental setup is explained in Chapter 3.

The main challenges and contributions of our research include devising a deep-learning solution that is both completely automatic and comparatively fast, while also presenting competitive results, without using any domain-specific knowledge. In the end, our system does not need any manual intervention, and runs 80× faster than the state of the art, on average. Our best model outperformed most of the CADDementia submissions, all of them using prior knowledge from the disease, such as specific regions of interest. The main reason for not using any information from the disease in our system is to empower it to automatically learn and extract relevant patterns from important regions of the brain, which can be used to support current diagnosis standards, and may even assist in new discoveries for different or new diseases.

Additionally, our generated ADNet and ADNet-DA models will be publicly available along with this work, including all supporting code to both use them or to train similar models on new data, which, to the best of our knowledge, has not been done before in this area. In Chapter 4, we provide more details of our results and the corresponding discussion, including a number of techniques related to explainable artificial intelligence (XAI), in order to to visualize and have a better understanding of what the CNN has learned and how it processes inputs, aiming at biological significance. Finally, we conclude and present possible further explorations in Chapter 5, with due acknowledgements at the end.

Chapter 2

Methodology

To the best of our knowledge, we are the first group to propose an end-to-end deep 3D CNN for the multiclass AD biomarker identification task. In terms of deep learning, a number of existing methods rely on a greedy layer-wise learning of stacked (regular or convolutional) autoencoders, and we believe this is mainly due to the high complexity in optimizing a very deep 3D CNN with limited data and hardware, even considering current standards.

In this chapter, we provide details of our pipeline, including image preprocessing, CNN architectures, and optimization techniques. In particular, it is important to highlight the reason we did not use images in their original space, and the need for a fixed brain size. Even though convolutional layers can operate on data with variable dimensions, optimizing a deep-learning system using images without any standard requires it to learn discriminative patterns invariant to a number of transformations, such as translation, scaling, and rotation. This would demand larger models, with increasing training times, and an even larger number of samples, with all expected variations. By registering our images to a standard template, we can expect similar structures to be roughly in the same spatial location, hence we can handle the entire image at once, and automatically determine the most important regions of interest.

2.1 Brain Extraction and Normalization

We used the Advanced Normalization Tools (ANTs) [5] version 2.1.0 to extract and normalize brain images. Since this is not the focus of our research, our pipeline was based on previously defined scripts[1] [6, 80], and we made use of the provided default parameters, including transformation types, sequence, and metrics. We refer the reader to the code repository for more details on these parameters. Essentially, our brain extraction and normalization pipeline comprises the following steps:

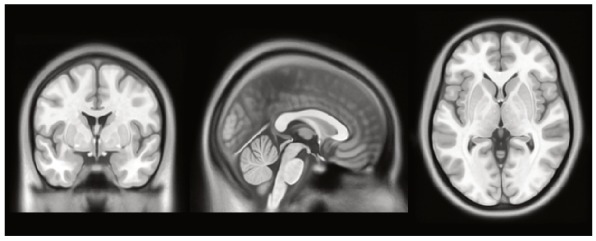
1. Winsorize image intensities on 1% and 99.9% quantiles
2. Bias field correction using N4 [79], a variant of the popular nonparametric nonuniform intensity normalization (N3) algorithm

[1] Specifically, scripts *antsBrainExtraction.sh* and *antsRegistrationSyNQuick.sh*

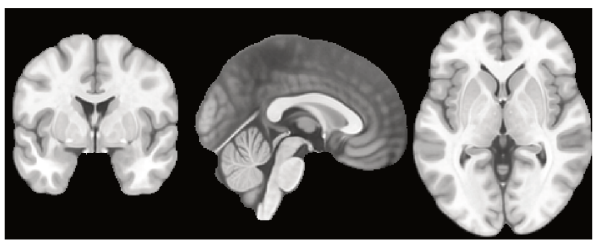
3. Winsorize image intensities on 0.5% and 99.5% quantiles
4. Translation alignment using center of mass
5. Rigid transform (rotation and translation)
6. Affine transform (shearing and scaling)
7. Deformable symmetric normalization (SyN) transform (non-linear)
8. Application of brain mask from atlas
9. Normalization of intensities to $[0, 1]$

Winsorizing is a statistical transformation that minimizes effect of outliers. This is achieved by replacing extreme values with the corresponding defined percentiles [16]. As we used registered brains in our research, we opted for a less rigid and less linear atlas, allowing some degree of variation during the registration process, and which also had a high-spatial resolution, so finer details would not be lost in the process. As such, the Montreal Neurological Institute (MNI) 152 International Consortium for Brain Mapping (ICBM) 2009c Nonlinear Asymmetric $1 \times 1 \times 1\ mm^3$ [12, 24, 25] atlas was chosen, and is illustrated in Figure 2.1. This is an unbiased standard MRI template brain volume from a normal population, meaning that even better registration results could be achieved by using an atlas specific for Alzheimer's population, *i.e.*, including subjects diagnosed with Alzheimer's, and elder controls. We display some intermediate results from our pipeline in Figure 2.2.

After the brain extraction and normalization process, the output image has the same dimensions as the atlas, *i.e.*, $193 \times 229 \times 193$. From all these 8 530 021 voxels, only 1 886 574 (22%) of them are not zero. Since the brain is enclosed in a smaller region inside the

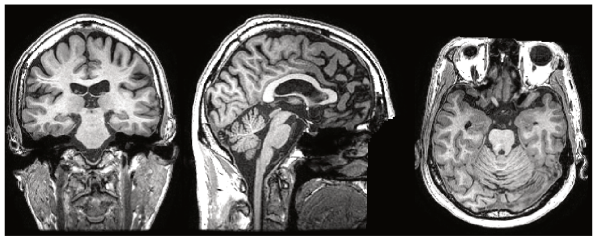


(a) Original template.

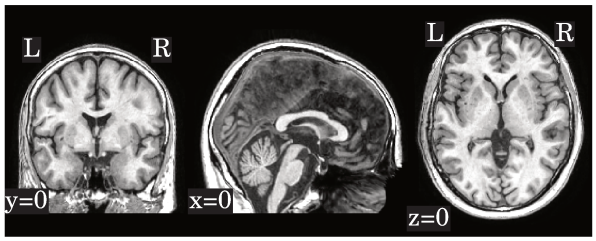


(b) Template after brain mask application.

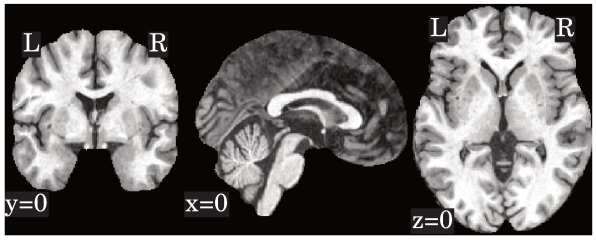
Figure 2.1: Anatomical planes of the atlas used for registration (MNI 152 ICBM 2009c Nonlinear Asymmetric).



(a) Original image (*train_ vumc_ 007*).



(b) Image after registration.



(c) Image after brain mask application.

Figure 2.2: Anatomical planes of a cognitively normal individual: (a) original image, (b) after registration, and (c) after brain mask application. Coordinates are in MNI space.

image, we removed the border dimensions that contained no information, resulting in a final image of $145 \times 182 \times 155$. This new space represents 48% of the original volume, reducing sparsity from 78% to 54%. Finally, we normalized the data to zero mean and unit variance, using the training set to compute these parameters, and then applying them to other sets. Given that the used datasets did not fit in main memory, we adopted a single-pass online mean and estimated variance algorithm [87].

2.2 Convolutional Neural Network

CNNs have been used for many different computer vision tasks, and often achieving state-of-the-art results. In face of this, we decided to explore this technique for Alzheimer's biomarker identification. To this end, we considered the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) [61] as reference. This contest is held yearly, for some computer vision tasks, such as general object recognition. Analyzing the techniques that achieved the best results in the last few years, we selected some of these distinct models, adapting them to our task. We additionally selected a smaller and older architecture to contrast with current larger and deeper models.

We describe here the general architecture and adopted modifications. We refer the

reader to the specific papers for a more detailed description. The most natural adaptation was to convert all 2D operations, such as convolution or pooling, to 3D ones, since these networks were originally designed for 2D color images, while we are dealing with 3D grayscale MRIs.

Given such adaptations, we were unable to directly employ a transfer learning approach [66] with the original networks. In short, the idea is to pre-train a network on a different problem with a different set of images, and then take advantage of this network on a new task, either as a feature extractor, or as a good starting point for another optimization, also known as fine-tuning, instead of using random weights. In principle, we could make an odd adaptation of the first convolutional layer from a traditional 2D CNN trained on color images, since it has a 3D shape; however, these filters will most likely be optimized to find 2D color patterns, instead of 3D patterns. Additionally, we could not find any publicly available trained model for the AD task. As such, with this work, we are releasing one of the first models ready to be used, encouraging open science and reproducible research, while also setting a starting point for researchers working with 3D MRIs.

Architecture	Layers	Parameters (in millions)
LeNet-5	7	0.3
VGG 2048	11	89.8
VGG 512	11	26.8
GoogLeNet	22	14.6
ResNet A	18	33.0
ResNet B	18	33.2

Table 2.1: Description of evaluated CNN architectures.

A common attribute to all considered architectures is that spatial dimension is reduced as information flows to deeper layers. This is usually achieved with max-pooling layers, or with larger strides in convolutional layers. In order to accommodate our different data shapes that were not necessarily divisible by two, we adopted an ad-hoc approach by zero-padding each layer as needed, so no information was lost. We also added batch normalization [35] to every convolutional and fully-connected layers. All activation functions were rectified linear units (ReLU) [55], defined as $f(x) = \max(0, x)$, except for the classification output, which was a softmax function. Finally, the exact number of layers or depth varied according to the adopted network standard. In Table 2.1, we took into account the original approach for each network, highlighting eventual differences along with their descriptions.

We started with a small network, based on the LeNet-5 [41, 42] architecture. Considering that this network was significantly older than the others, it went through the most modifications. Basically, it was composed of the following layers: convolution, subsampling, convolution, subsampling, fully connected (originally implemented as convolutional), fully connected, and output. As subsampling layers had learnable parameters, we converted them to convolutions, with filter (kernel) size and stride equal to $2 \times 2 \times 2$, thus

keeping the subsampling behavior. The main difference was in the connection scheme between S2 (second layer, subsampling) and C3 (third layer, convolutional). In the original work, these connections had a very particular arrangement, which we converted to a dropout layer with probability of 40%. Also, similarly to the original architecture, if we adapted the last convolutional layer (C5) to match the output size of the previous layer, it would have 120 feature maps with a kernel size of $34 \times 43 \times 36$, and this would be extremely huge in number of parameters. So we adapted C5 kernels to $5 \times 5 \times 5$, and added a global average pooling layer right after it, similarly to GoogLeNet [74] and Residual Net [32]. Naturally, the last layer contained three units (one for each class), with a softmax function activation. The resulting model contained five convolutional layers, followed by two fully-connected layers, and the output.

