Classification of multimodal MRI images using Deep Learning : Application to the diagnosis of Alzheimer's disease.

Karim Aderghal

THÈSE EN COTUTELLE PRÉSENTÉE
POUR OBTENIR LE GRADE DE

DOCTEUR DE

L'UNIVERSITÉ DE BORDEAUX

ET DE L'UNIVERSITÉ IBN ZOHR

.....

ÉCOLE DOCTORALE DE MATHÉMATIQUES ET D'INFORMATIQUE
SPÉCIALITÉ : Informatique
Par **Karim, ADERGHAL**

Classification des images IRM multimodales par l'apprentissage profond: Application au diagnostique de la maladie d'Alzheimer

---

Soutenue le: 26/02/2021

Membres du jury :

Pascal Desbarats	Professeur des universités	Université de Bordeaux	Président du jury
Christine Fernadez-Maloigne	Professeur des universités	Université de Poitiers	Rapportrice
Mohamed Sadgal	Professeur des universités	Université de Marrakech	Rapporteur
Mohamed El Adnani	Professeur des universités	Université de Marrakech	Rapporteur
Mohamed El Ansari	Professeur des universités	Université d'Agadir	Rapporteur
Su Ruan	Professeur des universités	Université de Rouen	Examinatrice
Jenny Benois-pineau	Professeur des universités	Université de Bordeaux	Directrice de thèse
Karim Afdel	Professeur des universités	Université d'Agadir	Directeur de thèse
Gwénaëlle Catheline	Maitre de conférences	Université de Bordeaux	Invité

Titre : Classification des images IRM multimodales par l'apprentissage profond: Application au diagnostic de la Maladie d'Alzheimer.

Résumé: Dans cette thèse, nous nous intéressons à la classification automatique des images IRM cérébrales pour le diagnostic de la maladie d'Alzheimer (MA). Nous cherchons à construire des modèles intelligents qui fournissent au clinicien des décisions sur l'état de la maladie d'un patient à partir de caractéristiques visuelles extraites d'images IRM. L'objectif consiste à classifier les patients (sujets) en trois catégories principales : sujets sains (NC), sujets atteints de troubles cognitifs légers (MCI), et sujets atteints de la maladie d'Alzheimer (AD). Nous utilisons des méthodes d'apprentissage profond (Deep learning), plus précisément les réseaux neuronaux convolutifs (CNN) basés sur des biomarqueurs visuels à partir d'images IRM multimodales (IRM structurelle et l'IRM de tenseur de diffusion - DTI), pour détecter les changements structurels dans le cerveau, en particulier dans la région hippocampique du cortex limbique. Nous proposons une approche appelée "2-D+ε" appliquée sur notre ROI (Region-of-Interest): hippocampe. Cette approche permet d'extraire des coupes 2D à partir de trois plans (sagittale, coronale et axiale) de notre région en préservant les dépendances spatiales entre les coupes adjacentes selon chaque dimension. Nous présentons une étude complète de différentes méthodes artificielles d'augmentation de données, ainsi que différentes approches d'équilibrage de données pour analyser l'impact de ces conditions sur nos modèles pendant la phase d'entraînement. Ensuite, nous proposons nos méthodes pour combiner des informations provenant de différentes sources (projections/modalités) avec notamment deux stratégies de fusion (fusion précoce et fusion tardive). Enfin, nous présentons des schémas d'apprentissage par transfert en introduisant trois cadres : (i) un schéma inter-modale (IRM structurelle et DTI), (ii) un schéma inter-domaine qui implique des données externes (MNIST), (iii) et un schéma hybride avec ces deux méthodes (i) et (ii). Les méthodes que nous proposons conviennent à l'utilisation des réseaux (CNN) peu profonds pour les images IRM multimodales. Elles donnent des résultats encourageants même si le modèle est entraîné sur de petits ensembles de données, ce qui est souvent le cas en analyse d'images médicales.

Mots clés : Maladie d'Alzheimer (MA), Imagerie par Résonance Magnétique (IRM), Apprentissage Profond, IRM structurel, IRM de tenseur de diffusion (DTI), Déficit cognitif léger (MCI), réseaux neuronaux convolutifs (CNN), Apprentissage profond (DL), Transfert d'apprentissage, Multimodalitie, Traitement d'image, Classification des images

Title : Classification of multimodal MRI images using Deep Learning: Application to the diagnosis of Alzheimer's disease.

Abstract: In this thesis, we are interested in the automatic classification of brain MRI images to diagnose Alzheimer's disease (AD). We aim to build intelligent models that provide decisions about a patient's disease state to the clinician based on visual features extracted from MRI images. The goal is to classify patients (subjects) into three main categories: healthy subjects (NC), subjects with mild cognitive impairment (MCI), and subjects with Alzheimer's disease (AD). We use deep learning methods, specifically convolutional neural networks (CNN) based on visual biomarkers from multimodal MRI images (structural MRI and DTI), to detect structural changes in the brain hippocampal region of the limbic cortex. We propose an approach called "2-D+ε" applied to our ROI (Region-of-Interest): the hippocampus. This approach allows extracting 2D slices from three planes (sagittal, coronal, and axial) of our region by preserving the spatial dependencies between adjacent slices according to each dimension. We present a complete study of different artificial data augmentation methods and different data balancing approaches to analyze the impact of these conditions on our models during the training phase. We propose our methods for combining information from different sources (projections/modalities), including two fusion strategies (early fusion and late fusion). Finally, we present transfer learning schemes by introducing three frameworks: (i) a cross-modal scheme (using sMRI and DTI), (ii) a cross-domain scheme that involves external data (MNIST), and (iii) a hybrid scheme with these two methods (i) and (ii). Our proposed methods are suitable for using shallow CNNs for multimodal MRI images. They give encouraging results even if the model is trained on small datasets, which is often the case in medical image analysis.

Keywords : Alzheimer's Disease (AD), Magnetic Resonance Imaging (MRI), Diffusion Tensor Imaging (DTI), Mild Cognitive Impairment (MCI), Convolutional Neural Network (CNN), Deep Learning, Transfer learning, Multi-modality, Image processing, Image classification

---

Unité de recherche

Univ. Bordeaux, CNRS, Bordeaux INP, LaBRI, UMR 5800, F-33400, Talence, France.

LabSIV, Faculty of Sciences, Department of Computer Science, Ibn Zohr University, Agadir, Morocco.

Univ. Bordeaux, CNRS, UMR 5287, Institut de Neurosciences Cognitives et Intégratives d'Aquitaine (INCIA), Bordeaux, France.

Acknowledgments

First of all, I would like to express my gratitude to my supervisors Pr. Jenny Benois-Pineau and Pr. Karim Afdel for providing this opportunity and their endless support and advice on this Ph. D. I would also like to thank Dr. Manuel Boissenin, Dr. Pierrick Coupé and Dr. Gwenaëlle Catheline for their kindly fruitful discussions, especially during the early stages of this Ph.D. We also tank Mr. Boris Mansencal for his kindly help in the technical platform.

Furthermore, My sincere gratitude goes to colleagues and professors in LaBRI. I would like to name professors Pascal Desbarats, Akka Zemmari, Carole Blanc, Marie Beurton-Aimar, Ahmed Toufik, Sylvain Lombardi, Simon Laurent and Denis Barthou. Ph.D. students Trang, Karim, Abraham, Tidiane, Pierre-Etienne, Paul, Jason, Chahrazed, Attila, Zakaria and many others who made this journey exceptional.

Many thanks also go to the administrative team in LaBRI, Maïté Labrousse, Cathy Roubineau, Sylvaine Granier, Elia Meyere, Sylvie Le-Laurain and Céline Michelot for their help and support.

A very special thanks to ADNI and its collaborators for their great efforts, large amounts of work and willingness to share their data, without which this thesis and the new work described herein would not be possible.

This research was supported by the Toubkal Project: AlClass N 34697TE, and The University of Bordeaux & Bordeaux-INP Enseirb-Matmeca.

I cannot be grateful enough to my family, who did everything for my success, my brothers and sisters.

Résumé

Introduction

Maladie d'alzheimer

La maladie d'Alzheimer (MA) est l'une des formes les plus courantes de démence pour laquelle il n'existe pas encore de remède ou de traitement efficace. Il s'agit d'une maladie dégénérative progressive et irréversible qui dévaste les cellules du cerveau humain et provoque la démence chez les personnes âgées, principalement celles de 65 ans ou plus. La maladie est une priorité mondiale majeure en matière de santé publique, qui s'est considérablement accrue au cours des dernières décennies.

Aujourd'hui, nous estimons que plus de 44 millions de personnes sont touchées par la maladie dans le monde, et il y a environ 7,7 millions de nouveaux cas chaque année. Selon les prévisions de l'organisation mondiale de la santé (OMS), ce nombre va presque doubler tous les 20 ans, pour atteindre 75 millions en 2030 et 131,5 millions en 2050. Autrement, toutes les 67 secondes, quelqu'un dans le monde développe (MA). La croissance des soins de santé de la (MA), outre le fait qu'elle constitue un important problème social et économique, est une source de préoccupation due au fait que la maladie dévaste non seulement les personnes touchées, mais également la famille et les aides-soignants. Ces derniers ont la lourde tâche de prendre soin du patient.

De nos jours, de nombreux projets de recherche s'intéressent à la détection automatique de la maladie, en particulier à son stade précoce, ce qui peut contribuer à retarder le développement et la progression de la maladie ou à conduire un meilleur traitement.

Les phases cliniques de la maladie d'Alzheimer

L'évaluation de la progression de (MA) montre que le patient passe par trois étapes différentes avant d'être converti en pathologie probable de la (MA). Cependant, nous pouvons diagnostiquer le patient en utilisant différentes méthodes et outils pour identifier le degré de gravité de la maladie. En effet, le diagnostic précoce de la maladie peut aider les cliniciens à prescrire des traitements pour aider les patients à préserver le fonctionnement quotidien ou réduire les risques de la maladie pendant un certain temps.
Nous pouvons définir les cas en trois phases cliniques de la maladie comme suit:

AD préclinique: Les personnes à ce stade ne signalent aucun symptôme de troubles cognitifs. Néanmoins, certains changements structurels peuvent se produire dans des régions spécifiques du cerveau, comme dans le sang et le liquide céphalo-rachidien (LCR). En effet, cette phase n'est

pas encore évidente à détecter car la dégénérescence des cellules pertinentes pour la (MA) peut commencer des années, voire des décennies, avant l'apparition des premiers symptômes.

MCI (le déficit cognitif léger): Avec l'âge de la population, certains sujets développent des difficultés de mémoire plus importantes que celles prévues pour leur âge. Dans la jungle de la (MA), ces personnes peuvent être atteintes de la maladie (MCI). La MCI est une phase de transition au cours de laquelle les sujets commencent à développer un certain déclin des fonctions cognitives avant de devenir atteints de la maladie d'Alzheimer. À ce stade, les symptômes liés à la capacité de mémoire et à la réflexion apparaissent progressivement chez les patients eux-mêmes, sans pour autant influencer leur vie quotidienne.

Diagnostic clinique de la MA: C'est le dernier stade du diagnostic de la maladie d'Alzheimer, où les sujets souffrent d'une diminution de leurs capacités de réflexion et de comportement. À ce stade, les symptômes sont déjà lucides et évidents en raison de la dégénérescence des cellules du cerveau, en particulier dans les zones considérées comme atteintes par la (MA).

Etat de l'art sur la classification automatique de la maladie d'alzheimer

Dans ce chapitre 1, nous présentons une étude bibliographique complète sur la problématique de la classification de la maladie d'alzheimer (MA) à l'aide des méthodes d'apprentissage automatique. Ce chapitre se compose de trois grandes partie:

Les biomarqueurs visual pour la détection de (MA):

Nous couvrons une liste exhaustive de tous les biomarqueurs visuels utilisés pour diagnostiquer la maladie (AD), telles que l'épaisseur de la corticale, l'élargissement des ventricules, et l'atrophie de l'hippocampe. Ces biomarqueurs visuels peuvent être classés en plusieurs groupes en fonction du type d'informations qu'ils fournissent: (i) Atrophie globale: morphométrie basée sur les voxels et les tenseurs, par exemple pour mesurer l'élargissement des ventricules). (ii) Atrophie diffusée: atrophie propagé de la matière grise le long du cortex cérébral (épaisseur corticale). (iii) Atrophie focale: morphométrie basée sur le volume dans des régions spécifiques (hippocampe, cortex entorhinal).

Les méthodes d'évaluation appliqués sur les images IRM:

Nous présentons les méthodes de mesure qui permettent de suivre les changements structurels à travers de l'imagerie IRM. Dans cette section, nous faisons un bref détour et examinons certaines des approches les plus utilisées dans le diagnostic de (MA), ainsi pour étudier les variations de toutes les régions vulnérables à la pathologie.

Les algorithmes de classification automatiques pour MA:

Nous présentons les travaux les plus pertinents utilisant les techniques d'apprentissage automatique pour la classification de la (MA). Nous couvrons différentes méthodes et approches, y compris les types conventionnels de classificateurs, les algorithmes d'extraction de caractéristiques et les méthodes de réseaux neuronaux profonds. Cependant, dans cette section, nous mettons l'accent sur seulement un aperçu général des méthodes utilisées à cet égard.

A la fin de ce chapitre, nous conclurons par une analyse comparative de toutes ces méthodes et approches (2D, 3D, etc ...) et les différents biomarqueurs visuels afin de choisir une piste de recherche qui peut nous amener à de meilleurs résultats de notre problème de classification.

Imagerie médical pour le diagnostic de la maladie d'alzheimer

Comme nous avons discuté dans l'introduction de cette thèse , nous pouvons diagnostiquer le patient en utilisant différentes méthodes et outils pour déterminer le degré de gravité de la pathologie : (1) Tests d'évaluation basés sur les scores : (MMSE, ADAS-Cog, CDR-SB etc...), (2) ou par l'analyse de l'imagerie cérébrale: (IRM, PET, DTI etc...).

Dans ce travail, nous utilisons uniquement les méthodes basées sur l'imagerie cérébrale. Le choix d'une modalité d'imagerie médicale pour un diagnostic d'une telle maladie nécessite une connaissance biologique préalable de la pathologie, et aussi une connaissance approfondie de l'imagerie médicale, ainsi dans la première partie de ce chapitre 2, nous introduisons les méthodes d'acquisition pour l'imagerie IRM en expliquant les deux types de modalité d'images utilisées dans ce travail : l'IRM structurelle et l'IRM en tenseur de diffusion (DTI), et la théorie qui les entourent, et pourquoi sont ils convient pour l'évaluation et le traitement de la maladie d'alzheimer. Ensuite, nous présentons une brève description des différentes bases de données publiques ainsi que leurs compositions. Dans la dernière partie, nous fournissons le processus de prétraitement de données et l'extraction automatique des régions d'intérêt (la région hippocampique). En effet, nous avons développé une bibliothèque qui permet de générer une base de données d'images IRM concrètes. Cette étape consistait à suivre une chaîne ordonnée des processus de traitement d'images que ce soit le pré-traitement et post-traitement de toute la base de données. Une sélection automatique des régions d'intérêt (ROI) avec une augmentation artificielle spécifique de données pour nos modèles de l'apprentissage.

Les méthode d'apprentissage profond pour la classification

Dans ce chapitre 3, nous fournissons une présentation complète des algorithmes d'apprentissage profond utilisés dans ce travail. Nous couvrons une analyse des concepts de bases des réseaux de

neurones, plus précisément la théorie derrière les réseaux de neurones convolutifs (CNN) avec leurs principaux composants et les méthodes d'optimisation telles que la méthode de descente du gradient et ses dérivés. En outre, nous abordons les contraintes de la limitation des données et introduisons des solutions alternatives adaptées pour surmonter et affronter le phénomène de sur-apprentissage (overfitting).

Modélisations des réseaux de neurones pour la classification

Dans ces trois chapitres, nous présentons les différents modèles de classification que nous avons développés durant les travaux de cette thèse.

Partie I: 2-D+epsilon

Après avoir étudié l'état de l'art de la classification de la maladie d'alzheimer (Chapitre 1), nous avons fait une synthèse qui compare les méthodes d'évaluation et de mesure. Nous avons entamé une piste de recherche en focalisant uniquement sur le région hippocampique (la région la plus vulnérable à la pathologie). Dans le chapitre 4, nous proposons une approche appelée "2-D+ε" appliquée sur notre région d'intérêt (ROI). Cette approche permet d'extraire des coupes 2D à partir de trois plans (sagittale, coronale et axiale) de notre région en préservant les dépendances spatiales entre les coupes adjacentes selon chaque dimension. Ensuite, nous explicitons notre modèle d'apprentissage (CNN) en proposant une architecture efficace et adéquate pour notre problème de la classification. Nous exposons les différentes démarches suivies pour trouver un modèle qui offre de meilleurs résultats. Nous exposons, par la suite, une étude complète de différentes méthodes d'augmentation artificielles de données, ainsi que différents paramètres d'équilibrage de données pour analyser l'impact de ces conditions sur nos modèles pendant la phase d'entraînement.

Partie II: Fusions et combinaisons des modèles

Dans le chapitre 5, nous décrivons l'extension de l'approche proposée auparavant concernant le modèle basé sur les CNN afin de rendre la classification plus performante. Dans ce sens, nous proposons nos méthodes en appliquant la fusion d'information à partir de diverses sources dans le but de rendre plus robuste le modèle proposé. Nous combinons les informations provenant de différentes sources (projections/modalités) avec deux stratégies de fusion (fusion précoce et fusion tardive). Nous exposons un cadre à flux multiple composé de trois modèles uniques tout en pratiquant l'apprentissage simultanément à partir de différentes projections et modalités.

Partie III: l'apprentissage par transfert

Dans ce dernier chapitre 6, nous exposons les démarches proposées pour l'apprentissage par transfert en introduisant trois cadres: (i) un schéma inter-modale (IRM structurelle et DTI): ici, nous appliquons le transfert d'apprentissage à partir d'un modèle pré-entraîné sur les données IRM structurelles (source) vers l'ensemble de données DTI (cible). (ii) un schéma inter-domaine (en utilisant un jeu de données externe - MNIST): nous introduisons une base de données externe du domaine pour étudier et évaluer la méthode via une base de données complètement différente de notre domaine. (iii) un schéma d'apprentissage hybride incluant ces deux méthodes (i) et (ii). Par conséquent, les méthodes intermodales et hybrides fournissent des résultats prometteurs. Elles convient pour travailler avec des réseaux CNN peu profond avec des images IRM à basse résolution et une donnée de données très limitée. Les résultats fournis témoignent des performances de la classification comparables aux méthodes utilisant un jeu de données volumineux.

Conclusion et perspectives

Enfin, nous présentons une conclusion qui récapitule les principales contributions, en outre nous exposons quelques perspectives et pistes pour les futurs travaux.

Contents

List of Tables

List of Figures

Introduction

In this section, we will briefly expose our research motivation in the field of Computer-Aided Diagnosis of Alzheimer's Disease.

0.1 Motivation

0.1.1 Dementia

As the median age of the population in developed countries increases, aged people are exposed to neurodegenerative diseases with dementia. Dementia is an overall term for a variety of mental illnesses that covers a wide range of specific cognitive pathologies. It is characterized by a decline in memory, loss of thinking and reasoning skills, and behavioral abilities that interfere with daily life. The clinical syndrome of dementia is very subtle, vague, and may not be evident in early-stage patients. Worldwide, it is estimated that 50 million people have dementia, and this number is projected to reach 82 million in 2030 and 152 million in 2050 [18]. Dementia can be caused by a variety of diseases, mainly those related to memory troubles.; the most common is Alzheimer's Disease (AD), followed by Parkinson's disease (PaD), Frontotemporal dementia (FD), Cerebrovascular disease (CDB), and Dementia with Lewy bodies (DLB). According to statistics, solely AD represents around 60% (see Figure 1). Therefore, it is necessary to lead efforts in Alzheimer's disease research to prevent and control its evolution.

0.1.2 Alzheimer's Disease

Alzheimer's Disease (AD) is one of the most common forms of dementia for which there is no known cure or effective treatment thus far. It is a progressive degenerative disease that devastates cells in the human brain and causes dementia for elderly individuals, mostly those aged 65 or older. The disease is a major worldwide public health priority that increased significantly over the last decades [20, 19]. The prevalence of AD and pain both increase with advancing age [9]. Nowadays, numerous research projects investigate on detecting the disease, especially in its early stage - this may help to achieve a delay in the disease's progression or lead to better treatment outcomes.

People affected by AD show several types of symptoms. The sharpness degree and the severity of the onset of these symptoms depend on the disease's level progression. However, clinical symptoms

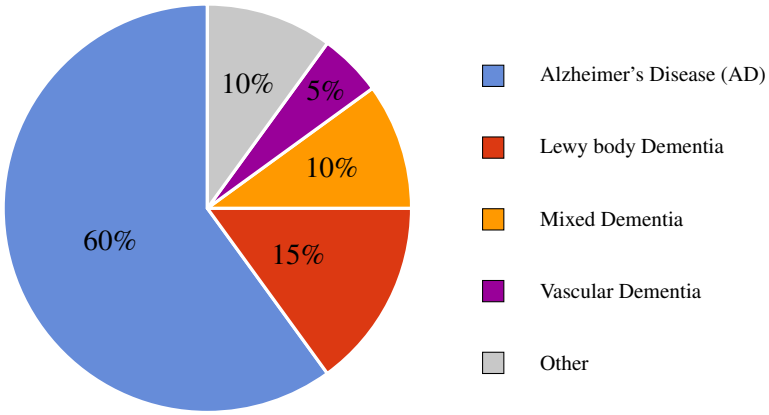


Figure 1: Pie-chart of the leading causes of dementia: patients who have Alzheimer's disease outline around 60% of the dementia [55].

can be classified into different types, related to physiology, psychology, cognitive functions, and behaviors.

Indeed (Today), about 44 million people are diagnosed with AD in the world, and there are 7.7 million new cases each year. According to World Health Organisation predictions, this number will almost double every 20 years, to reach 75 million in 2030 and 131.5 million in 2050; every 67 seconds, someone in the world is developing Alzheimer's disease [9]. In the United States (USA), 5.7 million individuals live with AD in 2018, and this number will approximately triplicate by 2050. Currently, 5.3 million of these people are over the age of 65 [20]. The AD costs around 150 Billion $ per year, 18 Billion hours of unpaid care, and a contribution to the nation valued budget at over 220.2 Billion $. See [18] for a more extended picture and details.

In Europe, in 2009 more than 7 million people are suffering directly from Alzheimer's Disease (AD), and the cost is evaluated to 71 Billion € of direct spendings and a prudent opportunity cost of 89 Billion €, which represents more than 22000 € per sufferer per year [102]. In recent statistical studies, about 900.000 persons are diagnosed with AD in France, and each year 225.000 new cases are identified [108]. In 2020, 3 million people will be affected and impacted by Alzheimer's disease (patients and caregivers); the costs are estimated at more than 14.3 Billion € (5.3 Billion € of medical and paramedical cost + 9 Billion € of social-medical cost). Furthermore, the opportunity costs are estimated at 14 Billion € [8]. In 2015, an estimated 119.000 people were living with Alzheimer's disease in Morocco. This number is expected to nearly quadruple to 460.000 people by the year 2050, and 30,000 new cases are diagnosed each year [160].

The growth of healthcare concern of the AD aside from being a significant social and economic issue is due to the fact that the disease devastates not only the affected people but also the family members and caregivers. The latter have the heavy duty of taking care of the patient. In the following section, we present the main clinical phases of AD, in which a patient may pass through before being

diagnosed with AD. Furthermore, a brief review of the biological transformations that occur in the brain substances will be presented.

0.2 Clinical phases of Alzheimer's Disease

The assessment of Alzheimer's disease progression shows that the patient goes through three different stages before being converted to probable AD condition [157]. However, we can diagnose the subject using different methods and tools to identify the disease's severity degree. Indeed, as stated above, the early diagnosis of the disease can aid the clinicians in prescribing treatment to help patients preserve daily functioning for a while. Medically speaking, AD patient has an accumulation of a protein called beta-amyloid in healthy neurons [64], this process makes neurons weaker and consequently neurons lose their ability to communicate with other neurons. Eventually, as the disease progresses over time, these neurons undermine and die. Thus, this gives rise to what is called atrophy caused by the loss of brain cells (shrinkage). Figure 2 illustrates the difference between a normal and an affected brain.

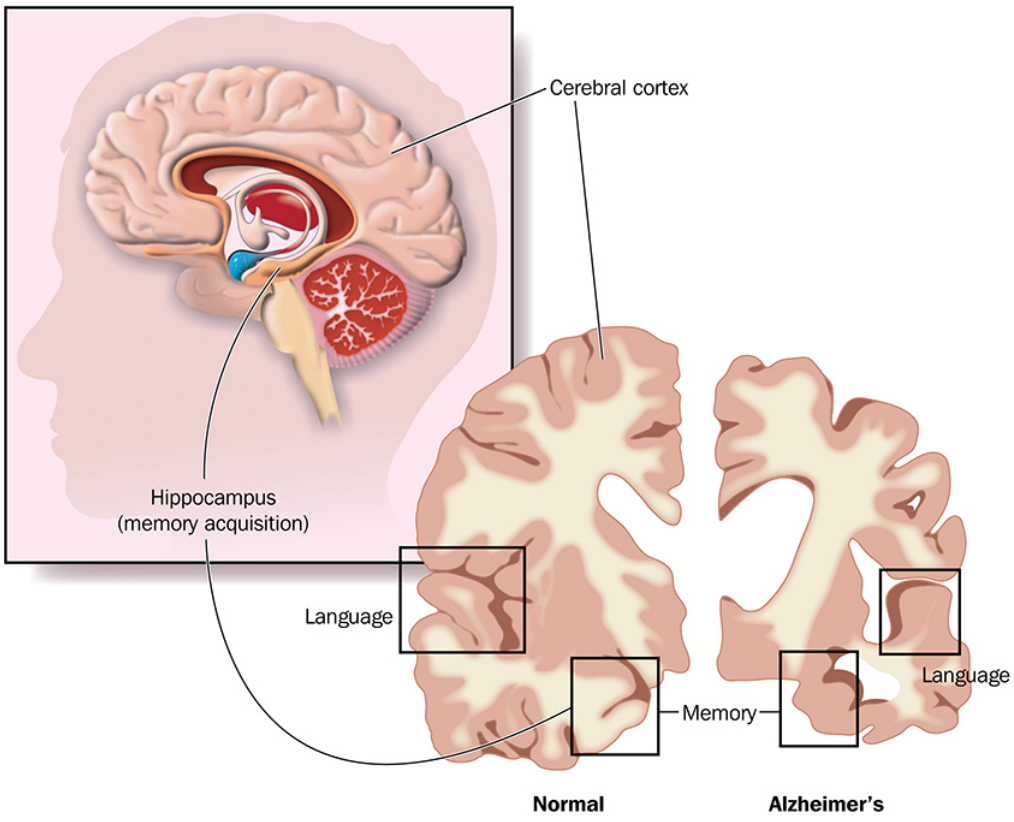


Figure 2: Brain shrinkage instance for two subjects: *(left)* - Normal control brain and *(right)* - brain with Alzheimer's disease.

We can summarize the cases in three clinical phases (stages) of AD as follows:

- **Preclinical AD:** People at this stage do not report any symptoms of cognitive troubles. Nevertheless, some structural changes may occur in specific brain regions, such as in blood and the Cerebro Spinal Fluid (CSF). Indeed, this phase is still not evident to be detected since the degeneration of the cells relevant to AD may start years or even decades before any earlier symptoms arise. Hence, it is a big challenge to interpret structural patterns or biological changes in analyzing the early stage of the AD.

- **Mild Cognitive Impairment (MCI):** With the growing age of the population, some subjects develop memory difficulties which are stronger than those expected for their age. In the AD jungle, these people may have the (MCI) condition. MCI is a transitional phase where subjects start developing some declines in cognitive functions before they become in AD condition. At this stage, symptoms related to the memory capacity and thinking ability arise incrementally for patients themselves, and yet do not influence their daily life. Indeed, longitudinal studies show that not all patients diagnosed in the MCI stage develop AD. This group of patients is considered stable MCI (s-MCI). It was estimated that only 10-15% of people with MCI might develop dementia and then convert over a while to AD (c-MCI). On the other hand, MCI is a challenging group - it contains two separate sub-categories of MCI: early MCI (e-MCI) and late MCI (l-MCI). Patients prone to the disease in the l-MCI stage are more likely to convert to AD [135] contrariwise to those in the e-MCI stage, where they have the earlier symptoms of cognitive troubles.

- **Clinically Diagnosed AD:** Is the last stage of the Alzheimer's Disease diagnosis, where subjects are suffering from decreased thinking and behavioral ability. In this stage, symptoms are already lucid and evident due to cell degeneration in the brain, particularly in areas considered to be stroked by AD. The hippocampal region is such one. In the advanced stage, the prevalence of cells decrying to other regions in the brain reaches a state where the patient becomes unable to complete the daily activities. At this level, the patient is considered (converted) to AD condition. Figure 3 shows current thinking about the evolution from healthy aging to AD condition.

In the next section, we will briefly expose the principal tools used to assess AD, namely the psychological evaluation tests and neuroimaging methods.

0.3 Diagnosis of Alzheimer's disease

Early diagnosis of Alzheimer's Disease (AD) is an important step to help patients getting appropriate treatment, care, and plans for the future. Self-reporting regarding the disease symptoms or information provided by nearby family members can be vital to the disease assessment. On the

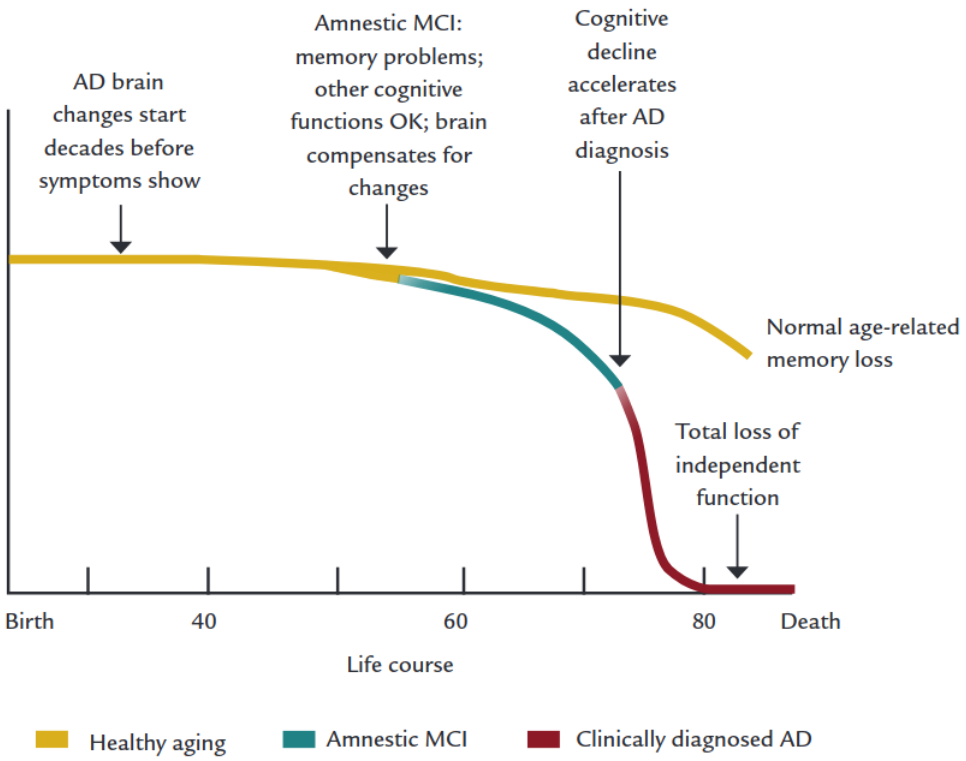


Figure 3: Clinical phases: Charting the course from healthy aging to AD condition [156].

other hand, the disease can be assessed using methods based on physiological signals, psychological evaluation, cognitive analysis, and imaging methods to evaluate abilities associated with Alzheimer's disease condition. However, Lab tests, imaging tools, and some other exams are designed to detect and control disease progression. In this section, we review solely two tools to diagnose the AD: Score-based tests and imaging methods.

0.3.1 Score-based evaluation tests

Mental status tests are conducted to evaluate some skills and abilities relevant to the disease, proffering an overall view of the diagnosed patient. These tests can be performed in different ways and forms to assign scores according to the severity of the disease. The evaluation process comprises methods based on psychological examinations and cognitive analysis. Mini Mental State Evaluation (MMSE) is most commonly used in the diagnostic of Alzheimer's disease, alongside other methods using cognitive tests such as the Severe Cognitive Impairment Rating Scale (SCIRS) and the Alzheimer's Disease Assessment Scale-Cognitive Subscale (ADAS-Cog), which focus on attention, orientation, language, executive functioning, and memory skills.

Below, we briefly describe the most frequently used tests:

- **Mini Mental State Evaluation (MMSE)**: The MMSE is the most widely used cognitive assessment tool designed to test and evaluate patients' cognitive state of a range of mental skills. It is composed of a series of questions asked by health specialists to assess the cognitive domains affected in Alzheimer's disease. The test is used both in clinical and research settings. The maximum score of the MMSE is 30 points. Patients are categorized accordingly to a range of intervals: A score of 20 to 24 indicates mild dementia, 13 to 20 indicates moderate dementia, and less than 12 indicates severe dementia [70].

- **Alzheimer's Disease Assessment Scale-Cognitive Subscale (ADAS-Cog)**: Was developed to measure the level of cognitive dysfunction in AD assessment. It is used in pre-dementia clinical studies where cognitive impairments are more serve. ADAS-cog can detect changes at earlier stages of AD progression [150]. It includes eleven tasks that assess the cognitive domains of memory, language, word recall, naming objects and fingers, commands, constructional praxis, ideational praxis, orientation, word recognition, and comprehension [158].

- **Clinical Dementia Rating–Sum of Boxes (CDR-SB)**: The CDR-SB is a widely used method to stage the severity of dementia using the patient's provided scores. It evaluates the patient's functioning in six domains commonly affected in Alzheimer's disease (AD): memory, orientation, judgment and problem solving, community affairs, home, hobbies, and personal care [137].

However, these clinical tests remain limited to the early diagnosis of Alzheimer's disease. No single test can discover whether if a person presents the disease or no. Despite the widespread utilization of these tools in clinical routine, it is complicated to arrive at a correct disease state diagnosis. Many conditions can exhibit symptoms resembling those in the early stage of AD. These constraints lead to the development of new methods determining specific symptoms and biological brain indications in the earlier stage before these impairments diminish patients' cognitive abilities.

0.3.2 Brain imaging analysis

Brain imaging exams are another way to assess physical, structural, and functional changes in the brain. Since AD is a gradual degeneration that damages brain cells and can occur in different regions, neuroimaging methods are promising tools for effectively studying the disease, particularly for the early detection of AD. The imagery methods provide the ability to observe and track variations in the brain's areas, which are supposed to be affected before arising of any cognitive symptoms. Hence, an early diagnosis may help prevent the disease's progression from expanding to other vulnerable regions. Unlike the mental status test, imaging methods would help health professionals to designate medication to slow the disease's progression in its more initial stage.

Nevertheless, using only scans to examine the patient's condition is not enough, since they can not provide the final stage of the disease. Here is an overlap in what doctors consider a standard age-related change in the brain and abnormal change. However, brain imaging tools can avert other causes, such as brain tumors or the distinguishing of different degenerative diseases, and provide a baseline of the degree of degeneration.

The brain-imaging technologies most often used are:

- Magnetic resonance imaging (MRI): MRI uses powerful radio waves and magnets to produce detailed images of the brain. It is considered to be among the safe and painless exams compared to other imaging methods.
- Computerized tomography (CT): CT scans use X-rays, which are aimed to generate cross-sectional images (or slices) of the brain. These slices are called tomographic images and are collected together to form 3D images.
- Diffusion-weighted imaging (DWI): DWI is a form of MR imaging that measures the random motion of water molecules found in different brain tissues. In general simplified terms, highly cellular tissues, or those with cellular swelling exhibit lower diffusion coefficients. Diffusion is particularly useful in tumor characterization and cerebral ischemia.
- Positron emission tomography (PET): scans have recently been developed that detect clusters of amyloid proteins (plaques), which are associated with Alzheimer's dementia; however, this type of PET scan is typically used in the research setting.

The human brain is a very vulnerable substance that can be harmed by high-power rays. Thus, the use of painless and non-invasive methods can produce relevant information without compromising the brain. MRI techniques use a powerful magnetic field and radio waves to produce high-resolution 3D detailed internal brain structures (commonly known as Head MRI or cranial MRI). These technologies are suitable for brain disease analysis since they do not use ionizing radiation (x-rays). MRI scans allow clinicians and researchers to investigate the brain structures and seamlessly identify AD disease stages by visual or automatic image analysis.

Therefore, in this thesis, we work with MRI modalities widely used for AD diagnosis, especially with the structural Magnetic Resonance Imaging (sMRI) and the Diffusion Tensor Imaging (DTI). Chapter 2 exhibits the acquisition methods for these two modalities and provides the reasons and facts on which we have based our choice for selecting suitable images for AD study.

0.4 Computer-aided diagnosis for AD classification

Computer-Aided Diagnosis (CAD) systems assist the clinicians and radiologists in detecting and analyzing diseases; it yields support and information that medical professionals necessitate understanding better the diseases and their evolution in a short time. CAD comprises many technologies that merge elements of artificial intelligence and computer vision alongside radiological and pathology image processing.

The needs of accurate and seamless AD diagnosis lead to integrating intelligent modules to manage and interpret medical information that helps professionals' health sustain reliable decisions. Conventional machine learning algorithms such as Support Vector Machines (SVM), Decision trees, linear and logistic regression, and Artificial Neural Network (ANN) have been extremely explored in investigating AD classification until the apparition of powerful computer units. Indeed, there are whole swaths of models that have been developed over the years for AD classification and detection using engineered-features algorithms [153]. This kind of method needs reduced and alleviated vectors of features to feed models, probably yielding loss information. Although these models provide competitive results in terms of accuracy and other scores, they recommend the pre-reduction of representative information since they cannot use substantial data dimensions. However, it will be more beneficial to introduce end-to-end solutions that include full information and provide robust brain diagnosis results.

With the advent of high-powered computers equipped with strong GPUs, the return to the application of deep methods is becoming ubiquitous. These methods have reached a level of maturity that allows them to be used for various intelligent diagnostics. Deep learning methods have an extreme number of parameters to train, which allows going insight and encompass more input data to better construct intelligent models, unlike conventional methods that are bounded, as they use a merely small vector of restricted information (features /signatures).

An improvement of the accuracy was reached nearly up to 10% on sMRI data, especially for AD/NC binary classification compared to conventional algorithms [62, 76]. Thus, these limitations and constraints lead to bring and integrate deep and robust models instead of implementing conventional methods for better disease diagnosis.

Therefore, in this thesis, we investigate the deep learning models for AD classification using multi-modal MRI due to its success and emergence.

0.5 Challenges and Objectives

MRI imaging provides comprehensive information concerning the anatomy of the brain for the diagnosis of AD. As noted in the previous section, AD can progress by infecting brain cells before the onset of any associated cognitive symptoms. Hence, the classification of the selected patterns

via the MRI images to be analyzed should focus on areas that are more at risk than others of being affected by AD degeneration.
This thesis aims to investigate the automatic AD classification problem through MRI multi-modal imaging, namely, structural MRI and DTI modalities. We identify solely reliable local visual biomarkers, e.g., the hippocampus region, of the AD rather than entirely working on the brain in order to provide valuable information for diagnosing the disease. However, categorizing brains into the appropriate classes (AD, MCI, and NC) using MRI images remains a challenge for intelligent classifier models. Indeed, the setting up of these models using deep learning methods needs a large dataset for training models - medical imaging datasets are not often available to deploy deep networks. Furthermore, since the research topic is interdisciplinary, it needs strong knowledge and background in the medical domain alongside computer science skills. In general, raw data needs robust preprocessing and segmentation processes; the latter is crucial if the treatment is focused on particular regions of the brain, for instance.
This research aims to design end-to-end classifiers that can recognize AD brains from MCI/NC, robust and less resource-consuming (time and complexity) on materials hardware, and achieve high classification accuracy. Hence, we attempt and investigate to design original concepts and models to answer these issues.

We can summarize the challenges of our work as follows:

- Using Deep learning methods for AD classification (small dataset - medical domain);
- Focusing only on a small region of the brain that requires domain knowledge for selecting the reliable Alzheimer's Disease biomarker (ROI);
- Restricted region (ROI-level method) results in a reduced amount of information and drives limited data to well-train AD classification models;
- Determining a suitable approach to deal with a small dataset (medical data are not often sufficient for deep learning models);
- Improve AD classification performance by combining models and data from multiple sources;
- Benefit from transfer learning to improve training behaviors to bypass overfitting phenomena;

0.6 Thesis contribution and organization

0.6.1 Summary of contributions

The work presented in this thesis includes the following contributions:

- We develop classification models for AD/MCI subjects separation from NC using 2D Multi-modal images rather than 3D MRI Volumes;
- We propose a specific data augmentation to tackle the problem of the limitation of the amount of data;
- We use the domain knowledge of the acquired MRI and Alzheimers disease characteristics to extract appropriate features from the most involved ROIs in AD: Hippocampus (HIPP);
- We propose an intermediate and late fusion of networks to improve the discriminating power of classifiers. We apply this approach not only to discriminate AD and NC but also to recognize the most difficult subject class (MCI);
- We introduce the Transfer Learning approach using cross-modal and cross-domain schemes: we evaluate the benefits of using this approach on medical and non-medical datasets;

Chapter	**Methods**	**Subjects**			**Modality**	**Accuracy**		
		AD	**MCI**	**NC**		**AD vs. NC**	**AD vs. MCI**	**MCI vs. NC**
Chapter (4)	M1	188	399	228	sMRI	82.8%	64.7%	61.8%
Chapter (5)	M1	188	399	228	sMRI	80.15%	66.40%	57.56%
	M2	188	399	228	sMRI	89.84%	63.28%	66.25%
	M3	188	399	228	sMRI	91.41%	69.53%	65.62%
Chapter (6)	M1	252	672	627	sMRI+DTI	92.11%	74.41%	73.91%
	M2	64	273	399	sMRI+DTI	86.83%	71.45%	69.85%
	M3	-	-	-	sMRI+DTI	92.30%	79.16%	78.48%

Table 1: Preview results: main binary AD classification results covered in this thesis.

0.6.2 Thesis outline

The remainder of the thesis is structured as follows:

- **Chapter 1:** In this chapter, we will summarize the related works. We will present the state-of-the-art methods and approaches relevant to the AD diagnosis, particularly classification and detection methods. First, we introduce the main visual biomarkers used for accurate Alzheimer's disease assessment. Second, we present a set of measurement approaches to different brain structures comprising different imaging modalities. Finally, we compare different studies using classifiers based on both conventional algorithms and deep learning methods.

- **Chapter 2:** This chapter consists of two main parts: first, we will present the acquisition method for both types of image modalities (sMRI and DTI). We will briefly explain the technology behind them, and why they are suitable for the diagnosis of AD, then we will describe the data set used in the elaboration of this thesis. Next, we cover the processes of data cleaning and preparation; lastly, we will present all the schemes concerning the extraction of the region-of-interest (ROI).

- **Chapter 3:** In this chapter, we are going to cover a whole presentation of the deep learning algorithms used in our AD classification implementation. It includes a theoretical analysis of CNN networks with their mains components and the optimization methods such as gradient descent and their derivatives. On the other hand, we will discuss the constraint of data limitation and introduce alternative AD-adapted solutions to overcome and confront the overfitting phenomenon.

- **Chapter 4:** In this chapter, we are going to present our effort in designing effective CNN model for AD classification. We will introduce and discuss the novel concept called "2-D+ε", which is well-suited for our problem. Besides, we will provide an overview of different settings for balancing datasets in order to analyze their result on accuracy metric. We will exhibit that the proposed framework has good results in terms of accuracy, and our experiments show that it achieves state-of-the-art performance.

- **Chapter 5:** In this chapter, we will further extend our proposed CNN-based model to consolidate the performances for AD classification. We will introduce different strategies to enhance the robustness of our models, applying the fusion from multiple sources. We will provide a multiple stream framework composed of a couple of single models and learn classification simultaneously from different projections and modalities.

- **Chapter 6:** In this chapter, we are going to introduce the transfer learning approach. We will evaluate various types of transfer learning through the following mechanisms: (i) cross-modal (sMRI and DTI) and (ii) cross-domain transfer learning (using external dataset- the MNIST dataset) (iii) a hybrid transfer learning through both of them.

Chapter 1

Alzheimer's Disease classification state-of-the-art: Background and Literature review

---

1.1 Introduction

In this chapter, we present a systematic overview of the state-of-the-art for Alzheimer's Disease (AD) classification. We provide an unbiased comparison of visual biomarkers relevant to AD on which methods focused on discriminating and classifying patients, such as the hippocampus and entorhinal cortex. We present the retrieval approaches used to measure and analyze brain AD, focusing on the most commonly used methods. Eventually, we survey current researches using machine learning algorithms. We pay particular attention to deep learning works that form the basis of this thesis. At the end, we briefly discuss the existing neural models for AD diagnosis to introduce novel methods for improving performances.

Highlights:

- We present a brief review of imagery (visual) biomarkers for Alzheimer's Disease diagnosis;
- We provide the state-of-the-art on the classification of (AD) using machine learning algorithms, by interesting on Deep learning works using MRI modalities;
- We summarize the different used approaches in the field applied either local region or whole brain for classification problem;
- We give a benchmark of methods using engineered features classifiers and Convolutional Neural Networks (CNN);

1.2 Visual biomarkers for Alzheimer's disease detection

Alzheimer's Disease (AD) is a neurodegenerative disease that causes structural changes in the brain's regions regarded as sensitive to memory and cognitive functions. These microscopic changes accompanied by progressive brain atrophy lead to continuous deterioration due to the death of neurons, followed by synaptic dysfunction. The use of MRI imaging technique has attracted considerable interest as a tool to observe the phenomenon in which we can measure and analyze the volumetric brain atrophy, notably in regions identified as biomarkers for AD. Several studies were carried out on the monitoring of atrophy or abnormal changes in different parts of the brain. We can review brain regions variations as follows:

- A global atrophy and ventricle expansion with increasing amounts of CSF;
- A spread gray matter atrophy along the cerebral cortex;
- Focal atrophy focused on the medial temporal lobe, particularly in the hippocampus and entorhinal cortex;

Numerous studies assessing structural brain variances have been reported in the literature, demonstrating the atrophy of AD and prodromal AD that is spatially distributed over many brain regions. The spread degeneration starts from the medial temporal lobe structures that include the entorhinal cortex and hippocampus to encompassing the whole cortex's brain, involving lateral and inferior temporal structures, anterior and posterior cingulate [36]. In this section, we will explore the imaging biomarkers used to diagnose Alzheimer's Disease, by which we can observe and measure the evolution stages in affected brains.

1.2.1 Cortical thickness

The cerebral cortex is a key element in many brain imaging studies. It is the outer covering of Gray Matter (GM) over the brain's hemispheres and is typically measured of 2-3mm. The cortex comprises areas of the posterior parietal lobe, the temporal lobe, and the anterior part of the occipital lobes responsible for essential functions such as memory, language, abstraction, creativity, judgment, emotion, and attention.

In AD studies, the cortex undergoes changes in thickness in a characteristic pattern during disease development and progression [185]. The measurement of these changes is a powerful approach that can expose worthy information about Alzheimer's disease progression. Indeed, various studies have reported that patients with AD, or who have cognitive impairment MCI, have restriction and atrophy of the cerebral cortex [38]. Therefore, cortical thickness estimation in vivo through MRI is an essential technique for diagnosing and understanding AD's progression. Different approaches have

been reported to automate this measurement of cortical thickness from (MRI) data [11]. Figure 1.1 illustrates a brain instance where the cortex is selected and measuring of the thickness is performed.