The Visual Geometry Group (VGG) analyzed very deep CNN architectures, achieving second place in the classification task at ILSVRC-2014 [68]. Basically, they designed very uniform networks, with a number of convolutional and max-pooling layers, followed by two fully-connected layers, and the output. Architectures ranged from 11 (configuration A) to 19 (configuration E) weight layers, *i.e.*, considering only convolutional and fully-connected layers. Due to its uniformity, mostly with filters of size 3×3, the VGG architecture is considerably large. In the first layers, all the original input dimensions are kept, making them consume plenty of GPU memory, while in the last layers, the dense connections generate several parameters. Since our input data were already quite large when compared to traditional 2D images, we adapted the VGG network configuration A by halving all numbers of filters in convolutional layers, and all numbers of units in fully-connected layers, while keeping filters sizes of $3 \times 3 \times 3$ and dropout rate at 50%. Even after reducing the network size, the first fully-connected layer of our adapted VGG-A, with 2048 units, accounted for 78 643 200 (88%) parameters. Similarly to our adaptation to LeNet-5, in later architectures, this was solved by adding a global average pooling layer after the last convolutional layer, and before the first fully-connected layer, drastically reducing the number of parameters, even with deeper architectures; however, we did not

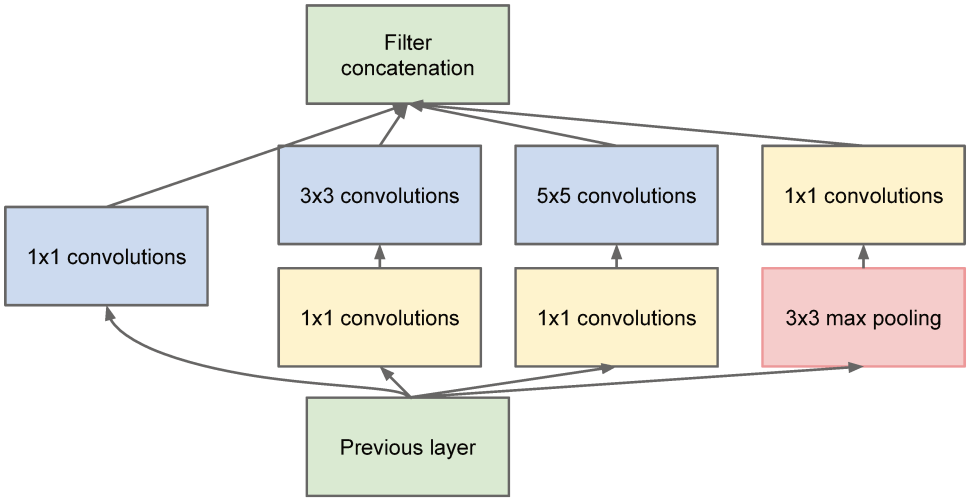


Figure 2.3: Inception module from GoogLeNet CNN architecture. Image from Szegedy et al. [74].

apply this technique to the VGG architecture. For comparison, in Table 2.1 we also include our VGG-A with 512 units in the fully-connected layers.

While VGG achieved second place in ILSVRC-2014, GoogLeNet secured the first place in the classification task [74], proposing a deep convolutional neural network architecture named Inception. The basic idea was to increase both depth and width, while keeping computational requirements constrained, which led to a deeper model, with fewer parameters, and better performance. We adapted directly from their GoogLeNet architecture, *i.e.*, only discarding the local response normalization [40] layer and the auxiliary networks. We also adjusted the last average pooling layer, following the output shape of the previous layer, and kept dropout rate at 40%. In this architecture, the number of layers actually came from depth, where single convolutional or fully-connected layers counted as one, while inception modules counted as two. However, each inception module internally had six individual convolutional layers, which is depicted in Figure 2.3.

In ILSVRC-2015, Residual Network [32] won first place for classification, localization, and detection tasks. Continuing analyses from VGG, the authors wanted to understand whether learning better networks meant simply to stack more layers. With this study, they found the degradation problem, where traditional models similar to VGG stopped improving performance after a certain number of layers, and even started getting worse afterwards. To overcome such problem, they proposed the residual function, which is the basic building block of a Residual Network (ResNet), presented in Figure 2.4. The idea was to create a shortcut connection between the input of a layer and the output of the following one, in a way that these layers could simply learn nothing, and the input would still be preserved, thus making it feasible to train very deep layers, even with more than a hundred layers, and diminishing the degradation problem. However, due to hardware constraints, we considered only smaller Residual Networks. We adapted ResNet directly from the non-bottleneck 18-layer architecture, in which shortcuts with increasing dimensions were either (A) identity shortcuts, *i.e.*, padding with zero, or (B) projection shortcuts, *i.e.*, convolutions with $1 \times 1 \times 1$ filter (kernel) size. Similarly to VGG, the number of layers came from convolutional and fully-connected layers, with projection convolutions not being considered in the layer count.

In summary, we adopted four main CNN architecture designs, namely, LeNet-5, VGG, GoogLeNet, and ResNet. LeNet-5 is considerably older and smaller, so it shall have a lower probability of overfitting. The VGG network is known for its uniformity, which makes it relatively simple to adapt, inspect and use for a number of different tasks; however, this

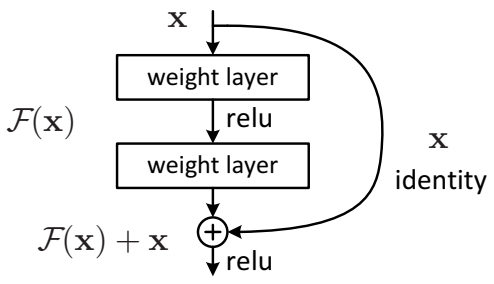


Figure 2.4: Building block of Residual Network (ResNet) CNN architecture. Image from He et al. [32].

characteristic also makes it large in number of parameters and in hardware requirements. These drawbacks were overcome in both GoogLeNet and ResNet architectures, which also adopted very specific building blocks, making it possible to extract more complex patterns from data, while also increasing the number of layers and reducing the number of parameters. The idea was to explore different architectures and understand how they would behave in the AD task.

To avoid overfitting, we adopted regularization with L1 and L2 norms. In L1, this effect is achieved by minimizing the absolute values of the weights, while in L2, this is done with their squared values. In principle, L2 norm tends to produce diffuse and small numbers, while L1 tends to produce sparse numbers. This property makes L1 particularly interesting to handle noisy data, acting as a feature selection algorithm, which could help us better visualize and explain what the CNN has actually learned. However, in general, L2 can be expected to provide superior results over L1.

All network architectures, and their optimization were implemented using upstream (*i.e.*, latest version from the code repository) Lasagne [15], which is a deep learning framework based on Theano [1]. At the time this research was performed, we used a development version of Lasagne 0.2, and a development version of Theano 0.9.0, with Python 2.7.6, CUDA 7.5, and CuDNN 5. We additionally used scikit-learn 0.18.1 [57] and numpy 1.11.3 [81].

2.3 Domain Adaptation

In addition to the brain processing and CNN pipelines, we also considered a domain adaptation approach. In our method, we trained a system using one dataset, and evaluated it on a different dataset (*i.e.*, CADDementia). Even though they are related, such difference means that the source data distribution could be different from the target data distribution. Thus, it should be possible to improve results further by adapting the previously trained system to the new dataset, even if using a small number of samples from this target domain. It comes as no surprise that the best methods in the CADDementia challenge, at some point, did use available data from both its training and test sets in their optimization pipeline.

In our domain adaptation approach, we started with our previously optimized CNN. Then, we used this CNN to extract features from the complete target dataset (*i.e.*, CADDementia), using one of the last layers in the network as output. After, we normalized these features to zero mean and unit variance, using only the target training set to compute the parameters, which more closely relates to a real-world scenario. With the normalized data, we optimized a one-versus-rest logistic regression [52] on the complete target training set. In order to find the best parameters for this classifier, we used grid search with leave-one-out cross-validation. Then, we finally had a system that was enhanced for the target domain, making it possible to output improved classification probabilities for each sample in the target domain. This pipeline is similar to a transfer-learning approach [66].

Chapter 3

Experimental Setup

Given that training a CNN from scratch usually requires massive amounts of data, we gathered as many different imaging sources as possible. To the best of our knowledge, we collected the largest AD sMRI dataset ever, comprising 23 165 images, which is orders of magnitude larger than commonly analyzed sets, as we discuss next. We additionally describe our optimization approach, including associated parameters.

3.1 Data

In our data collection process, we considered the following datasets:

- ADNI [54], including ADNI1, ADNIGO, ADNI2, and ADNIDOD [8] studies
- AIBL [19]
- CADDementia [9]
- MIRIAD [48]
- OASIS, including cross-sectional [49], and longitudinal [50] studies
- AddNeuroMed [46]

ADNI1 originally included three participant groups: CN, MCI and AD. Starting in ADNIGO, the MCI stage was split into two: early MCI (eMCI) and late MCI (lMCI). Later, in ADNI2, a significant memory concern (SMC) group was added. We refer the reader to Beckett et al. [8] for more details.

Similarly to ADNI1, both AIBL and CADDementia sets were composed of CN, MCI, and AD stages, whereas both MIRIAD and OASIS sets contained only CN and AD. Unfortunately, ADNIDOD did not have Alzheimer's diagnoses information, thus we did not include the corresponding images in our analyses. We also did not use AddNeuroMed due to agreement restrictions.

Since one of our main goals was achieving a good result in the CADDementia challenge, we adopted only equivalent diagnoses. As such, eMCI and lMCI stages were grouped along with MCI, and SMC was not considered. From these datasets, we downloaded all available

Dataset	Subjs.	Group	Images	Age (years)				Female (%)	1.5 T (%)
				Med	Avg ± Std	Min	Max		
Dset. 1	845	All	9 149	76.6	76.3 ± 6.9	54.6	93.0	42.2	82.2
		CN	2 701	76.7	77.2 ± 5.1	60.0	92.8	50.2	80.5
		MCI	4 845	76.5	76.0 ± 7.4	54.6	90.9	35.3	83.0
		AD	1 603	76.5	76.1 ± 7.9	55.2	93.0	49.5	82.5
Dset. 1 Train.	591	All	6 314	76.5	76.2 ± 6.9	54.6	93.0	43.4	82.6
		CN	1 809	77.2	77.3 ± 4.9	60.0	90.8	49.5	81.3
		MCI	3 399	76.1	75.7 ± 7.3	54.6	90.9	36.3	83.0
		AD	1 106	75.9	76.1 ± 7.9	55.2	93.0	55.3	83.5
Dset. 1 Val.	84	All	951	76.4	75.8 ± 6.8	56.2	89.2	40.5	82.8
		CN	301	75.7	76.5 ± 4.8	65.2	88.6	58.5	79.7
		MCI	501	78.2	76.7 ± 6.7	56.2	89.2	28.5	83.8
		AD	149	72.0	71.2 ± 8.6	56.5	85.0	44.3	85.2
Dset. 1 Test	170	All	1 884	77.2	77.0 ± 6.9	56.7	92.8	38.7	80.4
		CN	591	76.2	77.2 ± 5.6	63.3	92.8	47.9	78.5
		MCI	945	77.7	76.5 ± 7.8	56.7	90.9	35.1	82.4
		AD	348	79.7	78.0 ± 6.3	63.1	87.7	33.0	78.2
Dset. 2	1 503	All	15 885	75.8	75.4 ± 7.3	54.6	95.8	44.0	53.3
		CN	4 646	76.8	76.9 ± 5.8	56.3	95.8	50.0	56.5
		MCI	8 940	75.0	74.6 ± 7.7	54.6	93.5	40.0	50.5
		AD	2 299	76.4	75.8 ± 7.8	55.2	93.0	47.5	57.5
Dset. 3	1 715	All	18 303	75.8	75.5 ± 7.4	54.6	95.8	43.5	48.2
		CN	5 361	76.7	76.9 ± 6.0	56.3	95.8	50.0	52.5
		MCI	10 306	75.0	74.6 ± 7.7	54.6	93.6	39.5	45.5
		AD	2 636	76.2	75.8 ± 7.9	55.2	93.0	45.9	50.2
Dset. 4	2 984	All	23 165	75.0	73.5 ± 11.7	18.0	98.0	46.5	55.5
		CN	8 462	75.0	71.3 ± 16.1	18.0	97.0	53.9	62.8
		MCI	10 460	75.0	74.7 ± 7.7	54.6	96.0	39.6	45.1
		AD	4 243	75.4	75.3 ± 7.9	55.0	98.0	48.4	66.3
CADD. Train.	30	All	30	65.0	65.2 ± 6.9	54.0	80.0	43.3	0.0
		CN	12	62.0	62.3 ± 6.1	55.0	79.0	25.0	0.0
		MCI	9	68.0	68.0 ± 8.2	54.0	80.0	44.4	0.0
		AD	9	67.0	66.1 ± 5.0	57.0	75.0	66.7	0.0
CADD. Test	354	All	354	65.0	65.1 ± 7.8	46.0	88.0	39.8	0.0

Table 3.1: Datasets summaries: number of subjects, number of images, descriptive age statistics, image-wise percentage of females (*vs.* males), and image-wise percentage of 1.5 T field strength (*vs.* 3.0 T).

raw T_1-weighted sMRI scans associated with Alzheimer's, *i.e.*, we did not download any pre- or post-processed image.

To isolate possible confounding factors, we made a distinction between MP-RAGE and IR-SPGR/IR-FSPGR sequences, and aggregated different data sources and sequence techniques in steps. While all ADNI sets had both MP-RAGE and IR-SPGR/IR-FSPGR, AIBL and OASIS had only MP-RAGE, and MIRIAD had only IR-FSPGR. For more details on MP-RAGE and IR-SPGR/IR-FSPGR, we refer the reader to Jack et al. [36], Lin et al. [45]. The resulting datasets, summarized in Table 3.1, are:

- Dataset 1: ADNI1 (MP-RAGE only)
- Dataset 2: ADNI1, ADNIGO, and ADNI2 (MP-RAGE only)
- Dataset 3: ADNI1, ADNIGO, and ADNI2 (all)
- Dataset 4: ADNI1, ADNIGO, ADNI2, AIBL, MIRIAD, and OASIS (all)

For each dataset, we created training, validation, and test splits, which was done as follows. In Dataset 1, we randomly split the corresponding subjects, trying to keep the original age, sex, and diagnostic stratification across each set, with 70% of subjects for training, 10% for validation, and 20% for testing. In each subsequent dataset, we first assigned images from previous subjects to the respective set, then we proceeded with the stratified random split considering only new subjects. We also present split summaries for Dataset 1 in Table 3.1. It is important to note that we removed a few questionable images, for instance, with more than one image for the same identifier, mismatch between identifier and folder name, as well as corrupted images or without diagnosis information.

3.1.1 ADNI

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (`https://adni.loni.usc.edu`). The ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organizations, as a $60 million, 5-year public-private partnership. The primary goal of ADNI has been to test whether serial MRI, PET, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD. Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials.

The Principal Investigator of this initiative is Michael W. Weiner, MD, VA Medical Center and University of California - San Francisco. ADNI is the result of efforts of many co-investigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the U.S. and Canada. The initial goal of ADNI was to recruit 800 subjects but ADNI has been followed by ADNI-GO and ADNI-2. To date these three protocols have recruited over 1500 adults, ages 55

to 90, to participate in the research, consisting of cognitively normal older individuals, people with early or late MCI, and people with early AD. The follow up duration of each group is specified in the protocols for ADNI-1, ADNI-2 and ADNI-GO. Subjects originally recruited for ADNI-1 and ADNI-GO had the option to be followed in ADNI-2. For up-to-date information, see `http://www.adni-info.org`.

3.1.2 AIBL

Data was collected by the AIBL study group. AIBL study methodology has been reported previously [19].

3.2 Metrics and Optimization

To compare different methods, the CADDementia challenge adopted a number of metrics. The main evaluation measure was the traditional classification accuracy, which is basically the number of correctly classified samples divided by the number of all samples. Even though this performance value does not take into account class priors, the challenge organization considered that class sizes were not very different, regarding this metric as a better approach for the overall classification accuracy. Additionally, the receiver operating characteristic (ROC) curve and the respective area under the curve (AUC) were considered, as they provide metrics that are independent of the threshold chosen for classification. Also, since AUC does not traditionally depend on class sizes, the challenge adopted an AUC measure that does not rely on class priors. Finally, the true positive fraction (TPF) for each class was calculated, which is the number of correctly classified samples of a given class divided by the number of all samples from that class. According to the authors, TPFs for diseases (AD and MCI) can be interpreted as the two-class sensitivity, while TPF for CN corresponds to the two-class specificity. For more details on the challenge's metrics, we refer the reader to Bron et al. [9].

As we optimized and trained our networks, we compared them and selected the best ones using the average of TPFs, since it more closely relates to the accuracy, which was the main metric, and it does not depend on class priors. To perform the training process, we used Adam optimizer [38], with default parameters, *i.e.*, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$. With a small sample of images, we empirically decided to begin with a learning rate of $\alpha = 10^{-4}$, and settled to a batch size of three (for VGG architectures) or nine (for all the others), mainly due to GPU memory limitations, even though we only used GPUs with 12 GB of dedicated memory. Finally, we adopted Glorot uniform initialization [28] with scaling factor of $\sqrt{2}$, *i.e.*,

$$a = \sqrt{2} \cdot \sqrt{\frac{6}{fan_{in} + fan_{out}}} \tag{3.1}$$

$$W \sim U[-a, a] \tag{3.2}$$

where fan_{in} is the number of input units of the weight tensor and fan_{out} is the number of output units of the weight tensor.

Chapter 4

Results and Discussions

Accountability has become an important aspect in machine learning lately [30], and it is even more critical in the medicine area. Automatic systems that present impressive results no longer suffice, being necessary to also explain how and why they achieved such performances, and determine when they are applicable, thus actually helping specialists in their tasks.

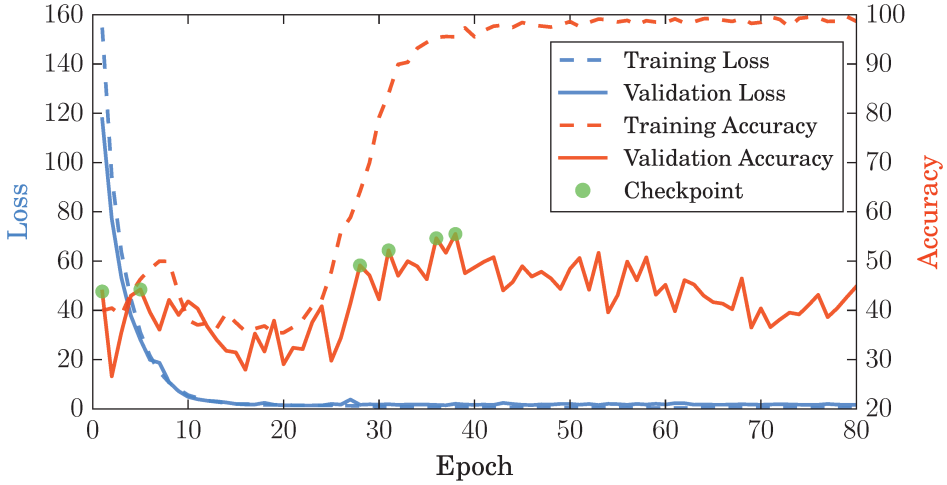
Given that, we now dive into more details of our study. We better describe our optimization process, specifying steps taken to handle overfitting problems. Then, we report performance results, including previously described metrics, along with efficiency measurements. Finally, we discuss our best CNN model, providing further insights into its functionality, and how it processes data to make predictions.

4.1 Optimization

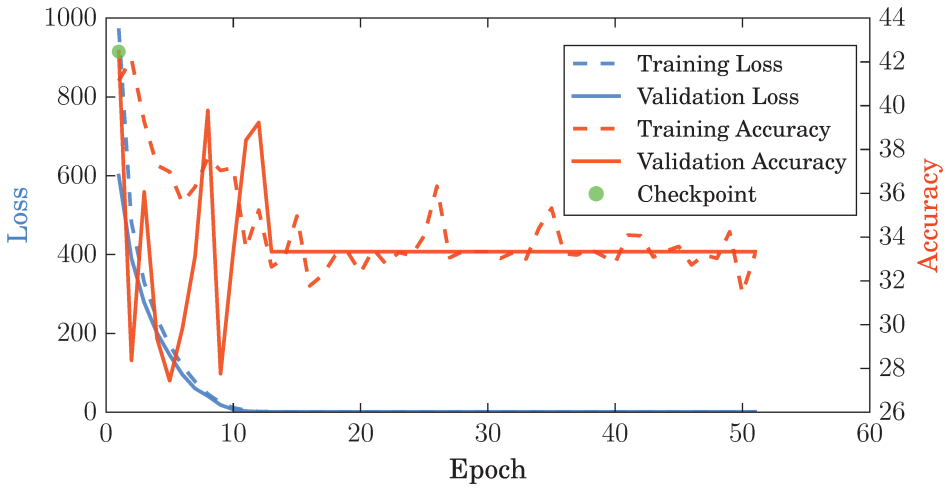
As stated earlier, we determined the initial learning rate of $\alpha = 10^{-4}$, and varied a number of configurations in each architecture, trying to achieve the best accuracy in the CADDementia training set. These options included regularization with L1 and L2 norms, regularization strength λ, number of units in fully-connected layers, dropout rates, and eventually the batch size, or multi-class hinge loss, instead of the traditional categorical cross-entropy loss.

The parameters for regularization strength, number of units, and dropout rates were also used for regularization, acting as trade-offs between model complexity and bias, thus managing the probability to overfit. This was a major concern for us due to the large size of our networks, and relatively small amount of data. The different batch size was an experiment to compare the behavior of all networks with the same batch size of three. Given that support vector machine (SVM) [13] classifiers usually present interesting results, and had also been successfully used for Alzheimer's biomarker identification before, we also experimented with the multi-class hinge loss.

In general, we varied regularization strength λ in powers of 10, between 10^{-5} and 10^{2}, number of units in fully-connected layers in powers of 2, between 32 and 2048, and dropout rate with steps of 10 percentage points, between 40% and 90%, including 95%, 99%, and 99.9%. Note that some networks had specific parameters, *i.e.*, these variations did not



(a) Overfitting with $\lambda = 10^{-2}$.



(b) Underfitting with $\lambda = 10^{-1}$.

Figure 4.1: Visualization of categorical cross-entropy loss and average TPF during optimization: overfitting *vs.* underfitting.

apply to all evaluated architectures. We followed a greedy approach, by first tuning regularization strength with L2 norm, followed by number of units, and then dropout rates. Next, we evaluated batches of size three for all networks, L1 norm, multi-class hinge loss, and, finally, larger datasets.

Regarding the batch size, we always selected the same number of samples from each class to form a batch, and we worked with a total batch size of either three or nine samples, depending on the network architecture. An epoch consisted of randomly sampling each class, limited by the class with fewer images, then the epoch was finished, and the next one started with a new random sampling.

We observed that, at some point, most networks would either underfit or overfit, presenting erratic metrics, with high variations between epochs. This is illustrated in Figure 4.1, where we plot our first VGG architecture, with 2048 units in fully-connected layers, and regularization strengths λ of 10^{-2} and 10^{-1} for L2 norm. To overcome this

Dataset	Median	Avg ± Std	Min	Max
1	11.6	11.6 ± 2.7	7.7	25.3
2	11.8	12.2 ± 3.6	4.3	122.5
3	11.9	12.2 ± 3.5	4.3	122.5
4	11.6	11.8 ± 3.3	4.3	122.5
CADD	11.9	13.4 ± 5.6	7.5	41.2

Table 4.1: Efficiency break down for brain extraction and normalization time, in minutes.

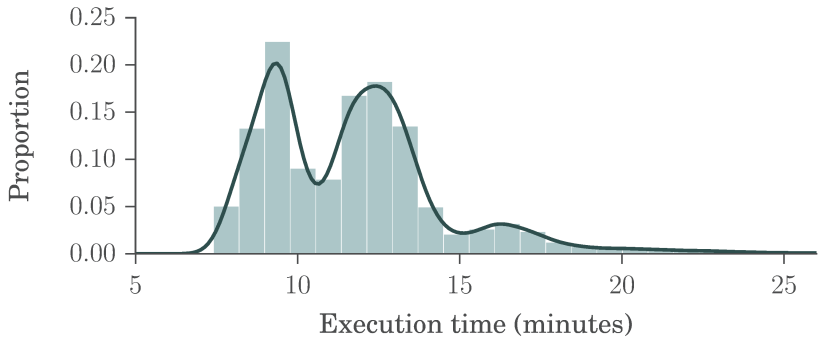


Figure 4.2: Histogram and kernel density estimation plots of brain extraction and normalization times for Dataset 4, in minutes.

situation, we saved only the model that presented the best average TPF in the respective validation set, making sure it would optimize for at least another 50 epochs without further improvement, with a hard limit of 200 epochs.