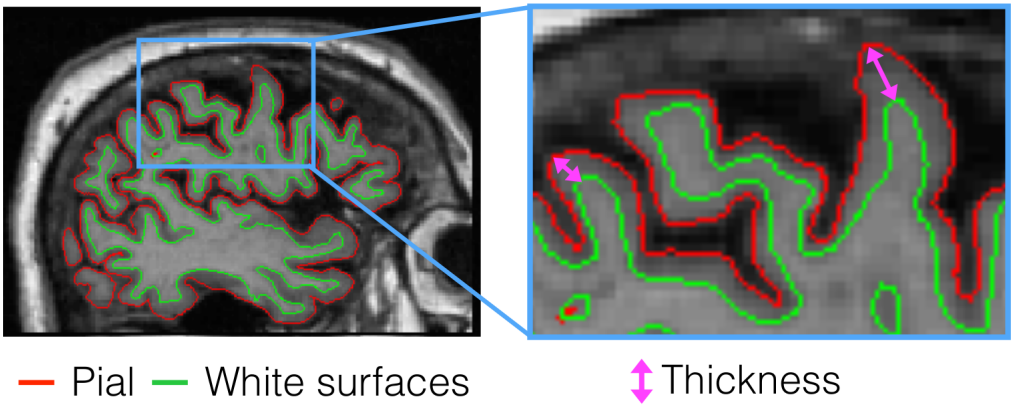


Figure 1.1: Illustration of Cortical thickness measurement through MRI imaging.

1.2.2 Hippocampus volume atrophy (loss)

Historically, the Hippocampus (HIPP) has enormous potential as a biomarker of AD pathology. It is the most discriminant in terms of diagnostic categorization [72, 73]. Hippocampus can be considered damaged when the first clinical symptom of Alzheimer's disease appears [34]. Indeed, the hippocampus is a structure within the temporal lobe that plays a significant role in long-term memory and cognitive information. Therefore, these two brain regions (left and right) hippocampus have become the primary target of many studies in the detection and recognition of AD.

Due to the high-resolution images that offer MRI technique, various studies have shown increased rates of hippocampal volume loss in patients diagnosed with Alzheimer's disease [103, 92, 93], and mild cognitive impairment (MCI) [89, 192] on average compared with those on normal control condition. Volume reduction of the hippocampal regions is the most common pronounced change that occurred in affected patients with AD, according to the study presented in [57]; There exists a relationship between hippocampal atrophy rate and memory impairment. The measure of the progression of hippocampal loss can be used as a potential surrogate for therapeutic interventions' efficacy measure.

The Quantification of the hippocampus volume helps establishing the diagnostic of the AD. Authors in [198, 165] demonstrate that over time, the hippocampal region changes the shape and looses volume at different rates, thus distancing the mild stage of AD from the normal state. Consequently, around 15-30% of the hippocampus volume is lost on average at the mild dementia stage, whereas 10-15% at the early stage of AD [170]. Figure 1.2 presents two instances of brains comparing the atrophy between a healthy brain and an AD diagnosed brain.

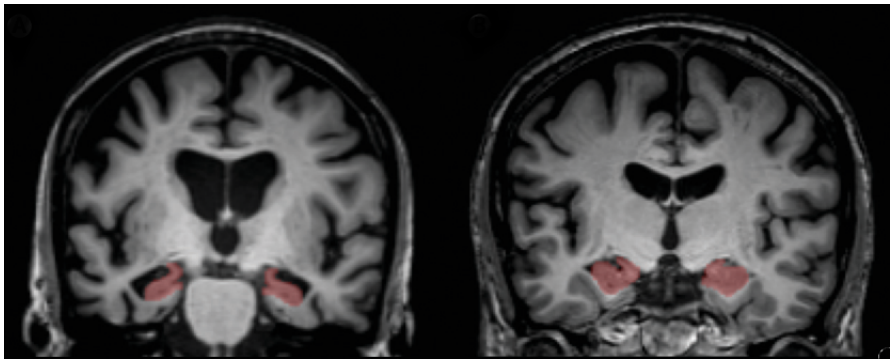


Figure 1.2: A comparison of the hippocampal region atrophy from two subjects (*left*) AD - (*right*) NC.

1.2.3 Entorhinal cortex (ERC)

Entorhinal Cortex (ERC) is a region located in the medial lobe structure alongside the Hippocampus (HIPP) that are vital systems responsible for attention, memory, and cognition. Like the hippocampus, it has been found that AD patients also have volume losses of ERC [61, 58, 207, 98, 33]. Indeed, many studies have demonstrated the medial temporal lobe structures are the first areas of the brain that undergo the earliest changes in AD, in particular, the ERC and HIPP [35]. The tracking of the degeneration progression using structural imaging shows atrophic changes of the entorhinal cortex from where it progresses to the hippocampus and then the limbic system, temporal and parietal lobes, and finally involves the frontal lobes in latest ages of AD [60, 99, 93, 56]. Volumetric analysis of MRI imaging indicates that in MCI stage, the entorhinal cortex and hippocampus show volume reductions of 20-25% relative to normal control.

1.2.4 Ventricles enlargement

Ventricular enlargement (expansion) is another useful visual biomarker known to be susceptible to Alzheimer's Disease alongside other related dementia neurodegeneration diseases. It represents a short-term marker of AD [139]. According to many studies using the MRI images, researchers demonstrate that Alzheimer's Disease causes an enlargement of ventricles in individuals' brain with MCI or AD compared to those measured in normal stage (NC) [4, 132, 155]. This construe the correlation between the volume change of lateral ventricle and AD over time. The level effect of the ventricles area depends on the stage of the disease progression. In fact, in the advanced stage of AD, a more significant expansion in this area can be observed; besides, prominent atrophy and shrinking of Hippocampus (HIPP) and Entorhinal Cortex (ERC), which can be recognized in MR scans [67]. However, the expansion of lateral ventricles may not be a sufficient measure to be specific to MCI or AD. This lead to incorporating other structural information (regions) of the brain, which could serve in diagnosing and identifying the related AD patterns, especially in the early stage. Figure 1.3 shows

typical MRI scans for subjects from the three stages of AD. It presents a coronal view with different enlargement of lateral ventricles.

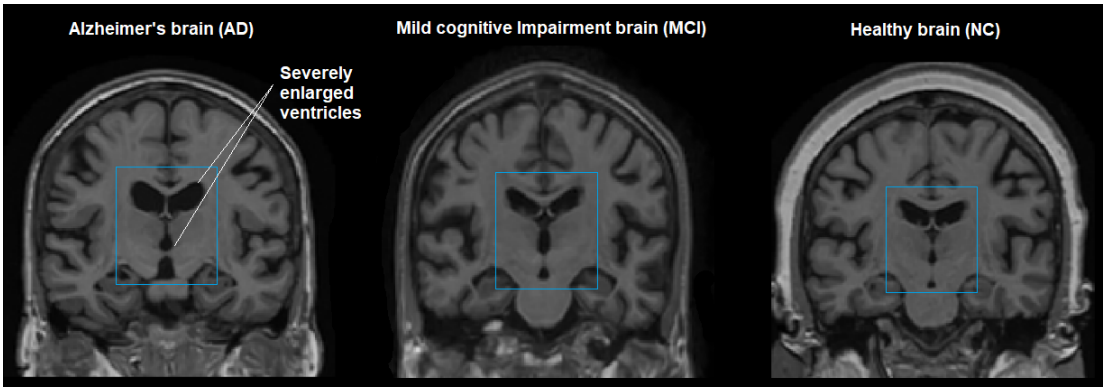


Figure 1.3: Ventricles enlargement severely over stages of AD brain.

1.2.5 Cerebro-spinal fluid (CSF)

Cerebro Spinal Fluid (CSF) is a fluid of the central nervous system that surrounds the brain and spinal cord. Besides, it is also located in the meninges and the central cavities of the brain. The fluid plays a vital role in protecting the brain from external shocks. CSF is qualified as a reliable biological marker of Alzheimer's disease in biofluids. Indeed, several studies have shown that proteins measured in the CSF are associated with Alzheimer's disease and that some features belong to it, such as the tau (T-tau), hyperphosphorylated tau (P-tau), and the 42-amino acid isoform of amyloid-beta (Aβ42) are considered biomarkers for the diagnosis and monitoring of AD progression [133]. Aβ is the most typical CSF measurement focused on the detection of AD.

In imagery analysis, the spread of CSF in some specific brain regions can be used as a visual biomarker for AD diagnosis. Indeed, since AD reveals extensive neuronal loss and atrophy, CSF can replace brain tissue affected by AD. In other words, AD causes degeneration of neuronal cells, leading to atrophy of sensitive region in the brain, particularly the hippocampus; moreover, an enlargement of the ventricles also allows the CSF filling the damaged tissue in these areas. The phenomenon shows a significant amount of fluid around the hippocampus, medial temporal, and ventricles in individuals with AD and MCI compared to those with NC. We can, for example, see and analyze the distribution of CSF fluid in structural MRI images represented by dark areas.

1.3 Methods and approaches for AD evaluation on brain images

Previously, we exposed the conventional imaging biomarkers of Alzheimer's disease. However, to analyze and evaluate the disease's progression, we need robust and effective measurement methods

to track structural changes using MRI imaging of the brain. In this section, we take a brief detour and consider some of the most widely used approaches in AD classification methods, to study the variations of all related-AD brain regions such as cortical thickness, ventricle enlargement, and hippocampus loss volume.

1.3.1 Volume-level methods

Features extraction can be fulfilled either at the whole brain or at some specific local regions. In AD detection literature, numerous methods using the full brain which rely on different approaches to extract representative and relevant volumetric features to the AD disease are reported. The methods consist of involving all regions to discriminate patients from AD/MCI stages. Many studies have been used volumetric frameworks for AD classification. Some of them are focused on the study of White Matter (WM), Gray Matter (GM), or CSF volumes for tracking the disease. Other works used the method to assess the correlation between a specific ROI and the whole brain to compare the classification process's performance. For example, [193] used a whole-brain voxel-based correlative approach to assess the relationships between hippocampal atrophy, WM, and GM. Overall, these methods remain heavy in terms of computation. Besides, the fact that the negative impact and effect of other regions considered is not informative for AD pathology.

1.3.2 ROI-level methods

ROI-based description methods focus on measuring anatomical volumes in predefined areas in the brain. As believed, the AD impacts regions that are related to the memory system and cognitive functions. Thus, it is intuitive to focus exclusively on regions identified as AD biomarkers; they can reflect the disease's stage, namely in the earlier stage. Many works used ROI-based methods in AD diagnosis involving different regions to study the disease progression, such as hippocampal volume [90, 91], cortical thickness [69, 54], and tissue density [209, 211] in specific brain regions using MRI imaging. However, designing frameworks to serve with ROI methods requires a priori knowledge of abnormal areas from a structural or/and function viewpoint to determine regions in the brain. In contrast, in practice, it can not always be inherent. For instance, the disease starts in certain regions in the early stage, and it may progress to span to other ROIs over time. Thus, focusing on a specific partition of the brain could provide outstanding results, although it may produce suboptimal learning performance. Nevertheless, these methods are still better than working on whole-brain methods from the viewpoint of complexity.

1.3.3 Slice-level methods

The use of the slice-based methods reduces volumetric data into a two-dimensional representation. There is undoubtedly a loss of information since the original tissue morphology is in a three-dimensional form. The benefit of the full slice-based methods is that they does not require tissue segmentation; they only take a 2D slice and ignore the rest. However, the selection methods differ from one to another. Many studies have used their specific way to extract 2D image slices from brain scans, whereas others consider standard projections of neuroimaging modalities; the axial, coronal, and sagittal planes. Nevertheless, none of the studies in this category performed a full brain analysis since a 2D image slice cannot include all the brain scan information.

In AD assessment works, the Axial plane is the most widely used, followed by the coronal. However, the latter covers the three most critical AD-related regions in the brain (hippocampal, cortex, and ventricles) [80]. Indeed, many studies employ Axial slice in their frameworks to analyze related AD brain regions, mostly GM structures. For example, the authors in [66] used the slice-based method on the GM volumes by analyzing the Axial view. In the same way, [191] used median axial slices from an MRI, [65] 166 axial slices of GM, and 43 axial slices of fMRI [101]. Moreover, this method is more suitable for measuring cortical thickness biomarkers since the cerebral cortex can be well presented.

1.3.4 Voxel-based morphometry

The voxel-based method is one of automated approaches used in various studies since it has been introduced [206, 15]; It is relatively easy to use and apply it to different types of scans (MRI or PET). It provides credible biological results. This approach measures local tissue of the brain through voxel-wise analysis for identifying pathological changes in discriminative regions in AD diagnosis. It uses the voxel intensity values from the whole brain volume or some tissue components from MRI data. However, it typically requires spatial alignment (coregistration) and normalization, where brain images are aligned to a common space. Most studies that imply this method perform full-brain research in either single-modality or multi-modality mode. In other works, tissue segmentation (such as GM, ROI) was performed on MRI images before feeding the classifier models. To implement classification models using the voxel-based method may require feature dimension reduction technique, especially methods using engineered visual features, but is not necessarily useful in deep structures. Nevertheless, various methods exist to overcome the problem of high feature dimensionality when using classical machine learning methods.

1.3.5 Summary

Many studies show that global brain atrophy is relatively low in discrimination compared to methods using target region atrophy. However, slice-based methods remain useful in the 2D analysis since it gives more representative information of some visual biomarkers, especially from specific dimensions or projections. For instance, focusing on the hippocampal region for AD analysis, the axial and coronal plans exhibit more information than the sagittal plane. Since the hippocampus is a symmetric substance (left and right regions), we can observe both hippocampal regions from these two projections. In contrast, the sagittal plane provides only one single region view. We open a comparison reference to the covered methods by presenting an overview of the advantages and inconveniences in the following Table 1.1.

Finally, we pay particular attention to all analysis methods based on the ROI-level and thickness of the cerebral cortex. These methods require reliable segmentation processes that are paramount to extract the studies' regions and depict the cerebral cortex's surface. Indeed, this segmentation process's achievement is a crucial and essential phase in the medical image processing area. Instead of using manual segmentation methods that need fastidious processing and a considerable amount of time to perform it on a more massive database, several algorithms are used to extract them automatically. The most popular we find those that use SPM [72], FSL [94] and FreeSurfer [68] tools, also online tools like VolBarin [130].

Methods	**advantage**	**drawback**
Volume-based	- Includes all brain regions - Does not require ROI extraction	- Bringing all regions risks confusing the discrimination - 3D information is heavy for computation
ROI-based	- Easily interpretable - Low feature dimension (specific region) - Fewer features can reflect the entire brain	- Produce limited knowledge about the brain regions involved in AD - Ignores detailed abnormalities (between regions)
Slice-based	- Avoids confronting with millions of parameters during training and results in simplified networks	- There is the loss of spatial dependencies in adjacent slices
Voxel-based	- Provide 3D information of a brain scan	- Contains high feature dimensionality and high computation load - Ignores the local information of the neuroimaging modalities as it treats each voxel independently

Table 1.1: A comparison of the advantages and disadvantages of different methods.

1.4 Automatic classification algorithms for Alzheimer's Disease

In this section, we review the most relevant works using machine learning techniques for AD classification. We cover different methods and approaches, including conventional types of classifiers,

feature extraction algorithms, and deep neural network methods. However, in this section, we point only a general overview of the methods used in this regard. Therefore, Chapters 4, 5, and 6, which are our contribution chapters, contain review of specific related works relevant to their contribution.

We present in this section two main types of methods: conventional classifiers using visual features and signature vectors such as SVM, and the second type of methods using deep learning approach, in particular CNN networks.

1.4.1 Methods using engineered Visual Features

Machine learning algorithms, for a long time, have used feature-based methods to reduce data dimensionality rather than working directly on entire data. Visual feature extraction is a crucial phase in decreasing the amount of data and retaining solely valuable information, especially when using heavy multimedia objects such as images and videos. Its principle is to derive enough information to represent the content of an object in feature vector format, and thus reducing computational complexity while preserving useful information. In the literature, a wide range of approved extraction and decomposition algorithms have been introduced in the clinical diagnosis process for AD detection, namely Bag-of-Visual-Words (BoVW), Circular Harmonic Functions (CHF), and ROI-wise based features. Indeed, there are different types of methods for designing visual features from images. However, some of them are more adaptable to the algorithms that are implemented in AD classification studies.

Back in the work of [105], the authors showed the performance of employing the SVM approach for automatic classification purpose of discriminating AD subjects from those with Normal control (NC). Their method was concentrated on the gray matter (GM) voxels. MRI scans were normalized into a standard anatomical space and then were segmented to extract gray matter area. They used the GM density map of the entire brain, together with SVM, and the results were promising. They proved a relatively low GM density in the hippocampal region in AD patients, indicating a strong relationship between the disease and the atrophy in that region [71]. This study was amongst the first ones to highlight the potential of computer-aided diagnostic methods as support where experts are scarce, or simply reduce time and effort for the diagnosis.

A comprehensive survey [154] yielded over 100 papers on MRI-based work for AD classification that compared various applied machine learning methods. These studies used multimodal MRI data, namely structural MRI, functional MRI, PET, and DTI. Half of the studies used only structural MRI. The most used classifiers include various kinds of SVM, Linear Discriminant Analysis, and Logistic Regression. Moreover, most of these methods are characterized by the following phases: feature extraction according to different atlases or Regions of Interest (ROIs), feature selection or dimensionality reduction, and finally, classification. In some cases, particular emphasis was given to

the process of finding ROIs; in others, it lays in the classification itself. Although there are numerous works, addressing the subject, we focus our attention only on some of them.

In a previous work, Ben Ahmed et al. [27, 6, 5] have computed visual features from sMRI scans for AD diagnosis. They have selected Hippocampus and Posterior Cingulate Cortex (PCC) regions as biomarkers. The originality of the work consisted of using a decomposition approach combined with features extractor to cut down information. However, instead of using traditional descriptors such as Sift and Surf, they implemented the Gauss-Laguerre Harmonic Functions (GL-CHFs) approach to capture local directions of the image signal involving the principal decomposition function of the original CHF algorithms. The method is appropriate for the gray-scale MRI modality due to the smooth contrasts it has. After that, they have calculated the signature vector from each projection of ROIs by using a Bag-of-Visual-Words Model (BoVWM) with a low-dimensional dictionary with 300 clusters. Thus for each image they obtained signature with a length of 1800. Afterwards, the signatures vectors were classified using SVM with RBF kernel and 10-fold cross-validation. The method achieves promising results in terms of accuracy, 87% for AD/NC, 78.2% for NC/MCI, and 72.23% for MCI/AD.

Wolz et al. [205] proposed multiple frameworks using both Linear Discriminant Analysis (LDA) and SVM linear classification methods. In order to increase classification accuracy, they establish a combination of different MR-based features extracted from different regions or/and brain structures. Their works are based on hippocampal volume, cortical thickness, tensor-based morphometry, and they used a novel technique based on manifold learning. The hippocampal region underwent the extraction process using the Atlas method. Indeed, they apply the multi-atlas label propagation to select the hippocampus segments for each brain MRI. From a pool composed of hippocampus atlas, they chose a set of them according to the image similarity to the query image. Then they generate a spatial prior from the multiple label maps to obtain a final segmentation of the hippocampus. In the case of cortical thickness, they practice the surface-based method on the MRI volume after registration and intensity normalization for each brain. The method consists of segmentation of the brain into white matter (WM), gray matter (GM), and cerebrospinal fluid (CSF). Moreover, they divide the brains automatically into two separate hemispheres; the cortex's inner and outer surfaces were extracted according to intersections between WM and GM (white matter surface, WMS) as well as GM and CSF (grey matter surface, GMS). Regarding the Tensor-based Morphometry (TBM) approach, they perform a multi-template approach by randomly selecting 30 images from NC, MCI, and AD subjects, where each template was non-rigidly registered to the study image. The determinant of the Jacobian matrix ('the Jacobian') of the deformation was used to measure the voxel-level morphometry. Finally, for classification methods, they employ the two classifiers, as mentioned above, performed either on individual features and or on their combination. As a result, for AD/NC, they achieved 81% accuracy for both HV and CTH, 87% for TBM, and 89% for combing all features

methods using LDA. At the same time, they obtained similar performance for SVM methods except for the combination method; they get lower results (3 points) compared to LDA method.

1.4.2 Methods using Deep learning approach

The arrival of powerful computing units, mainly (GPUs), has made it possible to develop sophisticated models based on machine learning algorithms, precisely on the deep learning methods. Indeed, deep learning methods have quickly become the more popular methodology for various medical domain applications. They encompass different types of models such as CNN, RNN, and Reinforcement Learning. Convolution neural networks (CNN) is one of the main sections of DL widely used for image classification, pattern recognition, and intelligent image segmentation. It has shown an inescapable performance to solve various complex problems, notably decision-making problems. For a complete overview of the CNN method, see Chapter 3.

Unlike those who need visual features, CNN-based architectures do not require engineering of features, since the first (convolutional) layers of CNNs serve as features extractors with trainable filters. The last layers of Deep CNNs represent fully connected Neural Network classifiers (MLP, see section 3.3.3). Hence the process of feature extraction and classification of them is implemented in a complete end-to-end architecture. However, the main constraint on using deep learning methods is that it requires abundant samples for training, which are not often available in the medical area. In this section, we review the most exciting works leveraging Deep learning methods. We present various implementation strategies and their performances for Alzheimer's disease detection, including volumetric and surface-based methods.

2D-CNN approach

Most of the published deep learning works for AD classification use CNNs either to perform slice-based or volumetric classification. There are considerable studies where the input data are composed of 2D slices extracted from 3D volume [163, 162, 164, 31, 173, 191, 115, 200, 42, 140, 66, 80, 83, 87]. The usual deep model here is a couple of convolutional layers paired with pooling layers and followed by fully-connected layers and a softmax layer. For example, 2D CNNs models were designed with several numbers of convolutional layers; two layers in [146], three layers in [184, 152] and five layers in [22]. All these works applied classification only on single-modality data. However, there are works employing 2D models with multimodality, such as [201], where the authors used five layers. Other examples using the 2D CNNs with two convolutional layers that use some MRI image slices from the coronal plan [80]; six convolutional layers taking only one sagittal MRI slice [199].

Some of the works take advantage of the existing CNNs, which had tremendous successes for image classification. The authors in [163, 162, 164] proposed an adaptive pipeline approach for 2-Way classification using structural and functional magnetic resonance imaging (sMRI, fMRI). They

designed a 2D architectures based on two popular networks; the lightest LeNet-5 [112] and GoogleNet [183], by converting images from 3D data to a stack of 2D slices. They resized the images to 28×28 pixels to match the input in the case of the adapted LeNet-5 network. In [31], Billones et al. proposed a modified version of the VGG-Net network [173], which they called DeMNet to classify sMRI images. The model takes 2D images as input with 224×24 resolution for both 2-Way and 3-Way classification. Their work classifies 20 slices separately, selected from the 3D volume. Another work used 2D-CNN approach, the authors in [191] used two networks, a baseline single-layer CNN (only one convolution layer), and a pre-trained ResNet network, they used a single 2D Axial slice per subject (median slice from the 3D volume) as input. They studied the impact of transfer learning from RestNet trained on ImageNet, besides, data augmentation approach. Lee et al. [115] used a modified AlexNet network known as a high-performance pre-trained model. They proposed a data permutation scheme with outlier rejection and slice selection methods, all 2D slices (obtained by permutation from axial, sagittal, and coronal planes) used for training the network.

In summary of these selected works, we found that some of them take advantage of existing CNNs models that resulted in findings successful in natural image classification. However, their approaches have limitations, mainly due to the fact that MRI is three-dimensional data, whereas 2D convolutional filters analyze all slices of a subject independently. Furthermore, there are many different ways to select slices to be used as input. On the other hand, the use of a full 2D slice may impact the performance of the classification task since the slices are not informative as they contain certain brain regions not considered as biomarkers for Alzheimer's Disease. Moreover, they loose spatial dependencies in adjacent slices.

3D subject-level CNN

Recent studies introduced the use of the 3D-based methods that integrate the whole brain. In this approach, the classification is performed on the complete data of the brain, where all regions (and spatial information) are fully included. However, building 3D-CNN models requires a larger quantity of parameters to train than in case of 2D-CNNs. This may lead to overfitting, especially when the dataset is relatively small.

Many works implemented 3D architecture with a couple of 3D convolution layers; four layers in [40, 88], five in [23], and a deep model with twelve layers was designed in [24]. In [116], Li et al. used a combination of multi-modal convolutional networks applied to the whole MRI brain. Their framework comprised two networks; a deep 3D-CNN for hierarchically feature extraction and a multi-scale 3D convolutions to learn features, which were combined with a fully connected layer and softmax layer for classification. In a related method, 3D CNNs were pre-trained with Auto-Encoders with multi convolutional layers, one in [148], or three in [85, 85]. Vu et al. [195] designed two

3D-CNNs networks, each with only one convolutional layer pre-trained with a sparse AE using two MRI modalities, fused within a fully connected layer.

Other studies defined original architectures [23, 24, 43, 85, 116, 195, 196, 197]. One crucial difference between these studies is in the preprocessing step: [23] used a non-linear registration whereas [43] did not perform registration. [116] proposed a more complex framework fusing the results of a CNN and three networks pre-trained with an AE.

Many studies take advantage of popular networks by adapting them to their classification problem. For instance, the authors in [106, 172] readapted the two classical architectures: the ResNet and VGGNet, to encompass the whole brain. Korolev et al, applied two different 3D-CNN approaches; the first, VoxCNN architecture with 21 layers derived from voxResNet, the second, a plain 3DvoxCNN network with four convolutional layers. In both works, their models perform quite well compared to other studies of the same section.

In 3D-subject level approach the number of samples is small compared to the number of parameters to optimize. Indeed, there is one sample per subject, typically a few hundreds to thousands of subjects in a dataset, thus increasing the risk of overfitting.

ROI-based CNNs

Using 3D full brain data remains a heavy approach since it encompasses all brain parts. First, it requires powerful computing resources; second, the whole brain includes surely non-informative regions. For instance, as shown before, 3D patch-level methods take data from the brain volume and slice them into small inputs. However, most of these inputs include some perturbing information, not a visual biomarker for the pathology. Methods based on regions of interest (ROI) overcome this issue by focusing only on specific brain regions [63]. However, using these methods implies apriori knowledge from a long-term experience of the disease studies to select proper ROIs. In this way, the framework's complexity can be decreased since a few inputs are used to feed and train the networks.

The authors in [177, 179] computed around 93 ROIs from MRI and PET data; they extract features only from GM tissue volume. Similarly, S. Liu et al. in [124, 126, 125] used 83 functional ROIs obtained from MRI (GM) and PET. Choi et al. [181] calculated GM tissue volumes of 93 ROIs, and next picked out regional abnormalities utilizing a single deep model of each region. Another work [118, 117, 129] extract 93 ROI-based volumetric features from MRI and the same number of PET features, here the authors applied (PCA) to reduce the dimensionality of data.

In [86, 97], 90 ROIs were extracted from fMRI images and the correlation coefficient between each possible pair of brain regions computed. Ortiz et al. [145], applying a voxel preselection process, they selected 98 ROIs (GM only) from both modalities, MRI and PET, and they conceived a deep architecture for each ROI.

Suk et al. [182] selected 116 ROIs from fMRI images and then trained a deep model on the mean intensities of each ROI; in this way, they found, in an unsupervised and hierarchical way, the non-linear relations between the ROIs. Together with patch-based features of GM and deformation magnitudes (DM) from MRI scans, Shi et al. extracted 113 ROI volumes [169].

Image patches were extracted in each of 62 ROIs of PET [210] or MRI images in [41], while 85 ROIs from PET [128], and 87 ROIs from PET and MRI (GM only) [127] were extracted, from which these ROIs were further used in a patch-based method. In another study, 90 ROIs were extracted, and then a brain network connectivity matrix calculated from multi-modal data [201].

Bhatkoti et al. [30] devised a patch-based representation of different brain sub-regions, including left and right hippocampus, mid-occipital, parahippocampus, vermis, and fusiform. Shakeri et al. extracted morphological features as 3D surface meshes from the hippocampus structure of MRIs [167]. Dolph et al. [59] extracted Fractal Dimension (FD) texture features, together with volumetric, cortical thickness, and surface area features of the segmented hippocampus, from MRIs and then calculated the statistical properties of the Gray-Level Co-Occurrence Matrix (GLCM) to describe the FD feature pattern. Collazos-Huertas et al. [47] used morphological measurements of different parts of MRI scans, including cortical and subcortical volumes, average thickness and standard deviation, and surface area. In another MRI study [53], the two hippocampus were segmented and a local 3D image patch was extracted from the center of each; a deep model was then used for classification. In [121] a non-linear registration was performed to obtain a voxel correspondence between the subjects, and the voxels belonging to the hippocampus 12 were identified after a segmentation implemented with MALP-EM [114]. 151 patches were extracted per image with sampling positions fixed during experiments. Each of them was made of the concatenation of three 2D slices along the three possible planes (sagittal, coronal and axial) originated at one voxel belonging to the hippocampus.

The main drawback of this methodology is that it studies only one (or a few) regions while AD alterations span over multiple brain areas. However, it may allow avoiding overfitting because the inputs are smaller (3000 voxels in our bibliography) and fewer than in methods allowing patch combinations.

Having analyzed the previous works presented above, focusing on specific "region-of-interest" that is known to be a reliable discriminator might be better able to diagnosis the disease, especially when field expertise is involved in the choice of ROI. Consequently, the complexity of the computing is cut down, due to the reason of the reduction of the dimension of the features, knowing that only small regions are used instead of the whole brain.

Author	Methods	Approach	Modalities	Dataset	Instances
Shi et al., [171]	DPN + SVM	ROI-based	MRI-PET	ADNI	202
Suk et al., [180]	DBM + SVM	Voxel-based/Patch-based	MRI-PET	ADNI	398
Li et al., [117]	PCA + RBMs + SVM	ROI-based	MRI - PET - CSF	ADNI	202
Suk et al., [178, 181]	SAE + SVM	ROI-based	MRI-PET-CSF	ADNI	2020
Ortiz et al., [145]	DBN + SVM	ROI-based	MRI - PET	ADNI	275
Sarraf et al., [164, 162]	CNN	Slice-based	fMRI	ADNI	144
Payan et al., [148]	Sparse AEs and 3D CNN	Voxel-based	MRI	ADNI	2265
Liu et al., [124, 126]	Stacked sparse AEs and a softmax layer	ROI-based	MRI - PET	ADNI	311
Liu et al., [123]	3D-CNN for Landmark	Patch-based	MRI	ADNI + MIRIAD	1526
Wang et al., [199]	2D-CNN	Slice-based	MRI	OASIS + Local Data	196
Suk et al., [179]	Sparse regression + 2D-CNN	ROI-based	MRI	ADNI	805
Hosseini et al.,[85]	3D-CNN + Stacked 3D Conv AEs	Voxel-based	MRI	ADNI	240
Lu et al., [127]	DNNs + Stacked AE + Softmax layer	Patch-based/ROI-based	MRI + PET	ADNI	1242
Korolov et al., [106]	3D-CNN - ResNet/VGG-Net	Voxel-based	MRI	ADNI	231
Choi et al., [46]	3D-CNN	Voxel-based	PET (FDG and AV-45)	ADNI	492
Gupta et al., [81]	Sparse AE + CNN	Patch-based	MRI	ADNI	755
Gupta et al. [81]	CNN	Slice-based	sMRI	ADNI	483
Billones et al. [31]	CNN - VGG-Net	Slice-based	sMRI	ADNI	531
Bumshik et al. [115]	CNN - Alexnet	Slice-based	sMRI	ADNI	819
	CNN - Alexnet	Slice-based	sMRI	OASIS	416
Valliani et al. [191]	CNN - ResNet	Slice-based	sMRI	ADNI	660
Cheng et al. [44]	CNN	3D-full brain	sMRI	ADNI	428
Glozman et al. [78]	CNN - AlexNet	Slice-based	sMRI	ADNI	553
Hon et al. [83]	CNN - VGG-Net	Slice-based	sMRI	ADNI	200
	CNN - Inception V4	Slice-based	sMRI	ADNI	200

Table 1.2: Some studies used both engineered features and deep learning method reported in the literature.

1.5 Conclusion

In this chapter, we have provided the state-of-art of MRI-based methods for the Alzheimer's Disease (AD) classification. We have covered the most used approaches treating the related research. However, analyzing and comparing these methods, it is evidenced that deep learning methods show an efficient power for AD classification compared to those associating feature extraction. They are now considered the trend tools that provide high-level learning. On the opposite side, the deep learning methods require a sufficient dataset, mostly when the network model is excessively deep. Furthermore, the ROI-based approaches still turn out to be a suitable method over others for the extraction of the relevant specific biomarker and taking the classification decision on it.

In the next chapter, we will introduce the acquisition methods for MRI imagery, and the flow of dataset preprocessing steps.

Chapter 2

Acquisition Methods and Neuroimaging Data preprocessing

2.1 Introduction

In the previous chapter, we reviewed the state-of-art of AD classification related research by highlighting the most recent works using deep learning methods. We have seen different methods and approaches with different imaging modalities, especially MRI, PET, and DTI; before that, we covered the most AD visual biomarkers on which diagnosis of AD is focused. In this chapter, we introduce MRI imaging formation, specifically structural MRI (sMRI) and diffusion tensor imaging (DTI) modalities used in the preparation of this thesis. We briefly present the fundamental concepts behind the acquisition and formation of these modalities. Next, we cite the most popular datasets by providing a general presentation of their composition. Besides, we give the flow scheme of our contribution to data preprocessing; in this part, we provide all performed steps that aim to obtain a clean and coherent dataset, followed by a post-processing step to produce an aligned and normalized dataset. Finally, we present the approach used to extract AD-related ROIs.

Highlights:

- We present the formation of the multimodality of MRI and the theory behind it;
- We provide a brief detour through the most popular AD datasets;
- We present the fundamental process of data preparation, including denoising, correction, and normalization;
- We outline all the post-processing steps for ROI extraction;

2.2 Magnetic Resonance Imaging (MRI) image formation

Magnetic Resonance Imaging (MRI) is a spectroscopic imaging method used in medical settings to acquire information from the interior of the human body without dissection (in vivo). First developed in 1973 [109], it has many advantages compared to other imaging techniques like conventional X-rays, Computed Tomography (CT), or similar diagnostic methods. MRI is a mature analytical modality used in non-invasive tests to diagnose diseases within clinical medicine and research. The method utilizes radio waves and powerful magnetic fields to produce contrast images that do not rely on ionizing radiation. Indeed, this magnetic field is sensitive to the hydrogen element (in water molecules), which is the main constituent of any biological organ. Therefore, by mapping the position of water molecules, we could detect the contrast in MRI images, as hydrogen atoms behave somehow differently in different tissues of the patient's body. Figure 2.1 presents a cutaway view of the major components of an MRI scanner system.

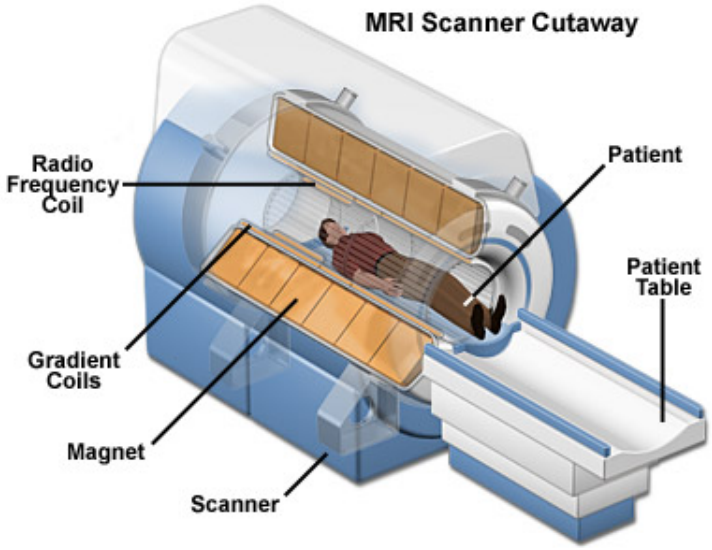


Figure 2.1: A cutaway view of the Magnetic Resonance Imaging (MRI) scanner system [52].

MRI units are based on sophisticated technology that uses radio waves sent into the patient's body to reorient the axes of spinning protons. It excites and detects the change in the direction of the rotational axis of protons found in the water, and thus serve to detect pathological changes deep within an organ [82].

Indeed, inside the MRI scanner system, there are three additional gradient magnets called x, y, and z; each oriented along with a different projection of the body, all of them far less powerful than the main magnet (as illustrated in Figure 2.2); they modify the magnetic field at selective points and work in conjunction with the RF pulses. When we place a patient inside a more powerful magnetic field, the radiofrequency current is pulsed through the patient's body, and every hydrogen atom is forced to align itself with this field. The atoms are stimulated to resist the attraction of the magnetic field.

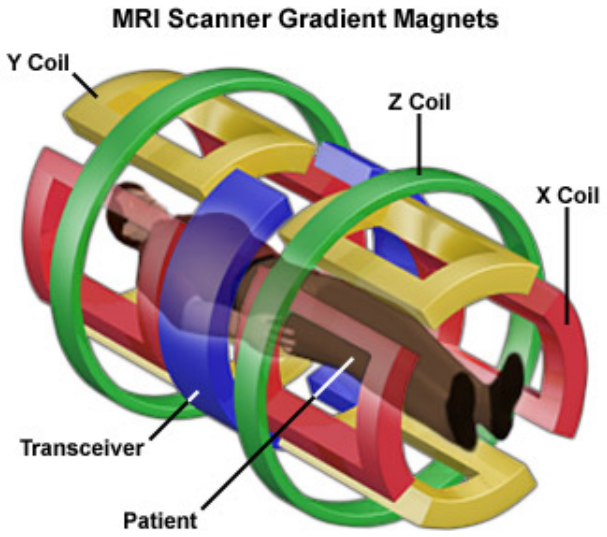


Figure 2.2: A coils view of the Magnetic Resonance Imaging (MRI) scanner system [52].

When the radio-frequency field is deactivated, the protons realign with the static magnetic field, and the MRI sensors detect and capture the energy released during the realignment [52], this depends on the time that protons take to realign with the magnetic field, as well as the amount of energy released. The x, y, and z gradients can be used in combination to generate image slices in any dimension and obtain 3D gray-scale images. We can also differentiate all parts of the body due to the variety of the environment and chemical nature of the molecules [37]. The resulting slices can be seen illustrated in Figure 2.3.

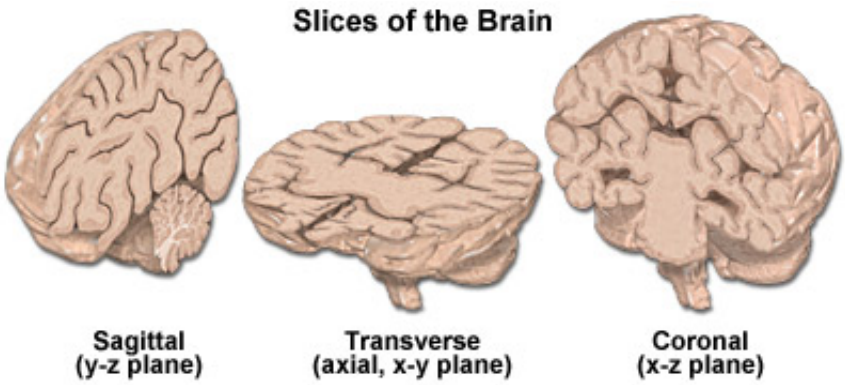


Figure 2.3: An example of human brain slices: *(Sagittal, Axial, and Coronal)* projections [52].

MRI systems are well-suited for working on the body's boneless or soft tissue parts since they do not emit the damaging ionizing radiation found in x-ray and Computed Tomography (CT) imaging, this makes the technology suitable for diagnosis or therapy, especially for the brain. Therefore, the MRI techniques are used in Alzheimer's Disease (AD) diagnosis, since it is a safety tool that produces

high-resolution slice images. It allows us to represent the brain on the axial, coronal, and sagittal planes.

2.2.1 Structural MRI (sMRI)

structural Magnetic Resonance Imaging (sMRI) is the most commonly used modality to monitor and observe structures development in the brain, amongst other modalities, such as PET and DTI. It permits tracking structural changes in the brain and measuring the inevitable atrophy caused by the neurodegenerative aspect of the AD pathology [73]. Moreover, sMRI provides helpful information with which we can affinely describe the shape, size, and integrity of GM and WM structures in the brain.

Here, we present the principle MRI imaging:

- **T1-weighted (T1; short TP and short TE):** The T1 Weighted (also known as the spin-lattice relaxation time T1) is one of the primary pulse sequences in MRI. It provides good contrast between gray matter (dark gray) and white matter (lighter gray) tissues, while Cerebrospinal Fluid (CSF) is void of signal (black). The contrast in these images allows for accurate differentiation of brain structures. We can reach the T1-weighted scans by using an opposite recovery sequence or by inserting short repetition time TR (<750ms) and echo time TE (<40ms) conditions in conventional spin-echo sequences.

- **T2-weighted (T2; long TR and long TE):** Unlike the T1-weighted image, the T2-Weighted MRI is built with long TE and long TR (TR > 2000ms, TE > 80ms). It provides a good contrast between (CSF) which is bright, and brain tissue (dark), GM is light gray, and WM is dark gray. The contrast in T2-weighted allows radiologists to see abnormalities within the ventricles and cerebral cortex better than on T1-weighted images due to the better measurement of water content. Moreover, white matter boundaries are not as clearly defined as in T1-weighted images.

2.2.2 Diffusion Tensor Imaging (DTI)

DTI Concept

Diffusion Tensor Imaging (DTI) is a relatively new imaging method that uses magnetic resonance technology, which was initially introduced by Peter Basser in 1994 [26, 25]. It is a powerful modality technique for inferring tissue structure largely used to map White Matter (WM) pathways in healthy and diseased brain [110, 17]. DTI is a sensitive probe of cellular structure that provides quantitative information, mainly used to measure the substance anisotropy by quantifying the isotropic and

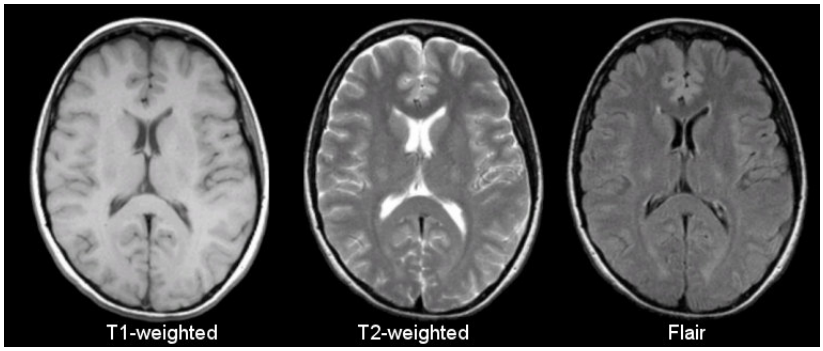


Figure 2.4: Example: Axial slices of T1-weighted (left), T2-weighted (center), and Flair (right) images of brain tissue.

anisotropic water diffusion. For instance, the diffusion in an equitable medium of pure water would be the same in all directions (isotropic), whereas, in an oriented tissue, along different directions (anisotropic) that describes the microscopic tissue heterogeneity [26, 147]. Thus, the method allows us to calculate and measure the distribution of the diffusion directions of water molecules at each spatial point. Indeed, quantifying the random motion of water molecules reflects significant characteristics of microstructural brain tissue. In other words, this new concept makes it possible to capture tissue microstructural properties through diffusion signals that were not possible with traditional anatomical MRI. Therefore, we indirectly obtain the position orientation of axons and the anisotropy of fibrous structures, particularly the white matter bundles of the brain. It should be taken into account that diffusion tensor is not able to entirely represent the crossing of the fiber tracts [204, 187].

DTI measurement

To measure water diffusion amount at different brain positions, a repetitive process of diffusion using the magnetic gradient field in several directions can capture molecules water behaviors at each point in the brain. This process yields an estimated three-dimensional model (called tensor) representing the degree of anisotropy in each voxel (volumetric pixels) in the 3D space. In other words, diffusion imaging introduces extra gradient pulses whose effect "cancels out" for stationary water molecules and causes a random phase shift for molecules that diffuse. Due to their random phase, the signal from diffusing molecules is lost and thus creates darker voxels. This means that white matter (WM) fiber tracts parallel to the gradient direction and thus appear dark in the diffusion-weighted image for that direction. However, the decreased signal (S_i) is compared to the original signal (S_0). The equation allows us to estimate the apparent diffusion coefficient (D) and describes the signal intensity at each voxel. Note that the diffusion is free and modeled by Gaussian diffusion. The measurement of the signal loss function is defined as follows:

$$S_i = S_0 \times e^{-b \times \hat{g}_i^T \times D \times \hat{g}_i} \; where \; b = \gamma^2 G^2 \delta^2 (\Delta - \frac{\delta}{3}) \tag{2.1}$$

Where S_0 is the original image intensity at the voxel (without the diffusion weighting) and S_i is the intensity measured after the application of the i-th diffusion gradient in the (unit) direction $\hat{g}_i$, γ is the proton gyromagnetic ratio, G is the strength of the diffusion gradient pulses, δ is the duration of the diffusion gradient pulses, and Δ is the time between diffusion gradient RF pulses [203].

The product $ADC_i = \hat{g}_i^T \times D \times \hat{g}_i$ represents the estimated diffusivity or the apparent diffusion coefficient (ADC) in direction $\hat{g}_i$.

$$\ln(S_i/S_0)/b = -ADC_i = -\hat{g}_i^T \times D \times \hat{g}_i \tag{2.2}$$

- D is the estimated diffusion tensor (a 3×3 matrix).

Geometrical interpretation

As presented above, the diffusion tensor (DT) is a three-dimensional diffusion object/model that describes the diffusion of water molecules. Mathematically speaking, the tensor is proportional to a symmetrical positive **D** matrix of degree 3 (a co-variance matrix) that models the diffusive flux represented by the variation of the Gaussian distribution. We define the tensor object as the 2D matrix as follows:

$$\mathbf{D} = \begin{bmatrix} D_{xx} & D_{xy} & D_{xz} \\ D_{yx} & D_{yy} & D_{yz} \\ D_{zx} & D_{zy} & D_{zz} \end{bmatrix} \tag{2.3}$$

The diagonal elements D_{xx}, D_{yy} and D_{zz} are the diffusion variances along the x, y and z axes, and the off-diagonal elements are the covariance terms and are symmetric about the diagonal ($D_{xz} = D_{zx}$).

The matrix **D** is a symmetric tensor that can be diagonalized to ran eigenvalues and eigenvectors as follow:

$$\mathbf{D} = \mathbf{E}\Lambda\mathbf{E}^{-1} \tag{2.4}$$

with

$$\mathbf{E} = [e_1, e_2, e_3] \;\; and \;\; \Lambda = \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix} \tag{2.5}$$

This yields the orthogonal eigenvectors e_1, e_2, e_3 and the diagonal matrix of eigenvalues λ_1, λ_2, and λ_3. The three eigenvalues λ_i correspond to the diffusivities along the principle axes of the diffusion tensor, and the three e_i are describing the orientation of these axes which are mutually orthogonal by definition. Consequently, the principal axes of the ellipsoidal isoprobability surface of

the diffusion tensor and their corresponding radii, are given by the eigenvectors e_i and the eigenvalues λ_i , respectively. By convention, the eigenvalues and their corresponding eigenvectors are sorted as follows: $\lambda_1 > \lambda_2 > \lambda_3$. Consequently, the first eigenvector e_1 describes the dominant diffusion direction and is also called the principal diffusion vector (PDV).