4.2 Performance

We first analyze the efficiency of our processing pipeline, divided in brain image, CNN, and domain adaptation stages. In Table 4.1, we present some statistics for the execution time of brain extraction and normalization steps. Each row represents a complete dataset, *e.g.*, Dataset 4 includes all previews computations, plus its own. CADD comprises both CADDementia training and test sets. In Figure 4.2, we plot a histogram and a kernel density estimation of this execution time for Dataset 4, which is our largest, and contains 23 165 volumes. It is interesting to note that only 151 (0.7%) images took longer than 25 minutes to process. Each process used two cores in a shared cluster of commodity hardware, such as Intel® Xeon® CPU E5645 at 2.40 GHz, and around 2 GB of RAM.

To train our CNNs, we used three different models of NVIDIA GPUs: GeForce GTX TITAN X (Maxwell microarchitecture), Tesla K40c, and Tesla K80. Training usually lasted for about 100 epochs, taking around 4 days to complete. We performed a total of 121 experiments. In Table 4.2, we present the processing time of our best network (VGG 512), which includes data loading time, considering all epochs from Dataset 1 validation set. In other words, to provide these values, we aggregated all experiments performed with the VGG 512 architecture, so we could have a better confidence. We divided the complete set time by the number of samples in the set, with 1 504 epochs on GeForce

GPU	Median	Avg ± Std	Min	Max
TITAN X	0.122	0.152 ± 0.100	0.098	1.011
Tesla K40c	0.298	0.348 ± 0.101	0.295	0.855

Table 4.2: Efficiency break down for our best network (VGG 512) processing time, in seconds.

GTX TITAN X, and 81 epochs on Tesla K40c. The grid search for domain adaptation took less than one minute to complete, while the classification of all 354 samples from CADDementia test set happened in about one millisecond.

In summary, our method is expected to provide an output for an input volume in less than 15 minutes, with extreme cases taking a little longer than 2 hours. To put it into perspective, this processing time contrasts with the current best method in CADDementia challenge, which needs 19 hours of computation [69]. In other words, our method is nearly 10× faster, considering the worst case scenario, or almost 80× faster, on average. We understand that, given the challenge date, such execution time comparison is not always fair, considering most methods of that time did not employ deep learning approaches. However, we also understand that all these techniques are solving essentially the same problem, *i.e.*, AD diagnosis, and, in this sense, such comparison remains valid, and it is part of method evolution.

Regarding performance metrics in terms of results, we present our best configuration for each network architecture in Table 4.3. The best VGG had 512 units in each fully-connected layer, and the best ResNet used the projection shortcut (B). We also include our main optimization metric, average TPF (avgTPF), for the training set of CADDementia, with our top value being 75.9%, which translated to 76.7% in accuracy. All these results were found while optimizing the networks with Dataset 1.

Architecture	avgTPF	Norm	λ	Dropout
LeNet-5	56.5%	L2	10^{-2}	40%
VGG 512	75.9%	L2	10^{-4}	50%
GoogLeNet	58.3%	L1	10^{-3}	80%
ResNet B	60.2%	L2	10^{-2}	–

Table 4.3: Performance results (average TPF) of our best CNN architectures, and respective configurations.

As initially expected, L2 norm performed better for almost all architectures. The best GoogLeNet using L2 achieved 57.4%, which is pretty close to the one using L1 (58.3%), while the L1 norm performed considerably worse for the other networks. ResNet with identity shortcuts (A) achieved 57.4%, which is slightly inferior to the projection shortcut (B), with 60.2%, being a similar difference found in the original work [32]. we hypothesize that deeper architectures did not achieve the highest scores because they tend to take advantage of larger datasets, which is not exactly our scenario.

A batch size of three, instead of nine, only produced significantly worse results, in-

Model	Dataset	Split	Accuracy	TPF			AUC			
				CN	MCI	AD	All	CN	MCI	AD
ADNet	Dset. 1	Train.	60.6	89.6	36.7	86.8	87.9	90.3	80.6	88.8
		Val.	44.1	71.1	22.4	62.4	68.9	72.2	56.9	72.5
		Test	43.6	67.3	21.1	64.7	68.0	73.9	57.0	68.9
ADNet	CADD	Train.	76.7	83.3	55.6	88.9	90.3	92.1	83.1	96.3
		Test	51.4	77.5	27.9	46.6	68.5	70.5	61.2	73.6
ADNet -DA	CADD	Train.*	76.7	75.0	55.6	100.0	88.5	90.7	79.4	95.8
		Train.	90.0	83.3	88.9	100.0	98.0	95.8	97.9	100.0
		Test	52.3	68.2	37.7	49.5	70.9	72.8	60.5	79.0

Table 4.4: Multiple performance results of our best CNN, in percentage. Train.* refers to leave-one-out cross-validation results.

dicating that our best VGG model could potentially achieve even better results, using GPUs with larger memory or a multi-GPU framework implementation. Similarly, multi-class hinge loss did not improve our results. Most surprisingly, Dataset 1, our smaller, presented the best performances, with Dataset 2 achieving as high as 72.2% on average TPF. We hypothesize that this happened due to the higher diversity of data sources and conditions in larger sets, indicating that a smaller but more cohesive dataset should be sufficient for optimization.

Considering our best network model (VGG 512), we present all performance metrics in Table 4.4. We named our CNN approach ADNet (Alzheimer's Disease Network), with the domain adaptation method being ADNet-DA, and submitted our prediction scores to the CADDementia challenge. As of this writing, there were 48 different submissions, including ours[1]. Similarly to reported results [9], we did expect a drop in results for the test set, when comparing to the training set. However, with so few samples to estimate accuracy, our evaluation was overly optimistic, even if optimizing our method on a completely different dataset.

In general, ADNet presented promising results in the CADDementia training set, with the exception of MCI TPF. However, the decrease in MCI and AD TPFs between training and test sets were higher than expected. As such, this method achieved an interesting two-class specificity, with a modest two-class sensitivity, meaning it performs better at determining healthy patients. Regarding accuracy in the test set, ADNet ranked in 25$^{\text{th}}$, tied with two other systems, meaning it outperformed 22 submissions. Also, this result is only statistically different, with a 95% confidence interval, from the first one, and the last three systems. Considering the fact that, to the best of our knowledge, we were the first group that did not use any domain specific information for this task, we can affirm that our CNN method did learn meaningful patterns automatically. The corresponding receiver operating characteristic (ROC) for CADDementia test set is displayed in Figure 4.3, and the respective confusion matrix is in Table 4.5.

As for the domain adaptation approach, we extracted 512 features from the second-

[1] `https://caddementia.grand-challenge.org/results_all/` [Online; accessed 2019-01-27]

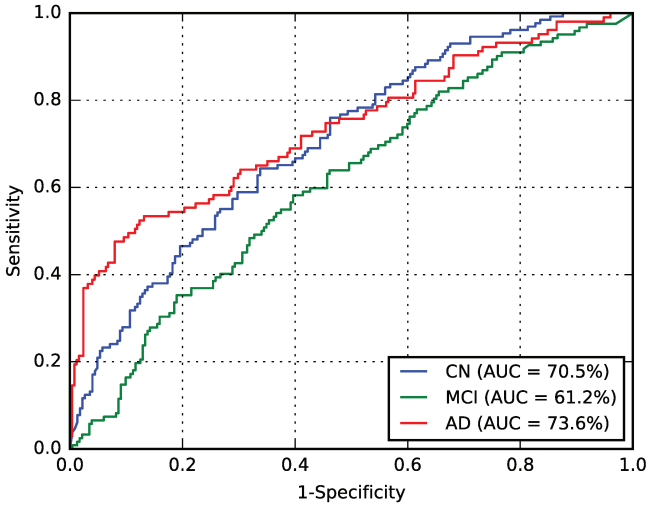


Figure 4.3: Receiver operating characteristic (ROC) curve for ADNet, provided by CADDementia.

		True		
		CN	MCI	AD
Pred.	CN	28.2	21.2	11.3
	MCI	5.6	9.6	4.2
	AD	2.5	3.7	13.6

Table 4.5: Confusion matrix (in percentage) for ADNet, provided by CADDementia.

to-last layer of ADNet, then we performed a grid search on the parameters of a logistic regression classifier. Using the best parameters found, most importantly, $C = 0.001$, we optimized this classifier on the complete training set, and applied it to output classification probabilities for each sample from the challenge. We also submitted these predictions to CADDementia, naming it ADNet-DA (ADNet with domain adaption), and the corresponding results are also indicated in Table 4.4. This method ranked in 21st, outperforming 27 submissions, with statistical difference from the first one, and the last four systems.

Considering this approach, we reported the leave-one-out cross-validation results in the training set while performing a grid search, and also the results in this same set after the last optimization with all training samples. As expected, developing and evaluating a system on the same data overestimates its generalization performance; however, even our cross-validation attempt did not significantly improved our estimations for the test set. In comparison with ADNet, ADNet-DA improves both MCI and AD TPFs, while decreasing CN TPF, with an overall improvement of almost one percentage point in accuracy, which shows that domain adaptation is indeed an important technique. The corresponding receiver operating characteristic (ROC) for CADDementia test set is displayed in Figure 4.4, and the respective confusion matrix is in Table 4.6.

In closing, to the best of our knowledge, we are the first group to propose end-to-end training a very deep 3D CNN for the multiclass AD biomarker identification task, and the

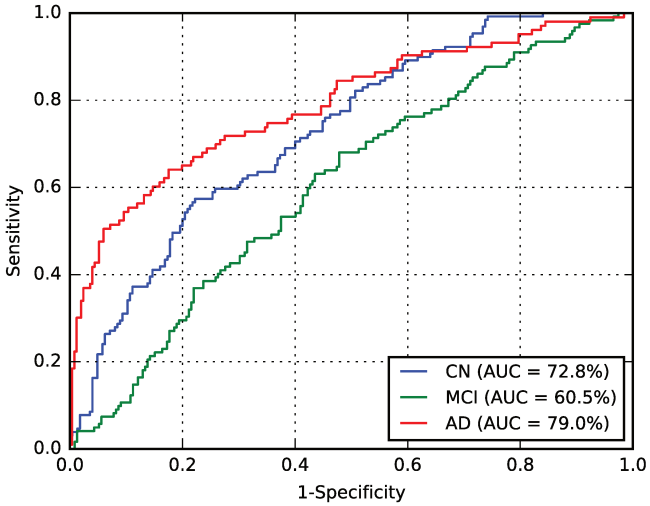


Figure 4.4: Receiver operating characteristic (ROC) curve for ADNet-DA, provided by CADDementia.

		True		
		CN	MCI	AD
Pred.	CN	24.9	17.8	8.5
	MCI	9.3	13.0	6.2
	AD	2.3	3.7	14.4

Table 4.6: Confusion matrix (in percentage) for ADNet-DA, provided by CADDementia.

first one to submit such an approach to CADDementia. However, we are the second to use deep learning on this challenge, with Dolph et al. [18] being the pioneers. One of their systems ranked 7$^{\text{th}}$, with 56.8% accuracy, while the other ranked 25$^{\text{th}}$, tied with ADNet on 51.4%. This shows that our ADNet-DA method was able to outperform a deep-learning system that uses domain-specific information, which demonstrates the effectiveness of the approach proposed in this work.

4.3 Accountability

Understanding the decision-making process of a machine-learning algorithm has become crucial lately, and even more so in medicine. To work in the real world, an algorithm must not only present impressive performance results, but it also needs to demonstrate how predictions are generated. This has become even more critical in recent years with rules such as the General Data Protection Regulation (GDPR), which also brought explainable artificial intelligence (XAI) to the spotlight.

Explaining what and how a neural network has learned is an open problem, with a rapidly evolving research field. In order to better understand what our model is analyzing in brain images and how it is done, we present here a number of visualization approaches, considering the most used techniques in accountable machine learning for neural networks. Some of these approaches were also recently explored by Rieke et al. [59].

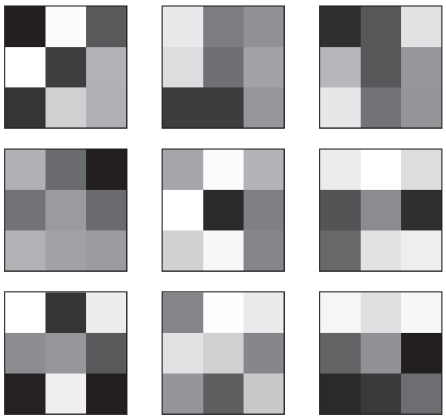


Figure 4.5: Weights visualization, with each row representing a complete filter of size $3 \times 3 \times 3$ from the first convolutional layer.

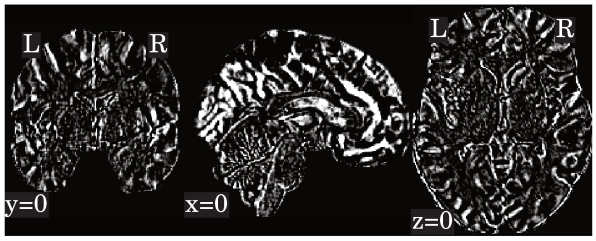
4.3.1 Weights

Similarly to Krizhevsky et al. [40], we plot 3 out of 32 filters from our first convolutional layer in Figure 4.5. While their kernels were of size $11 \times 11 \times 3$, presenting some interesting smooth and colorful patterns, our kernels are $3 \times 3 \times 3$ in grayscale, producing less than ideal images for visualization. Nonetheless, there are some three dimensional patterns in our filters, as can be observed in contrasting brighter and darker weights within and across planes. The spatial distribution of white and black voxels represent interesting 3D border structures that are emphasized by these filters.

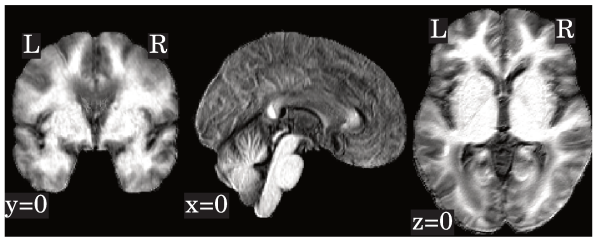
4.3.2 Activations

Another traditional approach for visualization is to show outputs of activation functions from the network, after processing an input. Activation is simply the result of a mathematical function. To this end, we selected one sample from each class in CADDementia training set that maximally activated the corresponding output probability. Specifically, they were *train_vumc_007* (CN), *train_emc_007* (MCI), and *train_emc_009* (AD). Considering the first convolutional layer, we had 32 outputs of dimensions $145 \times 182 \times 155$. For each input image, we selected the output with maximum total activation, out of 32, which were then zero-padded to the original atlas dimensions of $193 \times 229 \times 193$, and plotted in Figure 4.6. These outputs represent some of the initial patterns that the network learned to be the most relevant for this task, which are then non-linearly combined with additional and more complex patterns before the final classification. The idea is that brighter regions are more important to the network processing pipeline.

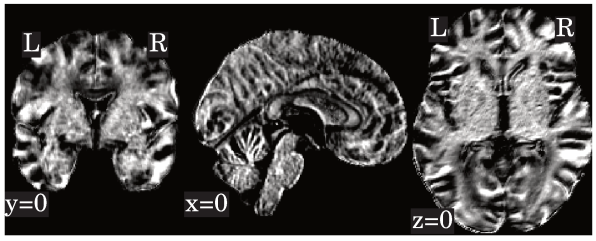
According to these images, it is possible to see that the regions considered significant for the CN processing pipeline seem to reveal larger clusters distributed across the gray matter of the brain, and randomly distributed small agglomerates in the white matter. Moreover, we can also note an edge enhancement in the interface between ventricles and gray and white matter. This result can be interpreted in the light of findings from the neuroimaging field: first, Alzheimer's is a disease originally known to mainly affect gray matter, and a widespread atrophy can be observed in AD brains; MCI subjects, in turn,



(a) Cognitively normal (*train_ vumc_ 007*).



(b) Mild cognitive impairment (*train_ emc_ 007*).



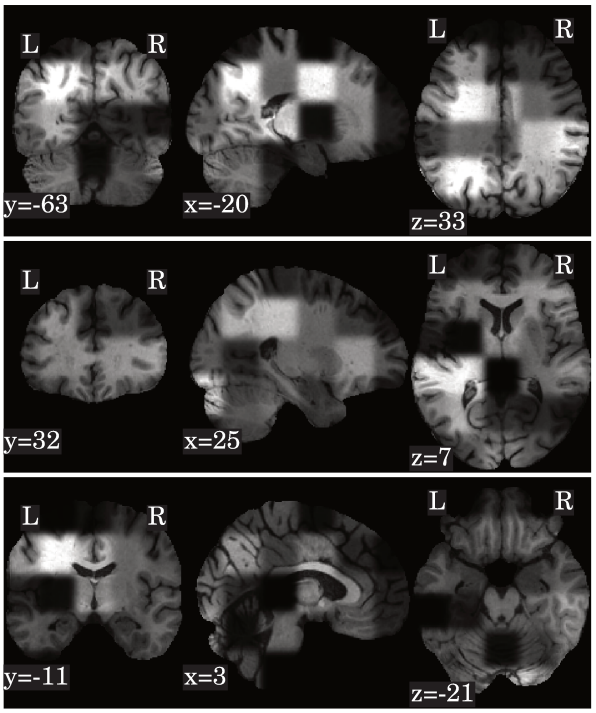
(c) Alzheimer's disease (*train_ emc_ 009*).

Figure 4.6: Anatomical planes of individuals from each group, considering the output with maximum total activation after the first convolutional layer. Coordinates are in MNI space.

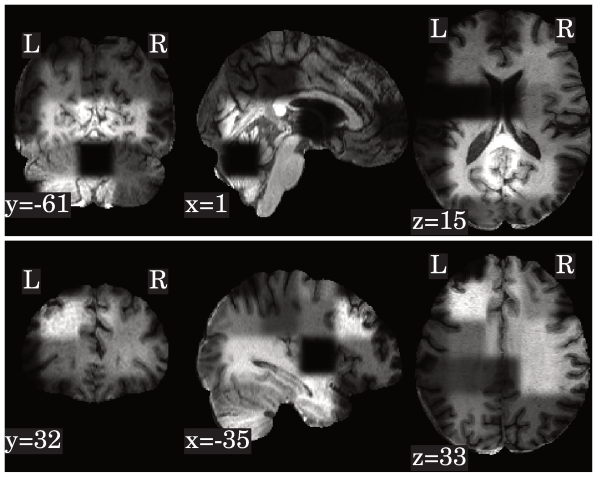
present a gray matter atrophy similarly distributed but less intense than that of AD group [7]. One could speculate that such a pattern of activation in the first convolutional layer for the CN processing perhaps reflects the preservation of gray matter for this group.

Interestingly, the regions with largest activation within these MCI and AD images include gray and white matter equally. This increased importance in white matter regions may indicate that abnormalities in this area may also play an important role in the pathogenesis and diagnosis of the disease. In fact, several neuroimaging studies have found that AD patients have extensive, more severe and widespread white matter damage than expected [11, 85], in which axonal demyelination might occur prior to the presence of amyloid β plaques and neurofibrillary tangles in the presymptomatic stages of AD [63]. In addition, small-vessel cerebrovascular alterations, observed as white matter hyperintensities in MRI, independently predicted AD diagnosis as much as amyloid burden measured by PET [58].

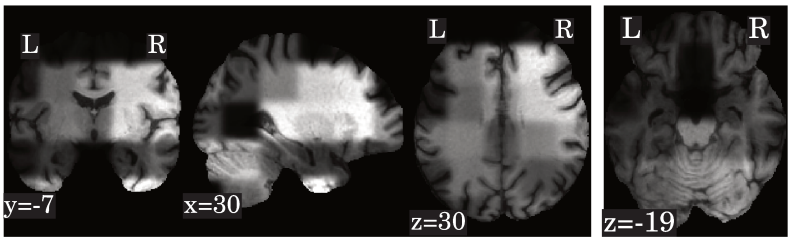
We also adopted this visualization approach to the last layer that still kept spatial information. In our network, this was the fifth pooling layer, which outputs 256 images with dimensions $5 \times 6 \times 5$. Since each voxel in this dimension corresponds roughly to a region of $30 \times 30 \times 30$ voxels in the original space, we projected this layer back to the



(a) Cognitively normal (*train_ vumc_ 007*).



(b) Mild cognitive impairment (*train_ emc_ 007*).



(c) Alzheimer's disease (*train_ emc_ 009*).

Figure 4.7: Anatomical planes of individuals from each group, considering the output with maximum total activation after the last pooling layer. Coordinates are in MNI space. Brighter regions represent areas with more activation, *i.e.*, more important to the network processing pipeline.

input image, so we could have a better visualization. To this end, we first resized the $5 \times 6 \times 5$ image to $145 \times 182 \times 155$ with nearest neighbor interpolation, then we applied a Gaussian filter with kernel size of $25 \times 25 \times 25$ and $\sigma = 3$, in order to reduce pixelation effects on voxel borders, thus improving visualization. Lastly, we multiplied this image with the original input image, so we could analyze how important voxels in this pooling layer roughly related to the brain regions, and then zero-padded to $193 \times 229 \times 193$.

We generated this visualization for the same previously selected patients, considering the output with maximum total activation, out of 256, and depicted resulting images in Figure 4.7. These outputs represent some of the final patterns that the network learned to be the most relevant for this task, which are still then non-linearly combined with additional and more complex patterns before the final classification. In this visualization, a brighter region indicate that one or more $3 \times 3 \times 3$ patterns within this region were considered relevant throughout the network processing pipeline.

From these images, we can observe that for the CN, there were activations in the white and grey matter of bilateral posterior, bilateral frontal, and left temporal regions; whereas for MCI, higher activations occurred in white and grey matter of the bilateral occipital, temporal and frontal regions, including the transverse cistern, the medulla and the lower pons. AD activations mainly occurred in the white and grey matter of the right frontal and temporal regions, as well as cerebellum.

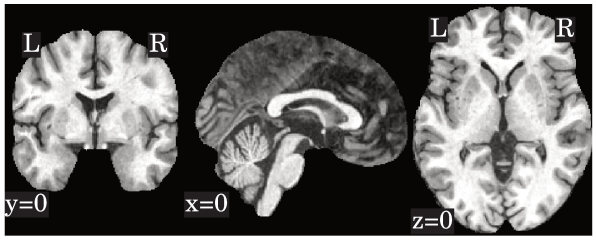
It is hard to speculate upon these findings, especially because the last convolutional layer is at a more abstract and complex level. Moreover, the grid pattern is not anatomically correlated with brain structures. Therefore, the partial volume effect definitely impacts the observations, precluding more precise discussions.