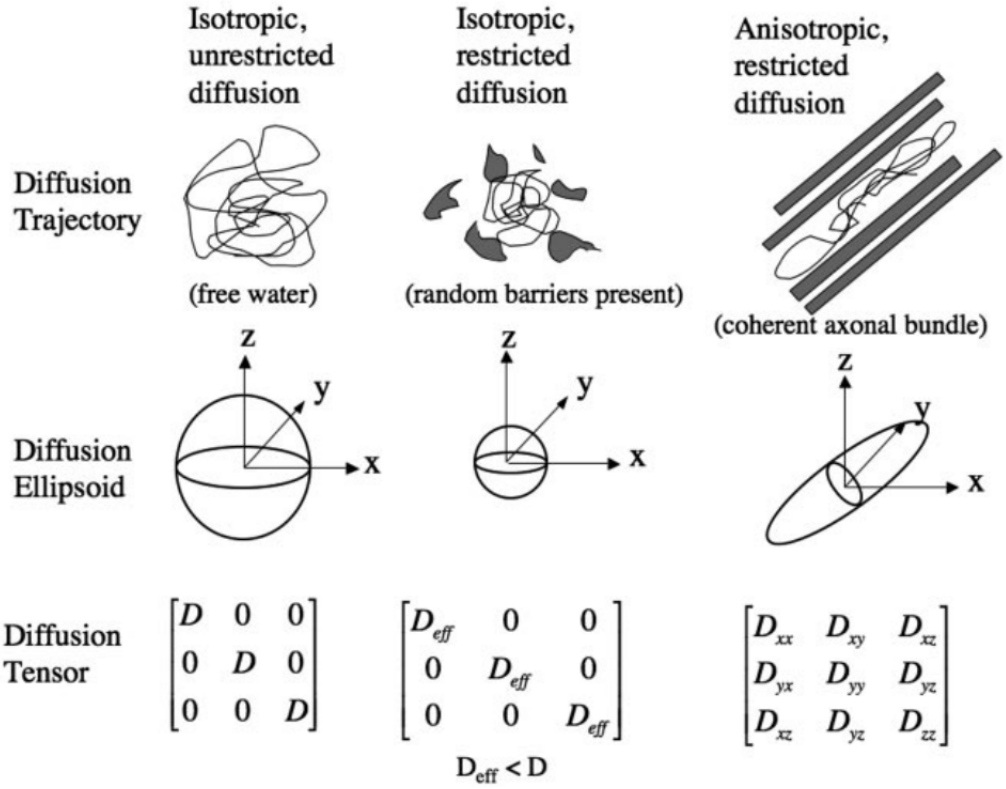


Figure 2.5: The diffusion ellipsoids and tensors for isotropic unrestricted diffusion, isotropic restricted diffusion, and anisotropic restricted diffusion are shown [138].

Scalar measures

From the eigenvalues, various DTI maps can be generated that characterize the diffusion profile such as fractional anisotropy (FA) and mean diffusivity (MD) providing representative and meaningful information. Indeed, throughout the whole volume of the brain and in each voxel, different scalars can be computed that yield different information characterizing the diffusion behaviors. As a result, image data are distilled into more straightforward scalar maps that are suitable to study pathology well.

Here we present the main scalar measures can be extracted:

- **Mean diffusivity:** Mean Diffusivity (MD) is an inverse measure of membrane density. It represents the average magnitude of molecular displacement by diffusion, as it informs on the microstructure of (WM) being sensitive to cell density, axon size, and quantities of water. MD

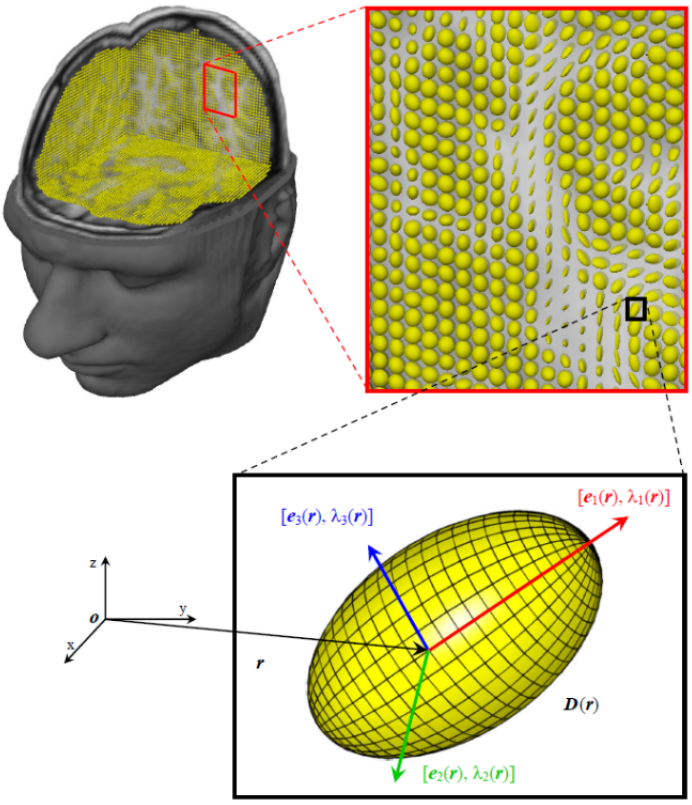


Figure 2.6: Isoprobability surfaces derived from the diffusion tensor field. Note that in each voxel the isoprobability surface is an ellipsoid which is uniquely defined by the tensors' eigenvectors and eigenvalues. Image courtesy of Alexander Leemans [95].

is the average of the three eigenvalues $(\lambda_1 + \lambda_2 + \lambda_3)/3$. A higher MD value reflects more isotropy of the tissue.

$$\mathbf{MD} = \widetilde{\lambda} = \frac{\lambda_1 + \lambda_2 + \lambda_3}{3} \tag{2.6}$$

- **Fractional anisotropy:** Fractional Anisotropy (FA) is a measure of the degree of diffusion anisotropy. It reflects the directions of molecular motion in a certain voxel. FA is calculated from the standard formula:

$$\mathbf{FA} = \sqrt{\frac{3}{2}} \sqrt{\frac{(\lambda_1 - \widetilde{\lambda})^2 + (\lambda_2 - \widetilde{\lambda})^2 + (\lambda_3 - \widetilde{\lambda})^2}{\sqrt{\lambda_1^2 + \lambda_2^2 + \lambda_3^2}}} \tag{2.7}$$

  Where $\widetilde{\lambda}$ is the mean diffusivity (MD). The values of FA vary between 0, which means the voxel space is totally isotropic, and 1, which means infinite anisotropic diffusion. In CSF, the value of FA is zero due to the equality of the diffusion in all directions.

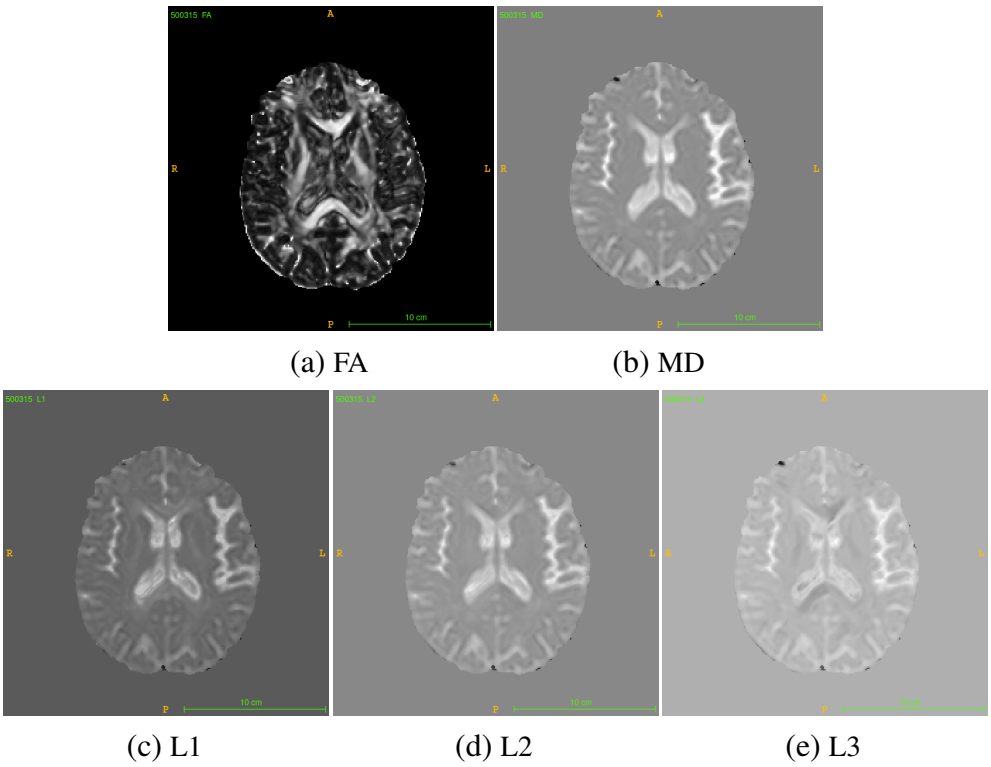
- **Axial diffusivity:** Axial diffusivity was defined as the primary (largest) eigenvalue (AxD = λ_3), and captures the longitudinal diffusivity, or the diffusivity parallel to axonal fibers (assuming

of course that the principal eigenvector is indeed following the dominant fiber direction, which may be unclear in regions with extensive fiber crossing).

$$\mathbf{AD} = \lambda_3 \tag{2.8}$$

- **Radial diffusivity:** Radial diffusivity (RD), which captures the average diffusivity perpendicular to axonal fibers, was calculated as the average of the two smaller eigen-values:

$$\mathbf{RD} = \frac{\lambda_2 + \lambda_3}{2} \tag{2.9}$$



(a) FA (b) MD

(c) L1 (d) L2 (e) L3

Figure 2.7: An example of the isoprobability surfaces derived from the diffusion tensor field.

2.3 Data sets and corrections for image analysis

In this section, we present the most known public datasets used for AD studies. We briefly give an overview of each of them, then confer a description of the dataset used in the current work. We describe the necessitated data correction for each modality afterwords.

2.3.1 Data sets

- **ADNI:** Data used in the preparation of this work were downloaded from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database [1]. The ADNI was launched in 2004 by the National Institute on Aging (NIA), the National Institute of Bio-medical Imaging and Bio-engineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organizations, as a $60 million, 5-year public-private partnership, it is a longitudinal multi-center study designed to develop clinical, imaging, genetic, and biochemical bio-markers for the early detection and tracking of Alzheimer's disease (AD). 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. Since its launch dates of more than a decade ago, the project has now three phases, the ADNI-1, ADNI-GO (Grand Opportunities), ADNI-2, and ADNI-3. 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. ADNI3 began in 2016 and involves scientists at 59 research centers in the United States and Canada. Between 1070-2000 participants will be enrolled: approximately 700-800 rollover participants from ADNI2 and 370-1200 newly enrolled subjects. Clinical, cognitive, imaging, biomarker and genetic characteristics will be assessed across three cohorts: Cognitively normal, MCI and mild AD dementia. For up-to-date information, see www.adni-info.org/.

- **AIBL:** Australian Imaging Biomarkers and Lifestyle (AIBL) [2]: lunched in 2006, is an Australian study project of over 2,000 people assessed over a long period (over ten years). It aims to discover and determine the biomarkers and cognitive characteristics of AD that can help diagnose the disease before symptoms appear. The objective is to provide new preventative treatments and find diet exercise interventions that can prevent or delay the disease. AIBL data is collected in two centers, where images are obtained from patients aged at least 60 years old. The dataset is composed of three categories of subjects; the CN for cognitively normal, MCI for mild impairment cognitive stage, and AD for subjects diagnosed with the disease. Similar to the ADNI project, AIBL provides a longitudinal study where data and information are collected over time; it contains the converters and stable subjects.

[1] http://adni.loni.usc.edu/
[2] https://aibl.csiro.au/

- **OASIS:** Open Access Series of Imaging Studies (OASIS) [3] is an open-access project distributing datasets of brains acquired from MRI studies freely available to the scientific community. It provides anatomical MRI data and clinical assessments of CN subjects and AD patients. Indeed, the project consists of three phases; OASIS-1 and OASIS-2 contain MRI data that represents subjects across the adult life span aged from 18 to 96 years, with numbers equal to 416 and 150 subjects, respectively. OASIS-3 has both MRI and PET modalities sessions from around 1098 subjects.

- **3-City Cohort:** The Three-City (3C) [4] study was established in 1999 to investigate the influence of vascular factors on the risk of dementia and cognitive impairment. The Three-City Study, a prospective French study designed to evaluate the risk of dementia in persons aged 65 years and older. Participants were recruited from three French cities: Bordeaux (South-West), Dijon (North-East), and Montpellier (South-East). The 9,294 eligible participants who participated in the baseline examination have since been invited to three waves of follow-up, 2001-2002, 2003-3004, and 2006-2007. At the time of the baseline examination, 60% of the participants were female, and they were, on average, 74 years old. The participants in a subset from the Bordeaux site of the Three-City (3C) study, a longitudinal multicenter population-based cohort designed to evaluate risk factors of dementia count 2104 subjects. Subjects were non-institutionalized individuals aged from 65 years old and older and were randomly recruited from electoral lists.

In this work, we use only MRI and DTI data from the ADNI project. Data were collected from three phases; ADNI-1, ADNI-2&Go, and ADNI-3. Pertaining to ADNI-1: it contains a total of 815 subjects that hold only sMRI data. Categorized into three classes, as presented in Table 2.1. In this phase, the images are standard 1.5 T screening baseline T1 weighted obtained using volumetric 3D MPRAGE protocol. However, ADNI-2&Go and ADNI-3 hold both modalities; for each subject, there are sMRI and the DTI with their derived maps MD, FA, RD, and AD maps. Table 2.1 provides the demographic description of the datasets, besides, their clinical informations.

2.3.2 Data correction

Noise correction (Denoising)

MRI scans are subject to interference and random noise generated during the acquisition process. The noise introduces some disturbance in the measurement of voxel intensities for further pathology analysis. There are several techniques to disbar or reduce these disturbances, such as averaging multiple MRI acquisition methods. The non-local mean filter (NLM) method initially proposed for

[3] http://www.oasis-brains.org/
[4] http://www.three-city-study.com/

	Classes	# Subjects	Age [range] / $\mu(\theta)$	Gender (#F/ #M)	MMSE [range] / $\mu(\theta)$
ADNI-1	AD	188	[55.18, 90.99] / 75.37 $\pm$ 7.52	99/89	23.3 $\pm$ 2.03
	MCI	399	[54.63, 89.38] / 74.89 $\pm$ 7.30	256/143	27.0 $\pm$ 1.78
	NC	228	[60.02, 89.74] / 75.98 $\pm$ 5.02	118/110	29.1 $\pm$ 1.00
ADNI-2/Go	AD	*48	[55.73, 90.87] / 75.60 $\pm$ 8.63	28/20	23.0 $\pm$ 2.42
	MCI	*108	[55.33, 93.62] / 74.40 $\pm$ 7.47	66/42	27.4 $\pm$ 1.99
	NC	*58	[59.91, 93.25] / 74.91 $\pm$ 5.90	28/30	28.9 $\pm$ 1.18
ADNI-3	AD	*16	[55.26, 86.10] / 74.63 $\pm$ 9.92	4/12	-
	MCI	*165	[55.88, 95.93] / 75.01 $\pm$ 7.91	71/94	-
	NC	*341	[55.79, 95.39] / 73.52 $\pm$ 7.82	209/132	-

Table 2.1: Demographic description of the ADNI dataset group. Values are reported as mean and $\pm$ standard deviation ($\star$ Subjects with both modalities).

denoising natural images was adapted to process MRI images. It was introduced by Buade et al. [39] to deal with images, and it is based on a non-local averaging of all pixels in the processed image. Methods that use the NLM approach aim to find similar regions and average them to overcome noise impact. These methods provide excellent results for denoising, whereas conserving the high frequency in the scans [50]. Figure 2.8 illustrates an example of MRI noise correction.

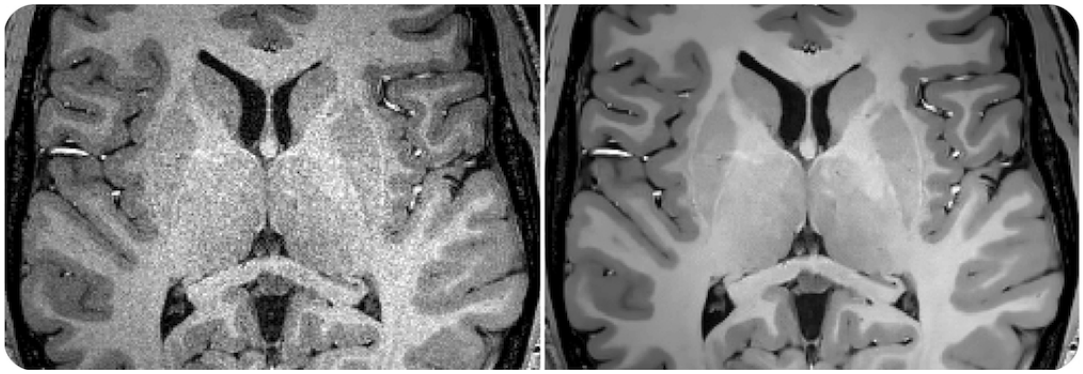


Figure 2.8: Illustration of denoising method, images at the left and right represent the data before and after the denoising process, respectively [50].

Bias field Correction

Correcting bias fields is one of the most fundamental correction methods that need to be applied to MRI data. Indeed, the MRI scans can suffer from artifacts caused by the intensity inhomogeneity. Technically speaking, during the image acquisition process, artifacts may come from two different sources, either from the deficiency produced by MRI devices or from the property of the patient himself due to his position, shape, and orientation inside the magnet field of the scanners. This variation in intensity can be seen as degradation throughout the image. Figure 2.9 shows an example of this variation between two examples (Original and corrected image). This phenomenon can negatively influence the intensity of tissue in some regions compared to others. However, to tackle

this problem, some correction algorithms can be used to minimize the influence of the signal that obscures the white/gray matter.

In this thesis, we have applied the "N4 bias field" algorithm of the ANTS package [189] on the whole structural MRI dataset. Indeed, N4 is a variant of the popular N3 bias correction algorithm (nonparametric, nonuniform normalization). Based on the assumption that the low-frequency bias field corruption can be modeled as a convolution of the intensity histogram by a Gaussian, the basic algorithmic protocol consists in iterating between the deconvolution of the intensity histogram by a Gaussian, the remapping of the intensities, and then the spatial smoothing of this result by B-spline modeling of the bias field itself. The modifications and improvements obtained over the original N3 algorithm are described in [189].

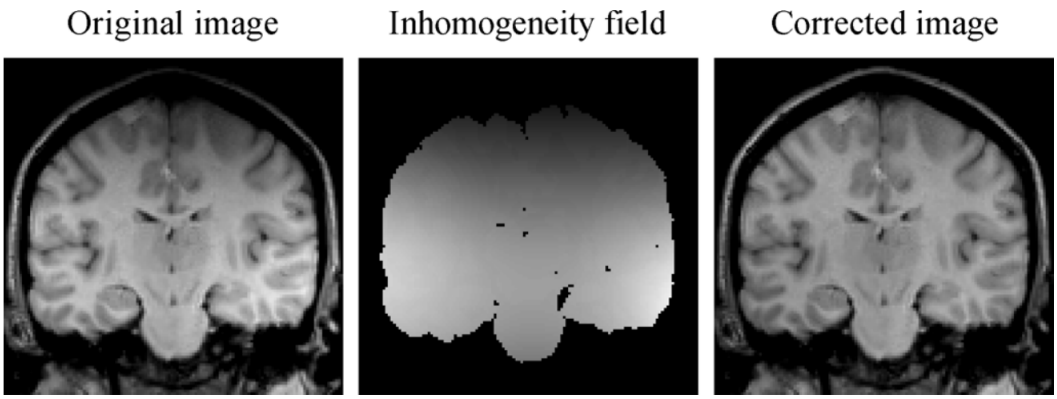


Figure 2.9: Intensity inhomogeneity in MR brain image [194].

Eddy current correction

Using DTI modality for the the detection and the classification of Alzheimer's disease (AD) involves an essential step of correction, which is the "eddy current" correction. Eddy current, also known as (Foucault currents) are parasites or distortions that occur due to the fast diffusion gradients. The source of these changes in the magnetic field can be either the imaging gradients or the radio frequency (RF) coils. Furthermore, it is widely accepted to take into account that parasites originally came from the patent's motion caused by himself. Indeed, the conductive material in which the eddy currents are induced can be any metallic component of the MR scanner (other coils, screens, tubes, wires, or devices inside), or even possibly on the patient himself. However, Eddy currents are undesirable because they generate their magnetic fields, which may oppose the first magnetic field via Lenz's law, and thus distort the spatial and the temporal performance of the desired overall magnetic field. Consequently, these distortions need to be corrected to carry out a reliable analysis, and many methodologies have been proposed to address these distortions. This correction, therefore, consists of compensating the non-linear susceptibility [175] and the eddy current distortions induced by the

movement of the head [10]. After that, computing the mean diffusivity (MD) maps of the scattering tensor to integrate DTI data into the classifier model [94]. In this work, we used the "eddy current" (Foucault currents) correction tool of FSL (Version 5.0, FMRIB, Oxford, UK, [5]) [174].

2.4 Data processing for region-of-interest (ROI) extraction

In this section, we present data-preprocessing methods, which include spatial normalization, multi-modal co-registration, and intensity normalization.

2.4.1 Spatial Normalization (Alignment)

The MNI template

In the literature, there are many templates used in data pre-processing, especially for the registration process. However, the most commonly used templates for spatial normalization are those developed at the Montreal Neurological Institute, known as the MNI templates, which are based on MRI imaging technology. These templates were developed to provide a very useful tool to perform the automatic registration process. Among these templates, we find MNI305, which was created from a set of images (305 images) earlier aligned with the Talairach atlas. Indeed, the principle is to compute an average image (template) with these images, then align each image of any database on this average image using a 12-parameter affine registration. Due to this process, we can spatially normalize a database. Subsequently, and with the limit of MNI305 at the resolution level, another template has been introduced with a high resolution, known as ICBM-152, developed with a number of unbiased non-linear averages from the MNI-152 database (AVG152). The MNI institution provides a different version of this template. In this thesis, we use the MNI-152-T1 version, built by averaging 152 scans of normal subjects. The Figure 2.10 presents it from different planes [72].

Affine transformation (registration)

Image registration (or image transformation) is one of the most fundamental and crucial methods in biomedical image processing. The technique is considered as an optimization method where the goal is to find the best spatial transformation (or deformation) parameters that align a source images to match the target image or template (Figure 2.11) [77]. Many of biomedical image processing tools and resources are based on the image registration techniques. Indeed, working with a heterogeneous dataset of medical images may have a great variety in terms of interest structure. Due to the full range of medical imaging applications, these limitations and constraints make the processing and the study of these images nearly impossible.

[5] http://www.fmrib.ox.ac.uk/fsl/

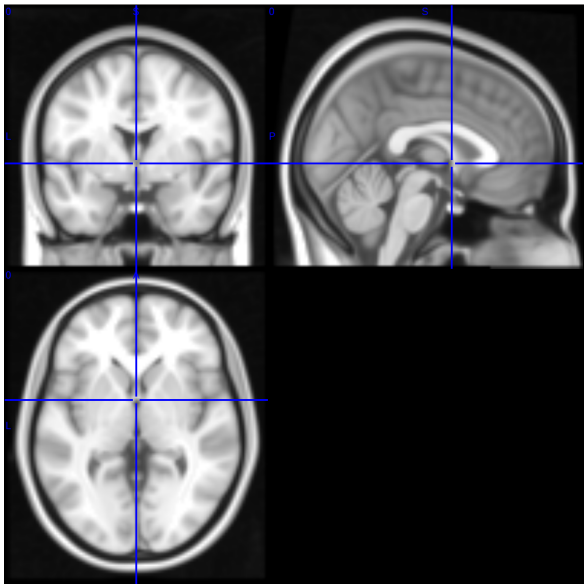


Figure 2.10: The MNI Template: MNI-152 example [72].

In the study of Alzheimer's disease, images of the brain come from different people. Besides, the different position of each subject during the process of acquiring MRI images, we have a variation in the form of internal structure from one person to another. Therefore, this heterogeneity requires the registration process. The purpose of registration methods is to align a set of images from a dataset to a common space to facilitate the processing and study of these images, and further for better performance and efficiency. The use of brain registration has been widely studied [21, 48, 12, 120, 131].

The registration methods can be classified into three kinds of transformation process [77]: i) A rigid transformation that preserves the distances between every pair of points, as well as lines and planes. The method involves translation and rotation operations. ii) Affine transformation is composed in addition to inherited operations from the rigid transformation, the shear, and scale modifications. Moreover, similar to the rigid method, affine transformation preserves also straight lines and planes. However, it does affect angles between lines and planes, besides, the distances between points which are not preserved. iii) Non-rigid registration methods, these types of transformation can align images with a proper transformation at each voxel. They allow us to get better alignment of anatomical structures, yet are less robust in comparison to the other methods [21]. In this thesis, we used only the affine transformation method, which is suitable for our AD studies owing to its performance and capabilities. The affine transformation has advantages over the rigid and non-rigid transformations. The method is very accurate in achieving successful results with reduced timing, especially compared to the rigid transformation, whereas the non-rigid method requires considerable time in the estimation process.

As mentioned in Section 2.3.1, in this thesis, we use the ADNI datasets which are composed of

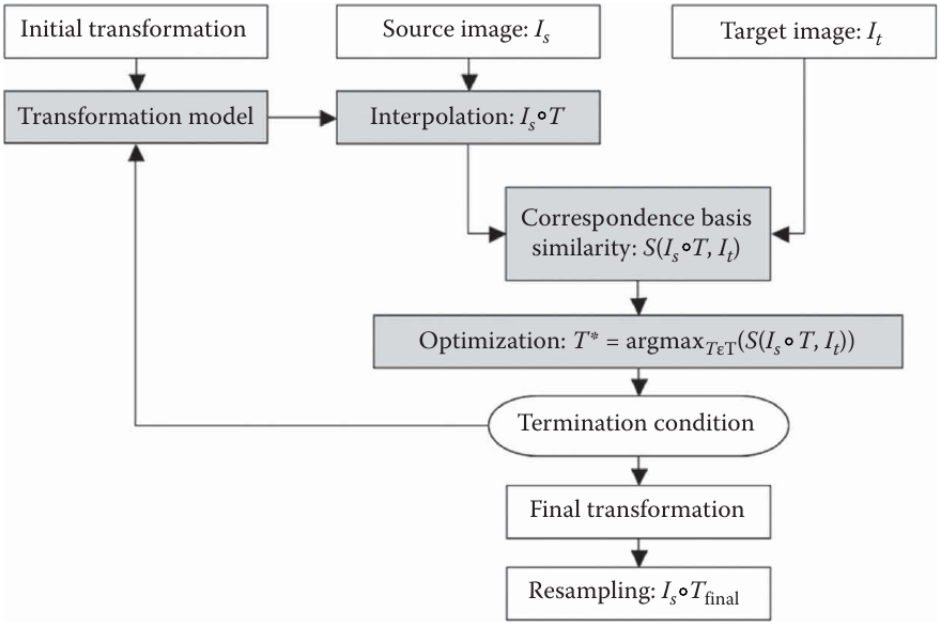


Figure 2.11: A typical registration algorithm consists of four main components: a transformation model, a correspondence basis, an optimization technique, and an interpolation method. The optimization problem can be carried out in a multiresolution or multiscale framework [77].

sMRI and DTI modalities. The preprocessing step for sMRI scans includes a couple of tasks to be completed, as illustrated in Figure 2.12. In the first step (1) of our scheme, we perform the 3D affine transformation to align sMRI images on the common template (MNI) 2.4.1 [1]. Affine transformation uses 12 parameters (m_1 to m_{12}) of the matrix $\mathbf{M}$ to perform deformation of sMRI images to the (MNI) standard space. The goal is to estimate these parameters for a given image (f) to fit a template image (g), including translation, rotation, scaling, and shearing deformations [16].

We can formally define the 3D affine transformation as a function $\mathbf{T}_{affine}$ that acts on a point (or vector) $x = (x_1, x_2, x_3)^\intercal$ in the real 3-dimensional vector space $\mathbb{R}^3$, and generates the transformed point (or vector) $y = (y_1, y_2, y_3)^\intercal$.

$$\mathbf{T}_{affine}(x) = y \tag{2.10}$$

The function uses the 12-parameters affine deformation (m_1 to m_{12}) as stated above, which are related to translation, rotation, scaling and shearing transformations. The registration algorithm uses an optimization method to estimate and find these 12-parameters by minimizing a cost function (e.g. mean square error criterion) or maximizing the similarity of the source images (f) and (g) target image, which means fit a given image (f) to the template image (g) [13].

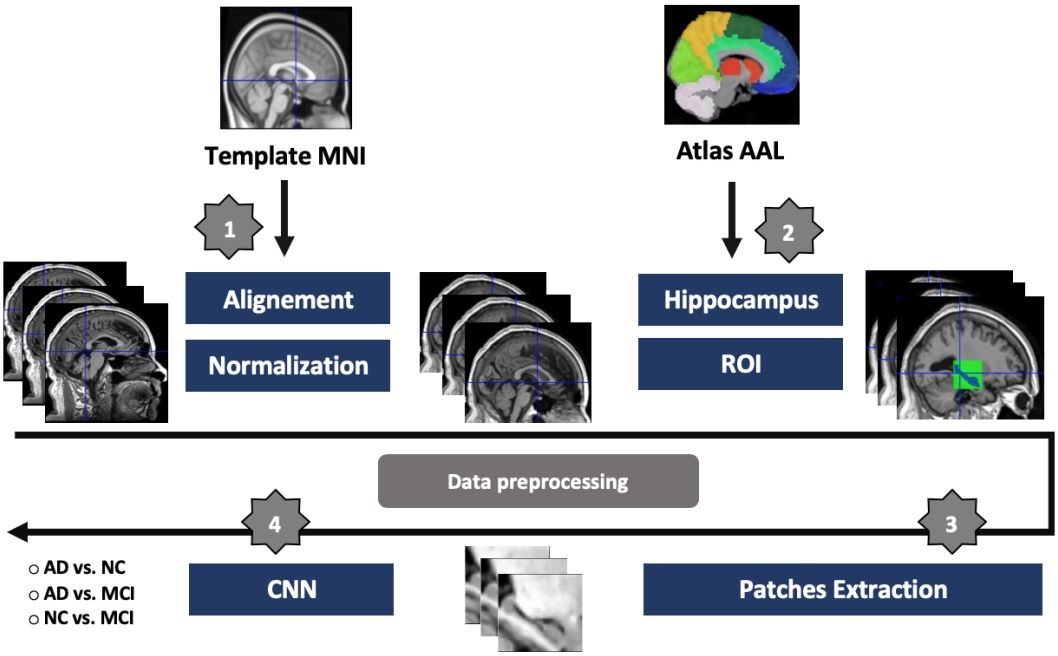


Figure 2.12: Schematic diagram of dataset preprocessing: i) registration of all MRI scans on MNI space, followed with intensity normalization. ii) ROI selection process using the Atlas AAL for both hippocampal regions. iii) 2D-slice extraction from selected 3D-volume. iv) feeding the CNN networks [3].

Hence, the affine transformation is defined as follows :

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ 1 \end{pmatrix} = \begin{pmatrix} m_1 & m_4 & m_7 & m_{10} \\ m_2 & m_5 & m_8 & m_{11} \\ m_3 & m_6 & m_9 & m_{12} \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ 1 \end{pmatrix} \tag{2.11}$$

We refer this mapping equation as $\mathbf{y} = \mathbf{M} \times \mathbf{x}$, where $\mathbf{M}$ is the transformation matrix and m_i elements are functions of parameters q_1 to q_{12} (the results of how to carry out the transformation).

The matrix $\mathbf{M}$ can be decomposed as a product of four matrices, translation, rotation, scaling, and shearing (eq. 2.12).

$$\mathbf{M} = M_{Translation} \times M_{Rotation} \times M_{Scaling} \times M_{Shearing} \tag{2.12}$$

The parameters q_1, q_2, and q_3 correspond to 3 translation parameters, q_4, q_5, and q_6 correspond to 3 rotations parameters. q_7, q_8, and q_9 to 3 zooms and finally q_{10}, q_{11}, and q_{12} are the 3 shear

parameters.

$$M_{Translation} = \begin{pmatrix} 1 & 0 & 0 & q_1 \\ 0 & 1 & 0 & q_2 \\ 0 & 0 & 1 & q_3 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{2.13}$$

$$M_{Rotation} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos(q_4) & \sin(q_4) & 0 \\ 0 & -\sin(q_4) & \cos(q_4) & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \times \begin{pmatrix} \cos(q_5) & 0 & \sin(q_5) & 0 \\ 0 & 1 & 0 & 0 \\ -\sin(q_5) & 0 & \cos(q_5) & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \times \begin{pmatrix} \cos(q_6) & \sin(q_6) & 0 & 0 \\ -\sin(q_6) & \cos(q_6) & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

$$M_{Scaling} = \begin{pmatrix} q_7 & 0 & 0 & 0 \\ 0 & q_8 & 0 & 0 \\ 0 & 0 & q_9 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{2.14}$$

$$M_{Shearing} = \begin{pmatrix} 1 & q_{10} & q_{11} & 0 \\ 0 & 1 & q_{12} & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{2.15}$$

The objective function to minimize is the sum of squared differences (SSD) between the subject (f) and template images (g). The optimization method is the Gauss-Newton algorithm [14, 75]. An additional parameter w is added to the function to correct the difference scale that can be produced in images. The function to minimize is then:

$$SSD(f,g) = \sum_{i=1}^{I} (f(\underbrace{M * x_i}_{y_i}) - wg(x_i))^2 \tag{2.16}$$

The process was done using the software SPM8 (Wellcome Trust Centre for Neuroimaging at UCL, London, UK) [6] to fulfill the registration and the normalization [74].

2.4.2 Multimodal co-registration for ROI Selection

The same subject's MRI and DTI modalities have to refer to the same physical structures in the brain. For this purpose, we need to coregister them. This step is essential in order to get correspondence between the regions of interest through these images. However, an important preliminary step is

[6] http://www.fil.ion.ucl.ac.uk/

required to be performed; the noise skull stripping, which has not been fulfilled on sMRI. Indeed, as we explain this further, we work only on biomarker ROI inside the brain volume on sMRI. Hence the skull stripping is not mandatory.

Skull stripping

As we have the DTI images without the skull, we need to extract the brain from sMRIby removing any other confused parts. In this step, the skull stripping task was performed to pull out only the brain from the sMRI modality. There are many methods and tools to perform skull stripping, such as FSL and BET. In our work, we used a method based on the segmentation of the brain. Indeed, the process was applied using the SPM12 toolbox with Matlab to segment the brain scans into Gray Matter (GM), White Matter (WM) and Cerebro Spinal Fluid (CSF), then merging these three maps we can subtract skull region from original sMRI scan. Figure 2.13 presents an example of sMRI skull removal.

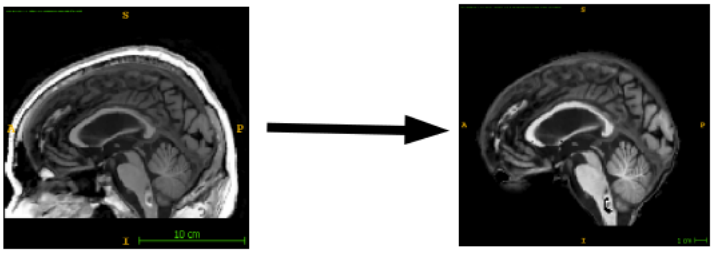


Figure 2.13: Example of sMRI skull stripping: - (left) an original brain scan - (right) the brain result after removal skull process.

Coregistration

After performing skull stripping step, we obtain a base of images of the brain without the skull. As we have different modalities of image acquisition, the co-registration step is essential to work with the multimodal images in our database [134, 45]. Indeed our objective is to extract the patches of our region of interest from the DTI-MDl map relative to the same region as in sMRI. The co-registration between sMRI and MD consists of estimating the transformation parameters using the criterion of mutual information in some specific areas to match the standard space (MNI) at the end. Thus, we routinely co-register the MD image in the corresponding sMRI coordinate system. The co-registration algorithm uses brightness of voxels. In order to avoid distortions which may be induced by bright skull voxels, the skullstripping step is performed as described above. Figure 2.14 illustrates the three main steps of data preprocessing, 1) the alignment of sMRI to the template MNI, as discussed in 2.4.1, 2) skull stripping, and the last step 3) the co-registration between sMRI and DTI image.

Finally, we obtain for each subject two MRI and DTI images as results. They are aligned to the MNI common space. These images have a similar resolution of $121 \times 145 \times 121$, and the voxel element measures $1.5 \times 1.5 \times 1.5\, mm^3$.

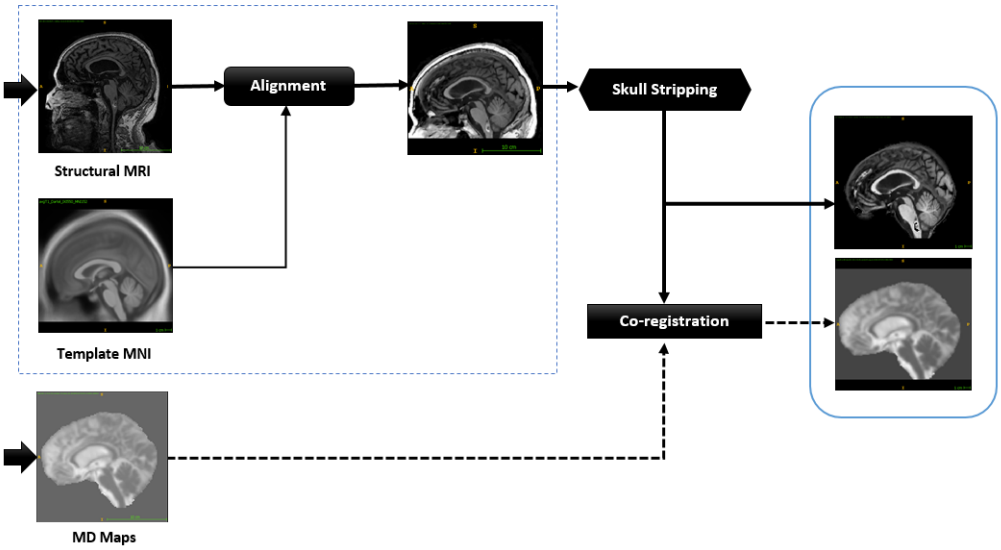


Figure 2.14: Illustration of the co-registration process includes spatial normalization and skull stripping.

2.4.3 Intensity Normalization

In the acquisition phase, the MRI images produced are acquired in arbitrary units. This arbitrariness develops a difference in the scale of the intensity provided by the scanners, which makes it difficult to analyze images together. Indeed, MRI images are not comparable across scanners, visits, or even sometimes when using the same protocol. This variability can negatively impact the performance of image processing and machine learning algorithms, i.e. methods of segmentation, detection, classification etc. The intensity normalization process is therefore an essential step to work on a set of images of a medical database. This task allows bringing the intensities to a common standard scale across all the elements of the database. Several image intensity normalization algorithms have been proposed by researchers in the field, among them, we find those that are adapted to brain images.

In this thesis work, we used an algorithm developed by "Nyul Laszlo and Udupa Jayaram" that we will call "Nyul and Udupa" for short [143]. This technique is based on the use of the histogram of the image. Indeed, the idea behind is to compute a standard histogram through the delimitation of landmarks of predefined interest on a specific database, then make a deformation of the histogram of each image to obtain a correspondence of intensities of this image with those of the standard histogram. In other words, the objective is to have with the same protocol and the same brain region, a similar intensity in all images of the database, which allows us to find the same tissue (See [142,

143] for more details).

In our case, we define landmarks as intensity percentiles at $1, 10, 20, \ldots, 90, 99$ percent (where the intensity values below 1% and above 99% are discarded as outliers).

As we must pre-define a standard intensity range, we set the interval as $[m^s_{\min}, m^s_{\max}]$ with $m^s_{\min} = 0$ and $m^s_{\max} = 255$.

Let us consider $\mathbf{I} = \{I_1, I_2, \ldots, I_K\}$ a set of K MRI brain images,

We calculate the following set of quantities m^i_1 and m^i_{99}, which are the 1% and 99% intensity values for the image $I_i \in \mathbf{I}$. We then map all the intensity values of I_i with the following linear mapping

$$\tilde{I}_i(\mathbf{x}) = \left(I_i(\mathbf{x}) - m^i_1 + m^s_{\min}\right)\left(\frac{m^s_{\max}}{m^i_{99}}\right) \tag{2.17}$$

which takes the intensities of I_i to the range $[m^s_m in, m^s_m ax]$ excluding outliers. Then we calculate the deciles for the new image $\tilde{I}_i$, i.e., the set $\tilde{m}^i_1 0, \tilde{m}^i_2 0, \ldots, \tilde{m}^i_9 0$ (note that $\tilde{m}^i_0 = m^s_m in$ and $\tilde{m}^i_1 00 = m^s_m ax$). This is done over every image $I_i \in I$ and the mean of each corresponding value is the learned landmark for the standard histogram. That is, for $n \in 10, 20, \ldots, 90$, we have:

$$m^s_n = \frac{1}{K}\sum_{i=1}^{K} \tilde{m}^i_n \tag{2.18}$$

and the standard scale landmarks is the set $\{m^s_{\min}, m^s_{10}, \ldots, m^s_{90}, m^s_{\max}\}$.

For a test image I, the transform for the normalization is done by first calculating the set of percentiles $\{m_1, m_{10}, m_{20}, \ldots, m_{90}, m_{99}\}$. These values are then used to segment the image into deciles, i.e., we define 10 non-overlapping sets of indices $D_{i,j} = \{\mathbf{x} \mid m_i \leq I(\mathbf{x}) < m_j\}$ where
$i, j \in \{1, 10, 20, \ldots, 90, 99\}$ and restricting j to equal the next value in the set greater than i. We then piecewise linearly map the intensities associated with these deciles to the corresponding decile on the standard scale landmarks. Noting that each $D_{i,j}$ is disjoint from the other, the normalized image is then defined as

$$I_{\text{nu}} = \bigcup_{i,j \in \{1,10,20,\ldots,90,99\} i \neq j, i \leq j+10} \left(\frac{I(D_{i,j}) - m_i}{m_j - m_i}\right)\left(m^s_j - m^s_i\right) + m^s_i. \tag{2.19}$$

2.4.4 ROI Selection using Automated Anatomical Labeling (AAL)

The alignment of sMRI brain scans to the common MNI space, followed by the co-registration with DTI images deforms the individual's morphology. Hence we do not perform fine- grained segmentation of the images. Instead of this our approach relies on selection of the ROI and generating patches encircling the area of a biomarker (hippocampus) Hence, we only need a locator that allows

us to select the ROI in the scans. Therefore, we have introduced a selection approach based on the use of atlases.

Atlas AAL

Before going any further to present the used Atlas, let us briefly define the Atlas meaning in the field and the difference between it and a template.

- An **atlas**: gives a pattern to the location of anatomical characteristics in coordinate space. It is useful for localization of activation and interpretation of results and avoiding performing the segmentation's fastidious process.
- A **template**: is a model or reference used as a representative of the atlas and provides a target to which individual images from a dataset can be registred to match this reference. It can hold an image from a single individual scan or an average of several individuals scans.

There exist many atlas that we can used to select region from brain, In this work, we used a brain Atlas called Automated Anatomical Labeling (AAL) [190] developed at the Institute of Neurodegenerative Diseases) IMN [7]. The AAL atlas is a single-subject atlas based on the MNI Colin27 T1 atlas. Figure 2.15 shows the standard AAL template (with different projections) which comprises 116 brain anatomical regions. After alignment of sMRI modality on MNI template and co-registration of MD-DTI, the ROI can be selected on the automated anatomical labeling Brain atlas (AAL) [190]. The Figure 2.16 is an example of the results of the selection of the hippocampal region.

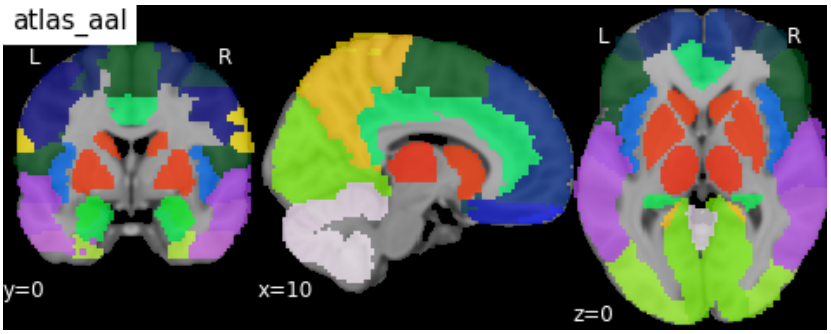


Figure 2.15: The AAL atlas views: (Left) coronal slice, (Center) Sagittal slice, and (Right) Axial slice. The regions are colored to identify region boundaries.

ROI and Patches extraction

Since, we have for each subject a sMRI and DTI images which are affinelly registered to the common standard space and interpolated in the same definition of the MNI template. The next step is identifying the region-of-interest.

[7] (IMN-UMR5293- CNRS, CEA, Université de Bordeaux

Hippocampus: is the region investigated in this study which is suggested by our medical partners, and is considered as a visual biomarker that is the first region affected by (AD) [57, 58] (See Chapter 1).

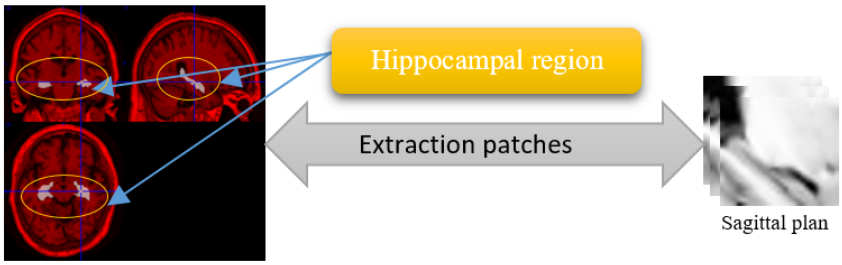


Figure 2.16: An illustration of the hippocampal region using the Atlas AAL.

The selection ROI process consists of superimposing geometrically the registred individual scan and the Atlas AAL, then extracting exclusively voxels that are identified as the hippocampal region in the Atlas. From the brain volume, we extract and compute Hippocampal regions and compute their 3D bounding boxes. We get a sub-volume of the whole 3D scan, which encircles the hippocampus in both modalities: sMRI and DTI. As we know that the hippocampus is a symmetrical anatomical structure in the brain consisting of two regions, we then get two 3D-bounding boxes for the two regions, the left hippocampal region (F_l) and the right (F_r) as illustrated in Figure 2.17.

Note: The resolution of normalized sMRI and MD volumes is quite low ($121\times145\times121$, see above), thus the hippocampal ROI occupies a small amount of voxels ($28\times28\times28$). Finally the data were converted to the lossless Portable Network Graphics (PNG) format for 2D studies, or were keeped as 3D volumes for 3D studies to feed the CNN classifier.

Formally, we can define the function that allows the extraction process of the ROIs with $\mathcal{L}_D$ as in (Equation 2.20) using the selected dataset D_{brain}, which return the two bounding boxes as follow:

$$H_l, H_r = f(D_{brain}) = \mathcal{L}_D(C_l, C_r, \theta, s) \;\; with \;\; H_l, H_r \in R^{d_1,d_2,d_3} \tag{2.20}$$

Where $C_l = [x_{l,(min,max)}, y_{(min,max)}, z_{(min,max)}]$, and $C_r = [x_{r,(min,max)}, y_{(min,max)}, z_{(min,max)}]$ are the coordinates that computed from the Atlas for x, y, and z dimensions. θ and s are supplementary parameters which used for the artificial augmentation process.

**In this work of thesis, we have focused on the use of ROI-based methods for AD classification. Therefore, we developed a library for ROI extraction suitable and adapted for all brain datasets, as presented in section 2.3.1. It provides seamless tools to select the investigated ROIs with the help of a compatible multi-label atlas and then generates a dataset of increased samples for model training and testing. Indeed, besides the simplicity that it offers for selecting regions, the tool comprises*

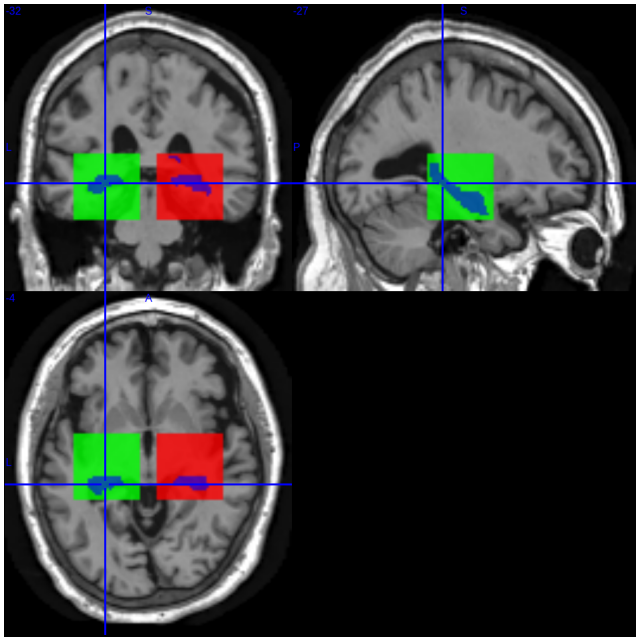


Figure 2.17: Two 3D Bounding Boxes include the Left (green) et the Right (red) Hippocampus ROIs in three projections.

domain specific data augmentation with geometrical and grey-level transformations we will describe in the following chapters. increases as well the dataset with further parameters predefined, and it includes a couple of methods for artificial data augmentation, such as translation, gaussian blurring, rotation, flipping, and others. See Annexe B (related paper in preparation).*

The tool and code-source are available in GitHub link below. "Brain Med Extraction" [8]

2.5 Conclusion

In this chapter, we have introduced the preprocessing pipeline of MRI and DTI data. We have also presented the MRI acquisition methods and the theory behind sMRI and DTI. Next, we provided an overview of the used dataset and its processing flow through a set of coupled methods to obtain a clean and coherent dataset. Finally, we presented the extraction module to achieve the final ROIs data.

In the next chapter, we will present the deep learning methods to build robust models for AD classification. We cover almost all basic concepts from simple neural network to deep CNN network, including the principle modules and functions to build architectures.

[8] https://github.com/kaderghal/ADNI_Data_processing/

Chapter 3

Deep learning methods for object classification

3.1 Introduction

Artificial Intelligence has been broadly applied to various domains. One of its branches, namely machine learning and specifically its sub-domain deep learning, has been developed extensively. Nowadays, computer-aided diagnostic (CAD) systems for neuroimaging have included advanced algorithms based on deep learning methods. Classification, detection, and segmentation are specific high-level tasks that have been engaged in research, and many studies are related to neurodegenerative diseases. Alzheimer's Disease (AD) is one such study. In this chapter, we present an overview of deep learning methods, which have been used in elaborating this thesis, particularly the Convolutional Neural Network (CNN) architecture. We can summarize this part in three main subparts: first, we briefly recall the traditional neural networks from the perceptron to the multi-layered perceptron, then we introduce the main element used in deep neural networks: the CNN approach. In the follow up we explore different standard components of CNN networks and present a short study of the most well-known activation functions. Besides, we overview optimization methods. Eventually, we contribute to the problem of data limitation in deep learning applications and submit alternative solutions.

3.2 Artificial Neural Networks

An Artificial Neural Network (ANN) is a computer system based on the functioning of the human brain, inspired by biological neural networks. These artificial neural networks have generated a great deal of excitement in the machine learning research and industry.

3.2.1 Formal Neuron (Perceptron)

The basic block in a neural network is the neuron (perceptron), sometimes called a node or a unit. It receives inputs from either some previous nodes or an external source, then computes and produces an output. Each input is associated with a weight (w), which is assigned on the basis of its relative importance compared to other inputs. The node applies a function (f) to the weighted sum of its inputs as shown in Formula 3.1.

$$\hat{y} = f(X) = f\left(\sum_{i=1}^{n} w_i x_i + bias\right) \tag{3.1}$$

The function f is the activation function that is used to activate or not the neurones (See 3.2.3). The idea and objective behind the activation function are to introduce non-linearity into the output of a neuron, emulating biological response.

The Figure 3.1, illustrates an example of perceptron, the latter takes numerical inputs (x_1, x_2, x_3) which has (w_1, w_2, w_3) the associated weights. Besides, another input parameter is b, which is used to adjust the output (called the Bias). The output $\hat{y}$ from the neuron is computed, as shown in the figure.

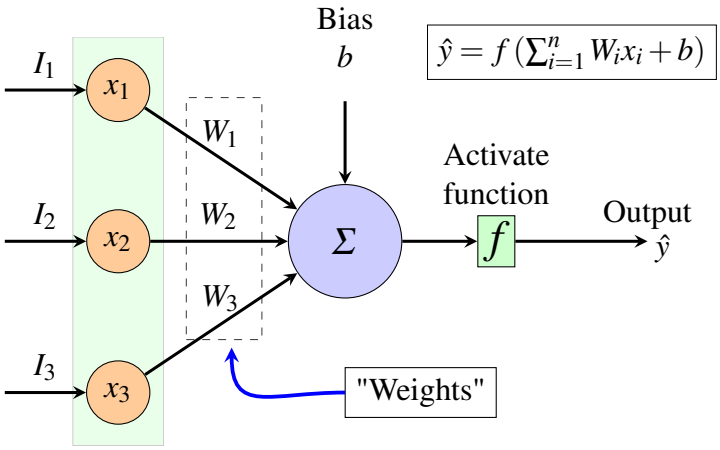


Figure 3.1: Formal Perceptron illustration: An example of single perceptron with tree inputs (I_1, I_2, and I_3).

3.2.2 Multi layer Neural Networks

The multi-layer perceptron (MLP) is a type of neural network that contains one or more hidden layers. A layer can be interpreted as a hyperplane that includes several simple perceptrons, it is used to enhance the separation capability of the network. If a (MLP) network contains more than one layer, it is called a deep neural network (DNN). The concept of back-propagation was introduced to train hidden layers in (MLP) networks. Indeed, the input layer contains, in addition to the bias, numerical values that depend on the database (I_1, I_2, ... I_n); they all feed nodes located in the first hidden layer. By performing the calculation process in the hidden layer, including an activation function ((f) see

3.2.3) we obtain local scores in this layer. The same process is repeated on each hidden layers until the last one which is the output layer. Figure 3.2 shows typical multi-layer perceptron network with only single hidden layer.

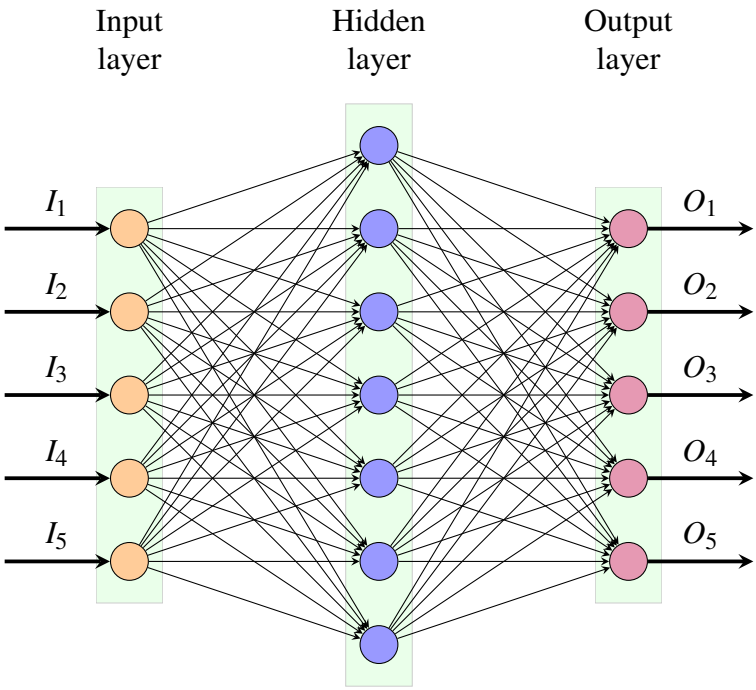


Figure 3.2: Multi Layer Neural Network: example of fully connected network with tree layers *(Input, Hidden and Output)*.

3.2.3 Activation functions

The output of a neuron is computed using an activation function which is a non-linear function as shown in Figure 3.1. There are several activation functions to model neural response [141].

Here we present some *common activation* functions mostly used in the deep learning area:

- **Sigmoid**: The function takes a real-valued input and squashes it to range between 0 and 1. The function also called (*logistic function*)

$$sigmoid(x) = \frac{1}{1+e^{-x}} \tag{3.2}$$

- **ReLU**: The *ReLU* function stands for Rectified Linear Unit. It takes a real-valued input and thresholds it at zero.

$$\text{ReLU}(x) = \max(0,x) \tag{3.3}$$

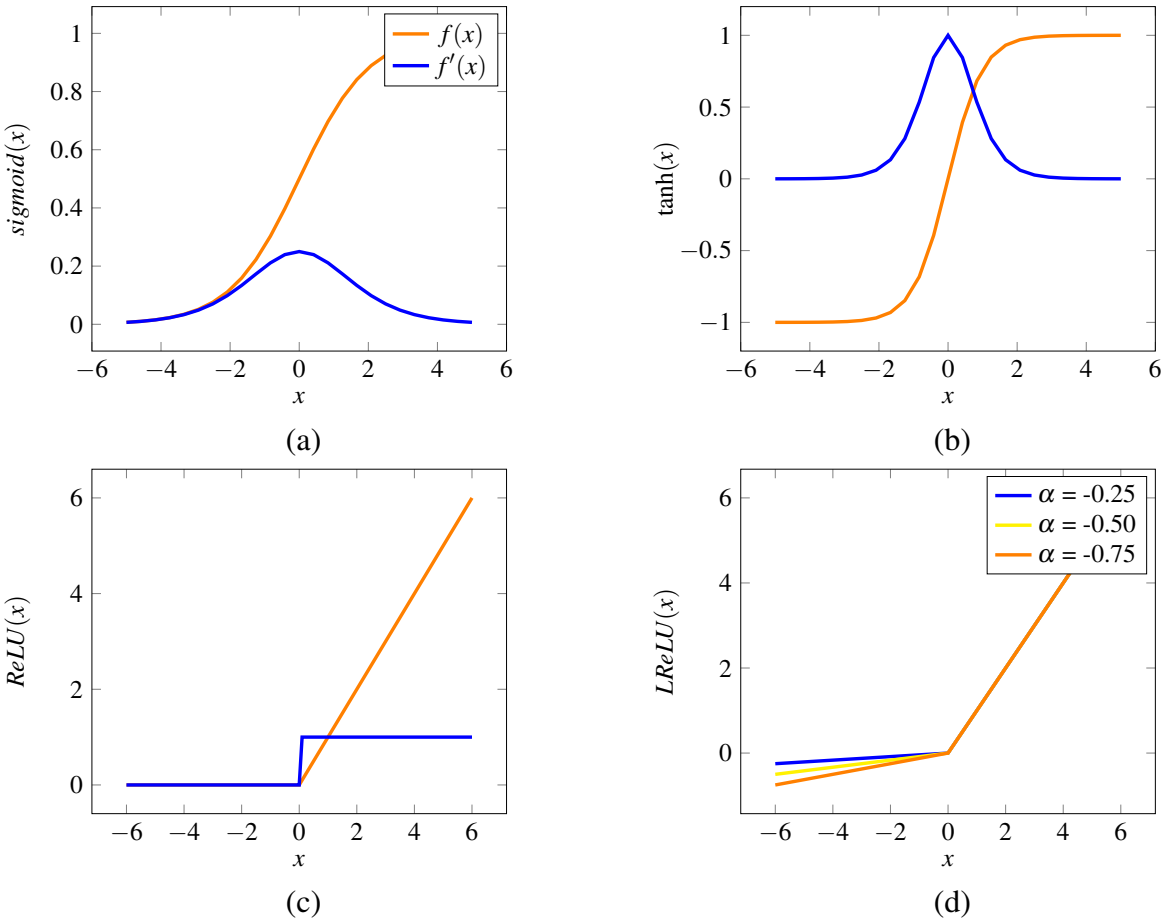


Figure 3.3: Activation functions graphs examples: (a) - the sigmoid function. (b) - the tanh function. (c) - the Rectified Linear Unit (Relu) function. (d) - the Leaky ReLU [7].

- **Leaky ReLU**: unlike ReLU, this function has a small slope for negative values, instead of a plain zero. It fixes the "dying ReLU" problem, as it does not have zero-slope parts, and it speeds up the training.

$$\text{LReLU}(x) = \max(\alpha x, x) \tag{3.4}$$

with $0 < \alpha \leq 1$

- **Tanh**: The function takes a real-valued input and squashes it to the range [-1, 1].

$$\tanh(x) = \frac{2}{1+e^{-2x}} - 1 = 2\sigma(2x) - 1. \tag{3.5}$$

Figure 3.3 plots curves of these functions. In this thesis, we use only the ReLU and sigmoid activation functions.

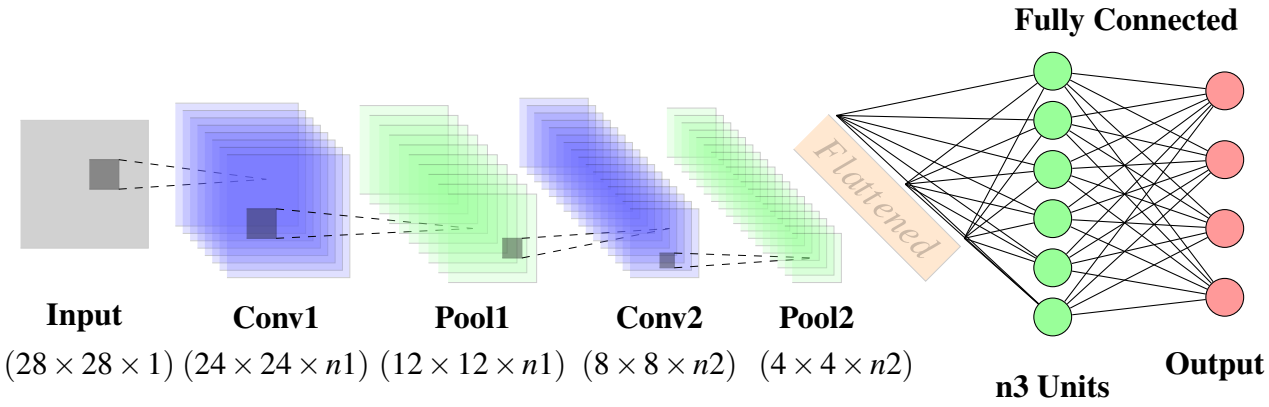


Figure 3.4: An example of a CNN architecture: model for handwritten digits classification.

3.3 Convolution Neural Networks (CNN)

A Convolutional Neural Network also called CNN or ConvNet, is a type of deep feed-forward artificial neural networks mostly applied in analyzing visual data. It is widely used in machine learning, especially in deep learning algorithms. Their functioning is based on the mathematical convolution operation. A CNN architecture is composed of a series of layers. Every layer of the network goes through a differentiable function to transform itself from one volume of activation to another. Many types of layers are used to build a CNN, among them we list: *i)* Convolutional layer, *ii)* Normalization layer, *iii)* Pooling layer, and *iv)* Fully-connected layer and so on. All these layers are stacked somehow together to form a full CNN model. In a more advanced architectures, CNNs may contain complicated blocks that are built from these layers or new innovated layers. The Figure below 3.4 is a complete example flow of CNN network LeNet to process an input image and classify the digits. It takes an image input of $1 \times 28 \times 28$ resolution from the MNIST [1] data set [112].

In CNN, there are generally more layers interspersed between these four basic layers:

- Convolution transformation.
- Pooling layer.
- Activation functions.
- Fully connected layer.

3.3.1 The Convolution transformation

The convolution layer is the core element of CNN networks, it allows us to extract the characteristics of an input image by specifying parameters such as the number of filters, kernel size etc. The choice

[1] http://yann.lecun.com/exdb/mnist/

of the size of the receptive field depends on the resolution of the input image, as well as on the nature of these images.

We can formally define the transformation (convolution) function $\mathbf{T}_{tr}(I)$ as a mapping of the input data $I \in \mathbb{R}^{H \times W \times C}$ where $I = [x^1, x^2, ..., x^C]$ and x^c represents a single channel of the input I, to feature maps $O \in \mathbb{R}^{H' \times W' \times C'}$ where $O = [y_1, y_2, ..., y_{C'}]$.

We denote $V = [v_1, v_2, ..., v_{C'}]$ the learned set of filter kernels, where v_i refers to i-th filter (one filter = one activation map). And each "learned kernel" has the same number of "channels" as the input I.
We can write $v_i = [v_i^1, v_i^2, ..., v_i^C]$. Equation 3.6 recaps the operation for each kernel (v_i).

$$y_c = v_c \circledast I = \sum_{s=1}^{C'} v_c^s \circledast x^s \tag{3.6}$$

Here $\circledast$ denotes the convolution operator, and v_c^s is a 2D single slice from learned filter.

The equation 3.7 represents the general rule to compute output size of the convolution. P is padding, F is receptive field size, and S is the stride step. The figure 3.5 illustrates an example of 2D convolution transformation.

$$W_{i+1} = (W_i - F + 2 \times P)/S + 1 \tag{3.7}$$

- An image matrix (Volume) of dimension $(H \times W \times C)$
- A filter $(f_h \times f_w \times C)$
- Outputs a volume dimension $(((I_h - f_h + 2 * P)/S + 1) \times ((I_w - f_w + 2 * P)/S + 1) \times 1)$

3.3.2 Pooling Layer (Pool)

The pooling layer is a quite a simple operation in a Convolutional Neural Network (CNN), this layer consists in reducing the spatial dimensions (**W** and **H**) while preserving the same depth as the previous layer (**C'**). Indeed, it takes some $k \times k$ region from a selected slice "feature map" and produces a single value. By using a stride step for x and y directions, we achieve "sub-sampled" feature maps in each feature channel. The application of this operation allows us to obtain a new block of feature maps with low resolution and same depth. The fundamental purpose of the Pooling layer is the gain of the computation performance to train in a neural network model ("less information means less parameters").

Different functions can be used in the pooling layer: we explore and evaluate two main functions, the average pooling (3.8) and the max-pooling (3.9) [166].

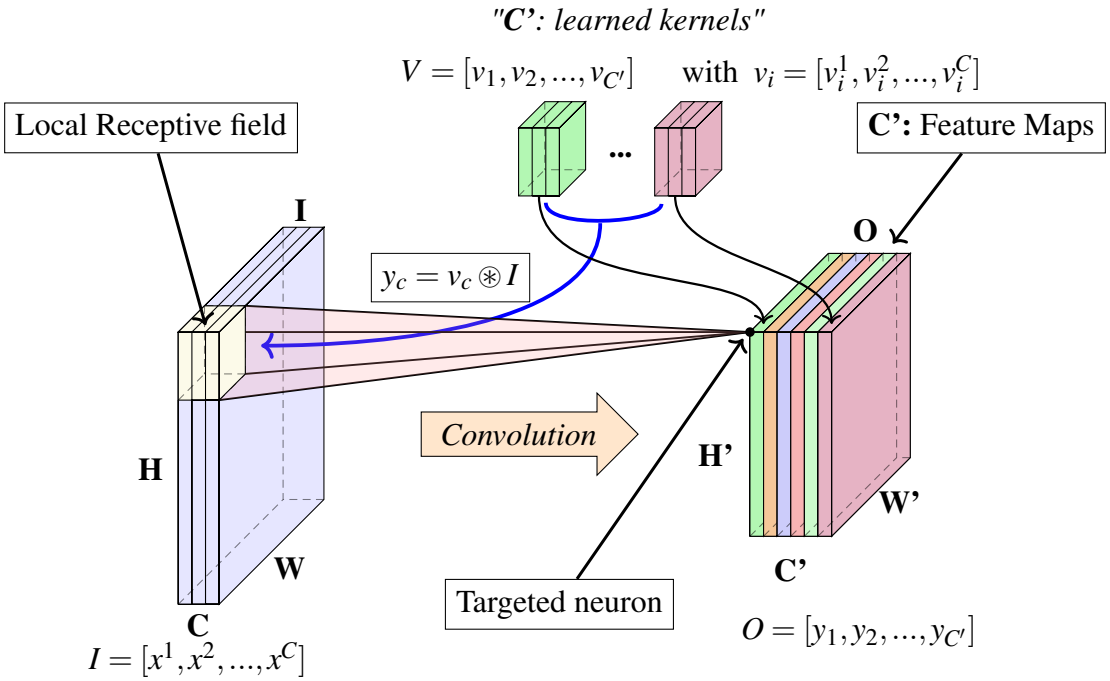


Figure 3.5: The convolution transformation: case of 2D input image .

- **Average pooling:** takes the average for each local window of *kxk* over features maps and produces new feature block with the same depth.

$$h_j^n(x,y) = 1/K \sum_{\bar{x} \in N(x), \bar{y} \in N(y)} h_j^{n-1}(\bar{x}, \bar{y}) \tag{3.8}$$

- **Max-Pooling:** this function retains the maximal value in a window *kxk* from the input feature map.

$$h_j^n(x,y) = \max_{\bar{x} \in N(x), \bar{y} \in N(y)} h_j^{n-1}(\bar{x}, \bar{y}) \tag{3.9}$$

We denote by N the *kxk* neighborhood of (x,y). In Figure 3.6 we illustrate an example of the Max-Pooling operation. In this example we choose a window with size of 2×2, and we assign 2 to the stride S which is the step of sliding the window on the feature map.

3.3.3 Fully connected layer

Fully connected Layer (or Inner Product Layer) is a layer where all neurons have full connections with the previous layer. As seen in (Section 3.2.2) each neuron in i^{th} Layer has full connection to all neurons in $i-1^{th}$ Layer, and we can compute its activation with a matrix of multiplication of values and weights of neurons in $i-1^{th}$ Layer (see Figure 3.2).

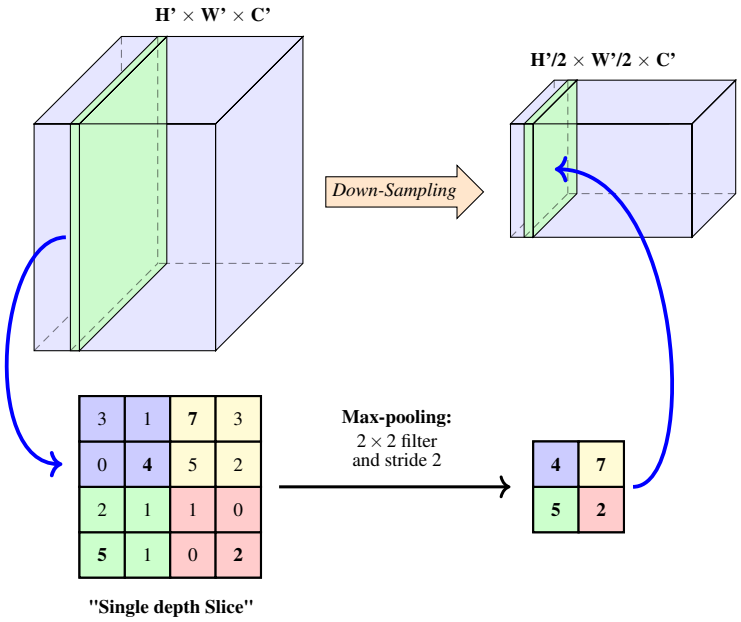


Figure 3.6: Sub-sampling illustration: case of Max-Pooling with a 2×2 filter and a stride with 2 steps.

3.4 Loss functions (Cost functions)

In the machine learning, all algorithms rely on optimizing an objective function or "criterion". Loss functions are a kind of function used to minimize the error between the true label "target" and the predicted element "output", this group of functions are also called cost functions [111]. CNN networks are supervised machine learning methods, they learn the prediction function from a training data set. The loss functions are used to measure how the model can predict the result compared with ground truth. Finding the minimum point with the help of the optimization methods (see Section 3.5), the CNN model converges to a state in which the error tends to zero or to the lowest possible value. In other form, we can summarize the definition of the loss function as a measure which quantifies the variation (error) of the output prediction $(\hat{y}_i)$ from the expected response (y_i) [168].

$$\mathrm{J}(W) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}_{(x_i,y_i)\in D_{train}}(f_w(x_i), y_i) \tag{3.10}$$

Formally, as the CNN networks are a typical supervised learning methods. let us consider $D_{train} = \{(x_i, y_i)\}$ with $i \in \{1,..,N\}$ a set of features (samples) that uses the model to make prediction, y_i corresponds to the label of x_i, and the $\hat{y}_i = f_w(x_i)$ is the function to optimize, with $w \in W$ its weight parameters. The optimization process utilizes dataset D_{train} in the training step.

By minimizing the loss function $\mathcal{L}(\hat{y}, y)$, the $f_w(x)$ model discovers "good" parameters (synaptic weights in the network) to make prediction the closest to the ground truth. Therefore the loss function is used to evaluate the error between input item/ground truth y and the predicted output $\hat{y}$.

$$\underset{W}{\text{minimize}}(\mathcal{L}_{(x,y)\in D_{train}}(f_w(x), y)) \tag{3.11}$$

Different loss functions give different errors for the same prediction task, we can broadly categorize them into two types: Classification and Regression Loss. In this chapter we tackle classification problems, that is our y variable is discrete. In the following we present the most popular loss functions.

The soft-max function: a desired function for multi-class classification

The Softmax function (3.12) is a function which aims to transform a real-valued vector Z of dimension K into a vector of real numbers P in range (0,1), which summs up to 1 (3.13):

$$p(Z)_i = \frac{e^{z_i}}{\sum_{j=1}^{k} e^{z_j}} \tag{3.12}$$

The function is used in the last layer of a given Multi-layer neural network, which takes real number scores from the previous layer $Z = [z_1, z_2, ..., z_k]$ with $z_i \in \mathbb{R}$ and $k \in \mathbb{N}$, then computes the estimated "probabilities" values $P = [p_1, p_2, ..., p_k]$ with $p_i \in [0,1]$.

$$\sum_{i=1}^{K} p(Z)_i = 1 \tag{3.13}$$

One hot encoding: is a method to represent a target vector into a binary vector where all values equal 0 except the target class which is equal to 1.

Consider K the number of classes and the dataset $D_{train} = \{(x_1, y_1), ..., (x_N, y_N)\}$, we introduce the K-binary vectors $L^i = [l_1^i, l_2^i, ..., l_K^i] \, \forall i \in [1..N]$ each mapped to an input label (y_i). Specifically, if $y_i = C_k$ then $l_k^i = 1$ and $\forall j \neq k \; l_j = 0$ where $k \in [1, ..., K]$.

3.4.1 L1 and L2 mean loss function

Mean Squared Error (L2 loss) and Mean Absolute Error (L1 loss) are the two standard loss functions, which produce a mean error on a selected train dataset.

- *Mean Squared Error (MSE)*: It is basically minimizing the sum of the square of the differences $\mathcal{L}(\hat{y}_i, y_i)$ between the target value y_i and the estimated values $\hat{y}_i$.

$$\mathcal{L}(\hat{y}, y) = \frac{1}{n} \sum_{i}^{n} (y_i - \hat{y}_i)^2 \tag{3.14}$$

- *Mean Absolute Error (MAE)*: It's used for minimizing the sum of the absolute differences $\mathcal{L}(\hat{y}_i, y_i)$ between the target value y_i and the estimated values $\hat{y}_i$.

$$\mathcal{L}(\hat{y}, y) = \frac{1}{n} \sum_{i}^{n} |y_i - \hat{y}_i| \tag{3.15}$$

3.4.2 Cross Entropy Loss (Log Loss)

The cross-entropy function is a particular loss function mostly used in classification problems that measures the cross-entropy between the predicted and the target class. The computation of this function consists in the closeness's comparison of the probability distribution of output scores produced from the model and the expected target class.

The equation 3.16 results only the output value corresponding the predicted class including its cost. l^i is the one hot encoding vector for i^{th} sample and p_j is the predicted output probability of each class.

$$\mathcal{L}(\hat{y}_i, y_i) = -\sum_{j=1}^{k} l_j^i \log(\rho_j) \tag{3.16}$$

To train the networks usual approach is to use a batch of n samples from the data set. Then the average cross-entropy function is computed, which is simply the average value over all the batch introduced in equation 3.17.

$$\mathcal{L}(\hat{y}, y) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} l_j^i \log(\rho_j) \tag{3.17}$$

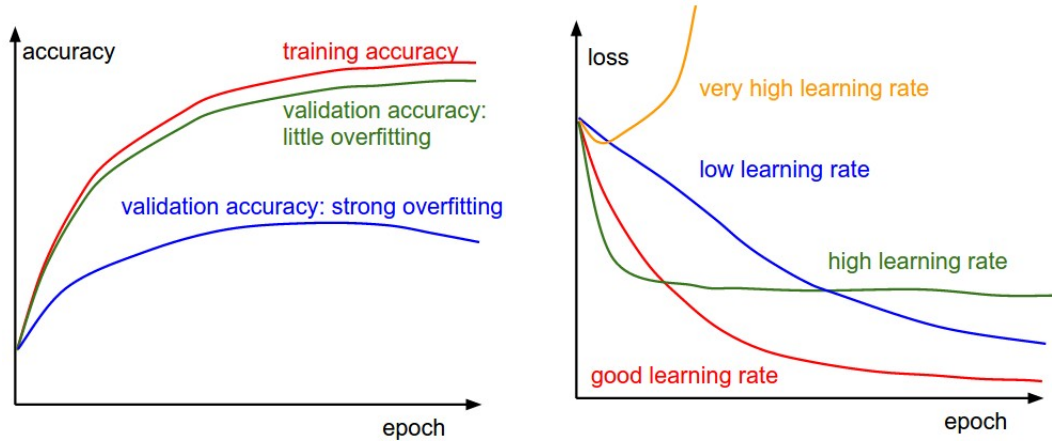
As seen before, machine learning models use the cost functions to estimate and minimize the error between y_i and $\hat{y}_i$, therefore the main objective is in finding the optimal parameters of the network to minimize the error.

In the next section we shortly review the optimization methods used for training Deep NNs.

3.5 Optimization Methods and policies for Model Training

In machine learning, finding the best parameters for a model is the most complicated process. Neural network models are non-linear, this means that they can learn complex non-linearity of classification resurfaces in data representation space. A downside of this flexibility is that they learn via a stochastic training algorithm. During the training phase, we try to minimize the variance of the cost function at each iteration "forward pass". As stated in Section 3.4, the aim is to reach/converge towards a state in which the model gives better results. The achievement of this purpose depends on several parameters, such as learning rate, learning policy etc. Hence the importance of optimization algorithms such as stochastic gradient descent and other algorithms from these family [7].

The principle of these algorithms is updating the model's parameters (weights), moving them in the direction opposite to the gradient of the loss function. These moves a regulated by the so-called learning rate, which also can be changed accordingly to different policies. 3.5.3). This algorithms follow iterative schemes of model parameter updates, regulated by learning rate. We schematize the behavior of such iterative schemes in Figure 3.7 below.



(a) training and validation accuracy curves. (b) Loss curve during training.

Figure 3.7: Different learning rate where training and validation of a Deep CNN [100].

Figure - (a) 3.7 presents curves of training and validation accuracies, the gap between the training and validation accuracy indicates the amount of overfitting (we will explicit this notion below). The second Figure - (b) shows an example of training and validation behaviors with different learning rate configurations.

As we can observe, choosing the learning rate parameter plays a very important role in model parameter optimization. Indeed, it is a way to avoid overfitting, that's why learning rate is considered as a regularization parameter 3.6.3.

In the following we present the most popular optimization algorithms used for DNN parameter optimization.

3.5.1 Gradient Descent

Gradient Descent (GD) is the fundamental algorithm for optimization of parameters in Deep Neural Networks. Once initialized,(randomly or in another way) the parameters-arguments W of the objective function $J(W)$ built with the loss function ($\mathcal{L}(f_w(x), y)$ 3.10) are moved into direction opposite to the direction of its gradient $\nabla_W \mathrm{J}(W)$ [7]. In other words, we follow the direction of the slope of the surface created by the objective function downhill until we reach a valley.

The iteration step for the gradient descent is given by:

$$W_{t+1} \leftarrow W_t - \eta \nabla_W \mathrm{J}(W) \tag{3.18}$$

The learning rate η determines the size of the steps we take to reach a "local" minimum. Without going very deep into the theory of Gradient descent algorithm and conditions of its convergence, we remark, that it is a well known mathematical fact that for the non-convex functions the optimization process may be stacked in a local minimum.

Figure 3.8 illustrates update process and convergency to (a) global minimum of a convex function, and (b) the possibility to still converge to a global and not local minimum of the objective function if the learning rate and thus induced step-size have been correctly chosen.

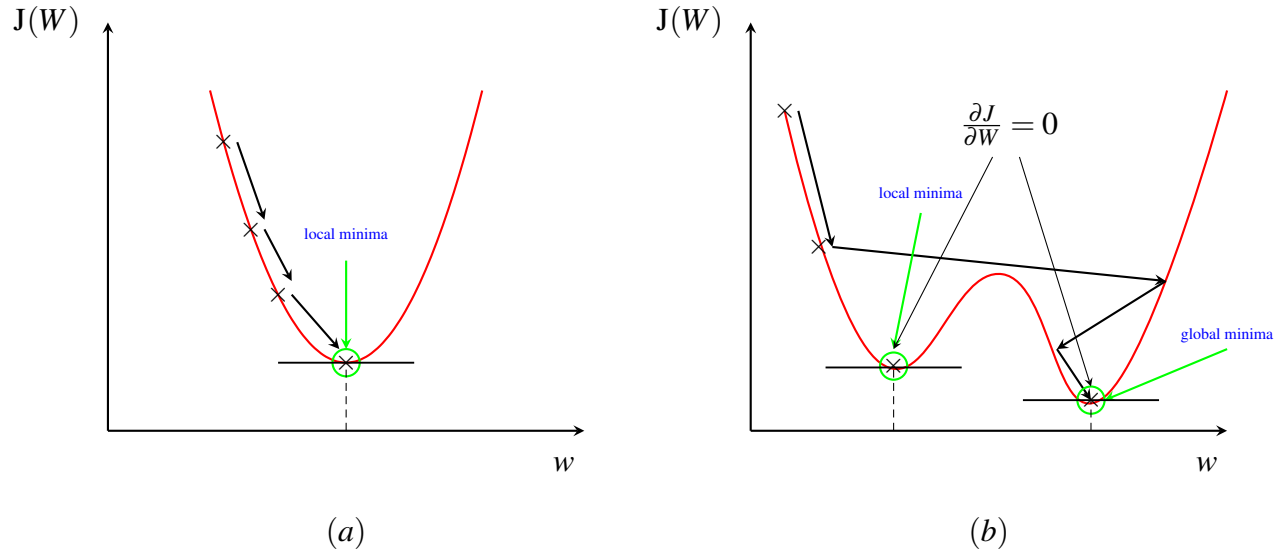


Figure 3.8: Gradient Descent examples: Two functions - (a) having global minima, and - (b) a non-convex function having a local minima and global minima.

- **Stochastic Gradient Descent Algorithm (SGD):**

  The Stochastic Gradient Descent (SGD) algorithm is another iterative variant of GD. Indeed, if we have a large database, practicing the GD algorithm needs a considerable time for only one pass on entire database before updating the parameter values, which makes training very time-consuming, even infeasible. Indeed we have to update the learning rate for each single sample of the training data set. Contrarily, in SGD, we use merely a single or subset of the training sample set to perform the update operation at each particular iteration. A sample subset is called "batch" on the field. It is a subset, randomly chosen of of the whole training dataset. SGD often converges fast compared to GD.

  As mentioned above, when we use a batch of data, for each iteration we compute the mean update of the gradient descent as in the equation 3.19. B_s design the batch-size, and s is the size.

$$W_{t+1} \leftarrow W_t - \eta \frac{1}{B_s} \sum_{j=1}^{s} \nabla_W \mathrm{J}\left(f_{w_t}(x_t^j), y_t^j\right) \tag{3.19}$$

  Remark: The relationship between epoch and iteration: One epoch is when an entire data set is passed forward and backward through the neural network (see 3.20).

$$\textbf{(1)}\ one\ epoch = \frac{"dataset"}{"batch-size"} = \sharp\ iteration(s) \tag{3.20}$$

3.5.2 Optimizing the Gradient Descent

Under some conditions, the stochastic gradient descent (SGD) algorithm can become very slow, e.g, when the gradient is consistently small. To tackle this cases, there are some acceleration methods which can improve the update rules. Momentum and Nesterov Momentum (also called Nesterov Accelerated Gradient) are slight variations of normal gradient descent that can speed up training and improve convergence significantly.

- **Momentum:**

  Momentum method can be used to accelerate learning compared to the plain Stochastic Gradient Descent. It is a technique that can accelerate gradient descent by taking into account previous gradients in the update rule at each iteration. This can be clearly observed in the update rule equation in every iteration.

  The iteration step for the momentum concept in the gradient descent update is given as follows:

$$\begin{aligned} v_{t+1} &\leftarrow \mu v_t - \eta \nabla_W \mathrm{J}(W_t) \\ W_{t+1} &\leftarrow v_{t+1} + W_t \end{aligned} \tag{3.21}$$

  Moreover, the algorithm is only guaranteed to converge to the global solution in the case where the function J is strictly convex. If that's not the case, the algorithm will not even be guaranteed to find a local minimal.

- **Nestrov accelerated gradient (NAG):**

  Nesterov Accelerated Gradient is another method that is related to Momentum. It is a simple change to normal momentum method where every update happens in two steps. First, the gradient term is computed from the position $(W_t + \mu v_t)$ in parameter space, and then the final update velocity is calculated as in the normal momentum method (see 3.22). If the momentum term points in the wrong direction or overshoots, the gradient can still "go back" and correct it in the same update step.

$$\begin{aligned} v_{t+1} &\leftarrow \mu v_t - \eta \nabla_W \mathrm{J}(W_t + \mu v_t) \\ W_{t+1} &\leftarrow v_{t+1} + W_t \end{aligned} \tag{3.22}$$

3.5.3 Adaptive Learning rate policy

The learning rate is a positive scalar that defines the step size during training [111]. It refers to the amount that the weights are updated and thus controls the speed at which the model learns. However, the learning rate is a configurable hyper-parameter and should be set prudently to well-train models.

Generally, a significant learning rate allows the model to learn faster by arriving on an optimal set of weights, nevertheless with a large learning rate there is a danger to bypass the "good " minimum of the objective function to minimize. In contrast, a lower learning rate may allow the model to learn more or less well; however, it takes a long time to train and converge and also does not prevent from stacking in a local minimum. To improve the training process and produce a model that converges quickly and efficiently, we can adapt the learning rate. We base this mechanism on different policies for updating the learning rate [7]. Here we recall some of them, which are widely used in the optimization of parameters of DNNs on the basis of gradient descent family of methods:

- Step decay: After each k epochs, multiply the learning rate by a constant $C < 1$.

- Polynomial decay: Set the learning rate as

$$\forall t \geq 0,\ \eta_t = \frac{a_0}{1 + b_0 t^n}, \qquad a_0, b_0 \in \mathbb{R}_+. \tag{3.23}$$

- Exponential decay: Set the learning rate as

$$\forall t \geq 0,\ \eta_t = a_0 e^{-b_0 t}, \qquad a_0, b_0 \in \mathbb{R}_+. \tag{3.24}$$

The sense of all these adaptation strategies consists in using a large learning rate at the beginning of the iterative process and little-by-little reducing it when approaching to the (hopefully) global minimum of the objective function.

3.6 Deep learning and Data limitation constraint

3.6.1 Motivation: Over-fitting Phenomena

Over-fitting or learning-by-heart means that resulting from the optimization process the model fits very well to the training data, but yield strong error on the unseen(test) data. One of the reasons of it is the small size of the used training dataset. Indeed, these methods require sufficient training samples to learn well and thus avoid the overfitting problem. To address the phenomenon in context of CNNs, there are specific techniques called regularization methods, such as artificial data augmentation and transfer learning, which may be employed to overcome the problem. In this section, we briefly introduce some of them.

3.6.2 Artificial data augmentation

Artificially increasing data is one of the most intuitive solutions to circumvent the over-fitting problem. However, different methods can be engaged in the image processing field to augment and expand dataset. Geometric transformation such as shearing, zooming rotation, and random mask, and application of noise such as gaussian blur, are the most popular methods. They are called "label-preserving transformations", as they do not change the label of the training data samples. In the medical field, the images' nature differs from those we perceive in daily life in both form and content. Hence a suitable domain-dependent data augmentation strategy has to be designed. In AD classification, we have applied the gaussian blur, and the flip operation, besides the translation along several dimensions. We avoided using some operations like the contrast change, the shearing, and the zooming since we already applied the spatial normalization that uses affine deformation, which includes these transformations to align images to a standard space. Recently data augmentation with Generative Adversarial Networks has become popular, when from a random noise image and some training example the network generates images similar to training examples. We do not apply this strategy in our work and thus do not present it here.

3.6.3 Regularization Methods

Regularization is a very important technique in machine learning to prevent over-fitting. First-of-all the way to prevent over-fitting is to stop the training process sufficiently early when the model fits well to the training data and still remains sufficiently good at the test data. This is called "early stopping"

Early Stopping of training

The question here is how many epochs we need to train a network. This is a major challenge in training neural networks. Too many training epochs can lead to an over-fitting of the model to the training data set, and consequently, a poor performance on the test set, while too few learning epochs means that the model will under-fit the training set. Wherefore, all standard neural network architectures are prone to over-fitting. To obtain good performance, early stopping is a method that allows us to process and fix the number of learning epochs, to specify an arbitrary number of epochs; while the model seems to get better results, we stop learning immediately when the performance on the unseen set is getting worse. Figure 3.9 shows an example of a learning task, in this figure we can notice that when we exceed the optimal capacity (epochs), the model over-fits (red curve).

L1 & L2 Regularization

L1 and L2 Regularization are a very important techniques to prevent learning algorithm from over-fitting. These methods add a penalty term to the cost (loss in our case) function to form the

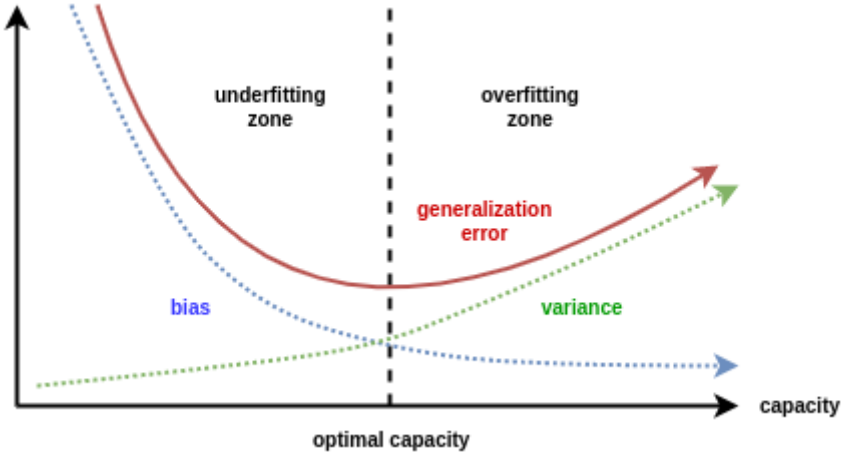


Figure 3.9: Early stopping for best generalization performance [100].

objective $J(W)$ and control the model complexity using that penalty term. The penalty term expresses a constraint on the parameters (weights of the network) to train. The difference between the L1 and L2 is just that L2 is the sum of the squares of the weights, while L1 is just the sum of the weights. As follows:

- L1 regularization in L1:

$$J(W) = \mathcal{L}_{(x,y)}(\sum_{i=1}^{k} f_{w_i}(x), y) + \lambda \sum_{i=1}^{k} |w_i| \tag{3.25}$$

- L2 regularization in L2:

$$J(W) = \mathcal{L}_{(x,y)}(\sum_{i=1}^{k} f_{w_i}(x), y) + \lambda \sum_{i=1}^{k} w_i^2 \tag{3.26}$$

The λ here is the regularization parameter. It is a hyper-parameter in the Neural Network training. It is usually fixed after several trials. Bisection method or other optimization methods could be used, everything depends on the computational resource available for training of the Neural Network.

Dropout Regularization

Dropout is another method that also aids to combat the problem of over-fitting. Basically, during training the method simulates a scattered activation of a given layer, which means that certain fractions of the neurons in that layer will be randomly deactivated or dropped out [176]. The effect of the dropout is to make the learning process noisy, forcing the nodes of a layer to take more or less responsibility for the inputs. The technique can be easily implemented. At training step, each neuron has a probability p of being disconnected. This improves generalization because it forces the layer

to learn the same “concept” with different neurons. Figure 3.10 presents an example of applying the dropout method.

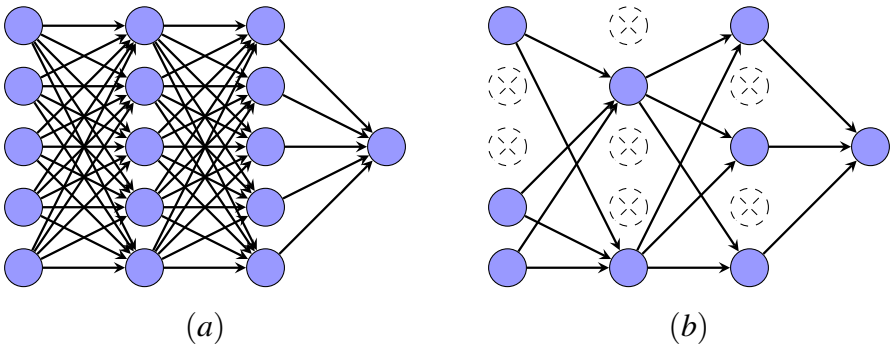


Figure 3.10: An example of two Neural networks: (a): Standard Neural Network - (b): After applying dropout method.

Batch Normalization

Batch normalization is another technique that may help, to prevent the over-fitting, It basically consists in the transformation of all data samples known in statistics as "whitening" of the data. It consists in subtraction of the mean and division by the standard deviation of the marginal distribution. This kind of approaches was proposed namely for the VGG-Net architecture [173].

3.6.4 Transfer Learning and Fine-tuning Approach

Generally, in the machine learning domain, we can represent a simple classifier as a function (f), which computes an output ($\hat{y}$) from a vector input (x), written as follow:

$$\hat{y} = f_W(x) \tag{3.27}$$

Where (W) is the weights parameters and (f) is the model - classifier.

Transfer learning is an optimization that allows for fast progress and improves performances when the model takes a very long time to converge "it is a shortcut to saving time", even more, it is used in cases where the size of the database is very limited to avoid the problem of over-fitting. This technique makes it possible to transfer the knowledge acquired on a "source" dataset to better process a new "target" dataset. The goal of transfer learning is to improve learning in the target task by leveraging knowledge from the source task. Transfer learning can be defined as a fine-tuning process [28, 208], this approach takes pre-trained model parameters and uses them as a starting point in other processing tasks. In fact, in real-life cases, learning a model from scratch is a relatively complicated task due to the depth size of the network, or sometimes in the case where the training data set is insufficient.

We can formally define the transfer learning strategy as follows:

$$\begin{cases} W_0 & \leftarrow W'_\phi \\ W_{i+1} & \leftarrow F(W_i) \end{cases} \tag{3.28}$$

where W'_ϕ is a trained model parameters on a "source" dataset. Therefore the initial W_0 receives these parameters, and fine-tune layers of the architecture. F is the optimization scheme as detailed in 3.5.

In general, it is not obvious that there will be a benefit from using transfer learning in the domain until the model has been developed and evaluated. Figure 3.11 illustrates the difference in performance measures, between two cases of training (with and without transfer).