4.3.3 Occlusions

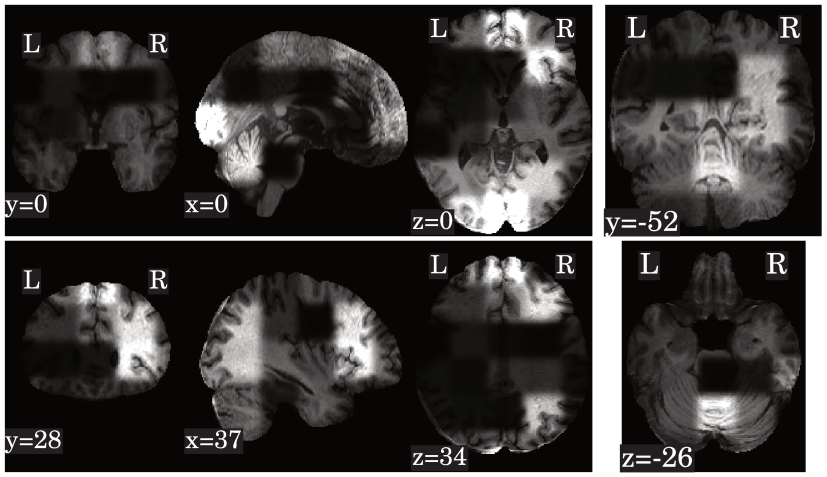
Occlusion is a technique to visualize how and where the input image affects the output of the network. The basic idea is to systematically hide (occlude) some regions of the input image, making the network not activate in these specific regions, and then storing the probabilities output. Given a class of interest, for instance AD, it is possible to create a heatmap with the corresponding prediction for each occluded region, where most important ones will present highest impact (with low probability), due to the occlusion. This was originally proposed by Zeiler and Fergus [89].

There is a number of ways to hide a region of the input image and try to avoid activations in a network. The simplest and most direct one would be to set input values to their respective averages, which, in our case, is zero. Considering images in a range from zero to 255, it is possible to occlude with the average value (gray), with zero (black), with 255 (white), and even more sophisticated approaches, such as different forms of noise.

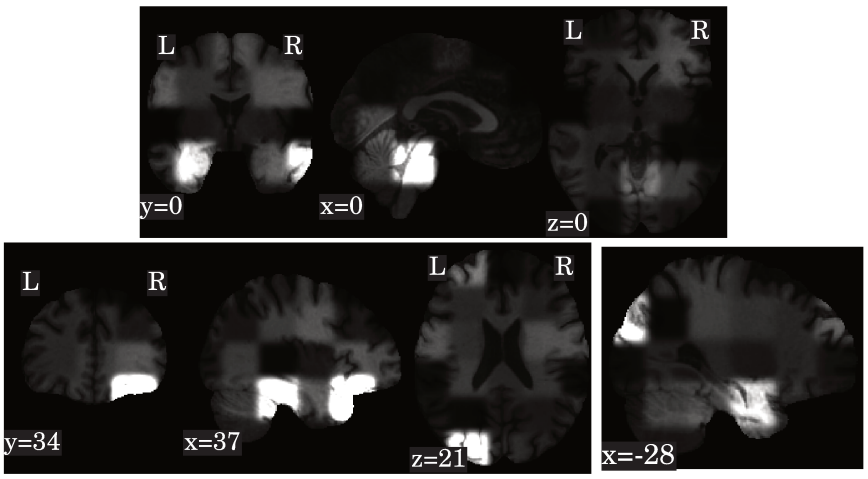
For each selected patient, we occluded (*i.e.*, set to zero) regions of $30 \times 30 \times 30$ voxels, with a stride of 30 between regions, and created a heatmap for each corresponding diagnosis class. Since lower probabilities represented higher impact, we inverted these values, and linearly normalized them between zero and one. Then, we proceeded with the same previous steps for smoothing, padding, and multiplication by the original input image. Resulting images are shown in Figure 4.8, where brighter regions imply higher influence,



(a) Cognitively normal (*train_ vumc_ 007*).



(b) Mild cognitive impairment (*train_ emc_ 007*).



(c) Alzheimer's disease (*train_ emc_ 009*).

Figure 4.8: Anatomical planes of individuals from each group, considering the occlusion heatmaps for the corresponding classes. Coordinates are in MNI space. Brighter regions imply higher influence, *i.e.*, when occluded, these regions caused more confusion to the classifier.

meaning that, when occluded, these regions caused more confusion to the classifier.

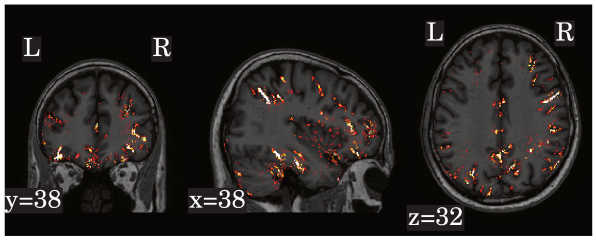
From the resulting images, we can see that the whole brain was equally activated in the CN case. This happened basically because our network interprets regions with zero as indicative of CN, thus only reassuring the previous classification, and producing a less than ideal visualization for occlusion. For MCI, we see higher activations in the occipital lobe, the medial superior region of the cerebellum, frontal lobe regions (*e.g.*, the prefrontal cortex and a in the right hemisphere comprising the inferior, medial and superior frontal cortex and their underlying white matter) and a posterior region of the right parietal cortex. For AD, we see that activation patterns were more symmetrical, with more intense activations in caudal parts of the brain such as the brainstem and ventral areas of the temporal and frontal lobes. It is important to highlight once again that the grid pattern has no direct correlation with anatomical structures, hindering further speculations regarding the biological meaning of such findings.

4.3.4 Backpropagation

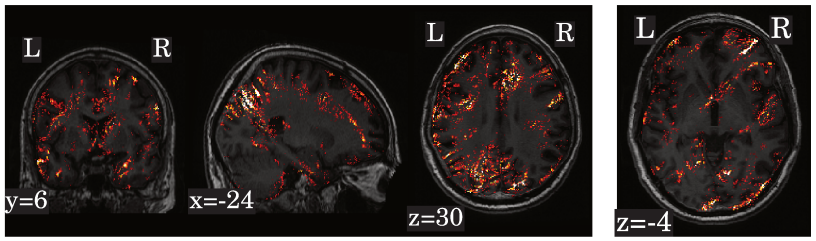
Finally, we investigated an approach that more closely related to the actual output decision of the network. Backpropagation [60] is a traditional technique to optimize neural networks. In short, we calculate the gradient of the network with respect to the input, which is used to update the network's internal parameters. These gradients may also be plotted, and interpreted as how much the output is affected by changes in input values; however, this simple approach produces rather noisy visualizations. Zeiler and Fergus [89] proposed an improvement to this technique, called deconvolution, which can be interpreted as reversing the operations performed by the network. Even though this is an interesting approach, the guided backpropagation method, originally proposed by Springenberg et al. [71], produces even sharper visualizations. Interestingly, guided backpropagation combines calculations from both backpropagation and deconvolution, resulting in more detailed images. In Figure 4.9, activations are shown in *hot* colormap overlaid on top of the respective registered MRI, where changes in brighter regions mean larger effect on the prediction output.

For the CN, we can see activations distributed in a diffuse pattern, but mainly restricted to cortex in the right temporal lobe (majorly in the medial temporal gyrus and the parahippocampal gyrus), the central portion of the occipital lobe, the posterior cingulum, and the posterior parietal cortex. For MCI, we can see activations in the left posterior parietal cortex, the right anterior cingulum and the right dorsolateral prefrontal cortex. For AD, larger activations were detected in the left posterior parietal cortex, right temporal pole, cerebellum and more diffusively in the spherical surface of the brain.

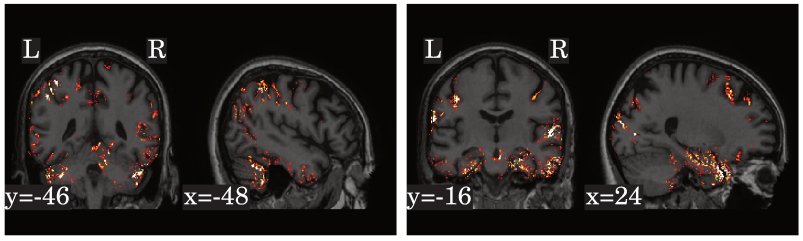
It is interesting to note the diffuse pattern of activations in all groups, but mainly in temporal and posterior regions of the brain. We can attempt to interpret such activations in the context of neuroimaging findings. Although no single structure is able to differentiate AD patients from CN subjects, atrophy in temporal regions is widely important in the context of AD. The medial temporal lobe regions might be the first ones affected in the course of the disease, presenting very early signs of neurodegeneration [37] and correlates well with clinical symptoms even in the prodromal stage, *i.e.*, MCI [26]. As in



(a) Cognitively normal (*train_ vumc_ 007*).



(b) Mild cognitive impairment (*train_ emc_ 007*).



(c) Alzheimer's disease (*train_ emc_ 009*).

Figure 4.9: Anatomical planes of individuals from each group, considering the outputs from guided backpropagation. Activations are displayed in *hot* colormap overlaid on top of the respective registered MRI. Coordinates are in MNI space. Changes in brighter regions, indicated in *hot* colormap, mean larger effect on the prediction output.

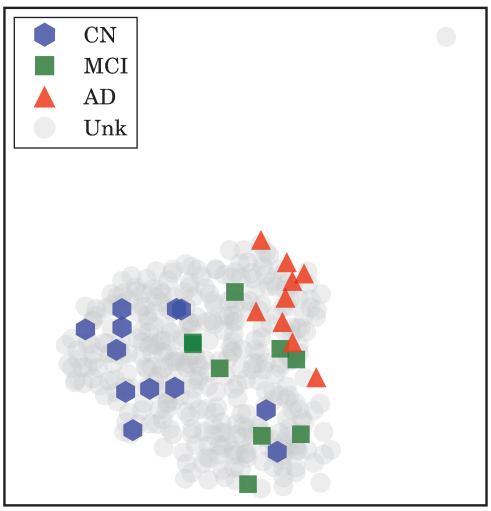
physiopathological aspects, the temporal regions mainly present intracellular aggregates of hyperphosphorylated tau protein, which associates with reduced grey matter density [77]. The other main signature of AD, extracellular amyloid β-protein (Aβ) deposition in form of plaques, are mainly observed in the midline regions (posterior cingulate and medial prefrontal cortices), and parietal areas. Longitudinal studies have shown that these areas not only are atrophic at the mild stage of AD [85], but they continue to degenerate at a rate of about 2% to 4% per year [43, 78].

4.3.5 Embeddings

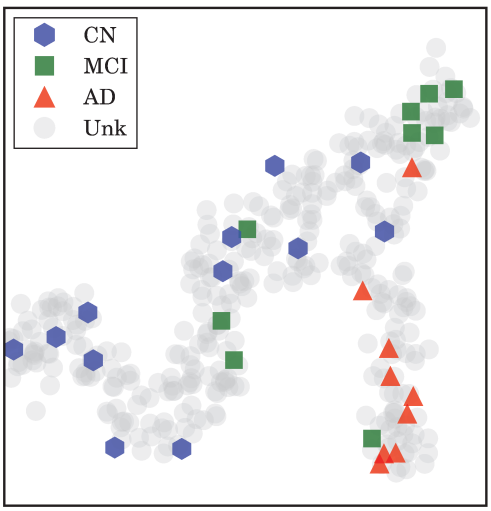
For our last visualization technique, our motivation was to understand how our data samples were spatially distributed within internal feature representations of our network, in order to determine whether these representations were really helpful to discriminate between each class. To plot our data from this high-dimensional space, we first projected them into two dimensions using the t-distributed stochastic neighbor embedding (t-SNE) [47], with principal component analysis (PCA) initialization. Considering the outputs from a specific layer of our network, we generated an embedding with all train-

ing and test data in CADDementia, and then colored training samples according to each respective class. It is important to remark that this projection approach did not use label information from training data, which was used solely to color our plots.