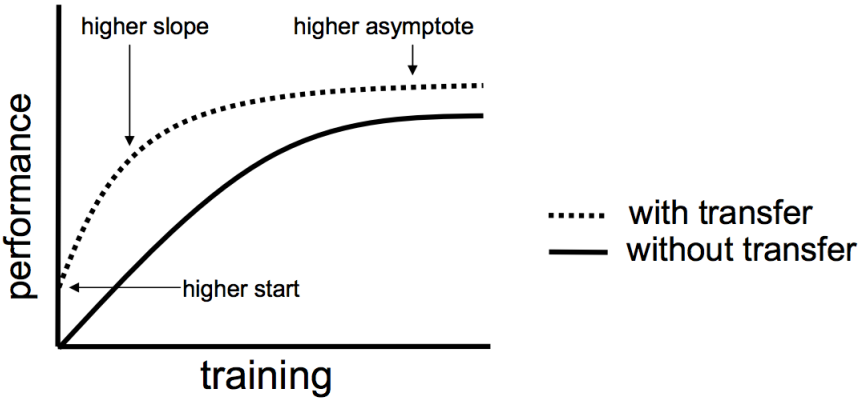


Figure 3.11: Three ways in which transfer might improve learning: a higher performance at the very beginning of learning, a steeper slope in the learning curve, or a higher asymptotic performance [144].

3.7 Conclusion

In this chapter, we presented an overview of the deep learning approach. We briefly exposed the fundamental theories of neural networks, from pure perception, through the definition of the multi-layer network, to deep convolutional networks. We also briefly reminded the fundamental concepts behind, which serve to optimize models in the machine learning domain and specifically in Deep Neural Networks classifiers. We then presented some techniques to address the over-fitting phenomenon, namely artificial data augmentation, regularization methods, and transfer learning. We do not pretend that our synthesis is exhaustive, nor we tried to present methods with all mathematical demonstrations. This chapter was a kind of reminder for introduction of our solutions in the problem of AD classification on MRI modalities.

In the next chapter, we will present our AD classification approaches and their implementations using CNN neural networks methods.

Chapter 4

The 2-D+ε Approach with Shallow Convolutional Neural Networks.

4.1 Introduction

In interactive health care systems, emerging models of convolutional neural networks (CNNs) have been integrated into various disease diagnostics applications, such as the classification of magnetic resonance imaging (MRI) scans for Alzheimer's disease assessment.

In this chapter, we present our contributions to the AD classification problem over the MRI imaging modality. We focus on the hippocampus morphology, which is known to be affected by the illness's progress by using the ROI-level methods. We use a subset of the ADNI (Alzheimer's Disease Neuroimaging Initiative) database, presented in chapter 2, to classify brains belonging to Alzheimer's disease (AD), mild cognitive impairment (MCI), and normal control (NC) classes.

As the number of images in such studies is somewhat limited regarding the needs of CNN models, we propose an analysis of different data augmentation strategies adapted to the specificity of sMRI scans. We further propose our "2-D+ε" approach, where only a very limited amount of consecutive slices are used for training and classification, besides a pragmatic investigation of engaging the well-suited network design to obtain encouraging classification results.

Highlights:

- We present the main target ROI which is a biomarker of Alzheimer disease and the proposed approach for its selection to be further used used for AD classification;
- We introduce the 2-D+ε approach as envisioned spatial data input and networks input layer architecture;
- We present an effective CNN network for AD classification;

- We study different implementation and settings for dataset augmentation;

4.2 Related work

Despite its recent tremendous successes, CNNs are only starting to be used for CAD in classification of brain images, and the literature does not provide many attempts to use CNN for diagnosis and prognosis of AD. Amongst them we have found two previous works [148, 85] closely related to our approach. In [148] MRI of the whole brain have been used, 150 features were first learned then used for the first and the only convolutional layer of their CNN. Features were using a sparse auto encoder on $5 \times 5 \times 5$ image patches. These features are not further trained during the last back-propagation training stage of the final network, the convolutional layers are followed by a max-pooling layer, a fully connected layer of 800 units and the output units. Results are promising: three- way classification: 89.47%, AD/NC: 95.39%, AD/MCI: 86.84%, NC/MCI: 92.11% on an ADNI dataset of 755 patients in each one of the three classes, for a total of 2,265 scans. In their experiments 3-D patches provided better classification results than 2-D patches. The main distinctions from our work is that we focus on a specific part of the brain while they considered the whole brain, we use more than one convolutional layer and we did not pre-train features.

Another study using 3-D CNN [85] confirms that the usage of CNN is a good choice for classifying MRI scans as belonging to NC/MCI/AD individuals. The 3-D CNN was used on the whole brain and initialised with convolutional auto-encoders, training was done on the CADementia database and the resulting CNN was tested on 210 scans of the ADNI database. Comparisons with other techniques using various image modalities confirm that both the choice of using sMRI and CNN is relevant. Nonetheless, several other previous attempts to use deep neural nets have already been made using multiple modalities. For instance, in [117] multiple indicators from multiple modalities: MRI, PET scans and CSF biomarkers were fused to evaluate the state of the patient. The inputs consist of 93-region-of-interest-based-volumetric features extracted from PET and MRI scans and completed with three bio-markers from the CSF for a total amount of 189 features. Principal Component Analysis (PCA) was then performed on these features, with selection of the discriminative ones. A deep belief network (DBN), consisting of three hidden layers with hidden units of 100-50-20, was trained and dropout was used in the multi-task learning (MTL) to fine-tune the network. The last layer was then used as a new feature representation on which SVM was applied to classify between AD vs. non-AD. Multiple learning schemes have been used and they evaluated their impact by using them or not and measuring the differences in the final classification accuracy. This method, dubbed impact evaluation, showed that dropout and MTL have the major impacts on their performance. The proposed method achieved 91.4%, 77.4%, 70.1%. accuracies for AD/NC, MCI/NC, AD/MCI. For the task of classification converters - to Alzheimer - cMCI vs. stable mild cognitive impaired - sMCI

the accuracy was 57.4%. We note that cMCI are also called progressive MCI (pMCI) and these classification task is very difficult, as the brain scans are not distinctive even for an experience human observer, such as medical doctor.

A similar study, combining multiple modalities and using artificial networks as an element of its learning algorithms is [181]. These studies [181, 186, 205, 51], focusing on prognosis, classify stable MCI (sMCI) vs. progressive MCI (pMCI) . In [205] an accuracy of 65% for sMCI/pMCI was obtained, 70% in [186] 74% in [51] and 83% accuracy for sMCI/cMCI in [181].

Despite many attempts made to use visual visual features extracted in patches, such as Scale Invariant Feature Transform (SIFT) or spectral representations [6], which have shown for a while state of the art results, CNN approaches are consistently superseding them, setting new standards for state of the art recognition. We will speak of filters in CNN instead of features in the remainder of our manuscript. This is often explained by the fact that filters relevant to the application are learnt and not designed and as such are more specific to the application domain on which they have been trained.

In the above cited studies various brain regions were used: the whole brain, the hippocampal region or the cingulate posterior cortex together with the hippocampus, etc. In our work we focus on the hippocampal region. Indeed, some parts of the brain present more modifications than others at the onset of AD. It was shown, using morphometric techniques on the hippocampus, in [49] that studying the hippocampus can give good prediction of the evolution from MCI to AD. The accuracy of these predictions appears to be comparable to classification methods operating at the whole brain level highlighting the relevance of this region. Pennanen et al. [149] showed that using stepwise discriminant function analysis the hippocampal volume can be used to classify AD vs. NC with an accuracy of 90.7% and AD vs. MCI with 82.3%. However, for MCI vs. NC, the volume of the entorhinal cortex provided better accuracy with 65.9% against 59.7% with the hippocampal volume. This 2003 study comprised 59 NC, 65 MCI and 48 AD subject from Finland. There could be a lateralization of the illness, also, as higher resolution scans of the hippocampus show that not all of its parts are affected equally. While the first element, if deemed useful, could be taken into account during the training of the network, CNN would naturally take into account the second element.

4.3 The hippocampal region and visual atrophy in AD diagnosis

In the diagnosis of Alzheimer's Disease (AD), different regions of the brain know structural changes resulting in global atrophy that encircling Gray Matter (GM) along the cerebral cortex. Besides, the changes also include the expansion of the lateral ventricle that is observable from the global view of the brain alongside a local narrowing centered on the medial temporal lobe.

The global brain atrophy analysis does not provide robust discrimination for AD diagnosis. Indeed, numerous researches showed that AD firstly affects some local regions in the brain even

before the relevant symptoms appear; the hippocampus and amygdala are the most common for their vulnerability to Alzheimer's disease.

In chapter 1, we have presented the state-of-art of AD diagnostics in which we discussed the different visual biomarkers used to analyze the AD pathology such as cortical thickness, Hippocampus, Entorhinal Cortex (ERC), and Ventricles enlargement over different MRI modalities scans (3D, 2D, and ROI methods).

Serving with full 3D brain scans can be considered as straightforward as it may seem at first glance. This method has advantages as it has drawbacks. As presented earlier in the chapter 1, using full brain methods can not be regarded as a reliable approach since i) the disease can be detected locally in certain regions instead of being analyzing the whole brain, as we know that the degeneration starts in some areas, and it spreads to the others, ii) some regions in the brain can provide confusing information if we integrate all regions.

For this reason, local ROI methods would be a useful alternative: first, lower quantity of features can be extracted that make studies more convenient; second, a small region can reflect the entire brain since the disease touch only some particular regions. However, ROI-based methods require expertise in the field to identify the proper regions despite extracting regions is a very complicated process. The method needs reliable segmentation techniques to select and specify interest regions from the whole brain scan.

Hippocampus volume is the emerging imaging biomarker to measure the severity of AD; it is considered as a robust discriminating region and well helps in the diagnosis of AD, especially in the early stages of the disease. Admittedly, the hippocampus substance undergoes shrinking in its volume due to cells' death. This phenomenon leads to CSF liquid to fill the resulted extra space that surrounds the hippocampus (black area) [193]. Figure 4.1 illustrates two examples of subjects; it shows the amount of the lost volume of hippocampus detectable with the black area.

We summarize possible different methods which can be employed as follows:

- **3D-level:** This approach provides information on 3D volume; it contains high characteristics. However, this method comprises some brain regions that are not necessarily appropriate, especially whether we want an early AD diagnosis. Furthermore, working with 3D volume requires high computing power (needs a strong GPU capability).

- **2D-slice-level:** Unlike the 3D method, the 2D slice-based approach avoids confronting with a massive number of parameters (millions) to train; the defect of this method is that we lose spatial dependencies in adjacent slices (brain morphology and connectivity in 3D representation).

- **ROI-based method:** The method is suitable for accurate analysis; it has a low feature dimension compared to the 3D volume. Moreover, changes in a small local region may reflect

the pathological degeneration of the entire brain, but has limited knowledge about other regions involved in AD.

Although various methodologies can measure the disease, the ROI-method,leads to plausible performance. Therefore, in this thesis, we focus only on the hippocampal region, which is considered our target ROI by carrying the analysis of all methods. Table 1.1 presents a review of the different introduced approaches (for more details, see chapter 1).

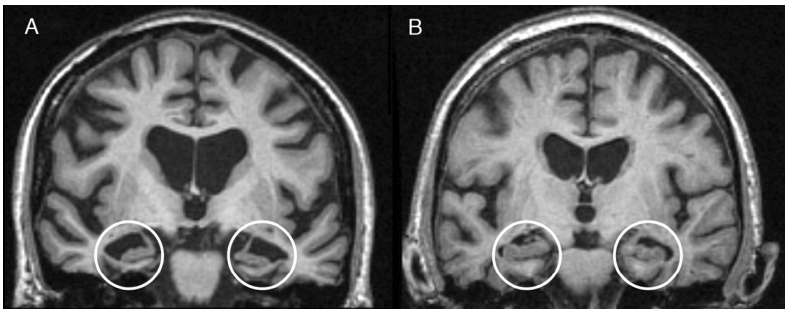


Figure 4.1: Example of the Hippocampus Atrophy: (A) Alzheimer's Disease subject - (B) Normal subject.

4.4 The 2-D+ε Approach

4.4.1 Problem formulation

In the Computer-Aided Diagnosis (CAD) field, the developed technologies can be used to improve diagnosis *in vivo* evaluation in hospitals and research centers. In this thesis, we work on the classification problems for Alzheimer's disease, and we are interested in predicting AD patients from those in MCI/NC stages. In other words, for each selected subject from a given dataset, we aim to estimate to which group he/she belongs. As we have seen in chapter , we can categorize the patients in three different main classes AD, MCI, and NC. The MCI class is composed of two sub-groups (e-MCI and l-MCI); nevertheless, in this work, we consider them only as a single class.

We summarize the process of the proposed approach, as illustrated in Figure 4.2.

As illustrated in figure 4.2, we have a couple of steps to achieve. We divide the problem into two main parts: First, the dataset preparation after data collections and cleaning - it includes several preprocessing procedures such as registration, normalization, and (ROI) extraction. Second, the step of searching and achieving a suitable network design to obtain influential classification scores.

4.4.2 The 2-D+ε concept

In order to design and build powerful models for AD classification, we have investigated various implementations of approaches to achieve better performance. We have confronted numerous

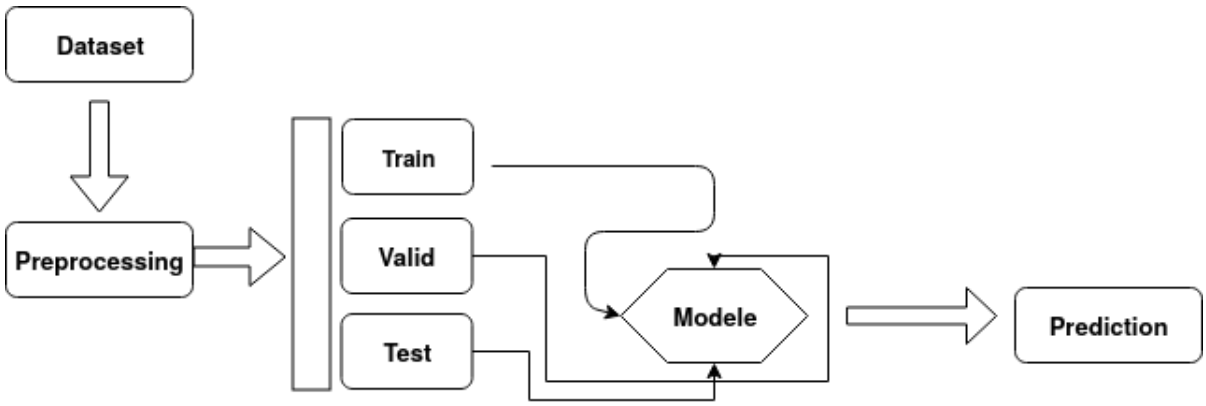


Figure 4.2: The global diagram of proposed approach for Alzheimer's Classification.

restrictions in solving the problem. We come up with two main issues here: 1) how to ensure that the input data were used to feed the model, and 2) how to design a proper network adapted to limited data of a low definition. As we saw in the previous section, the ROI method is targeted, which leads us to use only small data. Based on these constraints, we thought of producing a new approach using a few 2D slices selected from the ROI while preserving the inter-slices spatial dependency. We called it "2-D+ε". In the same vein, we take note that in the chapter 2, we presented data preprocessing, where we obtained a 3D-ROI of both hippocampal regions, and from which we were able to extract axial, sagittal and coronal slices.

2-D+ε Approach: Here is the explanation of the name of the technique "2-D+ε" instead of using only one sagittal slice of the hippocampus and taking into account inter-subject morphological differences we used three adjacent slices. This is not a full 3D approach, but still a 3D portion of the volume comprising a limited number of slices is used. Let us recall that inter-subject differences are kept intact during the alignment (registration) process, which was the rigid-body affine transformation (see chapter 2.4). For a given sagittal slice, from one subject to another, slightly different parts of the hippocampus can be captured. From an implementation perspective, the input layer of the network constituted of $28 \times 28 \times 3$ units and receives data from three 28×28 sagittal central slices of the hippocampal region. Additionally, during data augmentation, we considered translations of these three layers. The translation orthogonal to the sagittal plane led us to consider the two slices adjacent to the three above-mentioned central slices. As a result, the network is trained, considering the five sagittal slices at the center of the hippocampal region. Since we are not training our network directly on the 3-D region that encompasses the hippocampus, which would be of size $28 \times 28 \times 28$, and to distinguish the considered input from this more general input, we called the technique 2-D+ε.

The epsilon refers to the differences between the adjacent slices in a specific plane, such as sagittal. We denote the definition of the method as the following equation:

$$\Delta S^{i}_{Dim_k} = \| S^{i}_{2d} - S^{i+1}_{2d} \|_{Dim_k} = \varepsilon \; where \; k \in \{Sagittal, Coronal, Axial\} \tag{4.1}$$

Where k is the selected projection, and S^i_{2d} is the i-th slice along the axis projection in two-dimensions (2D). Figure 4.3 contains a geometrical illustration of the proposed approach.

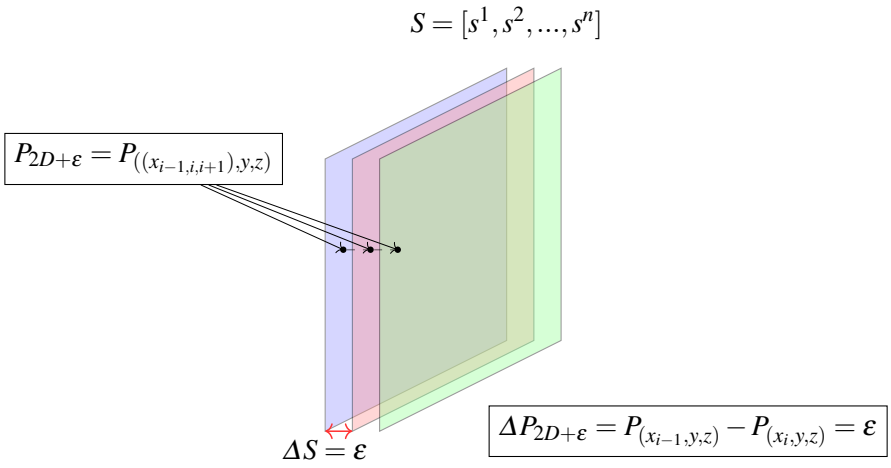


Figure 4.3: Geometric illustration of the 2-D+ε Approach.

The depletion of the computing platform's resources and the execution time are vital constraints; this implies us to think for reducing the computation time and resources allocated to the system instead of draining it. Therefore, using this approach ensures two fundamental pressures. It provides a slightly 3D representation of the target region, with tiny differences inter-slices (the epsilon), which is light compared to full ROI data.

Besides, it avoids to include disturbing information from outlier sides since the region of interest is a 3D form encompassed inside a bounding box (Figure 4.4). Thus taking the central slice and its neighbors lead to reduce and prevent the impact of the undesired surroundings of the hippocampus.

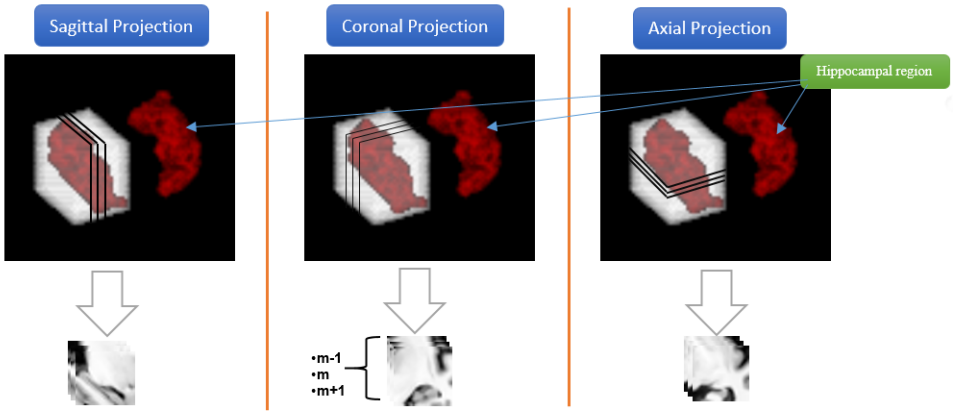


Figure 4.4: Example of the Hippocampus Region: Sagittal, Coronal, and Axial Projections.

4.5 Shallow Architecture design for AD classification

Building a effective classifier model needs in-depth studies and experiences. It requires investigation to find an adequate combination of different parameters for designing the network architecture, namely the number of layers such as the convolution layer, the quantity and the spatial dimension of the filters, and the data reduction factors.

Our dataset contains images of low resolution, the voxel element measures $1.5 \times 1.5 \times 1.5\, mm^3$ (see chapter 2 for further details), the full brain scans have $121 \times 145 \times 121$ dimension which is a relatively good definition. However, the volume of the ROI included in the cubic container has only $28 \times 28 \times 28$ definition. Therefore, we cannot reasonably propose an architecture that is too deep. Indeed, designing a good network depends not only on the input data's definition but also on the content and nature of the used dataset. In our case, it is a medical dataset. However, image data of the hippocampal region describe its pattern in gray-scale representation. We have conceived various implementations of networks with different depths: we have designed a couple of networks containing from one to five convolutional layers for both 2-way and 3-way classification problems. We realize that whether we increase the number of convolution layers, the model loses its power, and thus, an architecture containing three or more layers provides low resulting scores. In chapter 1, we presented some kinds of deep neural networks. The evolution of CNN architecture started from a small and straightforward network like LeNet [113] to very complex networks using a couple of blocks, ResNet is an example. However, before the adventure of these block-based networks, much deeper architectures were already introduced to improve performance. AlexNet is an example of a reasonably Deep simple architecture. Many works tried to overcome the limitation of some small networks by building profound architectures; however, through experiences, they proved that adding more layers does not systematically increase the performance of the network after several layers have been introduced. In our case, since we have a low resolution of the input data, we have studied various settings of architectures with different numbers of convolutional layers. Due to these preliminary experiments which we do not report in the manuscript, we design a shallow network whether we do not need to employ block-based networks. All the architectural principles of conventional deep neural networks are respected in it; specifically, we have convolution transformation designed to extract features, alongside using sub-sampling (pooling) and finally connected to the fully connected layers stacked with the softmax function.

However, we suggested a lightweight architecture to resemble the LeNET [113] in terms of the depth and number of layers. Nevertheless, these two networks are different from input data point of view; our architecture uses the "2-D+ε" input approach, whereas the LeNet uses single channels input data. The figure 4.5 presents the network, besides, different settings for each layer are given.

The architecture of our network consists of two convolutional and two max-pooling layers. Rectified linear units (ReLU) were used as activation functions. One fully connected layer is

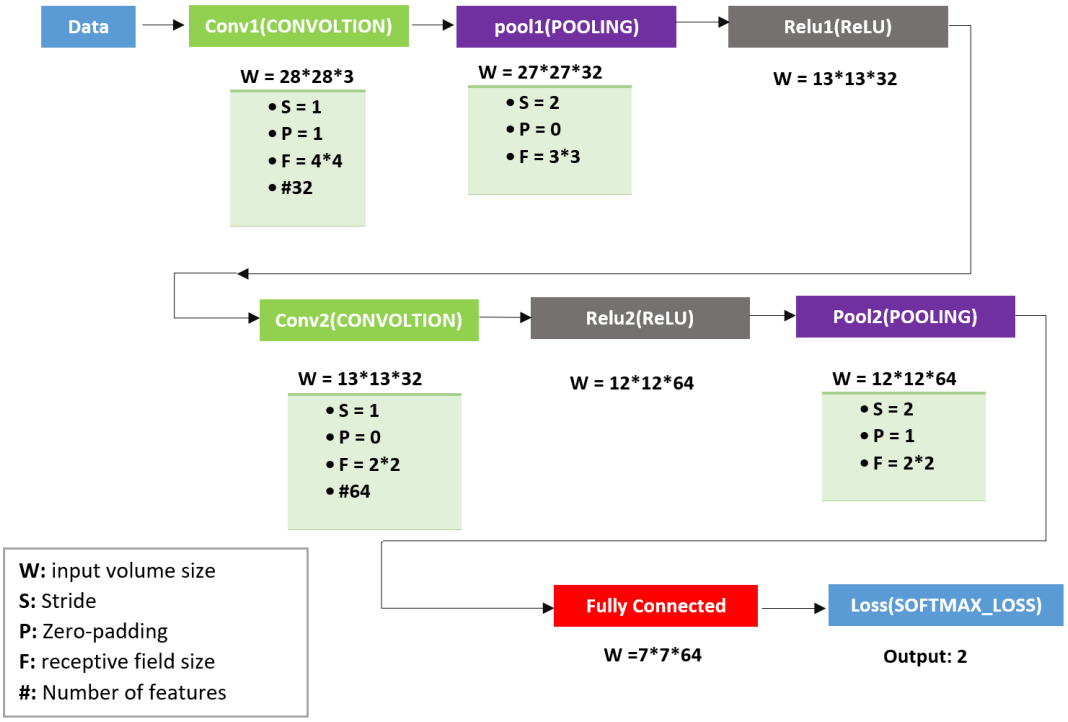


Figure 4.5: Architecture of our CNN: Shallow Network.

introduced before the output. The hyperparameters of our network are presented below and in the table 4.1.

- filter size at convolutional layers: F1 = 4x4, F2 = 3x3
- number of filters at the first conv. layer: N1 = 32
- number of filters at the second layer:N2 = 64
- stride at conv. layers: Sconv = 1
- stride at pooling layers : Spool = 2

4.6 Materials

In this section, we briefly present the dataset composition and its proper preprocessing steps for effectuating our presented approaches.

4.6.1 MRI processing

As described in the chapter 2, sMRI images were processed through a couple step of data preprocessing as shown in Figure 4.6. The preprocessing is based on: (a) a denoising step with

Layer (Type)	Layer (Params)	output shape	Trainable Params
Input data	-	[-1, 3, 28, 28]	-
Conv2D-1	F=4*4	[-1, 32, 27, 27]	1,568
Max-Pool-1	F=3*3	[-1, 32, 13, 13]	-
ReLU-1	-	[-1, 32, 13, 13]	0
Conv2D-2	F=2*2	[-1, 64, 12, 12]	8,256
ReLU-2	-	[-1, 64, 12, 12]	0
Max-Pool-2	F=2*2	[-1, 64, 7, 7]	0
Linear-2	-	[-1, 2]	242
	-	-	386506

Table 4.1: Details of the proposed architecture.

an adaptive non-local mean filter, (b) image alignment (affine registration [16]) in the MNI space [72], (c) image intensity normalization, (d) ROI selection and extraction using the AAL atlas [190].

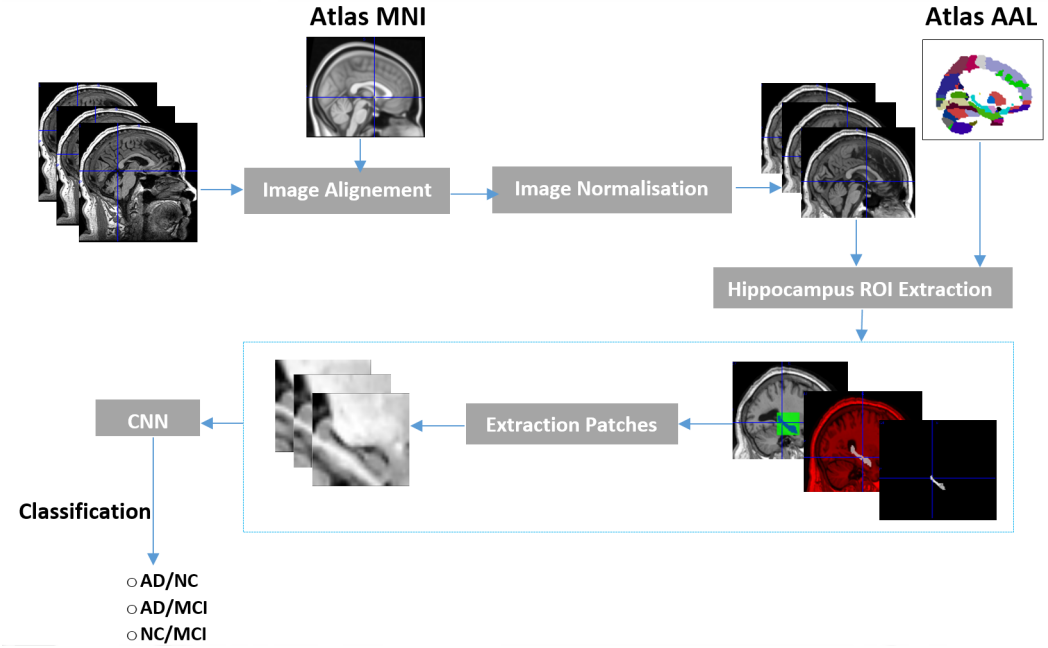


Figure 4.6: Schematic diagram of dataset preprocessing: i) registration of all MRI scans on MNI space, followed with intensity normalization. ii) ROI selection process using the Atlas AAL for both hippocampal regions. iii) 2D-slice extraction from selected 3D-volume. iv) feeding the CNN networks [1].

4.6.2 Data groups

Data used in this chapter contains 815 baseline structural MRIs from the ADNI-1 dataset with 188 Alzheimer's Disease (AD) patients, 228 cognitively normal (NC) and 399 Mild Cognitive Impairment (MCI) subjects. Images are standard 1.5T screening baseline T1-weighted obtained using volumetric 3D MPRAGE protocol. Demographic information about this group is given in Table 4.2.

	Classes	# Subjects	Age [range] / $\mu(\theta)$	Gender (#F/ #M)	MMSE [range] / $\mu(\theta)$
ADNI-1	AD	188	[55.18 90.99] / 75.37 $\pm$ 7.52	99/89	[18 27] / 23.3 ($\pm$ 2.03)
	MCI	399	[54.63 89.38] / 74.89 $\pm$ 7.30	256/143	[23 30] / 27.0 ($\pm$ 1.78)
	NC	228	[60.02 89.74] / 75.98 $\pm$ 5.02	118/110	[25 30] / 29.1 ($\pm$ 1.00)

Table 4.2: Demographic description of the ADNI dataset group. Values are reported as mean and $\pm$ standard deviation.

**Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD)*.*

4.7 Experiments and results

4.7.1 Evaluation metrics

To evaluate performance of our CNN classifiers in both single modality and fusion tasks we considered the metrics widely used in medical statistics. We denote *tp*, *tn*, *fp*, and *fn* respectively True positives, True negatives, False positives, and False negatives. The metrics used are as follows:

$$\text{Accuracy (Acc)} = \frac{tp+tn}{tp+tn+fn+fp} \tag{4.2}$$

$$\text{Sensitivity (Sen)} = \frac{tp}{tp+fn} \tag{4.3}$$

$$\text{Specificity (Spe)} = \frac{tn}{tn+fp} \tag{4.4}$$

Finally, balanced accuracy (BAcc) measure, which is the average of sensitivity and specificity is defined as:

$$\begin{aligned} \text{BAcc} &= \frac{1}{2}\left(Sen+Spe\right) \\ &= \frac{1}{2}\left(\frac{tp}{tp+fn}+\frac{tn}{tn+fp}\right) \end{aligned} \tag{4.5}$$

Here True Positives (TP) are AD patients correctly identified as AD, True Negatives (TN) are controls correctly classified as controls, False Negatives (FN) are AD patients incorrectly identified

as controls and False Positives (FP) are controls incorrectly identified as AD. Similar definition is hold for other binary classification problems NC/MCI and AD/MCI.

Hardware configuration: The experiments were conducted on a GPU-based high-performance computing platform featuring an Intel(R) Xeon(R) CPU E5-2680 v2 @2.80GHz processor, 187 Gb of RAM, equipped with two Nvidia TESLA P-100 graphics cards with 16GB dedicated memory. The computational time for one epoch at the training step with batch-size of 64 samples was 2.03 seconds in average.

The network was implemented with "Caffe" Deep learning Framework [96]. we have added a dropout layer to tackle the over-fitting phenomena as a method of regularization.

4.7.2 Specific data augmentation

The classification of medical images with Deep learning classifiers suffers from a lack of data. One of the direct ways to overcome this difficulty consists of data augmentation (DA) [28]. This is a process of generation of new data from existing data. Recently, such a generation is performed with Generative Adversary Networks (GANS) [79], but it is still a direct way to augment the data using domain knowledge remains plausible. For further details, (see chapter 2.4). The selected dataset has been split as follows: 60% for the training, 20% for validation, 20% to test the trained classifier. We proposed a domain-dependent data augmentation process compatible with the MRI acquisition.

The *first* way to augment the data consists in blurring it, as proposed for general-purpose images in [96]. This imitates possible contrast variations in original scans. The blurring was fulfilled with 3×3, 5×5, 7×7 Gaussian filters with (weak) scale parameter: $\theta = 0.7, 0.7, 0.6$ respectively. We illustrate the effect of blurring in figure 4.7

Our *second* augmentation technique is the translation of the hippocampal ROI by ± 1 pixel in each dimension, providing thus seven times more data than the original one. This considers possible variations due to alignment imprecisions of scans on the MNI template.

Finally, a "flipping" technique was used. Since the hippocampus is a symmetrical structure of the brain, for each scan twice as much information is obtained by flipping the left hippocampus to match the right one. This is summed up in Table 4.3.

Note, that the three DA techniques were applied only to the training and validation data. In our work we follow a scenario where no pre-processing is operated on a new brain scan submitted for classification; as such no DA is performed on the test dataset.

Table 4.4 presents early scores obtained on different metrics of the network on the original data with a ten-fold DA (additional blurred images, translated images and flipping). The over-sampling of the MCI category is noticeable for the sensitivity of AD/MCI and the specificity of MCI/NC. This suggests to balance the dataset.

	O	**G**(×3)	**T**(×6)	**(O+T+F)×2**
AD	188	564	1128	3760
MCI	399	1197	2394	7980
NC	228	684	1368	4560
Total	815			16300

O: number of original scans
G: number of Gaussian filtered scan
T: translated scans
(O+G+T)×2: total number of data input

Table 4.3: Data augmentation: "G" is the Gaussian blur, "T" is the translation, "F" is the flip.

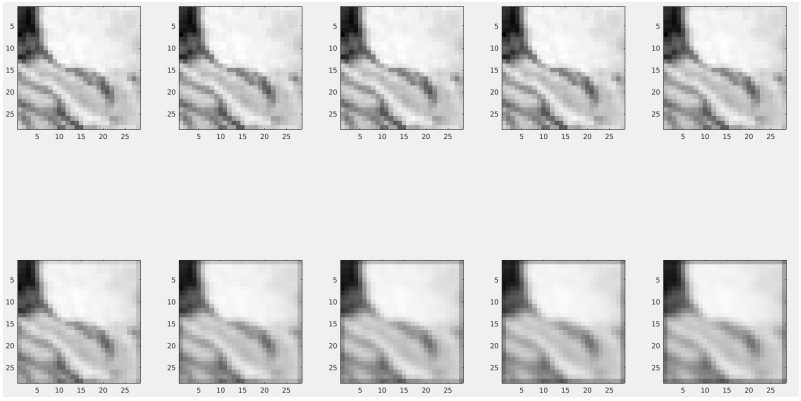


Figure 4.7: Multi-instance of the selected central slice (sagittal view) with different gaussian blur settings.

	AD/NC	AD/MCI	MCI/NC
Accuracy	83.7%	66.5%	64.9%
Sensitivity	79.16%	*36.76%*	76.9%
Specificity	87.2%	79.01%	*45.2%*
scans	AD(188), MCI(399), NC(228)		

Table 4.4: Binary classifications with augmented data (10x): flip, translation, blur.

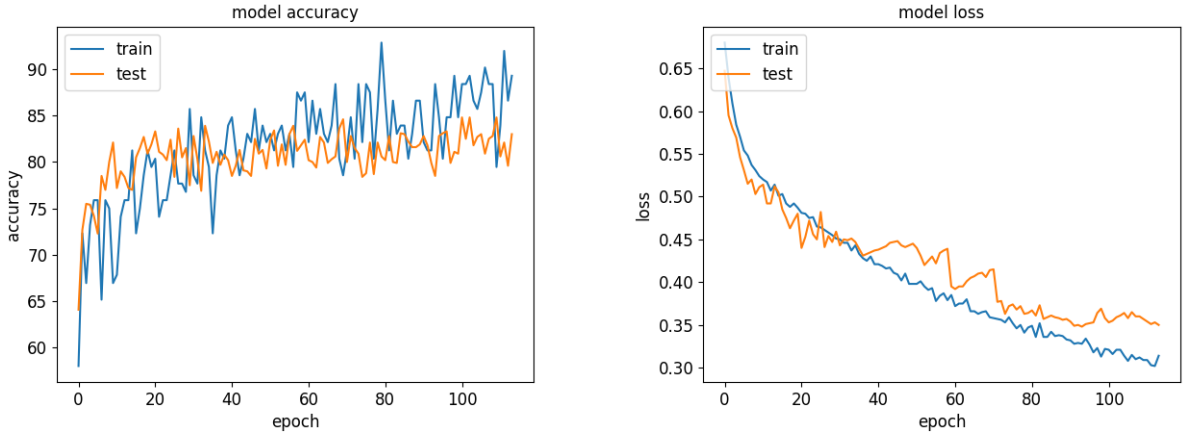


Figure 4.8: AD/NC: An example of accuracy and loss plots during training the network.

4.7.3 Evaluation

In this section we report the results we obtained when applying different training strategies with the designed shallow Network architecture and "2-D+ε" approach.

Balancing the Data:

Balancing the data consists of taking a similar number of samples for each category. We first have studied three balancing strategies on AD (188 scans) vs NC (228 scans) classification task, as these two classes are more easily separable. The three techniques of data balancing that we used are: i) the reduction of the number of original scans from the over-sampled category. Here 188 scans are taken for both AD and NC categories; ii) the random duplication of the original data from the under-sampled category. Here 228 scans were used for both AD and NC categories; iii) the random reduction of the augmented data from the category having more data, thus keeping the data variety of the dataset: 188 $\times$ 2 $\times$ 28 = 10528 scans. The results are given in table 4.5.

	1	2	3
Accuracy	82.8%	71.3%	79.3%
Sensitivity	79.68%	70.31%	82.81%
Specificity	85.93%	75%	75.86%

Table 4.5: Data balancing: (1) simple data reduction, (2) data augmentation by duplication of original scans, (3) randomized reduction of the augmented data.

From this table we can see, that the first data balancing technique by a random data reduction is the most effective. Indeed, in this case we preserve the original data and do not induce bias by data duplication. Hence we have applied it to the more challenging classification problems of AD/MCI and MCI/NC. Results are given in table 4.6. Despite a drop in accuracy, linked to the loss of training data, and showing the necessity of a data augmentation strategy, the three metrics are more balanced compared to results without data balancing.

	AD vs MCI		MCI vs NC	
	unbalanced	balanced	unbalanced	balanced
Accuracy	66.5%	63.22 %	64.9%	58.23%
Sensitivity	36.7%	60 %	76.9%	63.33%
Specificity	79.1%	67.14 %	45.2%	52.5 %
scans	AD(188), MCI(399)	(188), MCI(199)	MCI(399), NC(228)	MCI(199), (228)

Table 4.6: AD vs MCI and MCI vs NC with and without a roughly equilibrated number of scans (reduction balancing) with blurred images.

Blurring the Dataset:

To augment the dataset, a filter is used to blur scans. These newly generated scans are added to the training and validation datasets. Comparing the accuracy of the resulting trained network, all other parameters being kept unchanged, showed an increase in accuracy for AD/NC which seemed to validate the approach.

However, if we consider MCI/NC the results are opposite; for some reasons, probably related to small enough differences to be discarded by blurring, the blurring seems to affect negatively accuracy and specificity for this binary classification (table 4.7).

A similar effect appears for specificity in the result section when comparing the second and third part of table 4.11. This could be linked to the nature of the visual differences between these two categories and how it is affected by blurring.

	AD/NC		MCI/NC	
	0 blurring	+blurring	0 blurring	+blurring
Accuracy	80.7%	83.7%	69.1%	*58.23%*
Sensitivity	74.28%	79.1%	52%	52.5%
Specificity	86.45%	87.2%	84.21%	*63.33%*
scans	AD(188), NC(228)		MCI(199), NC(228)	

Table 4.7: AD/NC and MCI/NC with and without additional *blurred* images with reduction data balancing.

Applying blurring to all translated images the following results were obtained:

	AD/NC	AD/MCI	MCI/NC
Accuracy	83.5%	67.5%	65.1%
Sensitivity	81.92%	74.92%	77.64%
Specificity	86.74%	58.97%	52.94%
scans	AD(188), MCI(199), NC(228)		

Table 4.8: The results with *translated* and *blurred* images including the reduction data process for balancing.

Other tests on data balancing

In order to optimize the usage of the scarce available data we tried different balancing techniques. For AD duplicated in order to train the network, for each epoch equally on AD scans and MCI scans.

These three techniques have been used to categorize AD (188 scans) vs NC (228 scans):

1. reducing the number of scans, 188 scans for both categories, thus loosing part of the data.
2. randomly duplicating scans of the under-sampled category, 228 scans for both categories.
3. reducing the augmented data thus keeping the data variety of the dataset, 188*2*28=10528 augmented scans.

	1	2	3
Accuracy	82.8%	71.3%	79.3%
Sensitivity	79.68%	70.31%	82.81%
Specificity	85.93%	75%	75.86%
scans	A D(188), MCI(199), NC(228)		

Table 4.9: Data balancing, (1) simple data reduction (2) data augmentation by duplication (3) randomized reduction of the augmented data

The results appear in the following table (Table 4.9).

For (2) the idea was to train the network in such a way that for each epoch (see section 3.5.1), the data would be exposed to as many samples from both categories. Although more testing should be done it seems that the idea is misfunded.

Perhaps surprisingly, as there is a loss of original data, the simple reduction provided better results. Assuming generalization to other binary classification we opted for this approach to carry on tests on other binary classifications.

Ternary classification

Ternary classification lead to poor results (AD: 33.82%, MCI: 62.80%, NC: 52.12% of exact classification) the overall accuracy is 53.5%. mis-classification for AD and NC could be due to the oversampling of MCI subjects.

		Predicted classes			
		AD	MCI	NC	exact classification
Real	AD (68)	23	*42*	3	34%
classes	MCI (164)	26	35	103	21%
	NC (94)	3	*42*	49	52%
Prediction rates		44%	55%	56%	
Balanced prediction rates		63.98%	16.70%		

Table 4.10: Confusion matrix for 3-way classification

thus we dropped this avenue that might not be practical for clinicians anyway.

4.8 Discussion and comparison

4.8.1 Results of the method

The following table presents the test results of the CNN. We used a 56-fold Data Augmentation (DA) ($\times$ 7 from translations, $\times$ 4 from blurring and $\times$ 2 from symmetry) and reduction data balancing that balances sensitivity and specificity at the expense of a slight reduction in accuracy. Tests were performed on non augmented data first except for the "flipping" technique. Then translation was added

and finally blurring. In each binary classification task the number of original samples for training the two-class network was the same : 188 for AD vs NC, and 228 for MCI vs NC and 188 for AD vs MCI.

Wihtout data augmentation but for the flip (x2)

	AD vs NC	MCI vs NC	AD vs MCI
Accuracy	82.2 %	61.8 %	64.7 %
Sensitivity	88.3 %	53.0 %	70.0 %
Specificity	76.6 %	68.3 %	60.0 %

Data augmentation using translation and flipping (x14)

	AD vs NC	MCI vs NC	AD vs MCI
Accuracy	76.6 %	60.8 %	60.6 %
Sensitivity	73.4 %	60.6 %	57.8 %
Specificity	81.2 %	62.5 %	64.0 %

Trained with blurred, translated and flipped images (x56)

	AD vs NC	MCI vs NC	AD vs MCI
Accuracy	82.8 %	66.0 %	62.5 %
Sensitivity	79.6 %	73.7 %	60.0 %
Specificity	85.9 %	58.7 %	64.0 %

Table 4.11: Impact of the data augmentation on results

Contrary to what we expected, accuracies are reduced by the translation DA scheme, see part 1 and 2 of table 4.11. Because the stride is 1 for the receptive fields of the first convolution layer, translations, except for borders, have mostly no effects on the data over which filters are trained, additional tests with stride 2, that will provide a blindness effect compensated by the translations DA strategy are needed to have a finer interpretation of what is happening. The fact that border effects have such an impact on the formation of filters might rise some question about the validity of inter subject variability hypothesis, however from a creativity point of view it might also led to some innovation. Nonetheless all metrics, except specificity for MCI/NC, are boosted back by augmenting the dataset with blurred images (second and third part of table 4.11). It remains to determine which blurring parameters provide best results. It would also be interesting to isolate the effect of the blurring DA scheme without the translation DA scheme

4.8.2 Comparison

We assumed the real-life scenario of CAD interactive system where only real-world images obtained from MRI are submitted for classification without a preprocessing stage. Hence, in the test set we have not used any data augmentation. The scenario model being that a doctor submits the digital scan to the system which will return the probability for his patient to be in one of the three classes NC, AD and MCI.

To put our result into perspective we can use as baseline, studies on patch-based classification which, for the classification task of AD/NC, have an accuracy of 80%, sensitivity 81% and specificity 79% [186, 205]. In addition, it was shown [205] that an accuracy of 78% can be obtained for pMCI/NC with the same feature. These results can be compared to the 2 first columns for AD/NC and MCI/NC of the table 4.11.

In our study we used only five gray-value patches that are coarse segmentation of the hippocampus on a saggital projection as presented in the previous section. We can see, that our approach gives comparable and even slightly outperforming results for separation of these classes on nearly the same set of sMRI images.

The direct comparison of MCI classification is not possible, as they use more fine taxonomy of stable MCI and progressive MCI (Alzheimer converters). In [205] they use multiple features extracted from sMRI and by a Linear Discriminant Analysis (LDA) obtain rather good classification results for NC vs progressive MCI of 82%. In our work we used only grey-value patches and one projection. In [148] accuracies obtained are: 89.47% for 3-way classification, 95.39% for AD/NC 86.84% for AD/MCI and 92.11% for MCI/NC. Nevertheless the whole brain scan has been used to obtain these results. In [85] results in accuracy are respectively 89.1%, 97.6%, 95% and 90.8%, the classifier has also been trained on whole brain scans but this time on the CAD Dementia dataset and tested on 210 scans of the ADNI dataset. The base-line approaches [186], [205] do not use a CNN classifier. In our case, even if the architecture of the network is shallow, the quantity of available original scans for training is rather low. This can explain this difference. Furthermore, in our work we used only one, saggital projection of sMRI.

To interpret this, one has to keep in mind that morphological MCI features are likely to be closer to NC features than AD features, additionally the MCI range could be quite large it could then be difficult for specific MCI trained features to emerge stage classification from symptoms or from structure (MCI early stage, closer to NC farther from AD blurred line between AD/MCI plus a large range of MCI, feature training how blurring can affect them ? small sulcus could be masked by blurring disturbing the trained feature.

4.9 Conclusion

From this pilot study a few pathways are emerging. Since CNN discriminative power notoriously depends on the size of training data, we have artificially augmented the dataset by translating and blurring it. While the first DA technique by translation was not conclusive the DA technique using blurring has shown a strong potential of improvement for our type of classifier. Additionally, we measured the influence of a few data balancing techniques and showed that balancing the data by suppressing excess data from one category was providing best results. Despite using only a subset of

the hippocampal region, through the "2-D+ε" approach, encouraging accuracy results were obtained: AD/NC 82.2% MCI/NC 66.0% AD/MCI 62.5% that validate the approach and confirm independent and similar works on the topic [148, 85]. While figures for MCI binary tasks classification remain to be improved, this should naturally occur with the different following improvement perspectives. Considering the input of the networks the direct perspective of this research is to further this study with coronal and transversal slices. Combining these data could lead to a new type of feature/kernel. The more classical avenues of 3-D convolutional network and the usage of unsupervised pretrained filters could also be followed. From the point of view of the network architecture, a first improvement would be to add a fully connected layer of n units (n to be determined) just before the softmax layer. Finally, more investigations on the translations DA strategy has to be carried out. Tweaking the stride parameters of the network architecture could be the key.

Chapter 5

Data Fusion for Alzheimer's Disease Recognition on Brain Imaging.

5.1 Introduction

In the previous chapter, we introduced the "2-D+ε" as an approach for modeling input data. Besides, we designed a shallow network for the AD classification problem. Our proposed approach provides promising results, although it uses a restricted region with fewer slices. However, to enhance the classification performance, various methods and strategies can be used. The most efficient research works typically used various tracks, namely, ensemble methods, multiple algorithms, feature selection, fusion, and combining models and data [159, 151].

In our work, we seek to build strong classifier models by introducing multiple fusion approaches to design efficient architectures, broadening the classifier's capability from multiple data sources.

This chapter follows the same approach "2D+ε" as proposed in Chapter 4, where we accommodate multiple networks through different projections (Sagittal, Axial, and Coronal), moreover, implying different MRI modalities.

Highlights:

- We present the fusion framework through different mechanisms: early, intermediate and late fusion;
- We provide various aggregation methods in order to increase performance;
- We combine the selected approach by using data from different planes and MRI modalities;

5.2 Related work

In spite of its success in the classification problems tasks, CNNs are in their infancy to be used for decision making in brain medical image classification. Still their use is massively researched today completed with domain knowledge of AD phenomena in the brain. Here we will briefly discuss some of them. In [148] the authors have taken a two-stage approach on the whole MRI brain scans, firstly they used a sparse auto-encoder to learn filters for convolution operations, and secondly they built a 3D CNN whose first layer uses these learned filters. The auto-encoder was made with 150 hidden units, and was trained on a set of 3D patches of size $5\times5\times5$, extracted from the MRI scans. The 3D CNN architecture was made up of convolutional layers followed by a max-pooling, a fully-connected layer of 800 units and the output units. In a 3-class classification problem (AD, NC, MCI) they achieved the accuracy of 89.47% which was 4% higher than on a 2D projections. In case of pairwise binary classification problems they have achieved better accuracies AD vs. NC:95.39%, AD vs. MCI: 86.84%, NC vs. MCI: 92.11%, in the case of 3D convolutional networks on the ADNI dataset that consists of 755 patients in each one of the three classes (AD, MCI, and NC), for a total of 2,265 scans. The main distinction from our work is that we focus on a specific part of the brain while they considered the whole brain. We use more than one convolutional layer and we did not pre-train features. Another study using 3-D CNN [84] confirms that the usage of CNN is a good choice for classifying MRI scans as belonging to NC/MCI/AD individuals. The 3-D CNN was used on the whole brain and initialized with convolutional auto-encoders, training was done on the CA Dementia database and the resulting CNN was tested on 210 scans of the ADNI database. Comparisons with other techniques using various image modalities confirm that both the choice of using sMRI and CNN is relevant. The studies [186, 51, 205] are focusing on prognosis, the problem here is to classify stable MCI (sMCI) vs. progressive MCI (pMCI), also called MCI converters (cMCI), see chapter 4.1. In [205] an accuracy of 65% for sMCI vs. pMCI was obtained, 70% in [186], 74% in [51] and 83% accuracy for sMCI vs. cMCI in [181]. Multiple modalities have been used in [117], MRI, PET scans and CSF biomarkers were fused to classify subject state disease, 93 regions-of-interest were extracted from MRI and PET scans. A total of 189 features were used by adding 3 bio-markers from the CSF, and Principal Component Analysis (PCA) was applied. In a recent paper [179], the authors use MRI image segmentation into three tissue types of Gray Matter (GM), White Matter (WM), and Cerebrospinal Fluid (CSF). They then parcel them into 93 regions of interest (ROIs). Only the GM densities spatially normalized were considered in this work which is widely used in the field for AD classification problem. Their architecture named DeepESRNet was made up by two convolutional layers and a max-pooling layer, followed by two fully-connected layers. The proposed method achieved, for AD vs. NC 91.02%, 92.72%, and 89.94%, and for MCI vs. NC 73.02%, 77.60%, 68.22%, of Accuracy, Sensitivity, Specificity respectively. In the above cited studies various brain regions were used: the whole brain, the hippocampal region or the cingulate posterior cortex

together with the hippocampus, etc. In our work we focus on the hippocampal region as it is the strongest biomarker of Alzheimer disease and has not been sufficiently studied yet.

5.3 Fusion methods: From Single model to data and models combination

Aiming to improve models' performances, specific methods can be introduced to achieve better prediction capability. The primary purpose of all decision support systems (DSS) is to build a robust model that takes restricted or minimum input data and provides correct decisions. The fusion approach is such a method. It is an effective way to improve single-classifier results combining diverse data sources (features) or/and classifiers (models). This method remains a choice among available strategies to consolidate performance and integrate heterogeneous information from various representations. Indeed, it is inferred that if individual method performs well, multiple combinations of such of them may reduce overall classification errors and emphasize correct outputs. With its variances, this method has been widely introduced toward many computer vision and machine learning applications. However, the combination can be carried out on three levels of abstraction closely connected with the flow of the classification process: data-level fusion, feature level fusion, and classifier fusion [29]. There are two groups of fusion approaches; first, they generally operate on classifiers and emphasize the development of the classifier structures. Second, they operate mainly on classifiers output, and effectively the combination of classifier outputs, is calculated. Class labels, class ranking, and soft/fuzzy outputs are the main groups of classifier fusion systems.

In the previous chapter (4), we proposed an AD classification model based on a single network of data (uni-source) with solely a few patches (the "2-D+ε" input). However, this model design may encounter limitations of incoming data such as the lack of a complete representation of the region-of-interest; moving from 3D data to restricted 2D slices. Nevertheless, to overcome these drawbacks, different promising approaches using single or multiple characteristics over multi-source data resolve or alleviate this issue. Indeed, if the intention is to avoid considering the whole brain regions, a pragmatic combination of features coming from different views (planes) or areas (other ROIs) and even implying other imaging modalities, may yield complementary information of the target region. Here, the key point is to continue using the ROI-level method and add more useful information from other sources - joining features from alternative sources may produce better results. Thus, the combination methods can be considered suitable for studying the brain and its pathologies since the disease affects only some specific regions.

To remain with ROI-based method in a fusion framework is nevertheless justified since the disease affects only some specific regions. These approaches are based on different multi-projection and multi-modal features. On the other hand, it was observed that the disease-induced structural changes

also occur in several inter-related-regions; thus, the correlations between different brain regions could also be extracted for more accurate characterization of brain pathology. Still in our work we focus on only one region - hippocampal and will apply fusion methods on it.

However, if more irrelevant and noisy information is included in the feature set, the disease classification and interpretation could become very difficult due to the small number of training samples in the neuroimaging study. Although promising results have been reported for brain image analysis in the above studies, it is still potentially advantageous to investigate building and combining multiple classifiers for making full use of the rich imaging and structural information, to improve classification performance [122].

Here we can summarize three levels (or ways) to combine different useful sources:

- **The early fusion (low-level):** This strategy merges data in the input. Here, the raw data or pre-extracted features are combined from different sources and stacked to build a single input to the classifier. In the case of AD classification, the new feature vector becomes a single element with higher dimensionality and represents the single brain from different spaces. For instance, if the features are extracted from two independents ROIs, creating the new vector via concatenation yields the two vectors into a single new vector.

- **The intermediate fusion (mid-level):** As mentioned above, many fusion methods operate on the classifiers themselves rather than on their outputs. Here we study how to improve classification performance by advancing multiple classifiers in a single optimized structure. Indeed, each model can be combined with its counterpart through different internal mechanisms to share characteristics or scores in order to improve classification results. Concatenation is such an example of an intermediate method usually used to consolidate models.

- **The late fusion (late-level):** The classifiers produce the least amount of useful information for the combination process. Each model can receive multiple input data, and over its layers, it provides predicted class individually. Afterwards, the fusion lies at the decision level, which means taking the post-decision for every single model and applying a selected fusion method on classification scores to make the final decision. The two most representative methods used in this level are the generalized voting and Knowledge-Behavior methods [159].

5.4 Fusion application for AD classification

This section presents our fusion methods based on the same architecture previously showed in the chapter 4. We furthermore use the "2-D+ε" input approach, whereas we bring and integrate data from various sources. We serve on the sMRI data in the first step, and yet here; we only extract region

representation from different projection (planes). Next, we incorporate data from two modalities: sMRI and DTI, and tracking data from different view projections in the second step. However, we implement both strategies: Intermediate and late fusion, as seen earlier on either for single or both modalities.

5.4.1 Intermediate fusion designs

As seen above, fusion can be applied in different ways using three single networks - each image projection for each network. The only constraint is that the three input images must be taken from the same subject and the appropriate projection to feed the networks correctly. However, in this part, we consider the intermediate fusion, which consists of a concatenation layer at the three networks' FC layer. Indeed, as we have a 3D-ROI representation of our hippocampus region, we take the three projections sagittal, coronal, and axial as input, respectively, for each single network. Then, the networks are combined through the intermediate fusion by concatenation of the fully connected layers. Afterward, we get an FC layer merging the three combined layers into two output scores to reach the binary classification. An overview of the full architecture is illustrated in Figure 5.1. Each network for a single projection was implemented using the same architecture 2-D+ε approach made up of two convolutional layers, two pooling layers, followed by the ReLU activation function, a fully-connected layer as defined in the chapter 4.

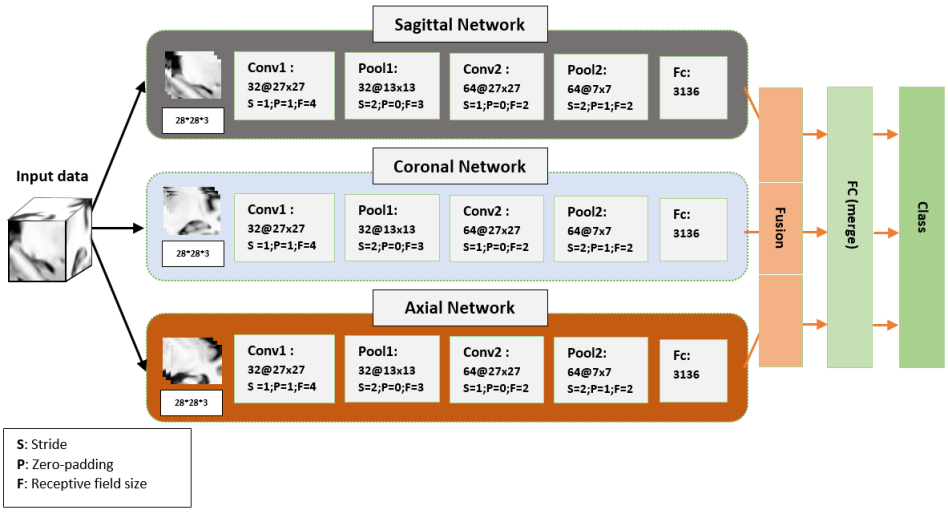


Figure 5.1: Intermediate fusion architecture: built for three the projection input data (Sagittal, Coronal, and Axial) of the sMRI modality.

5.4.2 Late fusion designs

Contrary to the intermediate fusion approach, which relies on fusion within models. The late fusion consists of applying some specific operations to the outputs of the last layers of each network. In

other words, we affect the operation either on the scores of the outputs of the lats layers or on the post-decision of each individual network to generate the final classification.
We distinct nevertheless two late fusion designs: (a) - we perform algebraic aggregation on the outputs such as mean, median, and max, followed by the softmax function to convert the scores to a probability decision. In (b) - we use the majority vote system on the post-decision of all models. Figure 5.2 illustrates the both architectures.

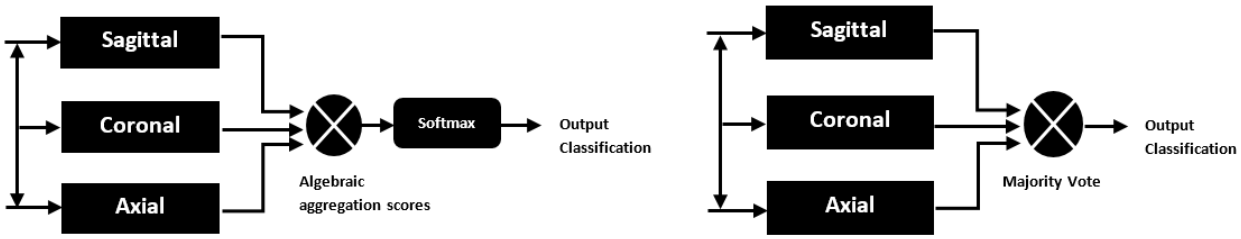


Figure 5.2: Late fusion-level using two strategies: (a) - different algebraic aggregation on scores (b) - Majority vote on final decisions.

Here we present the aggregation functions and the majority vote algorithm used in elaboration of this work:

- **Algebraic aggregation**

  Reminder: The "softmax function" (eq. 5.1) takes a vector of real values as input and convert them to range between 0 and 1. The sum of the latter values equal to 1 (See Chapter 3).

$$f_j(z) = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}}, \; with \; j = 1, ..., K \tag{5.1}$$

  Where K is the vector size of the input data, and j is the current element of the vector Z.

  In our case, as we have binary AD classification, the soft-max function takes a vector with only two real-value. The latter is the result values after applying the aggregation function on the output scores simultaneously for every single model. We define our condition as follows:

$$j = \{0,1\}, Z \in \mathbb{R}^2, k = \{0,1\}, f_j \in [\![0,1]\!] \tag{5.2}$$

  We proposed the following fusion functions to improve classification performance.

$$\alpha = max; \alpha_j = \max_{i=1..M} \left(f_j^i\right)$$
$$\alpha = mean; \alpha_j = \frac{1}{M} \sum_{i=1}^{M} \left(f_j^i\right) \quad \text{and } \underbrace{\mathbb{R}^2 * \cdots * \mathbb{R}^2}_{M}$$
$$\alpha = min; \alpha_j = \min_{i=1..M} \left(f_j^i\right)$$

Where M is the number of single networks, and j is j_{item}^{th} element of the output score for M network.

$$Z^{(i)} \begin{pmatrix} z_1^{(i)} \\ z_2^{(i)} \\ \cdots \\ z_K^{(i)} \end{pmatrix} \rightarrow^{softmax} f^{(i)} = \begin{pmatrix} f_1^{(i)} \\ f_2^{(i)} \\ \cdots \\ f_K^{(i)} \end{pmatrix} \tag{5.3}$$

- **Majority vote**:

  The second method is the majority vote system (see the pseudo algorithm below). In this approach, we implement the voting system on the post-decision for every single model in order to produce the final decision for a specific binary classification.

Algorithm 1 Majority vote algorithm pseudocode.

Input: A sequence **S**.
Output: Final decision.

```
1:  m ← α
2:  i ← 0
3:  for x of the sequence S do
4:      if i = 0 then
5:          m ← x
6:          i ← 1
7:      else
8:          if m = x then
9:              i ← i + 1
10:         else
11:             i ← i − 1
12: return m
```

Here we have some comments regarding the functioning of this system:

– Even when the input sequence has no majority, the algorithm will report one of the sequence elements as its result.

- It is possible to perform a second pass over the same input sequence in order to count the number of times the reported element occurs and determine whether it is actually a majority.
- In our case: the sequence consists of 3 and 6 elements, respectively, for the MRI projections and the MRI joined to the DTI-MD data. Both have outputs at intervals of [| 0, 1|].

5.4.3 Final multi-modal fusion architecture

In this part of our work, we apply the fusion approaches presented above. We use both modalities sMRI and DTI data. However, instead of using only three networks, we managed to enhance our models' design by modifying them to accept multiple data. Indeed, the design is realized by linking six networks of both modalities and all projections in each of them. We thus build a full siamese architecture presented in figure 5.3.

From left-to-right, we have the input of three slices of each projection for both modalities (sMRI and DTI-MD). Then the single branch network is designed and parameterized as that one presented in section 4.5. Finally the fusion layer consists in the two strategies presented before: The concatenation of the features (the intermediate fusion), and the aggregation functions besides majority vote as second application (the late fusion). The both approaches are applied on the six networks.

In the first design, we have implemented the intermediate fusion as introduced in the section 5.4.1, yet here it was modified to work in parallel over sMRI and DTI data. The output of each network has been concatenated to feed an FC layer, as depicted in Figure 5.3. In the second implementation, we followed the same way as in section 5.4.2; we have combined networks using late fusion. The two methods: the algebraic aggregation and the majority vote; nevertheless, in the case, we have designed to adapt multi-modal features.

We notice that this fusion scheme's implementation is provided in the following chapter (chapter 6). Furthermore, we present the transfer learning approach - which is not discussed in this chapter - in order to be able to use the DTI-MD modality associated with MRI data. In the next chapter, we present the two fusion methods using only MRI images. We present the materials and the results of our experiments.

5.5 Materials

In this chapter, we worked with the MRI materials associated with the previous chapter (see chapter 4.6). We use the same subject composition; however, we exclusively used the data balancing approach toward our augmentation process, which is considered the most suited setting for training our models, as proven by experiments.

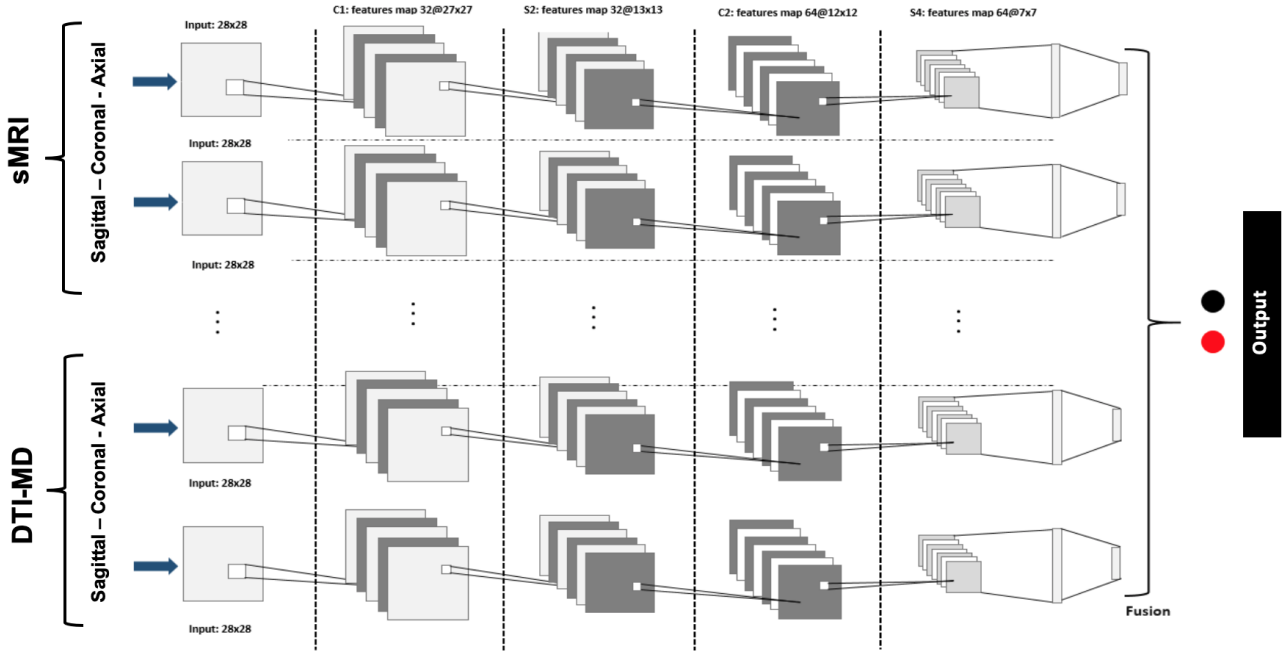


Figure 5.3: Multi-modal intermediate fusion architecture: Data come from sMRI and DTI scans, and the fusion is applied on six single network.

We retain the description of our ADNI-1 dataset (see table 5.1).

	Classes	# Subjects	Age [range] / $\mu(\theta)$	Gender (#F/ #M)	MMSE [range] / $\mu(\theta)$
ADNI-1	AD	188	[55.18, 90.99] / 75.37 $\pm$ 7.52	99/89	23.3 $\pm$ 2.03
	MCI	399	[54.63, 89.38] / 74.89 $\pm$ 7.30	256/143	27.0 $\pm$ 1.78
	NC	228	[60.02, 89.74] / 75.98 $\pm$ 5.02	118/110	29.1 $\pm$ 1.00

Table 5.1: Demographic description of the ADNI dataset group. Values are reported as mean and $\pm$ standard deviation.

5.6 Experiments and results

This section consists of the implementation and experiments of the fusion methods entirely for sMRI modality.

The experiments were conducted on a GPU-based high-performance computing platform featuring an Intel(R) Xeon(R) CPU E5-2680 v2 @2.80GHz processor, 187 Gb of RAM, equipped with two Nvidia TESLA P-100 graphics cards with 16GB dedicated memory. The networks were trained from scratch by stochastic gradient descent with Nesterov momentum (see chapter 3). The parameters used in the training phase were: 60.000 iterations which gives about 1000 epochs, the Learning rate: 0.0001 and policy: fixed; Momentum: 0.9 Batch-size: 256.

5.6.1 Single modality experiments

a) Data description used for single modality fusion:

We first prepared our sMRI dataset (Dataset-1) by adjusting the number of subjects for each class to generate a well-balanced dataset for all binary classification tasks. In chapter 4, we deduced from experiences that with an equivalent data set for our AD classification, we achieve better results than with an unbalanced data set. Similarly, we follow the same strategy in the ongoing work to build our dataset for our AD binary classification. The table 5.2 describes the composition of the used dataset after subjects balancing. We have reduced NC and MCI groups in two ways in order to match their pair classes. For instance, we have passed from NC/MCI (228/399) to AD/NC (228/228).

	Classes	**# Subjects**	**Age [range] / $\mu(\theta)$**	**Gender(#F/ #M)**	**MMSE [range] / $\mu(\theta)$**
ADNI-1	AD	188	[55 91] / 75.4 $\pm$ 7.52	99/89	[18 27] / 23.3 $\pm$ 2.03
	NC	188	[60 90] / 76.2 $\pm$ 7.18	98/90	[25 30] / 29.1 $\pm$ 2.03
	NC	228	[60 90] / 76.0 $\pm$ 5.02	118/110	[25 30] / 29.1 $\pm$ 1.00
	MCI	188	[57 89] / 74.9 $\pm$ 7.04	124/64	[23 30] / 27.0 $\pm$ 1.75
	MCI	228	[54 89] / 74.9 $\pm$ 7.16	148/80	[23 30] / 27.0 $\pm$ 1.74

Table 5.2: Demographic description of the ADNI screening 1.5T Images studied population (reduction subject details)

Similarly, we followed the same way for data augmentation and splitting settings, as practiced in the previous chapter. We divided the dataset into three folders, 20% for both validation and test sets, and the rest (60%) for the train set. Thus, we get an augmented and balanced dataset for our experiments. The table 5.3 illustrates the selected subjects alongside the augmentation details.

		Before Augmentation			**After Augmentation**		
		AD	**MCI**	**NC**	**AD** [a]	**MCI** [a]	**NC** [a]
Dataset 1	Train	112	112/136	112/136	6272	6272/7616	6272/7616
	Valid	38	38/46	38/46	2128	2128/2576	2128/2576
	Test	38	38/46	38/46	2128	2128/2576	2128/2576
		188	188/228	188/228	10528	10528/12768	10528/12768

Table 5.3: Number of subjects for each class, with its corresponding augmentation.

b) Classification results:

It is interesting to visualize the features to analyze and understand training process; In Figure 5.4 the two left images (a) result from processing of AD subject. In (b) in same figure a NC example is given. One can distinguish quite different structures in these outputs of the first Conv. Layer and the 2nd max pooling layer.

This part is composed of two series of experiments: The first (i), are single networks over the three projections scans. Here, we realize models for sagittal, coronal, and axial planes. The results are presented in the table 5.4. The second (ii) series are the proposed fusion approaches, we implemented

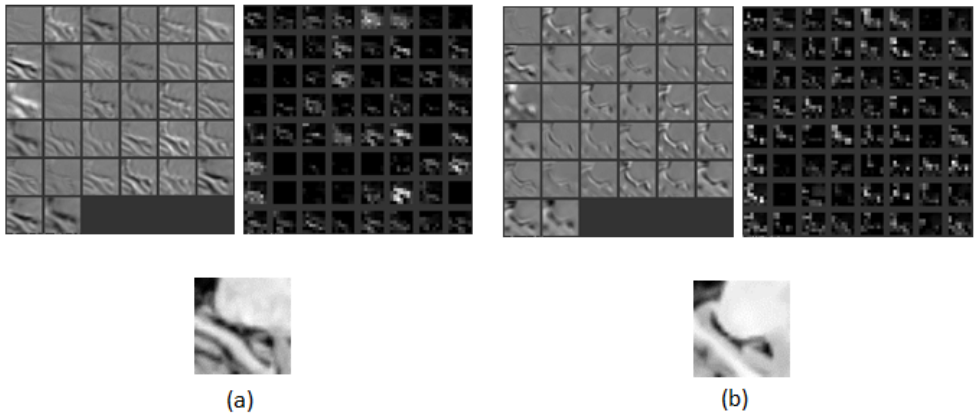


Figure 5.4: Features example patch of AD (a), NC (b) subjects and there features of conv1 and pool2 layers.

the two fusion methods and the results are given on the tables - (b) 5.5 for intermediate fusion, and (b) - 5.6 for late fusion with aggregation and majority vote.

	Sagittal			Coronal			Axial		
	AD vs. NC	**AD vs. MCI**	**MCI vs. NC**	**AD vs. NC**	**AD vs. MCI**	**MCI vs. NC**	**AD vs. NC**	**AD vs. MCI**	**MCI vs. NC**
Accuracy	**82.80%**	62.50%	**66.12%**	80.15%	**66.40%**	57.56%	79.69%	61.72%	61.25%
Specificity	**79.61%**	60.00%	**58.70%**	78.53%	57.89%	**58.71%**	78.12%	**68.75%**	55.00%
Sensitivity	**85.89%**	64.00%	**73.75%**	82.67%	**75.10%**	56.35%	81.25%	54.63%	67.00%

Table 5.4: MRI results: single-projection comparison.

	Fusion (intermediate)		
	AD vs. NC	**AD vs. MCI**	**MCI vs. NC**
Accuracy	**85.94%**	63.28%	65.61%
Specificity	**84.38%**	60.94%	**66.23%**
Sensitivity	**87.50%**	65.62%	65.12%

Table 5.5: MRI results: intermediate fusion.

5.7 Discussion and comparison

In the first series of experiments we identify the most discriminative projection for our binary classification tasks. In the Figure A.1 are shown three curves of accuracy and loss for single projection each. One can see that the sagittal projection ensures a little higher accuracies than coronal projection after stabilization. In Table 5.4 are given results for the three projections at the iteration ♯60.000 we selected after stabilization. Analyzing results of different projections, we state that the sagittal projection is the most discriminative. Indeed, in the most "clear" classification task from physiological point of view AD/NC, it performs the best in all three metrics. This is the case also for NC/MCI classification task. Nevertheless, in AD/MCI there is no consensus on metrics. Indeed AD/MCI is probably the most difficult classification task as it is difficult to trace the separation

	Max			Mean			Majority Vote		
	AD vs. NC	**AD vs. MCI**	**MCI vs. NC**	**AD vs. NC**	**AD vs. MCI**	**MCI vs. NC**	**AD vs. NC**	**AD vs. MCI**	**MCI vs. NC**
Accuracy	89.06%	59.38%	**66.25%**	89.84%	63.28%	**66.25%**	**91.41%**	**69.53%**	65.62%
Specificity	85.94%	57.81%	**71.25%**	85.94%	64.06%	68.75%	**89.06%**	67.49%	66.25%
Sensitivity	92.19%	60.94%	61.25%	**93.75%**	62.50%	63.75%	**93.75%**	71.88%	65.00%

Table 5.6: MRI results: Late fusion comparison (Max, Mean, and Majority Vote).

between Mild Cognitive Impairment (MCI) and already installed Alzheimer disease even for medical experts.

In the second series of experiments we explore if i) fusion of results from different projections in the same "2-D+ε" perspective improves the scores and ii) by which fusion method. The results of our intermediate fusion scheme on FC layer are presented in Figure A.2 in comparison with the best performing sagittal projection. The corresponding figures are presented in the Table 5.5. The proposed fusion scheme with fusion of FC layers performs the best in all three metrics Accuracy, Specificity and Sensitivity compared to the most discriminative sagittal projection. In order to benchmark this intermediate fusion scheme with a classical algebraic operators and other late fusion schemes, we performed three experiments with i) max, ii) mean and a majority vote. These outcomes are illustrated in Figure 5.2 and results are presented in Table 5.6 below. The max and mean fusion were done on the results of scores before their binarization, and the majority vote after the binarization of scores. The AD/NC classification gives the best results with Majority vote. This simple fusion scheme clearly outperforms intermediate fusion on Fully Connected layers, the improvement is of 5.5% in average. With regard to the single sagittal projection it is of 8.6%. In NC/MCI classification, max and mean fusion give better results than FC fusion and single sagittal, but the difference is very small (0.25%).

To position our approach with regard to the literature, we compare it with the most recent work in [179]. Note that a strict comparison is not possible as we use ADNI screening dataset and the authors of [179] do use baseline dataset. Both of them contain a strong intersection, but the screening dataset is larger. Nevertheless, the number of scans is similar. The difference is that we balance the number of subjects in all classification tasks. Particularly for AD/NC classification we use 188 scans for each class. In this paper the authors segment the brain into grey matter (GM), white mater (WM) and cerebrospinal fluid (CFS). Then they parcelate GM into 93 regions and use normalized densities of these regions as features as an input into a Deep ensemble sparse regression network. For the AD/NC classification problem we obtain the same figures in accuracy (91,41% vs. 91,02%), better results in sensitivity (93,75% vs. 92,72%) and we are nearly the same in specificity (89,06% vs. 89,94%).

5.8 Conclusion

In this chapter we continued elaborating the "2-D+ε" approach in the task of classification of MRI in a study of three groups of subjects NC, AD and MCI. In this classification, as well as in our previous works, we used the ROI in a brain, which is an Alzheimer biomarker that is Hippocampal region. We first studied the discriminative power of single projection data and stated that in accordance to the medical practice, the sagittal projection is more discriminative in terms of all metrics accuracy, specificity and sensitivity. To increase the classification power, we used two different fusion strategies. The first one – the intermediate fusion consists in a joint joint training of three Deep CNNs concatenating features in a FC layer. The second one consists in applying algebraic late fusion operators and a majority vote. The conclusion is indeed, that on the baseline classification problem, AD/NC, both fusion strategies achieve better performances. The winner is the majority vote, which results are comparable with the latest state-of-the-art methods which use much more complex approaches for input data preparation for Deep NNs. In the follow-up of this research, we will use new transfer leaning schemes and add more imaging modalities.

Chapter 6

Transfer Learning for Brain imaging classification with multiple sources

---

6.1 Introduction

In the previous chapter, we proposed the fusion methods for AD classification by introducing two principal fusion strategies. The results were promising, and we notice that this pathway is exciting, especially using the majority vote system. As seen in chapter 2, ADNI suggests the DTI-MD modality, which is further interesting to assess related AD-region shrinkage. The MD maps give extra information regarding our ROI. However, we strike to the amount of dataset element of this modality. Therefore, force us to investigate and go further to incorporate an innovating concept to grab this restraint,e.g., transfer learning in such one. In this chapter, we address the transfer learning methodology as an answer either for improving the performance results and tackling the constraint of data size limitation. Here, we transfer features previously learned on enough large dataset to a new network that will use a small dataset. Nevertheless, we introduce three principles schemes of transfer learning: (i) A cross-modal framework, which consists of using the transfer features from one modality to another. (ii) A cross-domain framework, here we include an external dataset, e.g., the MNIST - to study the model's behaviors and analyze the effect of using foreign features. (iii) A hybrid scheme where the transfer passes through two levels stages of the knowledge transfer (a mixture of the two proposed frameworks), in such case, we exhibit an evaluation of all results.

Highlights:

- We preserve the same shallow CNN method for Alzheimer's disease classification in a 2D approach;
- We combine multiple models through various data projections and MRI modalities;

- We introduce two transfer learning approaches: cross-modal and cross-domain transfer learning to improve performances;
- We empirically evaluate and report the improvements results over divers models combination and hybrid transfer learning settings;

6.2 Related work

An improved classification performance of such CNN models as AlexNet, VGGNet, GoogleNet, and ResNet with transfer learning has been reported in various applications in medical domain. There are some recent related works relevant to our methodologies using the 2D-based CNN classification instead of working with the 3D or 4D imaging. In the follow-up, we shortly review the works which use these known architecture and adapt them to the medical image classification, and on the application of transfer learning method as a solution for the limited volume of medical datasets.

6.2.1 Works based on a transfer learning approach

Overall, the more the architecture is deep the more it needs huge data, and consequently a considerable time for training. The crucial issue here is that, in general, data in medical area are not sufficiently available. For this reason, numerous research works have used transfer learning approach, whether based on popular networks or on novel methods to address the lack of data.

In [78] authors used AlexNet architecture pretrained on a general purpose large-scale ImageNet dataset to fine-tune last layers in the model on target sMRI and PET modalities. In the conclusion of their work, the authors state that as neuroimaging data differs significantly from the source domain data, such a transfer method is not optimal. Indeed, the accuracy values achieved for classification of AD/NC on sMRI modality are rather low (around 66%) and even worse when AD/MCI/NC classification problem is addressed. Another study [200] employs the transfer learning on the same kind of images, such as brain scans from the same modality (sMRI) but with the source database different from target data to deal with the limited target data to recognize MCI on MRI images. Here two different datasets OASIS [1] and LIDC [2] have been used for the pretraining stage. They achieved best performance with accuracy of 90.6% for MCI/NC. [42] integrates a method called Multi-Domain Transfer Feature Selection (MDTFS) to select discriminant features for classification of AD/NC. In their case, an auxiliary domain corresponds to classification problem on the same data but for different target classes. When classifying AD/NC they consider s-MCI/c-MCI and MCI/NC. Their experiments have been conducted on ADNI-1 sMRI and classification accuracies achieved were

[1] http://www.oasis-brains.org/
[2] https://wiki.cancerimagingarchive.net/display/Public/LIDC-IDRI/

of 95.2% and 82.1% for AD/NC, and MCI/NC respectively. The DTI image maps are often seen as a good modality for the detection of Alzheimer's disease. Thus the authors in [140] have compared the NC, AD and MC using MD and FA maps. Their results showed that MD was a better indicator of brain atrophy than FA.

With the same concept of reusing an efficient architecture, authors in [32] used a modified version of VGGNet [173] network called DemNet that takes 2D images as input. It was composed of 13 convolutional layers and of three fully connected layers with dropout after each pooling layer, to reduce the over-fitting.

The neural network took as an input a 2D slice from MRI data with 224×224 resolution for both 2-way and 3-way classification problems. The authors selected 20 slices for each brain, classified each of them and measured the accuracy. The results showed that the first and the last two slices (111, 129, and 130) had significantly lower accuracy than the average accuracy per slice. They have achieved an overall accuracy of 98.33% for AD/NC classification. In [191], the authors used two networks, a baseline single-layer CNN and a pretrained ResNet network, they used a single 2D axial slice per subject (median slice from the 3D volume) as an input, the baseline CNN network was composed of only one convolutional layer and of two FC layers. They studied the impact of transfer learning from RestNet trained on ImageNet, and the data augmentation in a real time, at training phase, they conclude that the ResNet architecture successfully fits to the MRI domain, and pretraining with data augmentation improves the prediction. In a recent study, Bumshi et al. [115] used a modified AlexNet network which is known as a high-performance pretrained model. The architecture is composed of five layers for convolution computation and a last FC layer with two outputs for binary classification AD/NC, or three outputs for 3-way classification (AD/MCI/NC). Due to the presence of some noise in the dataset, they proposed a data permutation scheme with outlier rejection, and slice selection methods by removing pixels to eliminate interfering data. All 2D slices (obtained by permutation from axial, sagittal, and coronal planes) were used for training the network, and finally the network was fine-tuned using OASIS and ADNI datasets. They achieved 98.74%, and 95.35% accuracy in AD/NC classification task, respectively on the OASIS and ADNI datasets. For 3-way classification their method achieved 98.06% on ADNI 3T dataset.

In a recent work, Ahsan et al. [188] proposed multiple deep 2D neural networks for binary AD/NC classification. They introduced two architectures that use the transfer learning approach based on the InceptionV3 and Xcepetion models whose weights are pretrained on Imagenet LSVRC. In addition, a custom CNN network is built with the help of separable convolutional layers. They used 96 central 2D-slices from each subject's brain by ignoring the first and the last 40 outer slices. They used three datasets with different settings from the OASIS project, which are composed of 416 T1-weighted MRI scans. The two datasets are respectively (i) a balanced with 180 and (ii) unbalanced sets with 114 subjects. The third dataset is the one used in the work of [83]. The authors used different

configurations of fold cross-validation over the three datasets for the experiments. They achieved an average accuracy of 64% on the first dataset, 82.79% on the second dataset, and 99.45% on the third dataset for AD/NC classification. The latter is doubtful as they used the dataset from [83], and it was mentioned that there could be a problem of leakage in this dataset [202]. They concluded that the transfer learning approaches outperform non-transfer learning-based approaches. This demonstrates the effectiveness of these approaches for the binary AD classification task.

Contrarily to all these approaches we propose a transfer learning scheme from one modality to another, on the same dataset. Indeed the sMRI modality shows good discrimination performance for AD diagnosis in brain atrophy analysis, it is our source modality. The supplementary modality is of the same nature, it is the DTI-MD modality which represents our target domain. Otherwise, in our previous chapter and in [2] we have shown that multimodal approach increases performances, hence it is interesting to explore if the fusion framework can be more efficient with a transfer learning. We remain using a shallow CNN-based architecture with only a small number of convolutional/pooling layers since the input region dimension is rather low ($28\times28\times28$).

The contributions of the chapter are as follows:

- we transfer knowledge between sMRI and DTI modalities using a shallow architecture specifically designed for our "2-D+ε" approach.
- we use a similar architecture, LeNet trained on a large set (60K) of images from MNIST[3] which has (28×28) resolution as input, i.e. the same size as of our Hippocampal region-of-interest, and then fine-tune this model on ADNI dataset.

Hence in both cases we perform the "cross" transfer. In the first case, it is a cross-modality transfer. The modalities are similar as the target pathology is expressed by the same image deformations yet in the opposite luminance. In the second case, it is a cross-domain transfer. The domains are different; character images (MNIST) have nothing to do with MRI scans except the fact that they are not coloured. In the next section we present our approaches.

6.3 Methodology and approach

Before introducing our study of transfer learning approach for brain image classification, we will briefly remind the architecture of our classification framework.

6.3.1 The 2D+ε Network Architecture

In this chapter we use our proposed architecture as in Chapter 4 wich we called it "2-D+ε" approach for ROI classification. We use a 2D convolution in a CNN architecture feeding it with

[3] http://yann.lecun.com/exdb/mnist/

three neighbouring slices for each projection Sagittal, Axial, and Coronal. The median slice of Hippocampal ROI and its two neighbours have been selected (see Figure. 6.1). Then the classification results of all three projections were fused, see figure 6.1. We remind that our CNN is relatively shallow due to the low resolution (28×28) of the ROI in each projection. It consists of two convolutional layers followed by a max pooling layer for each one, and a fully connected layer. We get three networks associated with each modality, for the Sagittal, Axial and Coronal projections. Different tested fusion schemes resulted in application of the (best) late fusion with majority vote fusion operator [2] on the six binary classification tasks AD/NC, NC/MCI, and MCI/AD on three projections. We use the same fusion scheme in the follow up of this chapter.

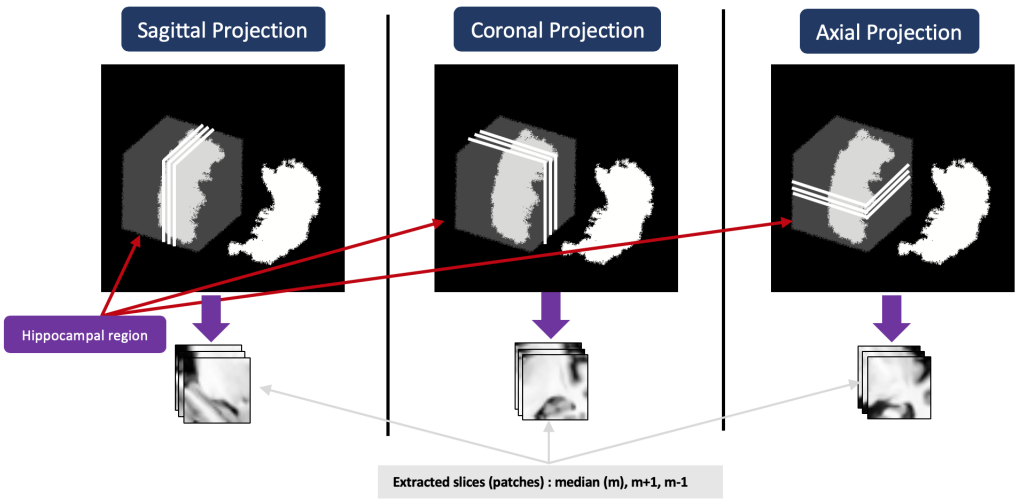


Figure 6.1: Illustration of the 2-D+ε Approach from each projection.

As we have stated, the amount of available training data in our Alzheimer studies remaining low, see chapter 4.7.2, the initialization of the training process is of crucial importance. This initialization is known as "transfer learning" we focus on it in the next section.

6.3.2 Transfer learning for brain image classification

Transfer learning [208] is a popular way of dealing with limited volumes of training datasets. Actually, the CNN models can be either learned *from scratch*, i.e; with random or arbitrary initialization of parameters or with the *fine-tuning* approach from pretrained models. From the domain knowledge of medical research, we retain that the shrinkage of hippocampal ROI which accompanies the development of AD is observable on both modalities sMRI and MD. Figure 6.2 illustrates this phenomenon, it presents two examples of subjects: for the *left* a normal control (NC) subject, and the *right* an (AD) subject, with both modalities; the (A) is the MD map, and (B) is sMRI scan. It shows the hippocampal region from different projection views. From the top to the bottom, also the Axial, Sagittal, and Coronal planes respectively.

As we can observe from the illustrative figure, the atrophy of the hippocampus can be recognized from both modalities by conserving the same shape, but in inverted representation. This means the signal spawned from CSF flows surrounding the Hippocampus portion can be interpreted by a dark area in sMRI scans, while it is bright in MD maps. For this reason, we could adopt a transfer learning strategy between these two types of data, from designing trained models in the source domain of sMRI towards the target domain MD called *cross-modal* transfer learning.

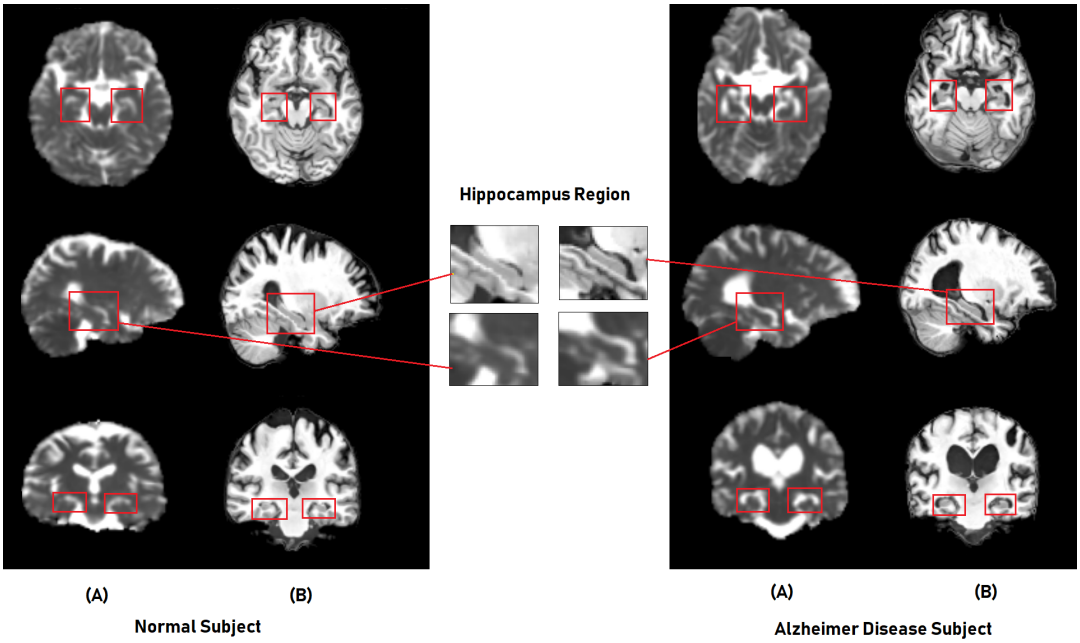


Figure 6.2: Example of the hippocampal region with different projections for two Subjects: (A) - MD and (B) - sMRI.

In the framework of learning CNN parameters, we can formally define the transfer learning strategy from the source modality (sMRI) to target MD as follows:

$$\begin{cases} W_0 & \leftarrow W'_\phi \\ W_{i+1} & \leftarrow F(W_i) \end{cases} \tag{6.1}$$

Where W'_ϕ is the best trained model on the large sMRI dataset, we initialize the training with the parameters of W'_ϕ, and fine-tune all or some layers of the used architecture. F is the optimization scheme.

As stated before, we manage to implement the cross-modal transfer learning approach from sMRI to MD-DTI modalities. We fine-tune only the last layer - the (FC) - in our architecture whereas we fix the two convolutional layers with the features extracted from the source dataset. Indeed, we deduced our investigation from the work of [28] regarding the transfer learning. They provided a pragmatic analysis of the transfer learning approach through various settings of the same architecture to well-understand the behaviors of their method. However, since we have two imaging modalities

of the same studied ROI, we used the same approach by fine-tuning only the FC layer. The other layers remained untouched as we have a pretty similar representation of our studied ROI.The transfer of the parameters in our CNN architecture is illustrated in Fig. 6.3. The arrows depict initialization of optimization process for each convolutional and fully connected layer.

For fine tunning we use Stochastic Gradient Descent with Nesterov momentum as in [107], and as a cost function $J(W_i)$ to minimize we used cross-entropy loss as in [1]. The weights update formula is defined as follows:

$$V_{i+1} \leftarrow \mu V_i - \alpha \nabla J(W_i + \mu V_i) \tag{6.2}$$
$$W_{i+1} \leftarrow W_i + V_{i+1}$$

where W_i are the parameters of each layer at iteration i, α is the learning rate, μ is the momentum coefficient and V_i is the velocity.

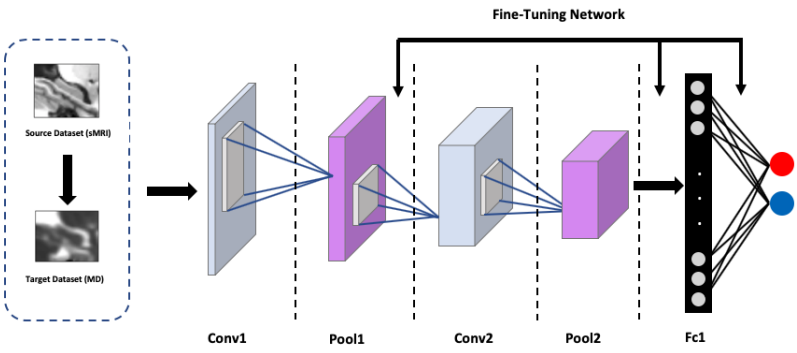


Figure 6.3: The scheme of Transfer Learning for parameters optimization from sMRI to MD-DTI modality. An example of the proposed architecture for 2-way classification.

6.3.3 Adapted cross-domain/cross-modal transfer learning schemes

It is believed that transfer learning improves classification performance, especially when the source and the target domains are close. In order to validate our proposed scheme of intra-domain, we compare the classification efficiency across two different domains by using both known pretrained model and dataset. We have selected the LeNet 6.4 network owing to similarity at the design level [112]. It takes the same input definition 28×28 as ours, and almost the same depth of layers except the fully connected layers. In this view, we take this model which is already pretrained on MNIST dataset, and apply it to our brain image data DTI-MD and sMRI. Nevertheless, the model has been modified in FC layers and adapted to 2-way classification problem instead of ten. We freeze the two first convolutional layers which already capture the universal features, and then we fine-tune it on the Alzheimer's disease dataset by optimizing weights only in the two FC layers of the model.