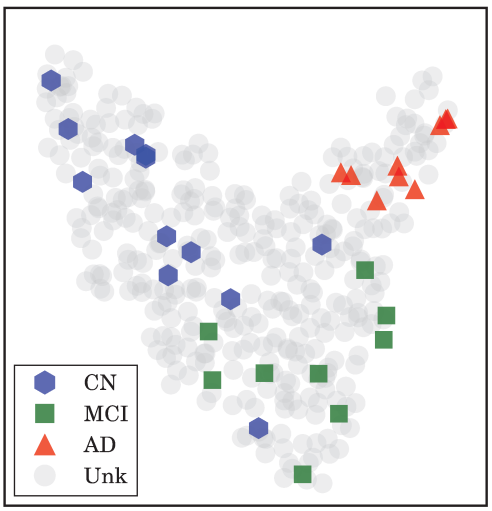
First, we extracted features from the second-to-last layer of our network, with 512 dimensions, which is a traditional layer used for transfer learning and domain adaptation. Then, we considered the final layer from ADNet, which outputs classification probabilities, with 3 dimensions, and also the probability outputs from ADNet-DA. Resulting embeddings are present in Figure 4.10. Considering ADNet, even though t-SNE [47] did not use any label information, training data points were better grouped in an internal feature representation space, rather than in the probability output space. This indicated that the softmax classifier used in the network did not perform as well as it could. From these plots, we can also see that probabilities from ADNet-DA are better distributed in comparison with ADNet, especially for AD group, while there was less confusion for MCI.



(a) Features from second-to-last layer.



(b) ADNet probabilities.



(c) ADNet-DA probabilities.

Figure 4.10: Features and probabilities visualizations with t-SNE projections.

Chapter 5

Conclusions

AD is a critical disorder, to which there is still no cure, killing more people than breast cancer and prostate cancer combined; between 2000 and 2015, deaths from AD have increased 123% [4]. Early diagnosis is currently the most fundamental hope for patients, benefiting their treatment and plans for the future. Magnetic resonance imaging is one such approach that could assist specialists to diagnose this disease as soon as possible, with the computer-aided diagnosis of dementia (CADDementia) [9] challenge launching a standardized evaluation protocol for this difficult task.

Using data from ADNI [54], we optimized a 3D convolutional neural network with the whole brain image as input, and the best accuracy was achieved with a network architecture based on VGG [68]. Our method, named ADNet, achieved interesting results, outperforming a number of other systems in prior art. Additionally, our method with domain adaptation, called ADNet-DA, reached 52.3% in accuracy on the CADDementia challenge test set, outperforming most submissions to this challenge, all of which using prior information from the disease. It is important to note that these approaches are completely automatic (*i.e.*, there is no need for manual intervention), and, in comparison to the state of the art, are also considerably fast.

In summary, ADNet is an adapted version of architecture VGG-A, with 11 weight layers. The main differences, besides 3D convolutions, were halving all numbers of filters in convolutional layers, and setting the numbers of units in fully-connected layers to 512. Considering all evaluated parameters, the best configuration used L2 norm, with regularization strength λ equal to 10^{-4}, and 50% of dropout rate. We hypothesize that this particular network architecture configuration surpassed all the others in our experiments mainly due to the fact that it tends to present strong results when evaluated alone, *i.e.*, not in an ensemble. Additionally, this was a network with 11 weight layers, while the best VGG configuration had 19 weight layers; for comparison, the best ResNet had 152 layers, and we only considered its shorter version, with 18 layers, so the smaller difference could also be a contributing factor. Further experiments are necessary to track down the root cause for this finding.

Since our method did not use any domain-specific knowledge from AD, we believe it could be directly applied to other disorders that could benefit from computer-aided diagnosis system using sMRI as input data. In this sense, we understand our approach is able to automatically determine meaningful patterns within data, and thus could corrob-

orate previous findings by specialists, assist in diagnosis scenarios, and eventually help with different or new diseases. This is supported by our explainable artificial intelligence (XAI) techniques, including accountability visualizations.

Similarly to our ADNet-DA approach, we believe that our publicly released learned model could be directly applied to different datasets even with very few samples. Additionally, outputs from either an internal representation space or the final probabilities could be used in combination with different approaches to further improve results, or even applied to novel problems, such as The Alzheimer's Disease Prediction Of Longitudinal Evolution (TADPOLE) challenge [51], which aims at predicting future evolution of individuals at risk.

In this research, we have learned that it is indeed possible to train a deep learning system to help with the challenging AD biomarker identification task. Even though we faced some difficulties with hardware, software and data limitations, our proposed solutions proved to be satisfactory. As such, we believe that our method can be actually employed in medical practice. While diagnosing patients, it is possible for specialists to use ADNet network to generate a diversity of explanatory visualizations for a given image, while ADNet-DA can assist with the diagnosis. This way, specialists can come up with a more informed decision and in less time.

Continuing our work, there is a number of possible improvements and alternative paths. In terms of accountability and visualization techniques, a more straightforward next step would be to explore additional patients, which could also be achieved with group aggregation approaches, such as calculating the mean or the median across a specific diagnostic group. Additionally, our CNN optimization pipeline could be used with a more recent deep learning framework, which enables multiple GPU support, for instance, so larger batches or even larger CNN architectures are possible. Another modification could be to use more recent network architectures, such as Inception-v4 or Inception-ResNet-v2 [75]. Moreover, since ADNet learned an internal embedding space, it would be possible to explicitly enforce it with a distance metric learning approach [86], for instance, using triplet loss. It would also be interesting to explore research paths that consider the native MRI as input, rather than going through a pre-processing pipeline that normalizes images to a standard dimension and, as such, could potentially alter important brain characteristics and structures.

Considering we have a relatively small dataset, when we think of data-driven methods, exploring alternatives to increase it is another promising research path. Simpler approaches could include additional information [20], such as age, sex, mini-mental state examination (MMSE), years of education, and genotypes. In a more complex option, it is also possible to generate synthetic data. Given that, in principle, AD symmetrically affects the brain [84], it is possible to simply flip input images [20]. Data augmentation could be done at a more abstract level, by extracting MRI features similarly to ADNet-DA, approximating these features to select distributions, and then sampling new data. It could also be done at input image level, using techniques such as generative adversarial networks (GANs) [67]. Finally, methods for learning with few samples are also promising [27]. ■

Acknowledgments

The authors appreciate the effort from researchers and patients in collecting and providing all the data used in our study. We additionally thank infrastructure support for running our experiments. Each group is explicitly acknowledged next, in alphabetical order. The authors have no competing interests to declare.

ADNI

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Alzheimer's Association; Alzheimer's Drug Discovery Foundation; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; GE Healthcare; Innogenetics, N.V.; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (`https://fnih.org/`). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California. This research was also supported by NIH grants P30 AG010129 and K01 AG030514.

CADDementia

Data used in the preparation of this article were obtained from the CADDementia challenge (`https://caddementia.grand-challenge.org/`).

Microsoft Azure

Cloud computing resources were provided by a Microsoft Azure for Research award (https://www.microsoft.com/en-us/research/academic-program/microsoft-azure-for-research/).

MIRIAD

Data used in the preparation of this article were obtained from the MIRIAD database (http://miriad.drc.ion.ucl.ac.uk). The MIRIAD investigators did not participate in analysis or writing of this report. The MIRIAD dataset is made available through the support of the UK Alzheimer's Society (https://www.alzheimers.org.uk/) (Grant RF116). The original data collection was funded through an unrestricted educational grant from GlaxoSmithKline (Grant 6GKC).

NVIDIA

The Tesla K40 used for this research was donated by the NVIDIA Corporation.

OASIS

Grant numbers P50 AG05681, P01 AG03991, R01 AG021910, P50 MH071616, U24 RR021382, R01 MH56584.

Bibliography

[1] R. Al-Rfou, G. Alain, A. Almahairi, et al. Theano: A Python framework for fast computation of mathematical expressions. *ArXiv e-prints*, May 2016.

[2] M. S. Albert, S. T. DeKosky, D. Dickson, et al. The diagnosis of mild cognitive impairment due to Alzheimer's disease: Recommendations from the national institute on aging-Alzheimer's association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia*, 7(3):270–279, 2011.

[3] G. I. Allen, N. Amoroso, C. Anghel, et al. Crowdsourced estimation of cognitive decline and resilience in Alzheimer's disease. *Alzheimer's & Dementia*, 12(6):645–653, 2016.

[4] Alzheimer's Association. 2018 Alzheimer's disease facts and figures. *Alzheimer's & Dementia*, 14(3):367–429, 2018.

[5] B. B. Avants, N. Tustison, and G. Song. Advanced normalization tools (ants). *Insight J*, 2:1–35, 2009.

[6] B. B. Avants, N. J. Tustison, G. Song, et al. A reproducible evaluation of ANTs similarity metric performance in brain image registration. *NeuroImage*, 54(3):2033–2044, 2011.

[7] M. L. F. Balthazar, C. L. Yasuda, F. R. Pereira, et al. Differences in grey and white matter atrophy in amnestic mild cognitive impairment and mild Alzheimer's disease. *European Journal of Neurology*, 16(4):468–474, 2009.

[8] L. A. Beckett, M. C. Donohue, C. Wang, et al. The Alzheimer's disease neuroimaging initiative phase 2: Increasing the length, breadth, and depth of our understanding. *Alzheimer's & Dementia*, 11(7):823–831, 2015.

[9] E. E. Bron, M. Smits, W. M. van der Flier, et al. Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural mri: The CADDementia challenge. *NeuroImage*, 111:562–579, 2015.

[10] T. Brosch and R. Tam. Efficient training of convolutional deep belief networks in the frequency domain for application to high-resolution 2d and 3d images. *Neural Computation*, 27(1):211–227, 2015. PMID: 25380341.

[11] F. Caso, F. Agosta, D. Mattavelli, et al. White matter degeneration in atypical Alzheimer disease. *Radiology*, 277(1):162–172, 2015. PMID: 26018810.

[12] D. L. Collins, A. P. Zijdenbos, W. F. C. Baaré, et al. Animal+insect: Improved cortical structure segmentation. In A. Kuba, M. Šáamal, and A. Todd-Pokropek, editors, *International Conference on Information Processing in Medical Imaging*, pages 210–223. Springer, 1999.

[13] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, Sep 1995.

[14] R. Cuingnet, E. Gerardin, J. Tessieras, et al. Automatic classification of patients with Alzheimer's disease from structural MRI: A comparison of ten methods using the ADNI database. *NeuroImage*, 56(2):766–781, 2011. Multivariate Decoding and Brain Reading.

[15] S. Dieleman, J. Schlüter, C. Raffel, et al. Lasagne: First release., August 2015.

[16] W. J. Dixon. Simplified estimation from censored normal samples. *The Annals of Mathematical Statistics*, 31(2):385–391, 1960.

[17] K. Doi. Computer-aided diagnosis in medical imaging: Historical review, current status and future potential. *Computerized Medical Imaging and Graphics*, 31(4):198–211, 2007. Computer-aided Diagnosis (CAD) and Image-guided Decision Support.

[18] C. V. Dolph, M. Alam, Z. Shboul, et al. Deep learning of texture and structural features for multiclass Alzheimer's disease classification. In *International Joint Conference on Neural Networks (IJCNN)*, pages 2259–2266, May 2017.

[19] K. A. Ellis, A. I. Bush, D. Darby, et al. The australian imaging, biomarkers and lifestyle (AIBL) study of aging: methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of Alzheimer's disease. *International Psychogeriatrics*, 21(4):672–687, August 2009.

[20] S. Esmaeilzadeh, D. I. Belivanis, K. M. Pohl, et al. End-to-end Alzheimer's disease diagnosis and biomarker identification. In Y. Shi, H.-I. Suk, and M. Liu, editors, *Machine Learning in Medical Imaging*, pages 337–345, Cham, 2018. Springer International Publishing.

[21] D. A. Evans, H. Funkenstein, M. S. Albert, et al. Prevalence of Alzheimer's disease in a community population of older persons: Higher than previously reported. *JAMA*, 262(18):2551–2556, 1989.