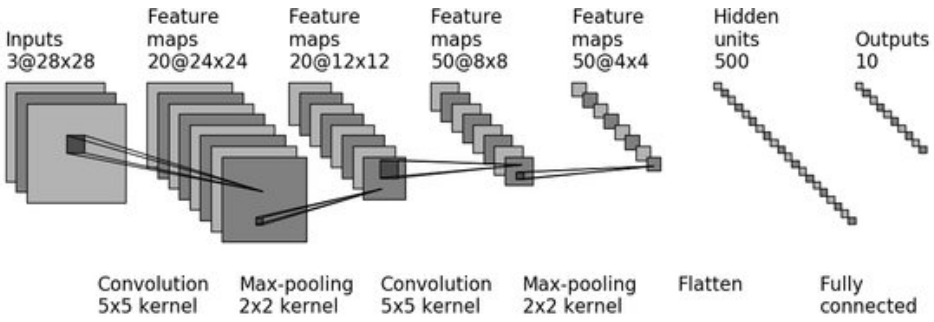


Figure 6.4: LeNET-5 design: A modified version the original LeNET which takes data of 28×28 resolution from MNIST.

We evaluate the approach through two mechanisms as follows:

(i) One-level transfer scheme: In this first approach, we realize the transfer using our LeNet-like model for both modalities from MNIST to sMRI and from MNIST to DTI-MD followed by our fusion scheme (Majority vote).
(ii) Two-level transfer scheme: In this second approach, we multiply the transfers; it is applied as MNIST-DTI-MD through sMRI. LeNet-like model trained on MNIST dataset is first used as the basis for training of sMRI classification model. Next from sMRI model we transfer to the DTI-MD images of the same domain. After that, we build our fusion framework from this model combined with the model we obtained from the cross-modal transfer (sMRI to DTI).

6.4 Experiments and results

As described in the chapter 2, sMRI images have undergone some pre-processing steps, illustrated in Figure 6.5. DTI images have also been pre-processed (See section 2.3.2 for further details). We briefly recap the data preprocessing: (a) a denoising step with an adaptive non-local mean filter (See 2), (b) image alignment (affine registration) in the MNI space, (c) image intensity normalization, (d) ROI selection and extraction using the AAL atlas. Alongside (d) skull striping, and (c) co-registration for DTI scans.

6.4.1 Dataset description and learning setup parameters

Since we work with multimodal imaging, we consider two subsets of the whole data from ADNI. The first one is the data that were selected from the ADNI-1 screening baseline with only anatomical MRI T1-weighted sequences, in this set all subjects underwent whole-brain MRI scanning on 1.5 Tesla at 14 acquisition sites. It is the same dataset as used in [1]. With the same demographic information for each of the diagnosis groups (NC, AD and MCI), The data sample consists of 815 structural MRIs including 188 Alzheimer's Disease (AD) patients, 228 cognitively normal (NC) and 399 subjects with

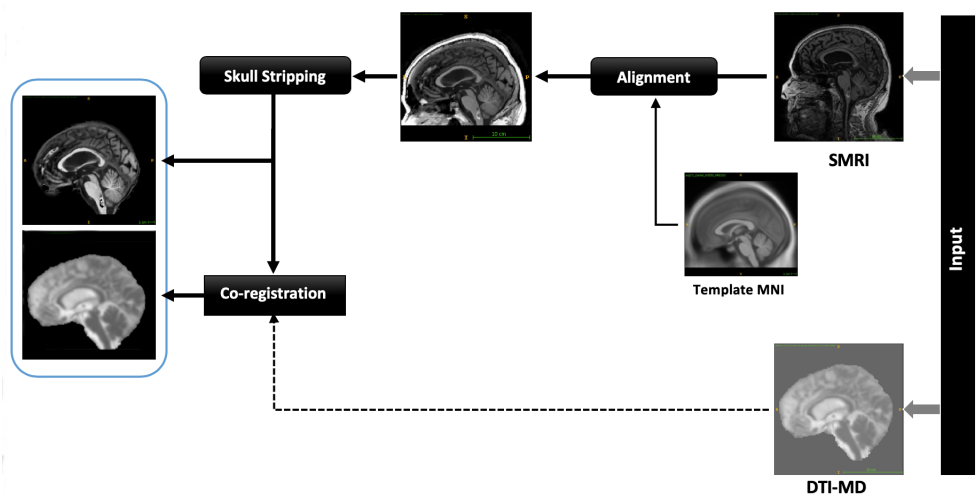


Figure 6.5: Schematic diagram of dataset preprocessing: i) registration of all MRI scans on MNI space, followed with intensity normalization. ii) ROI selection process using the Atlas AAL for both hippocampal regions. iii) 2D-slice extraction from selected 3D-volume. iv) feeding the CNN networks [1].

Mild Cognitive Impairment (MCI). The second subset includes images of subjects screened with both structural MRI and DTI modalities. It is a union of data from the ADNI-2&Go, and ADNI-3. The table 6.1 presents demographic characteristics of subjects, including age, gender, and the Mini Mental State Examination (MMSE) score. The age of different groups ranges between 54 and 95 years old, and the proportions of male and female are close in AD/NC groups while the proportions of male are higher than female in MCI groups. We have visually checked all T1-weighted MR images and DTI maps for quality assurance, to exclude scans with excessive motion and/or artifacts. We note that the MMSE score is not mentioned for the ADNI-3 phase (missed from metadata of all subjects of this phase). However, this lack of data does not affect our method as we do not use these features thus far.

	Classes	**# Subjects**	**Age [range] / $\mu(\theta)$**	**Gender (#F/ #M)**	**MMSE [range] / $\mu(\theta)$**
ADNI-1	AD	188	[55.18, 90.99] / 75.37 $\pm$ 7.52	99/89	23.3 $\pm$ 2.03
	MCI	399	[54.63, 89.38] / 74.89 $\pm$ 7.30	256/143	27.0 $\pm$ 1.78
	NC	228	[60.02, 89.74] / 75.98 $\pm$ 5.02	118/110	29.1 $\pm$ 1.00
ADNI-2/Go	AD	*48	[55.73, 90.87] / 75.60 $\pm$ 8.63	28/20	23.0 $\pm$ 2.42
	MCI	*108	[55.33, 93.62] / 74.40 $\pm$ 7.47	66/42	27.4 $\pm$ 1.99
	NC	*58	[59.91, 93.25] / 74.91 $\pm$ 5.90	28/30	28.9 $\pm$ 1.18
ADNI-3	AD	*16	[55.26, 86.10] / 74.63 $\pm$ 9.92	4/12	-
	MCI	*165	[55.88, 95.93] / 75.01 $\pm$ 7.91	71/94	-
	NC	*341	[55.79, 95.39] / 73.52 $\pm$ 7.82	209/132	-

Table 6.1: Demographic description of the ADNI dataset group. Values are reported as mean and $\pm$ standard deviation (* Subjects with both modalities).

Data Augmentation parameters: In order to sufficiently increase our dataset size, we applied the data augmentation strategy as presented in Chapter 2. Thus, we have set an augmentation factor F upon which the calculations are based. Indeed, we have proposed an approach to increase the data in an equitable manner. The method consists in setting the factor (a multiplication coefficient) for

the most represented class, and by multiplying this factor to the cardinal of this class we obtain large enough class (this one). The next step of the approach is increasing all other classes to reach the same size of the first one. In this way we obtain a balanced dataset.

Hence, we defined the factor F to 100 for the both datasets: The subset "1" and subset "2". However, the factor was set for MCI class in subset "1" since it is the most represented class, in the same way, NC class was selected for subset "2" as well. The max shift was set to 2 slices (note that two slices of sagittal axis for example represent about 7.4% of the Hippocampus 3D Bounding Box) and the maximum scale parameter of smoothing Gaussian Blur was set to 1.2 (See Algorithm 1). Indeed, the original signal on both modalities is blurred and a stronger blurring would destroy the structure of the ROI. The parameters were generated randomly and selected to avoid similar augmentation for the same brain scan. Table 6.2 describes the split of samples before and after the augmentation process. Data are divided into Training, Validation, and Test subsets.

Algorithm 1 DA pseudo algorithm

Input: $\mathcal{D}$ dataset, F for augmentation factor, S for Max-Shift, and σ for Max-Blur.
Output: augmented and balanced dataset.

procedure PROCESS() //* Function to generate samples. *//
  $N = (\max\{card(\#AD), card(\#MCI), card(\#NC)\}) \times \mathrm{F}$
  while $N \neq 0$ **do**
    $i, j, k, x \leftarrow$ *random_generate_parameters()* //* $(i, j, k) \in [\![\text{-S}, \text{S}]\!]$ and $x \in [0, \sigma]$ *//
    Compute: *element* $\leftarrow$ *augmentaion_function(i, j, k, x)* //* for a given scan. *//
    Compute: *roi* $\leftarrow$ *Hippocampus_cube(element)* //* return the mean of the left and right of the ROI. *//
    Compute: *patches_extraction()* //* Extraction of 2D patches. *//
    $N \leftarrow N - 1$

		Before Augmentation			After Augmentation		
		AD	**MCI**	**NC**	**AD** [a]	**MCI** [a]	**NC** [a]
Dataset 1	Train	146	**482**	446	48200	48200	48200
	Valid	42	**126**	117	12600	12600	12600
	Test	64	64	64	640	640	640
		252	672	627	61440	61440	61440
Dataset 2 [*]	Train	31	198	**299**	29900	29900	29900
	Valid	13	55	**80**	8000	8000	8000
	Test	20	20	20	200	200	200
		64	273	399	38100	38100	38100

Table 6.2: Number of subjects for each class, with its corresponding augmentation, ([*] Both modalities).

Hardware configuration: The experiments were conducted on a GPU-based high-performance computing platform featuring an Intel(R) Xeon(R) CPU E5-2680 v2 @2.80GHz processor, 187 Gb

of RAM, equipped with two Nvidia TESLA P-100 graphics cards with 16GB dedicated memory. The computational time for one epoch at the training step with batch-size of 64 samples was 2.03 seconds in average.

Optimization settings: To pick the best learning parameters, we have trained our basic "2-D+ε" network on the sMRI dataset, considering that the ideal results differ from a specific model to others, and also depend on the nature of the dataset. We have explored how the learning rate, the various optimization methods and the parameters affect model training for each learning policy. The Figure 6.6 presents examples of training behaviors with different configurations.

The choice of appropriate hyper-parameters was realized in various training trials, and the "Stochastic Gradient Descent" optimizer method was selected as the best configuration setting to minimize the cost function $\mathcal{L}(\boldsymbol{W}_i)$. The weights update formulas are as follows:

$$\begin{aligned} \boldsymbol{V}_{i+1} &\leftarrow \mu \boldsymbol{V}_i - \alpha \nabla \mathcal{L}(\boldsymbol{W}_i + \mu \boldsymbol{V}_i) \\ \boldsymbol{W}_{i+1} &\leftarrow \boldsymbol{W}_i + \boldsymbol{V}_{i+1} \end{aligned} \tag{6.3}$$

where $\boldsymbol{W}_i$ are the parameters of each layer at iteration i, α is the learning rate, μ the momentum ($\mu = 0.9$) and $\boldsymbol{V}_i$ is the velocity.

We use the exponential learning rate decay policy, $\alpha = \alpha_0 \cdot \gamma^i$, where α_0 is the initial value of learning rate, and $\gamma \in [0,1]$. In our case we set γ to 0.95, and α_0 = 0.0001. We set the batch size as 64. After several iterations, stabilization of the training is observed around the 30th epoch.

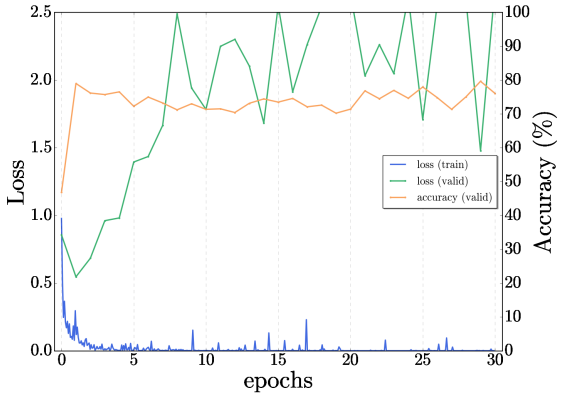
The network was implemented with "Caffe" Deep learning Framework [96]. we have added a dropout layer to tackle the over-fitting phenomena as a method of regularization.

In order to evaluate the efficiency of our method, different adaptive transfer learning schemes were adopted here, in this section we provide experimental results on the two proposed approaches: the cross-modal and cross-domain.

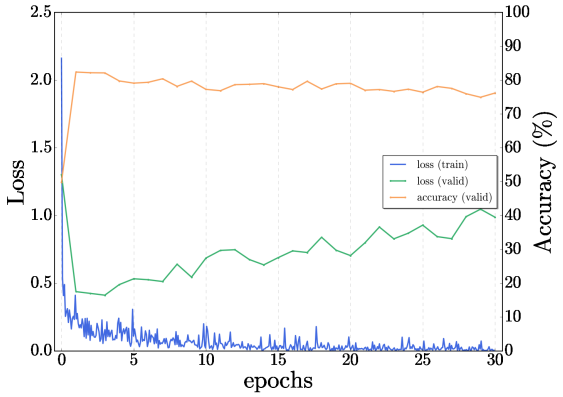
6.4.2 2-D+ε single and fusion architecture

We use the "2-D+ε" network (see Section 6.3.1) we proposed on sMRI data. It was shown that further improvements can be achieved through the fusion approach in particular the majority vote. Likewise, in this section we use only sMRI data but with a larger number of subjects compared to the previous results in this manuscript. In addition, the best fusion method the "Majority Vote" as used in [2] is adopted here. We base our classification on two different models: single network for each projection (Axial, Coronal, and Sagittal), and a late fusion which is designed to improve and enhance classification performances.

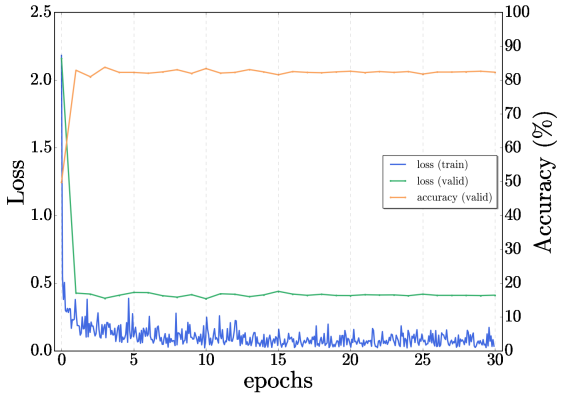
We have designed two architectures to perform the classification: (i) 3-way classification (AD/MCI/NC) and (ii) 2-way classification (AD/MCI, AD/NC, and MCI/NC) as the most works in the



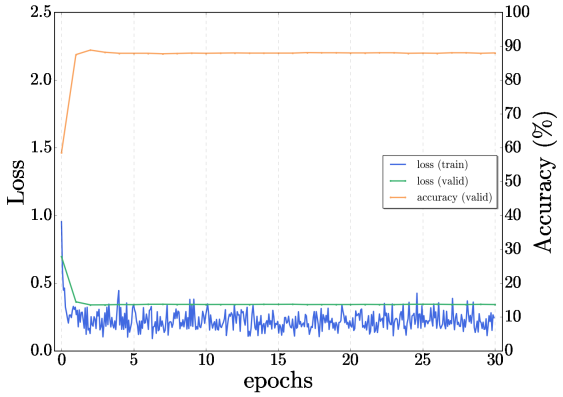
a) ADAM optimizer algorithm.



(b) Nesterov accelerated gradient (NAG)



(a) Stochastic gradient descent (SGD) with Gamma 0.95



(b) Stochastic gradient descent (SGD) with Gamma 0.75

Figure 6.6: Learning parameters for training comparison on the SMRI dataset.

literature. We built a 3-way classification baseline model, and accuracies of 60.23%, 58.71%, 56.84%, and 66.49% for Sagittal, Coronal, Axial, and fusion, respectively, were obtained. We found that the 3-way model performs somewhat faintly. However, we considered only the 2-way classification in our work since the application domain requires only to test positive or negative for the AD diagnosis. Besides, most of the related works provide only binary classification results with which we can make the comparison. Hence, we performed our method as presented above, and Table 6.3 presents an overview of the 2-Way classification results.

On average, the single network on the sagittal projection shows better results than other projections, but the fusion method with majority vote achieves the best results for each classification tasks in terms of accuracy which is coheret with results we obtained in earlier chapters. For example, for the AD/NC classification we obtain 82.92% of accuracy for the single sagittal network, while with the fusion we get 91.86%. For the classification tasks involving MCI class (e.g., AD/MCI, and MCI/NC), we can notice that the accuracy is lower than AD/NC, we have 69.65% for AD/MCI, and 68.52% for MCI/NC. As we have already stated, this class is special, as it includes two subclasses:

the early MCI (e-MCI), and the late MCI (l-MCI). l-MCI have more similarity of the atrophy with AD in our ROI, making it difficult to distinguish between subjects in MCI and AD classes. The same conclusion can be derived from the MCI/NC classification.

Tasks	Projection	Acc (%)	Sen (%)	Spe (%)	BAcc (%)
AD vs. NC	Sagittal	82.92%	85.72%	79.84%	82.78%
	Coronal	81.04%	83.20%	78.63%	80.41%
	Axial	79.81%	81.31%	77.65%	79.48%
	fusion *	**91.86%**	**93,90%**	**89.88%**	**91.89%**
AD vs. MCI	Sagittal	66.73%	68.52%	63.91%	66.21%
	Coronal	67.61%	71.25%	61.88%	66.56%
	Axial	65.55%	66.60%	61.57%	64.08%
	fusion *	**69.95%**	**73.41%**	**68.22%**	**70.81%**
MCI vs. NC	Sagittal	65.51%	61.64%	69.48%	65.56%
	Coronal	66.45%	60.27%	65.11%	62.69%
	Axial	63.89%	59.15%	64.57%	61.86%
	fusion *	**68.52%**	**65.59%**	**70.15%**	**67.87%**

Table 6.3: Classification results for each single projection and fusion by majority vote on sMRI dataset.

6.4.3 Evaluation of transfer Learning.

In this section we provide experimental results on the two proposed approaches: the cross-modal and cross-domain trasfer learning.

Transfer learning from sMRI to DTI-MD with the 2D+ε approach

With the similarity between the structural MRI and DTI-MD, we proposed a cross-modal transfer learning from sMRI dataset as a source to the DTI-MD dataset, which is considered the target dataset. The model was first trained on the sMRI dataset and then fine-tuned with the DTI-MD dataset. Obviously, with the experiments, the cross-modal method yields slightly better results than the training from scratch (by random initialization of network parameters). We can see the difference of the behavior in training (loss) and validation (loss and accuracy) in Figures 6.7, and 6.8 for AD/NC and AD/MCI respectively. We get improved accuracy at the final 30th epoch with transfer learning, and the loss is lower on both sets along training epochs. Figure 6.9 illustrates that with the transfer from sMRI to DTI-MD the overfitting is slightly reduced (a) compared to training from scratch (b). Table 6.4 presents the final results of cross-modal transfer learning for each projection and with the late fusion by majority vote. Compared with the results from table 6.3, we have clearly augmentation of all metrics up to 5% for the most challenging classes AD/MCI and MCI/NC.

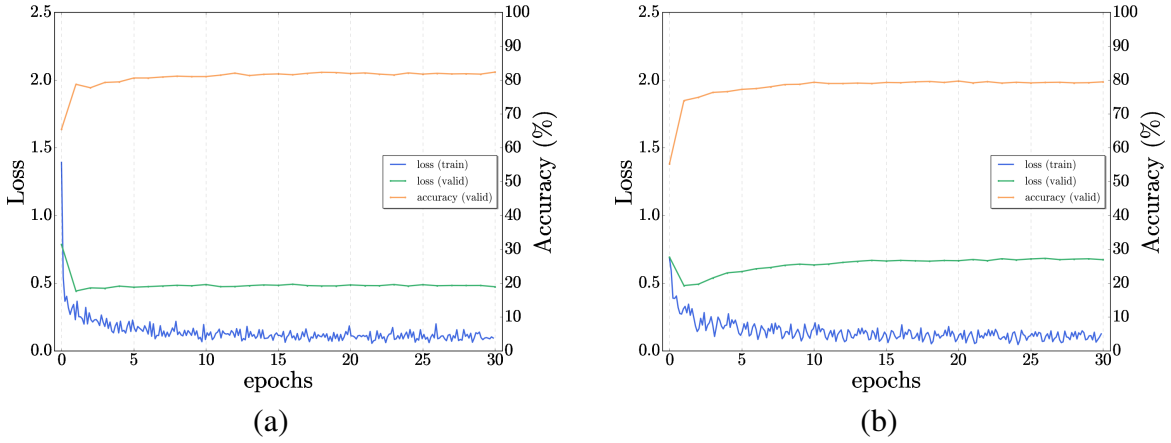


Figure 6.7: Example of Transfer learning for single network - comparison of AD/NC: a) Transfer from sMRI to MD-DTI, b) Training from scratch on MD-DTI Dataset.

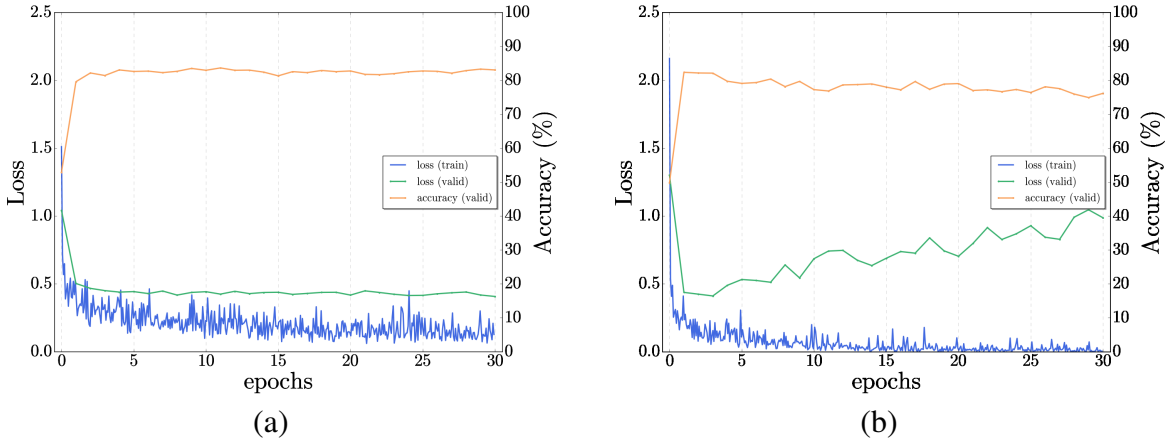


Figure 6.8: Example of Transfer learning - comparison of AD/MCI: a) Transfer from sMRI to MD-DTI, b) Training from scratch on MD-DTI Dataset.

From MNIST knowledge to sMRI and DTI cross-domain learning

In this part of the experiment, we apply our second proposed method for cross-domain transfer learning. We build the LeNet-like model, and perform the experiments as follows:

- **One-level transfer scheme**: The transfer is realized for both modalities from MNIST to sMRI and from MNIST to DTI-MD followed by our fusion scheme. Table 6.5 presents an overview of results for both experiments and also results for the fusion. Analyzing the table, we obtain accuracies around 3% in average lower than the previous results for each classification task. For AD/NC as example, we passed from 82.92% to 80.02% for sMRI, and from 84.93 % to 81.85% for DTI-MD on the sagittal projection. For the other classification tasks, the situation

Tasks	Modalities	Projection	ACC (%)	SEN (%)	SPE (%)	BAC (%)
AD vs. NC	MD	Sagittal	84.93%	86.07%	81.23%	83.65%
		Coronal	80.62%	81.15%	79.75%	80.45%
		Axial	79.50%	81.91%	78.04%	79.97%
	Fusion (*)		**92.11%**	**94.53%**	**90.02%**	**92.27%**
AD vs. MCI	MD	Sagittal	65.12%	72.25%	68.44%	70.34%
		Coronal	72.87%	76.58%	71.93%	74.25%
		Axial	64.79%	69.14%	66.28%	67.71%
	Fusion (*)		**74.41%**	**80.13%**	**76.02%**	**78.07%**
MCI vs. NC	MD	Sagittal	65.59%	66.48%	69.32%	67.90%
		Coronal	69.14%	67.97%	70.82%	69.39%
		Axial	64.98%	67.71%	71.06%	69.38%
	Fusion (*)		**73.91%**	**76.79%**	**79.63%**	**78.21%**

Table 6.4: Binary classification results with Transfer Learning from sMRI to MD-DTI data and fusion (* both modalities).

is pretty much the same. The results are poorer than those for cross-modal transfer in the same domain, see table 6.4.

- **Two-level transfer scheme**: In this setting we perform the experiments using the scheme as explained in 6.3.3. Table 6.6 presents the results. We can notice that the use of the Two-level transfer scheme, may clearly give better results which we will analyze in the following section.

6.5 Discussion and Comparison with literature review

Hence, we experienced three knowledge transfer types: cross-modal with LeNet-like designed architecture and cross-domain one-level and two-level transfer using LeNet Architecture. Comparing the results presented in Tables 6.4, 6.5, and 6.6, we can conclude the following. The cross-domain

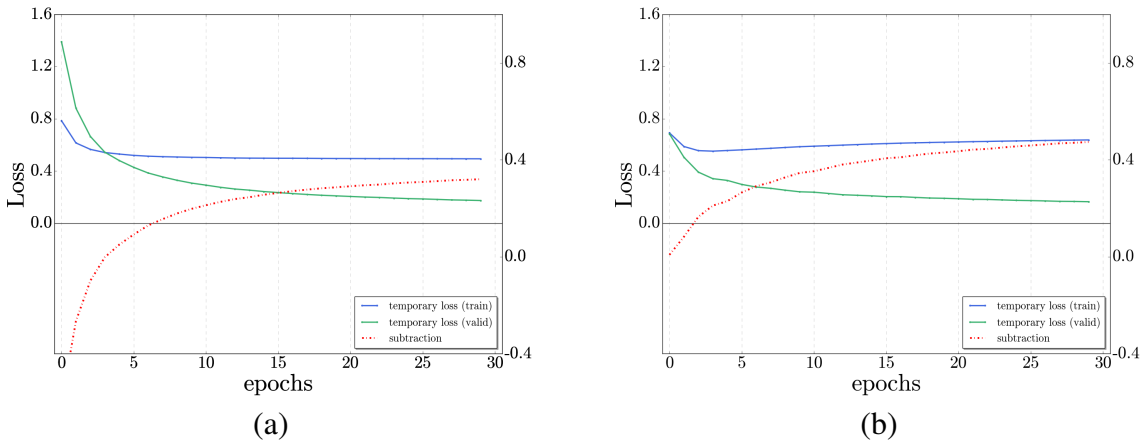


Figure 6.9: Temporal loss curves comparison: a) From sMRI to MD-DTI transfer learning with reduced over-fitting - b) Training from scratch with small over-fitting.

Tasks	Modalities	Projection	Acc (%)	Sen (%)	Spe (%)	BAcc (%)
AD vs. NC	SMRI	Sagittal	80.02%	81.95%	79.26%	80.60%
		Coronal	79.94%	80.59%	78.18%	79.38%
		Axial	79.11%	81.05%	79.42%	80.23%
	MD	Sagittal	81.85%	83.24%	79.49%	81.36%
		Coronal	79.22%	83.01%	78.56%	80.78%
		Axial	78.69%	82.44%	79.71%	81.07%
	Fusion (*)		**86.83**%	**90.94**%	**87.14**%	**89.04**%
AD vs. MCI	SMRI	Sagittal	65.32%	66.81%	64.52%	65.66%
		Coronal	64.57%	65.63%	63.74%	64.68%
		Axial	61.74%	63.05%	59.88%	61.46%
	MD	Sagittal	64.95%	70.19%	66.45%	68.32%
		Coronal	68.60%	72.51%	67.96%	70.23%
		Axial	62.36%	68.24%	62.47%	65.35%
	Fusion (*)		**71.45**%	**78.66**%	**73.16**%	**75.91**%
MCI vs. NC	SMRI	Sagittal	64.75%	62.35%	66.72%	64.53%
		Coronal	60.49%	58.62%	63.40%	61.01%
		Axial	60.15%	59.14%	62.63%	60.88%
	MD	Sagittal	63.59%	63.18%	66.93%	65.05%
		Coronal	67.14%	64.24%	69.86%	67.05%
		Axial	64.98%	63.91%	68.55%	66.23%
	Fusion (*)		**69.85**%	**70.46**%	**75.73**%	**73.10**%

Table 6.5: Classification results with One-level scheme Transfer Learning: From MNIST to SMRI & From MNIST to DTI-MD data.

transfer, which is a very popular transfer learning scheme, performs the worst even on very distinct classes such as AD and NC. Hence definitely, the cross-modal transfer in the same domain (sMRI and DTI in our case) is a better solution. When multiplying transfers such as in the two-level cross-domain transfer scheme, we manage to get slightly better results for the most difficult classification tasks.

Indeed with the transfer from MNIST to sMRI and then DTI, we get a nearly 5% accuracy increase in the classification AD/MCI and MCI/NC. We note that in other metrics, such as Specificity, Sensitivity, and BAcc, the methods perform similarly. Thus, the two-level transfer increases the metrics by more than 5% for the most difficult classification tasks, which is an interesting result. Indeed, the transfer from a pre-trained model does not cost too much; hence the first step of it can be done from a different domain using publicly available trained models such as LeNet on MNIST in our case. Although further transfer in the same domain is needed to improve the result.

Now we will compare our best results with methods from literature, see table 6.7. We have to note that an exact comparison in the medical image domain is not possible, as different ADNI databases are used in each work. To illustrate this, we show the number of analyzed brain scans for AD, MCI, and NC subjects in the first three columns of this table. The authors of [148, 44, 104] use 3D convolutions. The authors of [161, 81, 32, 191] use the whole brain scans. Our method remains "light" in the sense that we focus only on one ROI, which is the biomarker of AD, the Hippocampal ROI. Afterwards, we do not use the 3D volume entirely, but only a light version of it, such as three slices. Even with this lightweight method, we get quite decent results, namely in the separation of AD/NC.

Tasks	Modalities	Projection	Acc (%)	Sen (%)	Spe (%)	BAcc (%)
AD vs. NC	MD	Sagittal	85.14%	87.95%	84.14%	86.04%
		Coronal	82.57%	84.55%	80.84%	82.69%
		Axial	81.21%	84.26%	81.10%	82.68%
	Fusion (*)		**92.30%**	**93.95%**	**90.65%**	**92.30%**
AD vs. MCI	MD	Sagittal	70.84%	77.25%	73.51%	75.38%
		Coronal	76.53%	78.39%	76.64%	77.51%
		Axial	69.21%	73.08%	68.52%	70.8%
	Fusion (*)		**79.16%**	**82.72%**	**78.36%**	**80.54%**
MCI vs. NC	MD	Sagittal	71.09%	70.15%	74.95%	72.55%
		Coronal	75.34%	72.41%	76.39%	74.40%
		Axial	70.21%	69.10%	73.64%	71.37%
	Fusion (*)		**78.48%**	**77.72%**	**81.44%**	**79.58%**

Table 6.6: classification results with Two-level scheme Transfer Learning: From MNIST to DTI-MD crossed sMRI data.

However, for the most challenging classifications MCI/NC and AD/MCI, even though we earn some accuracy points, the results remain slightly weaker compared to AD/NC. This leads to the investigation of other brain structures and regions or even implies the entire brain for this study. Indeed, as our models focus on the hippocampus atrophy for the discrimination task, working on the hippocampus's limbus can involve more advantages to improve performances, where it encompasses the outermost surface of the hippocampus, which seems to be the most affected by the passage of the MCI stage. Going back to our approach "2-D+ε", we take only three slices where only a fraction of that surface, i.e., "the limbus" intersects, which could explain why MCI/AD and NC/MCI discrimination scores relatively low compared to a method that would be full 3D. Therefore, at the resolution at which we operate, the disease's characterization could be better determined at the hippocampus limbus level than with its whole internal structure [136].

Despite our classification results still need to be improved by introducing other regions of interest or additional information, as mentioned above, the proposed cross-modal transfer learning definitely yields increased performances. It can thus be re-used in combination with other methods of classification, as those using whole brain or full 3D information. Specifically, in such a field as medical image analysis and classification where large corpora of annotated data are not available, proposed transfer learning will help in circumventing the lack of training data. The cross-domain transfer learning method presents good results in applications on natural images. On the contrary, cross-domain transfer from natural images to medical image domain remains limited as our results show. This is due to the large difference in terms of content between natural and medical images. In this chapter, we have shown the efficiency of implementing a cross-modal transfer in medical diagnostic applications. We hope that this finding will be successfully used by the research community for medical image classification tasks.

Study	Subjects			Classifer	Modality	Approach	Accuracy		
	AD	MCI	NC				AD vs. NC	AD vs. MCI	MCI vs. NC
Sarraf et al. [161]	52	-	92	CNN - LeNET-5	sMRI	2D slice-level	97.88%	-	-
	211	-	91	CNN - GoogleNet	sMRI	2D slice-level	98.74%	-	-
khvostikov et al. [104]	53	228	250	CNN	sMRI+DTI	3D ROI-based	93.3%	86.7%	73.3%
Gupta et al. [81]	200	411	232	CNN	sMRI	2D slice-level	93.80%	86.30%	83.30%
Billones et al. [32]	53	228	250	CNN - VGG-Net	sMRI	2D slice-level	98.33%	93.89%	91.67%
Bumshik et al. [115]	192	398	229	CNN - Alexnet	sMRI	2D slice-level	98.74%	-	-
	100	-	316	CNN - Alexnet	sMRI	2D slice-level	95.35%	-	-
Valliani et al. [191]	188	243	229	CNN - ResNet	sMRI	2D slice-level	81.3%	-	-
Cheng et al. [44]	199	-	229	CNN	sMRI	3D subject-level	83.88%	-	-
Glozman et al. [78]	200	132	221	CNN - AlexNet	sMRI	2D slice-level	66.51%	-	-
Hon et al. [83]	100	-	100	CNN - VGG-Net	sMRI	2D slice-level	92.30%	-	-
		-		CNN - Inception V4	sMRI	2D slice-level	96.25%	-	-
payan et al. [148]	755	755	755	CNN	sMRI	3D subject-level	95.39%	86.84%	92.13%
Lian et al. [119]	358	-	429	H-FCN	sMRI	3D patch-level	90,00%	-	-
Proposed cross-modal transfer (1)	252	672	627	CNN	sMRI+DTI	2D ROI-based	92.11%	74.41%	73.91%
Cross-domain One-level transfer (2)	64	273	399	CNN - LeNet	sMRI+DTI	2D ROI-based	86.83%	71.45%	69.85%
Proposed Two-level transfer (3)	64	273	399	CNN - LeNet	sMRI+DTI	2D ROI-based	92.30%	79.16%	78.48%

Table 6.7: Comparison of classification performances reported in the literature.

6.6 Conclusion

In this chapter, we have shown that intelligently initializing the network parameters, through transfer learning, allows to obtain better classification of AD stages by more that 5 points in some classification tasks (MCI/AD and NC/MCI). We compared various transfer learning schemes: cross-modal transfer learning using sMRI and DTI-MD brain images, cross-domain transfer learning from non-medical data to medical brain scans and a combination of both using a shallow LeNet network. Our approach remains light-weight in the sense that we used the "2D+ε" scheme we previously developed on the hippocampal region, avoiding both 3D convolutions and full-brain usage. As an interpretation of our results, filters trained on a modality have similar geometrical characteristics that need a small adaptation when transferred to another modality. We think that this is due to the fact that underlying structures of the hippocampus present similar geometrical patterns, visual markers, on both modalities that characterize the progression of the disease. We think that proposed multi-modal transfer learning approach can be useful in other classification tasks on medical images.

General conclusion and future work

In this dissertation, we studied the challenge of Alzheimer's disease classification using the MRI multimodal imaging technique. These imaging tools could accurately diagnose Alzheimer's disease condition and its evolution for a while, whether in clinical assessment or research settings.

Machine learning algorithms, especially its "Deep Learning" sub-domain, have produced emerging results for various domain applications. The medical field is one of them. At this stage, there is no doubt about the advantages and benefits of using models (Deep Learning) for AD pattern recognition. Thus, in the current Ph.D. work, we proposed a CAD approaches for AD classification that integrates high automatic classification modules to help clinicians understand AD pathology and make accurate and quick decisions. These methods consist of developing robust end-to-end supervised architectures that do not require any prior feature extraction or signature generation process and adopt them for brain MRI imaging. We specifically considered three categories in disease assessment, and the purpose is to classify subjects with (AD) from those in normal condition (NC) or Mild Cognitive Impairment (MCI).

During this work, our goal has been to show the effectiveness of such approaches combined with domain knowledge in supporting AD diagnosis. We engaged both MRI modalities images - the structural MRI and the DTI-derived map - in developing intelligent models for AD classification. However, the deep methods necessitate large datasets for training models; overall, they consist of millions of parameters. This is a big challenge in the medical field, while datasets are not often readily available. Thus, we addressed two main crucial points. How to deal with the constraint of limitation of the available dataset, and how to conceive adaptable architecture designs for our AD classification challenge.

In chapter 2, we firstly introduced the acquisition methods for MRI scans by proving the theory and the concept behind these techniques. Besides a complete description of the used datasets. Then, we presented the neuro-imaging flow-diagram with preprocessing steps of data-preparation. It mainly consists of three main levels, (i) Data correction and denoising processes, which are applied to the raw incoming MRI scans. (ii) Specific processing in order to select our target region on the brain, we followed successive operations that contains the spatial normalization and intensity normalization for sMRI scans, and skull stripping, co-registration for DTI scans as well. (iii) This is the final step, which the ROI-extraction. Here we developed an automated selection framework that provides a final dataset properly extracted and augmented. The result is settings folders that are ready to feed models for training and testing purposes.

In chapter 3, we covered a detailed presentation on the deep learning algorithms used in our AD classification problem. It includes a theoretical analysis of CNN methods with their mains components. We also presented optimization methods such as gradient descent and their derivatives. On the other hand, it discusses the constraint of data limitation and introduces alternative AD-adapted solutions to overcome and confront the problem of overfitting.

In chapter 4, we proposed our "2-D+ε" approach using the ROI-level method. We investigated and designed an adapted architecture suitable for our challenging problem by taking into account the complexity of computing that needs to be low, whereas increasing the accuracy of results. We made various analyses with different settings in order to study the impact of data augmentation mechanism, namely (i) simple data augmentation (ii) data augmentation with duplication of original scans (ii) randomized reduction of the augmented dataset.

In chapter 5, we further extended our proposed CNN-based model by introducing the fusion methods. We developed multiple parallel single networks to include different projections and different modalities. We took advantage of the combination methods using different settings and strategies to enhance our models' robustness. We showed the effectiveness of using multiple data taken from various representations of the target ROI. Moreover, we exhibited the improved results provided by the majority vote system applied to post-decisions.

In chapter 6, we investigated the application of the transfer learning approach. Here, we proposed two innovative approaches: (i) A cross-modal transfer learning where we pre-trained our models on one MRI modality dataset to the other. (ii) A cross-domain transfer learning by incorporating an external dataset and a LeNET-like model. Besides these two mechanisms, we also introduced a hybrid scheme that combines two-levels of transfer knowledge using (i) and (ii) methods. We showed that the cross-modal method provides adequate results, and it is suitable for working with a shallow CNN network for low-resolution MRI scans. It yields significant results even if the model is trained on small datasets, which is often the case in medical image analysis.

The obtained results demonstrate promising classification performance and simplicity compared to the state-of-the-art, even for full-volumetric-level ou ROI-level AD diagnosis methods.

Lastly, we conclude this thesis by identifying and discussing potential future research lines that can be envisaged either for improving performances by bringing extensions to the proposed models preserving the ROI-level approach, or going further to investigate the AD prediction problem.

Turning to the proposed approaches in chapter 5. Combining information from varied sources obviously enhances performances, as exhibited by using structural MRI and DTI modalities. However, more complementary modalities used in Alzheimer's disease studies, namely PET imagery may be integrated. In fact, as we know, the hallmarks of AD include tissue lesions such as the amyloid plaques and neurofibrillary tangles (degeneration), besides neuronal and synaptic loss. PET

imaging may add more information about amyloid-β (Aβ) for plaques and tau since it utilizes specific radiotracers to visualize them in the brain at various stages. On the other hand, investigation of other ROIs related to AD alongside hippocampus regions would be more discriminative together for diagnosis.

Furthermore, the classification MCI category remains the more challenging class, whether for the two-binary classification - AD/MCI or MCI/NC - or itself as it contains two sub-class: the e-MCI and l-MCI, as stated in . Still, the ROI-level method that we applied in this thesis remains appropriate to deal with this challenge. However, it would be necessary for future work to reinforce models with other sources of information, and extended brains structures, probably the whole brain.

One interesting research line that we have not covered in this thesis is the AD prediction over time, i.e., predicting subject conversion to AD rather than recognizing its class. This is an exciting pathway that can help clinicians prescribe medical treatment and even present cognitive therapy to prevent patients from converting to AD. Siamese networks can be employed, for instance, to compare different states of a patient using MRI scans to predict its stability condition. Another approach would be interesting to investigate, which is the recurrent neural networks (R-NN). The latter has been introduced in many research applications that necessitate recursive recognition processes such as text treatment and speech recognition. This would develop more comprehensible and valuable systems that bring the concept of sequence or time dimensions for well-estimating the brain variation of more sensitive AD regions over time. Long Short-Term Memory (LSTM) and its variations methods, for example, seem a promising way of addressing this challenge.

Publication list & Workshop

1. **Journals and Conferences**

- Karim Aderghal, Karim Afdel, Jenny Benois-Pineau, Gwenaelle Catheline: "Improving Alzheimer's stage categorization with Convolutional Neural Network using transfer learning and different magnetic resonance imaging modalities", Heliyon, Elsevier 2020, 6 (12), pp.e05652.
  DOI: 10.1016/j.heliyon.2020.e05652.

- Karim Aderghal, Manuel Boissenin, Jenny Benois-Pineau, Gwénaëlle Catheline, Karim Afdel: "Classification of sMRI for AD Diagnosis with Convolutional Neuronal Networks: A Pilot 2-D+ε Study on ADNI." *International Conference MultiMedia Modeling (ICMMM), Reykjavik, Iceland, 2017: pp 690-701, Lecture Notes in Computer Science, vol 10132. Springer, Cham*
  DOI: 10.1007/978-3-319-51811-4_56.

- Karim Aderghal, Jenny Benois-Pineau, Karim Afdel: "Classification of sMRI for Alzheimer's disease Diagnosis with CNN: Single Siamese Networks with 2D+ε Approach and Fusion on ADNI." *International Conference on Multimedia Retrieval (ICMR), Bucharest, Romania, 2017: pp 494-498*
  DOI: 10.1145/3078971.3079010.
  Recipient of ACM/SIGMM student grants.

- Karim Aderghal, Jenny Benois-Pineau, Karim Afdel, Gwénaëlle Catheline: "FuseMe: Classification of sMRI images by fusion of Deep CNNs in 2D+ε projections." *International Conference Content-Based Multimedia Indexing (CBMI), Florence, Italy, 2017: pp 34:1-34:7*
  DOI: 10.1145/3095713.3095749.

- Karim Aderghal, Alexander Khvostikov, Andrei Krylov, Jenny Benois-Pineau, Karim Afdel and Gwenaelle Catheline: "Classification of Alzheimer Disease on imaging modalities with Deep CNNs using cross-modal transfer learning." *International Symposium on Computer-Based Medical Systems (CBMS), Karlstad, Sweden, 2018: pp 345-350*
  DOI: 10.1109/CBMS.2018.00067.
  IEEE Best Student Paper Award.

2. **Pre-publication**

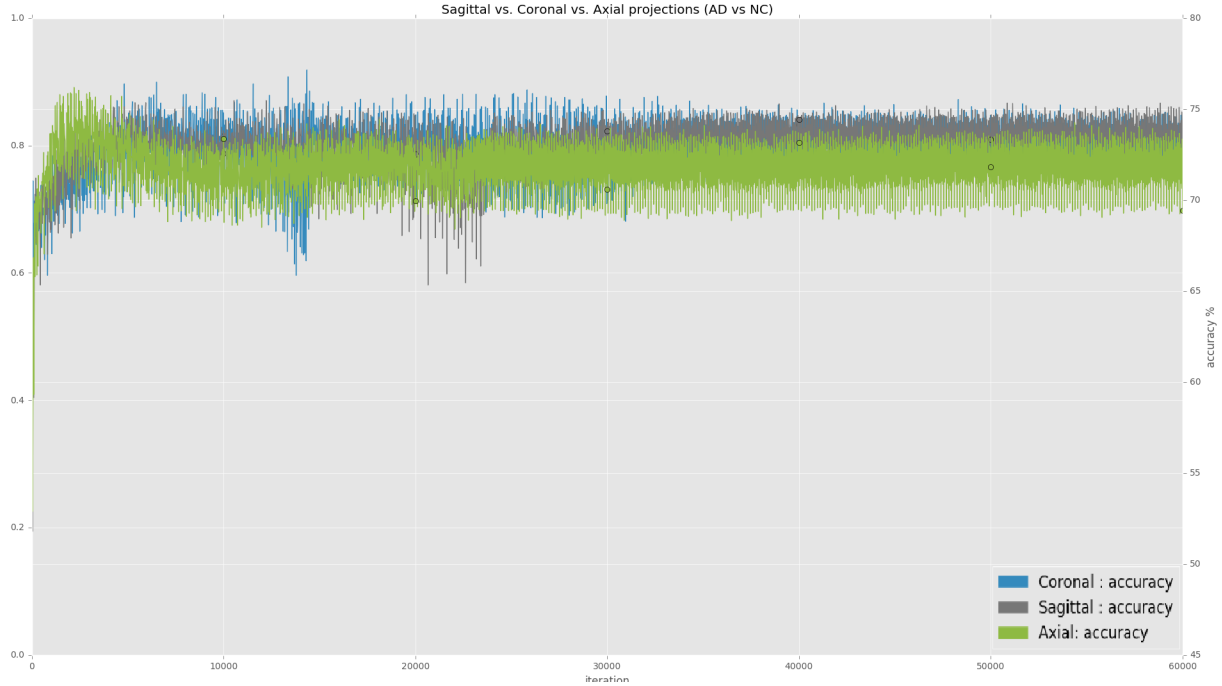
   - Alexander Khvostikov, Karim Aderghal, Jenny Benois-Pineau, Andrey Krylov, Gwenaelle Catheline:"3D CNN-based classification using sMRI and MD-DTI images for Alzheimer disease studies", arXiv preprint: 1801.05968v1.
   - Alexander Khvostikov, Karim Aderghal, Andrey Krylov, Gwenaelle Catheline, Jenny Benois-Pineau: "3D Inception-based CNN with sMRI and MD-DTI data fusion for Alzheimer's Disease diagnostics", arXiv preprint: 1809.03972.