[22] F. Falahati, E. Westman, and A. Simmons. Multivariate data analysis and machine learning in Alzheimer's disease with a focus on structural magnetic resonance imaging. *Journal of Alzheimer's Disease*, 41(3):685–708, 2014.

[23] C. P. Ferri, M. Prince, C. Brayne, et al. Global prevalence of dementia: a delphi consensus study. *The Lancet*, 366(9503):2112–2117, Dec 2005.

[24] V. Fonov, A. C. Evans, K. Botteron, et al. Unbiased average age-appropriate atlases for pediatric studies. *NeuroImage*, 54(1):313–327, 2011.

[25] V. Fonov, A. Evans, R. McKinstry, et al. Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage*, 47, Supplement 1:S102, 2009. Organization for Human Brain Mapping 2009 Annual Meeting.

[26] G. B. Frisoni, N. C. Fox, C. R. Jack Jr, et al. The clinical use of structural MRI in Alzheimer disease. *Nature Reviews Neurology*, 6:67–77, Feb 2010. Review Article.

[27] Y. Fu, T. Xiang, Y.-G. Jiang, et al. Recent advances in zero-shot recognition. *ArXiv e-prints*, October 2017.

[28] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In Y. W. Teh and M. Titterington, editors, *International Conference on Artificial Intelligence and Statistics*, pages 249–256. PMLR, 13–15 May 2010.

[29] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.

[30] B. Goodman and S. Flaxman. European Union regulations on algorithmic decision-making and a "right to explanation". *ArXiv e-prints*, June 2016.

[31] K. H. Gylys, J. A. Fein, F. Yang, et al. Synaptic changes in Alzheimer's disease: Increased amyloid-β and gliosis in surviving terminals is accompanied by decreased psd-95 fluorescence. *The American Journal of Pathology*, 165(5):1809–1817, 2004.

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

[33] E. J. Herrera, P. Caramelli, A. S. B. Silveira, et al. Epidemiologic survey of dementia in a community-dwelling brazilian population. *Alzheimer Disease & Associated Disorders*, 16(2):103–108, 2002.

[34] E. Hosseini-Asl, M. Ghazal, A. Mahmoud, et al. Alzheimer's disease diagnostics by a 3d deeply supervised adaptable convolutional network. *Frontiers in bioscience (Landmark edition)*, 23:584–596, January 2018.

[35] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In D. Blei and F. Bach, editors, *International Conference on Machine Learning (ICML)*, pages 448–456, 2015.

[36] C. R. Jack, M. A. Bernstein, N. C. Fox, et al. The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. *Journal of Magnetic Resonance Imaging*, 27(4): 685–691, 2008.

[37] G. Karas, P. Scheltens, S. Rombouts, et al. Global and local gray matter loss in mild cognitive impairment and Alzheimer's disease. *NeuroImage*, 23(2):708–716, 2004.

[38] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *ArXiv e-prints*, December 2014.

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

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

[41] Y. LeCun, B. Boser, J. S. Denker, et al. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, December 1989.

[42] Y. LeCun, L. Bottou, Y. Bengio, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, November 1998.

[43] A. D. Leow, I. Yanovsky, N. Parikshak, et al. Alzheimer's disease neuroimaging initiative: A one-year follow up study using tensor-based morphometry correlating degenerative rates, biomarkers and cognition. *NeuroImage*, 45(3):645–655, 2009.

[44] R. Li, W. Zhang, H.-I. Suk, et al. Deep learning based imaging data completion for improved brain disease diagnosis. In P. Golland, N. Hata, C. Barillot, J. Hornegger, and R. Howe, editors, *Medical Image Computing and Computer-Assisted Intervention (MICCAI 2014)*, pages 305–312. Springer, 2014.

[45] C. Lin, R. Watson, H. Ward, et al. MP-RAGE compared to 3D IR SPGR for optimal T1 contrast and image quality in the brain at 3T. *International Society for Magnetic Resonance in Medicine (ISMRM)*, 14:981, 2006.

[46] S. Lovestone, P. Francis, I. Kloszewska, et al. Addneuromed–the european collaboration for the discovery of novel biomarkers for Alzheimer's disease. *Annals of the New York Academy of Sciences*, 1180(1):36–46, 2009.

[47] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

[48] I. B. Malone, D. Cash, G. R. Ridgway, et al. MIRIAD–Public release of a multiple time point Alzheimer's MR imaging dataset. *NeuroImage*, 70:33–36, 2013.

[49] D. S. Marcus, T. H. Wang, J. Parker, et al. Open access series of imaging studies (oasis): Cross-sectional mri data in young, middle aged, nondemented, and demented older adults. *Journal of Cognitive Neuroscience*, 19(9):1498–1507, August 2007.

[50] D. S. Marcus, A. F. Fotenos, J. G. Csernansky, et al. Open access series of imaging studies: Longitudinal mri data in nondemented and demented older adults. *Journal of Cognitive Neuroscience*, 22(12):2677–2684, November 2009.

[51] R. V. Marinescu, N. P. Oxtoby, A. L. Young, et al. TADPOLE challenge: Prediction of longitudinal evolution in Alzheimer's disease. *ArXiv e-prints*, May 2018.

[52] P. McCullagh. Generalized linear models. *European Journal of Operational Research*, 16(3):285–292, 1984.

[53] G. M. McKhann, D. S. Knopman, H. Chertkow, et al. The diagnosis of dementia due to Alzheimer's disease: Recommendations from the national institute on aging-Alzheimer's association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia*, 7(3):263–269, 2011.

[54] S. G. Mueller, M. W. Weiner, L. J. Thal, et al. Ways toward an early diagnosis in Alzheimer's disease: The Alzheimer's disease neuroimaging initiative (ADNI). *Alzheimer's & Dementia*, 1(1):55–66, 2005.

[55] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In J. Fürnkranz and T. Joachims, editors, *International Conference on Machine Learning (ICML)*, pages 807–814. Omnipress, 2010.

[56] A. Payan and G. Montana. Predicting Alzheimer's disease: a neuroimaging study with 3D convolutional neural networks. *ArXiv e-prints*, February 2015.

[57] F. Pedregosa, G. Varoquaux, A. Gramfort, et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(Oct):2825–2830, 2011.

[58] F. Provenzano, J. Muraskin, G. Tosto, et al. White matter hyperintensities and cerebral amyloidosis: Necessary and sufficient for clinical expression of Alzheimer disease? *JAMA Neurology*, 70(4):455–461, 2013.

[59] J. Rieke, F. Eitel, M. Weygandt, et al. Visualizing convolutional networks for mri-based diagnosis of Alzheimer's disease. *ArXiv e-prints*, August 2018.

[60] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 323:533–536, Oct 1986.

[61] O. Russakovsky, J. Deng, H. Su, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[62] M. R. Sabuncu and E. Konukoglu. Clinical prediction from structural brain mri scans: A large-scale empirical study. *Neuroinformatics*, 13(1):31–46, 2015.

[63] P. S. Sachdev, L. Zhuang, N. Braidy, et al. Is Alzheimer's a disease of the white matter? *Current Opinion in Psychiatry*, 26(3):244–251, 2013.

[64] S. Sarraf, J. Anderson, and G. Tofighi. Deepad: Alzheimer's disease classification via deep convolutional neural networks using mri and fmri. *bioRxiv*, 2016.

[65] D. J. Selkoe. Alzheimer's disease: Genes, proteins, and therapy. *Physiological Reviews*, 81(2):741–766, 2001. PMID: 11274343.

[66] A. Sharif Razavian, H. Azizpour, J. Sullivan, et al. Cnn features off-the-shelf: An astounding baseline for recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 806–813, June 2014.

[67] H.-C. Shin, N. A. Tenenholtz, J. K. Rogers, et al. Medical image synthesis for data augmentation and anonymization using generative adversarial networks. *ArXiv e-prints*, July 2018.

[68] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *ArXiv e-prints*, September 2014.

[69] L. Sørensen, C. Igel, A. Pai, et al. Differential diagnosis of mild cognitive impairment and Alzheimer's disease using structural MRI cortical thickness, hippocampal shape, hippocampal texture, and volumetry. *NeuroImage: Clinical*, 13:470–482, 2017.

[70] R. A. Sperling, P. S. Aisen, L. A. Beckett, et al. 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 & Dementia*, 7(3):280–292, 2011.

[71] J. T. Springenberg, A. Dosovitskiy, T. Brox, et al. Striving for simplicity: The all convolutional net. *ArXiv e-prints*, December 2014.

[72] H.-I. Suk and D. Shen. Deep learning-based feature representation for ad/mci classification. In K. Mori, I. Sakuma, Y. Sato, C. Barillot, and N. Navab, editors, *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 583–590. Springer, 2013.

[73] H.-I. Suk, S.-W. Lee, and D. Shen. Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis. *NeuroImage*, 101:569–582, 2014.

[74] C. Szegedy, W. Liu, Y. Jia, et al. Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, June 2015.

[75] C. Szegedy, S. Ioffe, V. Vanhoucke, et al. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI Conference on Artificial Intelligence*, 2017.

[76] L. J. Thal, K. Kantarci, E. M. Reiman, et al. The role of biomarkers in clinical trials for Alzheimer disease. *Alzheimer Dis Assoc Disord*, 20(1):6–15, 2006. 16493230[pmid].

[77] P. A. Thomann, E. Kaiser, P. Schönknecht, et al. Association of total tau and phosphorylated tau 181 protein levels in cerebrospinal fluid with cerebral atrophy in mild cognitive impairment and Alzheimer disease. *J Psychiatry Neurosci*, 34(2): 136–142, Mar 2009. 0001585-200903000-00007[PII].

[78] P. M. Thompson, K. M. Hayashi, G. de Zubicaray, et al. Dynamics of gray matter loss in Alzheimer's disease. *Journal of Neuroscience*, 23(3):994–1005, 2003.

[79] N. J. Tustison, B. B. Avants, P. A. Cook, et al. N4itk: Improved n3 bias correction. *IEEE Transactions on Medical Imaging*, 29(6):1310–1320, June 2010.

[80] N. Tustison and B. Avants. Explicit b-spline regularization in diffeomorphic image registration. *Frontiers in Neuroinformatics*, 7:39, 2013.

[81] S. van der Walt, S. C. Colbert, and G. Varoquaux. The numpy array: A structure for efficient numerical computation. *Computing in Science Engineering*, 13(2):22–30, March 2011.

[82] C. Wachinger and M. Reuter. Domain adaptation for Alzheimer's disease diagnostics. *NeuroImage*, 139:470–479, 2016.

[83] A. Ward, S. Tardiff, C. Dye, et al. Rate of conversion from prodromal Alzheimer's disease to Alzheimer's dementia: A systematic review of the literature. *Dementia and Geriatric Cognitive Disorders Extra*, 3(1):320–332, 2013.

[84] M. Weiler, F. Agosta, E. Canu, et al. Following the spreading of brain structural changes in Alzheimer's disease: A longitudinal, multimodal MRI study. *Journal of Alzheimer's Disease*, 47(4):995–1007, Aug 2015.

[85] M. Weiler, F. Agosta, E. Canu, et al. Following the spreading of brain structural changes in Alzheimer's disease: a longitudinal, multimodal mri study. *Journal of Alzheimer's Disease*, 47(4):995–1007, 2015.

[86] K. Q. Weinberger, J. Blitzer, and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. In Y. Weiss, P. B. Schölkopf, and J. C. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 1473–1480. MIT Press, 2006.

[87] B. P. Welford. Note on a method for calculating corrected sums of squares and products. *Technometrics*, 4(3):419–420, 1962.

[88] B. T. Wyman, D. J. Harvey, K. Crawford, et al. Standardization of analysis sets for reporting results from ADNI MRI data. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, 9(3):332–337, May 2013.

[89] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision (ECCV)*, pages 818–833. Springer, 2014.