3. **Workshop**

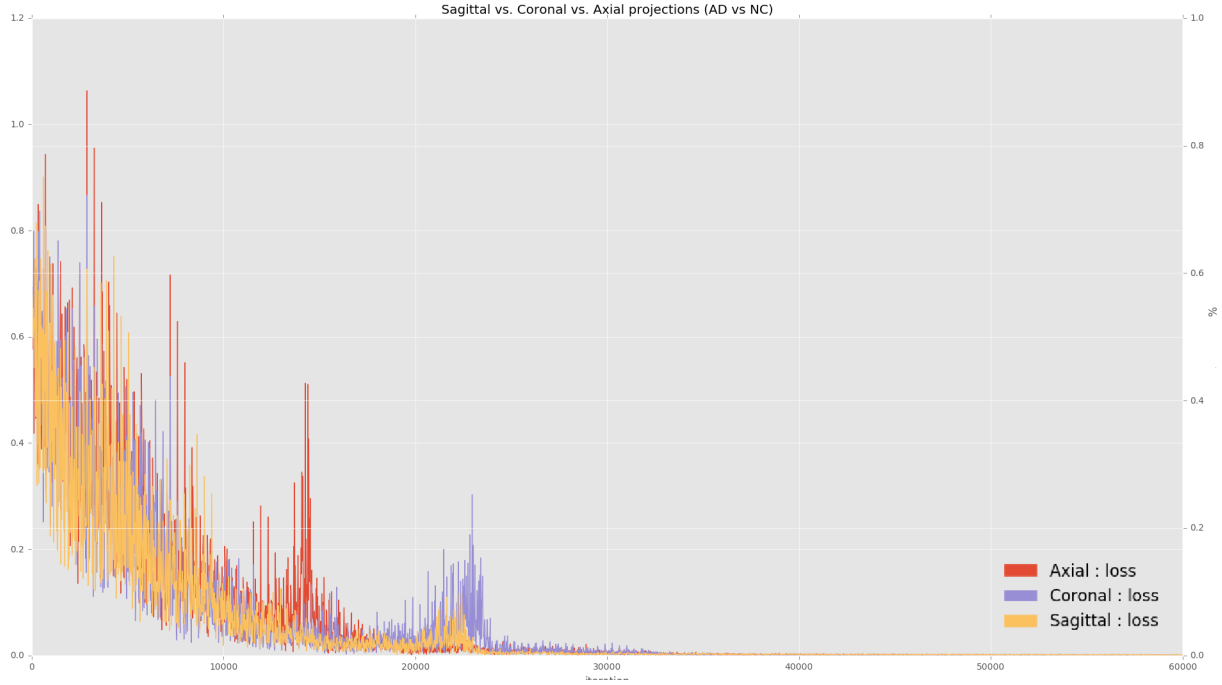
   - Karim Aderghal, Alexander Khvostikov, Andrei Krylov, Jenny Benois-Pineau, Karim Afdel and Gwenaelle Catheline: "Classification of Alzheimer Disease on imaging modalities with Deep CNNs using cross-modal transfer learning.", French-German Workshop: "Transfer Learning: From Theory to Applications", ENS Paris-Saclay, Cachan, France.

Appendix A

Figures

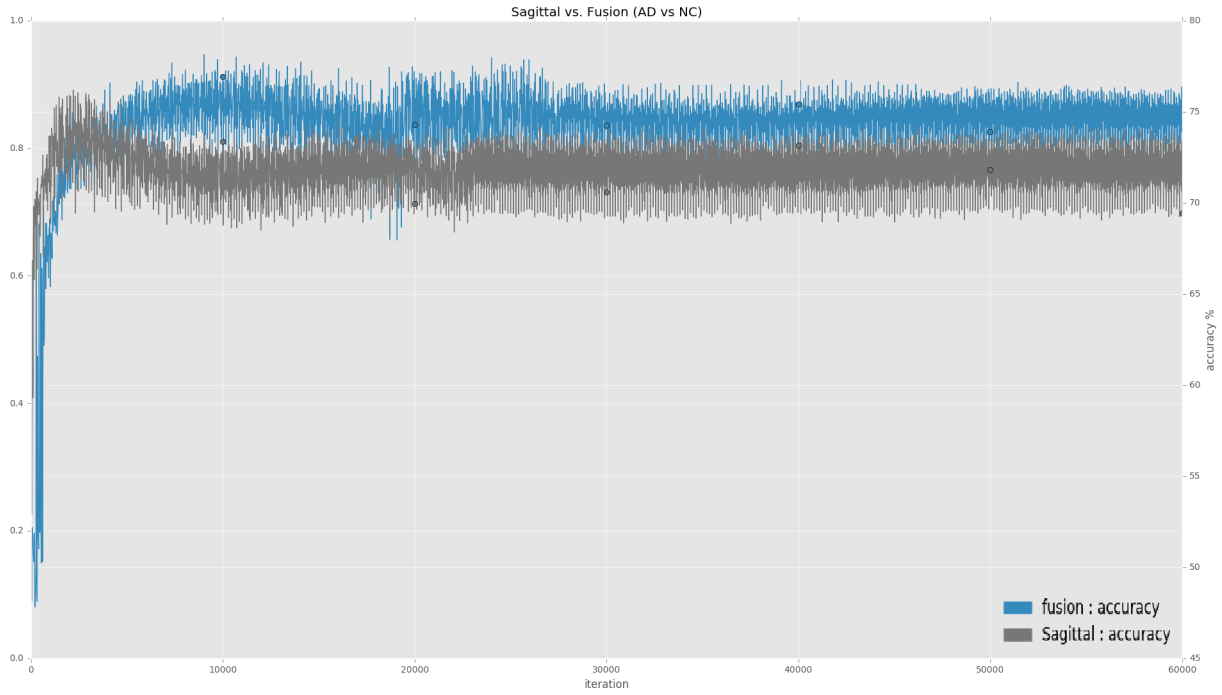


(a)

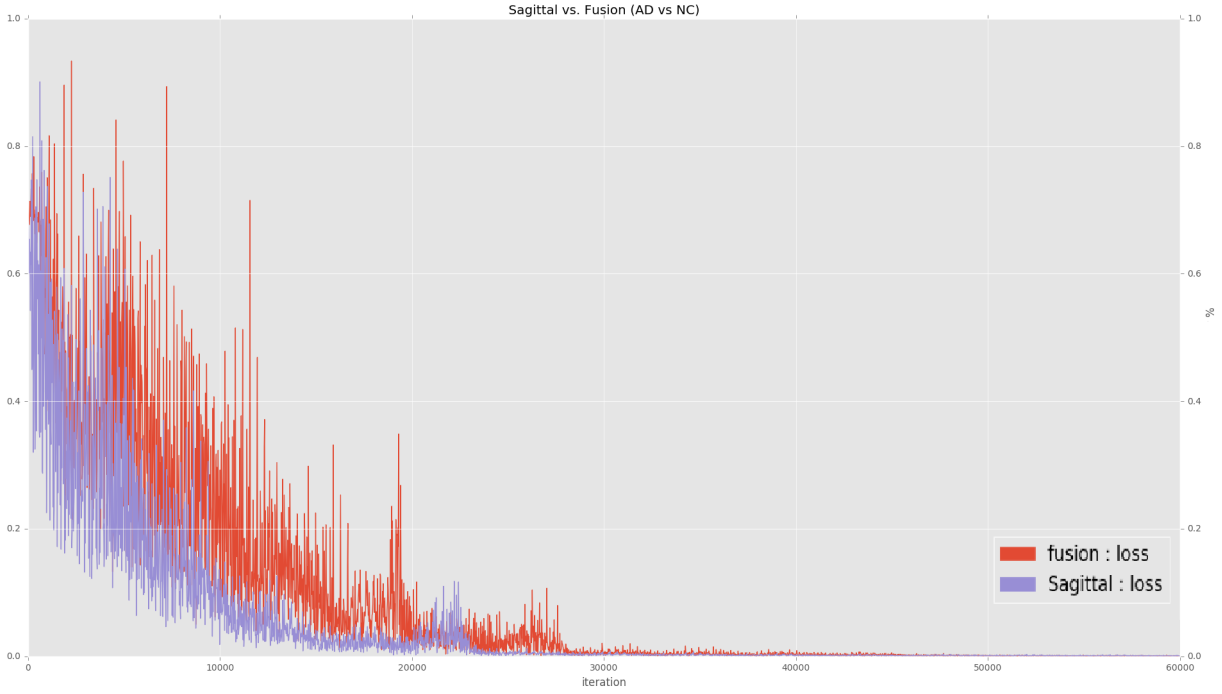


(b)

Figure A.1: AD/NC: Comparison of the three single projections curves (Accuracy and Loss).



(a)



(b)

Figure A.2: AD/NC: Comparison of Intermediate Fusion and Sagittal projection only (Accuracy and Loss).

References

[1] Karim Aderghal et al. “Classification of sMRI for AD Diagnosis with Convolutional Neuronal Networks: A Pilot 2D+ε Study on ADNI”. In: *International Conference on Multimedia Modeling*. Springer. 2017, pp. 690–701.

[2] Karim Aderghal et al. “FuseMe: Classification of sMRI images by fusion of Deep CNNs in 2D+ε projections”. In: *Proceedings of the 15th International Workshop on Content-Based Indexing*. ACM. 2017, p. 34.

[3] Karim Aderghal et al. “Improving Alzheimer’s stage categorization with Convolutional Neural Network using transfer learning and different magnetic resonance imaging modalities”. In: *Heliyon* 6.12 (2020), e05652.

[4] G Aël Chetelat and Jean-Claude Baron. “Early diagnosis of Alzheimer’s disease: contribution of structural neuroimaging”. In: *Neuroimage* 18.2 (2003), pp. 525–541.

[5] Olfa Ben Ahmed et al. “Alzheimer’s disease diagnosis on structural MR images using circular harmonic functions descriptors on hippocampus and posterior cingulate cortex”. In: *Computerized Medical Imaging and Graphics* 44 (2015), pp. 13–25.

[6] Olfa Ben Ahmed et al. “Classification of Alzheimer’s disease subjects from MRI using hippocampal visual features”. In: *Multimedia Tools and Applications* 74.4 (2015), pp. 1249–1266.

[7] Zemmari Akka and Jenny Benois-Pineau. *Deep Learning in Mining of Visual Content*. Vol. 1. 1. Springer International Publishing, 2020.

[8] Bérard Alain et al. “"Combien coûte la maladie d’Alzheimer ?"”. In: *Fondation Médéric Alzheimer* (2015).

[9] *Alzheimer’s Disease International (AZ) World Alzheimer Report 2016, London UK : Alzheimer’s Disease International*. https : / / www . alz . co . uk / research / WorldAlzheimerReport2016.pdf/. [Accessed December 12, 2017]. 2016.

[10] Jesper LR Andersson and Stamatios N Sotiropoulos. “An integrated approach to correction for off-resonance effects and subject movement in diffusion MR imaging”. In: *Neuroimage* 125 (2016), pp. 1063–1078.

[11] Steven E Arnold et al. “The topographical and neuroanatomical distribution of neurofibrillary tangles and neuritic plaques in the cerebral cortex of patients with Alzheimer’s disease”. In: *Cerebral cortex* 1.1 (1991), pp. 103–116.

[12] John Ashburner. “A fast diffeomorphic image registration algorithm”. In: *Neuroimage* 38.1 (2007), pp. 95–113.

[13] John Ashburner and K Friston. “Multimodal image coregistration and partitioning—a unified framework”. In: *Neuroimage* 6.3 (1997), pp. 209–217.

[14] John Ashburner and Karl J Friston. “Nonlinear spatial normalization using basis functions”. In: *Human brain mapping* 7.4 (1999), pp. 254–266.

[15] John Ashburner and Karl J Friston. “Voxel-based morphometry—the methods”. In: *Neuroimage* 11.6 (2000), pp. 805–821.

[16] John Ashburner et al. “Incorporating prior knowledge into image registration”. In: *Neuroimage* 6.4 (1997), pp. 344–352.

[17] Yaniv Assaf et al. “The CONNECT project: combining macro-and micro-structure”. In: *Neuroimage* 80 (2013), pp. 273–282.

[18] Alzheimer’s Association et al. “2016 Alzheimer’s disease facts and figures”. In: *Alzheimer’s & Dementia* 12.4 (2016), pp. 459–509.

[19] Alzheimer’s Association et al. “2017 Alzheimer’s disease facts and figures”. In: *Alzheimer’s & Dementia* 13.4 (2017), pp. 325–373.

[20] Alzheimer’s Association et al. “2018 Alzheimer’s disease facts and figures”. In: *Alzheimer’s & Dementia* 14.3 (2018), pp. 367–429.

[21] Brian B Avants et al. “A reproducible evaluation of ANTs similarity metric performance in brain image registration”. In: *Neuroimage* 54.3 (2011), pp. 2033–2044.

[22] Gururaj Awate et al. “Detection of Alzheimers Disease from MRI using Convolutional Neural Network with Tensorflow”. In: *arXiv preprint arXiv:1806.10170* (2018).

[23] Karl Bäckström et al. “An efficient 3D deep convolutional network for Alzheimer’s disease diagnosis using MR images”. In: *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE. 2018, pp. 149–153.

[24] Silvia Basaia et al. “Automated classification of Alzheimer’s disease and mild cognitive impairment using a single MRI and deep neural networks”. In: *NeuroImage: Clinical* 21 (2019), p. 101645.

[25] Peter J Basser, James Mattiello, and Denis LeBihan. “Estimation of the effective self-diffusion tensor from the NMR spin echo”. In: *Journal of Magnetic Resonance, Series B* 103.3 (1994), pp. 247–254.

[26] Peter J Basser, James Mattiello, and Denis LeBihan. “MR diffusion tensor spectroscopy and imaging”. In: *Biophysical journal* 66.1 (1994), pp. 259–267.

[27] Olfa Ben Ahmed. “Classification des IRM par les descripteurs de contenu : application au diagnostic précoce de la maladie d’Alzheimer”. Theses. Université de Bordeaux ; Université de Sfax (Tunisie), Jan. 2015. URL: https://tel.archives-ouvertes.fr/tel-01424145.

[28] Y Bengio. “Deep learning of representations for unsupervised and transfer learning, JMLR: Workshop Conf”. In: *Proc*. Vol. 7. 2011, pp. 1–20.

[29] James C Bezdek et al. *Fuzzy models and algorithms for pattern recognition and image processing*. Vol. 4. Springer Science & Business Media, 1999.

[30] Pushkar Bhatkoti and Manoranjan Paul. “Early diagnosis of Alzheimer’s disease: A multi-class deep learning framework with modified k-sparse autoencoder classification”. In: *2016 International Conference on Image and Vision Computing New Zealand (IVCNZ)*. IEEE. 2016, pp. 1–5.

[31] Ciprian D Billones et al. “DemNet: A Convolutional Neural Network for the detection of Alzheimer’s Disease and Mild Cognitive Impairment”. In: (2016), pp. 3724–3727.

[32] Ciprian D Billones et al. “DemNet: A Convolutional Neural Network for the detection of Alzheimer’s Disease and Mild Cognitive Impairment”. In: (2016), pp. 3724–3727.

[33] Maciek Bobinski et al. “MRI of entorhinal cortex in mild Alzheimer’s disease”. In: *The Lancet* 353.9146 (1999), pp. 38–40.

[34] H Braak and E Braak. “Evolution of neuronal changes in the course of Alzheimer’s disease”. In: *Ageing and dementia*. Springer, 1998, pp. 127–140.

[35] Heiko Braak and Eva Braak. “Neuropathological stageing of Alzheimer-related changes”. In: *Acta neuropathologica* 82.4 (1991), pp. 239–259.

[36] Heiko Braak, Eva Braak, and Jürgen Bohl. “Staging of Alzheimer-related cortical destruction”. In: *European neurology* 33.6 (1993), pp. 403–408.

[37] Robert W Brown et al. *Magnetic resonance imaging: physical principles and sequence design*. John Wiley & Sons, 2014.

[38] A Brun and L Gustafson. “Distribution of cerebral degeneration in Alzheimer’s disease”. In: *Archiv für Psychiatrie und Nervenkrankheiten* 223.1 (1976), pp. 15–33.

[39] Antoni Buades, Bartomeu Coll, and J-M Morel. "A non-local algorithm for image denoising". In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. Vol. 2. IEEE. 2005, pp. 60–65.

[40] Lei Cai et al. "Deep adversarial learning for multi-modality missing data completion". In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2018, pp. 1158–1166.

[41] Yuanyuan Chen et al. "Early Identification of Alzheimer's Disease Using an Ensemble of 3D Convolutional Neural Networks and Magnetic Resonance Imaging". In: *International Conference on Brain Inspired Cognitive Systems*. Springer. 2018, pp. 303–311.

[42] Bo Cheng et al. "Multi-domain transfer learning for early diagnosis of alzheimer's disease". In: *Neuroinformatics* 15.2 (2017), pp. 115–132.

[43] Danni Cheng and Manhua Liu. "CNNs based multi-modality classification for AD diagnosis". In: *2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*. IEEE. 2017, pp. 1–5.

[44] Danni Cheng et al. "Classification of MR brain images by combination of multi-CNNs for AD diagnosis". In: *Ninth International Conference on Digital Image Processing (ICDIP 2017)*. Vol. 10420. International Society for Optics and Photonics. 2017, p. 1042042.

[45] Andrea Cherubini et al. "Combined volumetry and DTI in subcortical structures of mild cognitive impairment and Alzheimer's disease patients". In: *Journal of Alzheimer's Disease* 19.4 (2010), pp. 1273–1282.

[46] Hongyoon Choi, Kyong Hwan Jin, Alzheimer's Disease Neuroimaging Initiative, et al. "Predicting cognitive decline with deep learning of brain metabolism and amyloid imaging". In: *Behavioural brain research* 344 (2018), pp. 103–109.

[47] D Collazos-Huertas et al. "MRI-Based Feature Extraction Using Supervised General Stochastic Networks in Dementia Diagnosis". In: *International Work-Conference on the Interplay Between Natural and Artificial Computation*. Springer. 2017, pp. 363–373.

[48] D Louis Collins et al. "Automatic 3D intersubject registration of MR volumetric data in standardized Talairach space." In: *Journal of computer assisted tomography* 18.2 (1994), pp. 192–205.

[49] Sergi G Costafreda et al. "Automated hippocampal shape analysis predicts the onset of dementia in mild cognitive impairment". In: *Neuroimage* 56.1 (2011), pp. 212–219.

[50] Pierrick Coupé et al. "An optimized blockwise nonlocal means denoising filter for 3-D magnetic resonance images". In: *IEEE transactions on medical imaging* 27.4 (2008), pp. 425–441.

[51] Pierrick Coupé et al. “Scoring by nonlocal image patch estimator for early detection of Alzheimer’s disease”. In: *NeuroImage: clinical* 1.1 (2012), pp. 141–152.

[52] Kristen Coyne. “MRI: A guided tour, National High Magnetic Field Lab, Florida State University, Tallahassee, FL”. In: *Internet:http://www.magnet.fsu.edu/* (2012).

[53] Ruoxuan Cui and Manhua Liu. “Hippocampus analysis based on 3D CNN for Alzheimer’s disease diagnosis”. In: *Tenth International Conference on Digital Image Processing (ICDIP 2018)*. Vol. 10806. International Society for Optics and Photonics. 2018, 108065O.

[54] Rémi Cuingnet et al. “Automatic classification of patients with Alzheimer’s disease from structural MRI: a comparison of ten methods using the ADNI database”. In: *neuroimage* 56.2 (2011), pp. 766–781.

[55] E. L. Cunningham et al. “Dementia”. In: *Ulster Med J* 84.2 (2015), pp. 79–87.

[56] MJ De Leon et al. “MRI and CSF studies in the early diagnosis of Alzheimer’s disease”. In: *Journal of internal medicine* 256.3 (2004), pp. 205–223.

[57] Bernard Deweer et al. “Memory disorders in probable Alzheimer’s disease: the role of hippocampal atrophy as shown with MRI.” In: *Journal of Neurology, Neurosurgery & Psychiatry* 58.5 (1995), pp. 590–597.

[58] Bradford C Dickerson et al. “MRI-derived entorhinal and hippocampal atrophy in incipient and very mild Alzheimer’s disease”. In: *Neurobiology of aging* 22.5 (2001), pp. 747–754.

[59] Chester V Dolph et al. “Deep learning of texture and structural features for multiclass Alzheimer’s disease classification”. In: *2017 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2017, pp. 2259–2266.

[60] AT Du et al. “Higher atrophy rate of entorhinal cortex than hippocampus in AD”. In: *Neurology* 62.3 (2004), pp. 422–427.

[61] AT Du et al. “Magnetic resonance imaging of the entorhinal cortex and hippocampus in mild cognitive impairment and Alzheimer’s disease”. In: *Journal of Neurology, Neurosurgery & Psychiatry* 71.4 (2001), pp. 441–447.

[62] Mr Amir Ebrahimighahnavieh and Raymond Chiong. “Deep Learning to Detect Alzheimer’s Disease from Neuroimaging: A Systematic Literature Review”. In: *Computer Methods and Programs in Biomedicine* (2019), p. 105242.

[63] Mr Amir Ebrahimighahnavieh, Suhuai Luo, and Raymond Chiong. “Deep learning to detect Alzheimer’s disease from neuroimaging: A systematic literature review”. In: *Computer methods and programs in biomedicine* 187 (2020), p. 105242.

[64] MM Esiri et al. “"Ageing and dementia. In: Graham DI, Lantos PL (eds.) Greenfield’s Neuropathology (6th ed.) Arnold"”. In: *Ageing and dementia* 2 (1997), pp. 153–233.

[65] Ammara Farooq et al. “Artificial intelligence based smart diagnosis of alzheimer’s disease and mild cognitive impairment”. In: *2017 International Smart cities conference (ISC2)*. IEEE. 2017, pp. 1–4.

[66] Ammarah Farooq et al. “A deep CNN based multi-class classification of Alzheimer’s disease using MRI”. In: *2017 IEEE International Conference on Imaging systems and techniques (IST)*. IEEE. 2017, pp. 1–6.

[67] Luca Ferrarini et al. “Shape differences of the brain ventricles in Alzheimer’s disease”. In: *Neuroimage* 32.3 (2006), pp. 1060–1069.

[68] Bruce Fischl. “FreeSurfer”. In: *Neuroimage* 62.2 (2012), pp. 774–781.

[69] Bruce Fischl and Anders M Dale. “Measuring the thickness of the human cerebral cortex from magnetic resonance images”. In: *Proceedings of the National Academy of Sciences* 97.20 (2000), pp. 11050–11055.

[70] Marshal F Folstein, Susan E Folstein, and Paul R McHugh. ““Mini-mental state”: a practical method for grading the cognitive state of patients for the clinician”. In: *Journal of psychiatric research* 12.3 (1975), pp. 189–198.

[71] GB Frisoni et al. “Detection of grey matter loss in mild Alzheimer’s disease with voxel based morphometry”. In: *Journal of Neurology, Neurosurgery & Psychiatry* 73.6 (2002), pp. 657–664.

[72] GB Frisoni et al. “Structural correlates of early and late onset Alzheimer’s disease: voxel based morphometric study”. In: *Journal of Neurology, Neurosurgery & Psychiatry* 76.1 (2005), pp. 112–114.

[73] Giovanni B Frisoni et al. “The clinical use of structural MRI in Alzheimer disease”. In: *Nature Reviews Neurology* 6.2 (2010), p. 67.

[74] Karl J Friston. “Statistical parametric mapping and other analyses of functional imaging data”. In: *Brain mapping: The methods* (1996).

[75] Karl J Friston et al. “Spatial registration and normalization of images”. In: *Human brain mapping* 3.3 (1995), pp. 165–189.

[76] Ritu Gautam and Manik Sharma. “Prevalence and Diagnosis of Neurological Disorders Using Different Deep Learning Techniques: A Meta-Analysis”. In: *Journal of Medical Systems* 44.2 (2020), p. 49.

[77] Ali Gholipour et al. “Brain functional localization: a survey of image registration techniques”. In: *IEEE transactions on medical imaging* 26.4 (2007), pp. 427–451.

[78] Tanya Glozman and Orly Liba. “Hidden Cues: Deep Learning for Alzheimer’s Disease Classification CS331B project final report”. In: (2016).

[79] Ian J. Goodfellow et al. *Generative Adversarial Networks*. 2014. arXiv: 1406 . 2661 [stat.ML].

[80] KANNP Gunawardena, RN Rajapakse, and ND Kodikara. "Applying convolutional neural networks for pre-detection of alzheimer's disease from structural MRI data". In: *2017 24th International Conference on Mechatronics and Machine Vision in Practice (M2VIP)*. IEEE. 2017, pp. 1–7.

[81] Ashish Gupta, Murat Ayhan, and Anthony Maida. "Natural image bases to represent neuroimaging data". In: *International conference on machine learning*. 2013, pp. 987–994.

[82] John R Hansen. "Pulsed NMR study of water mobility in muscle and brain tissue". In: *Biochimica et Biophysica Acta (BBA)-General Subjects* 230.3 (1971), pp. 482–486.

[83] Marcia Hon and Naimul Mefraz Khan. "Towards Alzheimer's disease classification through transfer learning". In: *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2017, pp. 1166–1169.

[84] Ehsan Hosseini-Asl, Robert Keynto, and Ayman El-Baz. "Alzheimer's disease diagnostics by adaptation of 3D convolutional network". In: *arXiv preprint arXiv:1607.00455* (2016).

[85] Ehsan Hosseini-Asl, Robert Keynton, and Ayman El-Baz. "Alzheimer's disease diagnostics by adaptation of 3D convolutional network". In: *2016 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2016, pp. 126–130.

[86] Chenhui Hu et al. "Clinical decision support for Alzheimer's disease based on deep learning and brain network". In: *2016 IEEE International Conference on Communications (ICC)*. IEEE. 2016, pp. 1–6.

[87] Jyoti Islam and Yanqing Zhang. "A novel deep learning based multi-class classification method for Alzheimer's disease detection using brain MRI data". In: *International Conference on Brain Informatics*. Springer. 2017, pp. 213–222.

[88] Emimal Jabason, M Omair Ahmad, and MN S Swamy. "Shearlet based Stacked Convolutional Network for Multiclass Diagnosis of Alzheimer's Disease using the Florbetapir PET Amyloid Imaging Data". In: *2018 16th IEEE International New Circuits and Systems Conference (NEWCAS)*. IEEE. 2018, pp. 344–347.

[89] Clifford R Jack et al. "Brain atrophy rates predict subsequent clinical conversion in normal elderly and amnestic MCI". In: *Neurology* 65.8 (2005), pp. 1227–1231.

[90] Clifford R Jack et al. "MR-based hippocampal volumetry in the diagnosis of Alzheimer's disease". In: *Neurology* 42.1 (1992), pp. 183–183.

[91] Clifford R Jack et al. "Prediction of AD with MRI-based hippocampal volume in mild cognitive impairment". In: *Neurology* 52.7 (1999), pp. 1397–1397.

[92] Clifford R Jack et al. “Rates of hippocampal atrophy correlate with change in clinical status in aging and AD”. In: *Neurology* 55.4 (2000), pp. 484–490.

[93] CR Jack et al. “Comparison of different MRI brain atrophy rate measures with clinical disease progression in AD”. In: *Neurology* 62.4 (2004), pp. 591–600.

[94] Mark Jenkinson et al. “Fsl”. In: *Neuroimage* 62.2 (2012), pp. 782–790.

[95] Ben Jeurissen. “Improved analysis of brain connectivity using high angular resolution diffusion MRI”. PhD thesis. Universiteit Antwerpen (Belgium), 2012.

[96] Yangqing Jia et al. “Caffe: Convolutional architecture for fast feature embedding”. In: *Proceedings of the 22nd ACM international conference on Multimedia*. 2014, pp. 675–678.

[97] Ronghui Ju, Chenhui Hu, Quanzheng Li, et al. “Early diagnosis of Alzheimer’s disease based on resting-state brain networks and deep learning”. In: *IEEE/ACM transactions on computational biology and bioinformatics* 16.1 (2017), pp. 244–257.

[98] K Juottonen et al. “Volumes of the entorhinal and perirhinal cortices in Alzheimer’s disease”. In: *Neurobiology of aging* 19.1 (1998), pp. 15–22.

[99] GB Karas et al. “Global and local gray matter loss in mild cognitive impairment and Alzheimer’s disease”. In: *Neuroimage* 23.2 (2004), pp. 708–716.

[100] Andrej Karpathy. “Stanford university cs231n: Convolutional neural networks for visual recognition”. In: *URL: http://cs231n.stanford.edu/syllabus.html* (2018).

[101] Yosra Kazemi and Sheridan Houghten. “A deep learning pipeline to classify different stages of Alzheimer’s disease from fMRI data”. In: *2018 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*. IEEE. 2018, pp. 1–8.

[102] Paul-Ariel Kenigsberg et al. “Impact socio-économique de la maladie d’Alzheimer et des maladies apparentées en Europe”. In: *Gérontologie et société* 32.1 (2009), pp. 297–318.

[103] J Patrick Kesslak, Orhan Nalcioglu, and Carl W Cotman. “Quantification of magnetic resonance scans for hippocampal and parahippocampal atrophy in Alzheimer’s disease”. In: *Neurology* 41.1 (1991), pp. 51–51.

[104] Alexander Khvostikov et al. “3D Inception-based CNN with sMRI and MD-DTI data fusion for Alzheimer’s Disease diagnostics”. In: *arXiv preprint arXiv:1809.03972* (2018).

[105] Stefan Klöppel et al. “Accuracy of dementia diagnosis—a direct comparison between radiologists and a computerized method”. In: *Brain* 131.11 (2008), pp. 2969–2974.

[106] Sergey Korolev et al. “Residual and plain convolutional neural networks for 3D brain MRI classification”. In: *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*. IEEE. 2017, pp. 835–838.

[107] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In: vol. 25. 2012, pp. 1097–1105.

[108] C. A. Lane, J. Hardy, and J. M. Schott. "Alzheimer's disease". In: *European Journal of Neurology* 25.1 (2017), pp. 59–70.

[109] Paul C Lauterbur et al. "Image formation by induced local interactions: examples employing nuclear magnetic resonance". In: (1973).

[110] Denis Le Bihan and Heidi Johansen-Berg. "Diffusion MRI at 25: exploring brain tissue structure and function". In: *Neuroimage* 61.2 (2012), pp. 324–341.

[111] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning". In: *nature* 521.7553 (2015), p. 436.

[112] Yann LeCun et al. "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324.

[113] Yann LeCun et al. "LeNet-5, convolutional neural networks". In: ().

[114] Christian Ledig et al. "Robust whole-brain segmentation: application to traumatic brain injury". In: *Medical image analysis* 21.1 (2015), pp. 40–58.

[115] Bumshik LEE, Waqas ELLAHI, and Jae Young CHOI. "Using Deep CNN with Data Permutation Scheme for Classification of Alzheimer's Disease in Structural Magnetic Resonance Imaging (sMRI)". In: *IEICE Transactions on Information and Systems* E102.D.7 (2019), pp. 1384–1395.

[116] Fan Li, Danni Cheng, and Manhua Liu. "Alzheimer's disease classification based on combination of multi-model convolutional networks". In: *2017 IEEE International Conference on Imaging Systems and Techniques (IST)*. IEEE. 2017, pp. 1–5.

[117] Feng Li et al. "A robust deep model for improved classification of AD/MCI patients". In: *IEEE journal of biomedical and health informatics* 19.5 (2015), pp. 1610–1616.

[118] Feng Li et al. "Robust deep learning for improved classification of AD/MCI patients". In: *International Workshop on Machine Learning in Medical Imaging*. Springer. 2014, pp. 240–247.

[119] Chunfeng Lian et al. "Hierarchical fully convolutional network for joint atrophy localization and Alzheimer's disease diagnosis using structural MRI". In: *IEEE transactions on pattern analysis and machine intelligence* (2018).

[120] Shu Liao and Albert CS Chung. "Feature based nonrigid brain MR image registration with symmetric alpha stable filters". In: *IEEE transactions on medical imaging* 29.1 (2009), pp. 106–119.

[121] Weiming Lin et al. “Convolutional neural networks-based MRI image analysis for the Alzheimer’s disease prediction from mild cognitive impairment”. In: *Frontiers in neuroscience* 12 (2018), p. 777.

[122] Manhua Liu et al. “Hierarchical fusion of features and classifier decisions for Alzheimer’s disease diagnosis”. In: *Human brain mapping* 35.4 (2014), pp. 1305–1319.

[123] Mingxia Liu et al. “Landmark-based deep multi-instance learning for brain disease diagnosis”. In: *Medical image analysis* 43 (2018), pp. 157–168.

[124] Siqi Liu et al. “Early diagnosis of Alzheimer’s disease with deep learning”. In: *2014 IEEE 11th international symposium on biomedical imaging (ISBI)*. IEEE. 2014, pp. 1015–1018.

[125] Siqi Liu et al. “Multi-phase feature representation learning for neurodegenerative disease diagnosis”. In: *Australasian Conference on Artificial Life and Computational Intelligence*. Springer. 2015, pp. 350–359.

[126] Siqi Liu et al. “Multimodal neuroimaging feature learning for multiclass diagnosis of Alzheimer’s disease”. In: *IEEE Transactions on Biomedical Engineering* 62.4 (2014), pp. 1132–1140.

[127] Donghuan Lu et al. “Multimodal and multiscale deep neural networks for the early diagnosis of Alzheimer’s disease using structural MR and FDG-PET images”. In: *Scientific reports* 8.1 (2018), pp. 1–13.

[128] Donghuan Lu et al. “Multiscale deep neural network based analysis of FDG-PET images for the early diagnosis of Alzheimer’s disease”. In: *Medical image analysis* 46 (2018), pp. 26–34.

[129] Angshul Majumdar and Vanika Singhal. “Noisy deep dictionary learning: Application to Alzheimer’s Disease classification”. In: *2017 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2017, pp. 2679–2683.

[130] José V Manjón and Pierrick Coupé. “volBrain: an online MRI brain volumetry system”. In: *Frontiers in neuroinformatics* 10 (2016), p. 30.

[131] Kasper Marstal et al. “SimpleElastix: A user-friendly, multi-lingual library for medical image registration”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 2016, pp. 134–142.

[132] Guy McKhann et al. “Clinical diagnosis of Alzheimer’s disease: Report of the NINCDS-ADRDA Work Group* under the auspices of Department of Health and Human Services Task Force on Alzheimer’s Disease”. In: *Neurology* 34.7 (1984), pp. 939–939.

[133] Kelsey E Melah et al. “Cerebrospinal fluid markers of Alzheimer’s disease pathology and microglial activation are associated with altered white matter microstructure in asymptomatic adults at risk for Alzheimer’s disease”. In: *Journal of Alzheimer’s Disease* 50.3 (2016), pp. 873–886.

[134] Lilia Mesrob et al. “DTI and structural MRI classification in Alzheimer’s disease”. In: *Advances in molecular imaging* 2.02 (2012), p. 12.

[135] Ludovico Minati et al. “Reviews: current concepts in Alzheimer’s disease: a multidisciplinary review”. In: *American Journal of Alzheimer’s Disease & Other Dementias®* 24.2 (2009), pp. 95–121.

[136] Jonathan H Morra et al. “Automated 3D mapping of hippocampal atrophy and its clinical correlates in 400 subjects with Alzheimer’s disease, mild cognitive impairment, and elderly controls”. In: *Human brain mapping* 30.9 (2009), pp. 2766–2788.

[137] John C Morris et al. “Validation of clinical diagnostic criteria for Alzheimer’s disease”. In: *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society* 24.1 (1988), pp. 17–22.

[138] Pratik Mukherjee et al. “Diffusion tensor MR imaging and fiber tractography: theoretic underpinnings”. In: *American journal of neuroradiology* 29.4 (2008), pp. 632–641.

[139] Sean M Nestor et al. “Ventricular enlargement as a possible measure of Alzheimer’s disease progression validated using the Alzheimer’s disease neuroimaging initiative database”. In: *Brain* 131.9 (2008), pp. 2443–2454.

[140] Milap A Nowrangi et al. “Longitudinal, region-specific course of diffusion tensor imaging measures in mild cognitive impairment and Alzheimer’s disease”. In: *Alzheimer’s & dementia: the journal of the Alzheimer’s Association* 9.5 (2013), pp. 519–528.

[141] Chigozie Nwankpa et al. “Activation functions: Comparison of trends in practice and research for deep learning”. In: *arXiv preprint arXiv:1811.03378* (2018).

[142] László G Nyúl and Jayaram K Udupa. “On standardizing the MR image intensity scale”. In: *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* 42.6 (1999), pp. 1072–1081.

[143] László G Nyúl, Jayaram K Udupa, and Xuan Zhang. “New variants of a method of MRI scale standardization”. In: *IEEE transactions on medical imaging* 19.2 (2000), pp. 143–150.

[144] Emilio Soria Olivas et al. “Handbook Of Research On Machine Learning Applications and Trends: Algorithms, Methods and Techniques-2 Volumes”. In: (2009).

[145] Andres Ortiz et al. “Ensembles of deep learning architectures for the early diagnosis of the Alzheimer’s disease”. In: *International journal of neural systems* 26.07 (2016), p. 1650025.

[146] Juan M Ortiz-Suárez, Raúl Ramos-Pollán, and Eduardo Romero. "Exploring Alzheimer's anatomical patterns through convolutional networks". In: *12th International Symposium on Medical Information Processing and Analysis*. Vol. 10160. International Society for Optics and Photonics. 2017, 101600Z.

[147] Lauren J O'Donnell and Carl-Fredrik Westin. "An introduction to diffusion tensor image analysis". In: *Neurosurgery Clinics* 22.2 (2011), pp. 185–196.

[148] Adrien Payan and Giovanni Montana. "Predicting Alzheimer's disease: a neuroimaging study with 3D convolutional neural networks". In: *arXiv preprint arXiv:1502.02506* (2015).

[149] Corina Pennanen et al. "Hippocampus and entorhinal cortex in mild cognitive impairment and early AD". In: *Neurobiology of aging* 25.3 (2004), pp. 303–310.

[150] Jana Podhorna et al. "Alzheimer's Disease Assessment Scale–Cognitive subscale variants in mild cognitive impairment and mild Alzheimer's disease: change over time and the effect of enrichment strategies". In: *Alzheimer's research & therapy* 8.1 (2016), p. 8.

[151] Moacir P Ponti Jr. "Combining classifiers: from the creation of ensembles to the decision fusion". In: *2011 24th SIBGRAPI Conference on Graphics, Patterns, and Images Tutorials*. IEEE. 2011, pp. 1–10.

[152] Jianping Qiao et al. "Multivariate Deep Learning Classification of Alzheimer's Disease Based on Hierarchical Partner Matching Independent Component Analysis". In: *Frontiers in aging neuroscience* 10 (2018), p. 417.

[153] Saima Rathore et al. "A review on neuroimaging-based classification studies and associated feature extraction methods for Alzheimer's disease and its prodromal stages". In: *NeuroImage* 155 (2017), pp. 530–548.

[154] Saima Rathore et al. "A review on neuroimaging-based classification studies and associated feature extraction methods for Alzheimer's disease and its prodromal stages". In: *NeuroImage* 155 (2017), pp. 530–548.

[155] Basil H Ridha et al. "Volumetric MRI and cognitive measures in Alzheimer disease". In: *Journal of neurology* 255.4 (2008), pp. 567–574.

[156] Anne Brown Rodgers. *Alzheimer's disease: unraveling the mystery*. Vol. 1. National Institutes of Health, 2002.

[157] Christoffer Rosén et al. "Fluid biomarkers in Alzheimer's disease–current concepts". In: *Molecular neurodegeneration* 8.1 (2013), p. 20.

[158] Wilma G Rosen, Richard C Mohs, and Kenneth L Davis. "A new rating scale for Alzheimer's disease." In: *The American journal of psychiatry* (1984).

[159] Dymitr Ruta and Bogdan Gabrys. “An overview of classifier fusion methods”. In: *Computing and Information systems* 7.1 (2000), pp. 1–10.

[160] Nicaise Salomé and Palermiti Federico. “"Alzheimer and the mediterranean Report 2016"”. In: *Monegasque Association for research on Alzheimer’s disease* (2016).

[161] Saman Sarraf, Ghassem Tofighi, and. “DeepAD: Alzheimer’s Disease Classification via Deep Convolutional Neural Networks using MRI and fMRI”. In: *bioRxiv* ().

[162] Saman Sarraf and Ghassem Tofighi. “Classification of alzheimer’s disease structural MRI data by deep learning convolutional neural networks”. In: *arXiv preprint arXiv:1607.06583* (2016).

[163] Saman Sarraf and Ghassem Tofighi. “Deep learning-based pipeline to recognize Alzheimer’s disease using fMRI data”. In: *2016 Future Technologies Conference (FTC)*. IEEE. 2016, pp. 816–820.

[164] Saman Sarraf, Ghassem Tofighi, et al. “DeepAD: Alzheimer’s disease classification via deep convolutional neural networks using MRI and fMRI”. In: *BioRxiv* (2016), p. 070441.

[165] Ann I Scher et al. “Hippocampal shape analysis in Alzheimer’s disease: a population-based study”. In: *Neuroimage* 36.1 (2007), pp. 8–18.

[166] Dominik Scherer, Andreas Müller, and Sven Behnke. “Evaluation of pooling operations in convolutional architectures for object recognition”. In: *International conference on artificial neural networks*. Springer. 2010, pp. 92–101.

[167] Mahsa Shakeri et al. “Deep spectral-based shape features for Alzheimer’s disease classification”. In: *International Workshop on Spectral and Shape Analysis in Medical Imaging*. Springer. 2016, pp. 15–24.

[168] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[169] Bibo Shi et al. “Nonlinear feature transformation and deep fusion for Alzheimer’s Disease staging analysis”. In: *Pattern recognition* 63 (2017), pp. 487–498.

[170] Feng Shi et al. “Hippocampal volume and asymmetry in mild cognitive impairment and Alzheimer’s disease: Meta-analyses of MRI studies”. In: *Hippocampus* 19.11 (2009), pp. 1055–1064.

[171] Jun Shi et al. “Multimodal neuroimaging feature learning with multimodal stacked deep polynomial networks for diagnosis of Alzheimer’s disease”. In: *IEEE journal of biomedical and health informatics* 22.1 (2017), pp. 173–183.

[172] Yaroslav Shmulev, Mikhail Belyaev, Alzheimer's Disease Neuroimaging Initiative, et al. "Predicting conversion of mild cognitive impairments to alzheimer's disease and exploring impact of neuroimaging". In: *Graphs in Biomedical Image Analysis and Integrating Medical Imaging and Non-Imaging Modalities*. Springer, 2018, pp. 83–91.

[173] Karen Simonyan and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556* (2014).

[174] Stephen M Smith. "Fast robust automated brain extraction". In: *Human brain mapping* 17.3 (2002), pp. 143–155.

[175] Stephen M Smith et al. "Advances in functional and structural MR image analysis and implementation as FSL". In: *Neuroimage* 23 (2004), S208–S219.

[176] Nitish Srivastava et al. "Dropout: a simple way to prevent neural networks from overfitting". In: *The journal of machine learning research* 15.1 (2014), pp. 1929–1958.

[177] Heung-Il Suk and Dinggang Shen. "Deep ensemble sparse regression network for Alzheimer's disease diagnosis". In: *International Workshop on Machine Learning in Medical Imaging*. Springer. 2016, pp. 113–121.

[178] Heung-Il Suk and Dinggang Shen. "Deep learning-based feature representation for AD/MCI classification". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2013, pp. 583–590.

[179] Heung-Il Suk et al. "Deep ensemble learning of sparse regression models for brain disease diagnosis". In: *Medical image analysis* 37 (2017), pp. 101–113.

[180] Heung-Il Suk et al. "Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis". In: *NeuroImage* 101 (2014), pp. 569–582.

[181] Heung-Il Suk et al. "Latent feature representation with stacked auto-encoder for AD/MCI diagnosis". In: *Brain Structure and Function* 220.2 (2015), pp. 841–859.

[182] Heung-Il Suk et al. "State-space model with deep learning for functional dynamics estimation in resting-state fMRI". In: *NeuroImage* 129 (2016), pp. 292–307.

[183] Christian Szegedy et al. "Going deeper with convolutions". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 1–9.

[184] Arwa Mohammed Taqi et al. "The impact of multi-optimizers and data augmentation on TensorFlow convolutional neural network performance". In: *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*. IEEE. 2018, pp. 140–145.

[185] Paul M Thompson et al. "Mapping cortical change in Alzheimer's disease, brain development, and schizophrenia". In: *Neuroimage* 23 (2004), S2–S18.

[186] Tong Tong et al. “Multiple instance learning for classification of dementia in brain MRI”. In: *Medical image analysis* 18.5 (2014), pp. 808–818.

[187] David S Tuch et al. “High angular resolution diffusion imaging reveals intravoxel white matter fiber heterogeneity”. In: *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* 48.4 (2002), pp. 577–582.

[188] Ahsan Bin Tufail, Yong-Kui Ma, and Qiu-Na Zhang. “Binary classification of Alzheimer’s disease using sMRI imaging modality and deep learning”. In: *Journal of digital imaging* 33.5 (2020), pp. 1073–1090.

[189] Nicholas J Tustison et al. “N4ITK: improved N3 bias correction”. In: *IEEE transactions on medical imaging* 29.6 (2010), pp. 1310–1320.

[190] Nathalie Tzourio-Mazoyer et al. “Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain”. In: *Neuroimage* 15.1 (2002), pp. 273–289.

[191] Aly Valliani and Ameet Soni. “Deep Residual Nets for Improved Alzheimer’s Diagnosis.” In: *BCB*. 2017, p. 615.

[192] LA Van De Pol et al. “Baseline predictors of rates of hippocampal atrophy in mild cognitive impairment”. In: *Neurology* 69.15 (2007), pp. 1491–1497.

[193] Nicolas Villain et al. “Relationships between hippocampal atrophy, white matter disruption, and gray matter hypometabolism in Alzheimer’s disease”. In: *Journal of Neuroscience* 28.24 (2008), pp. 6174–6181.

[194] Uro Vovk, Franjo Pernus, and Botjan Likar. “A review of methods for correction of intensity inhomogeneity in MRI”. In: *IEEE transactions on medical imaging* 26.3 (2007), pp. 405–421.

[195] Tien Duong Vu et al. “Multimodal learning using convolution neural network and Sparse Autoencoder”. In: *2017 IEEE International Conference on Big Data and Smart Computing (BigComp)*. IEEE. 2017, pp. 309–312.

[196] Tien-Duong Vu et al. “Non-white matter tissue extraction and deep convolutional neural network for Alzheimer’s disease detection”. In: *Soft Computing* 22.20 (2018), pp. 6825–6833.

[197] Hongfei Wang et al. “Ensemble of 3D densely connected convolutional network for diagnosis of mild cognitive impairment and Alzheimer’s disease”. In: *Neurocomputing* 333 (2019), pp. 145–156.

[198] Lei Wang et al. “Changes in hippocampal volume and shape across time distinguish dementia of the Alzheimer type from healthy aging”. In: *Neuroimage* 20.2 (2003), pp. 667–682.

[199] Shui-Hua Wang et al. “Classification of Alzheimer’s disease based on eight-layer convolutional neural network with leaky rectified linear unit and max pooling”. In: *Journal of medical systems* 42.5 (2018), p. 85.

[200] Shuqiang Wang et al. “Automatic Recognition of Mild Cognitive Impairment from MRI Images Using Expedited Convolutional Neural Networks”. In: *International Conference on Artificial Neural Networks*. Springer. 2017, pp. 373–380.

[201] Yan Wang et al. “A Novel Multimodal MRI Analysis for Alzheimer’s Disease Based on Convolutional Neural Network”. In: *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE. 2018, pp. 754–757.

[202] Junhao Wen et al. “Convolutional neural networks for classification of Alzheimer’s disease: Overview and reproducible evaluation”. In: *Medical image analysis* 63 (2020), p. 101694.

[203] C-F Westin et al. “Processing and visualization for diffusion tensor MRI”. In: *Medical image analysis* 6.2 (2002), pp. 93–108.

[204] Mette R Wiegell, Henrik BW Larsson, and Van J Wedeen. “Fiber crossing in human brain depicted with diffusion tensor MR imaging”. In: *Radiology* 217.3 (2000), pp. 897–903.

[205] Robin Wolz et al. “Multi-method analysis of MRI images in early diagnostics of Alzheimer’s disease”. In: *PloS one* 6.10 (2011), e25446.

[206] IC Wright et al. “A voxel-based method for the statistical analysis of gray and white matter density applied to schizophrenia”. In: *Neuroimage* 2.4 (1995), pp. 244–252.

[207] Y Xu et al. “Usefulness of MRI measures of entorhinal cortex versus hippocampus in AD”. In: *Neurology* 54.9 (2000), pp. 1760–1767.

[208] Jason Yosinski et al. “How transferable are features in deep neural networks?” In: *Advances in neural information processing systems*. 2014, pp. 3320–3328.

[209] Lei Yuan et al. “Multi-source feature learning for joint analysis of incomplete multiple heterogeneous neuroimaging data”. In: *NeuroImage* 61.3 (2012), pp. 622–632.

[210] Chuanchuan Zheng et al. “Early Diagnosis of Alzheimer’s Disease by Ensemble Deep Learning Using FDG-PET”. In: *International Conference on Intelligent Science and Big Data Engineering*. Springer. 2018, pp. 614–622.

[211] Xiaofeng Zhu, Heung-Il Suk, and Dinggang Shen. “A novel matrix-similarity based loss function for joint regression and classification in AD diagnosis”. In: *NeuroImage* 100 (2014), pp. 91–105.