Politecnico di Torino

Master's Degree in Biomedical Engineering

Development and evaluation of a deep learning approach for the assessment of healthy aging and neurodegenerative-induced brain structural changes

Supervisors

Prof. Filippo MOLINARI

Prof. Hugo FERREIRA

Candidate

Edoardo FILIPPI

July 2023

Abstract

This study investigates the feasibility of using learning models as alternative methods for brain analysis, particularly in evaluating damage from neurodegenerative diseases. The main objective is to compare their effectiveness and assess the impact of results on future feature utilization. A comparative analysis highlights the strengths and weaknesses of deep learning versus traditional methods. Additionally, a practical experiment is conducted using both approaches to test the software in a real-world setting. This research aims to advance brain analysis techniques and provide insights for accurate assessment of neurodegenerative disease-induced damage.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Dementia is estimated to be very common in the modern world, in 2010, more than 35 million people worldwide were estimated to be living with it. It is a progressive global cognitive impairment syndrome. It does not always evolve in the same manner, in fact while some people with mild cognitive impairment (MCI) will progress to dementia, some will recover or remain stable. Finding good predictors of dementia is thus of great interest.[1]

AD accounts for 60% to 70% of cases of progressive cognitive impairment in elderly patients. The prevalence of AD doubles every 5 years after the age of 60 increasing from 1% at the age of 60 to 40% at the age of 85 or older. The disease is more common in women, with a ratio of 1.2 to 1.5. The population of patients with AD is growing rapidly, with the associated growth of direct and indirect costs of patient care to hospitals and society. [2]

The onset of the symptoms usually appears after significant damage to the brain tissue developed, and the amount of damage that corresponds to the onset of the illness varies between individuals. Modern tools for the analysis of brain medical images can be of great help in the identification of the signs of a developing dementia. But conventional methods have a very long processing time. The use of technologies

based on deep learning can tackle this issue by offering alternatives that drastically reduce processing time, making the process of image analysis more effective.

This work focuses on testing the standard methods against deep learning methods to assess the performances of the two methods when dealing with a clinical problem.

The process to do this starts by assessing the state of the art in the field with a bibliographic research, the choice of the most promising methods to then test against an established software for brain image analysis. Many subjects' MRI volumes collected by different available datasets will be processed with both methods and the results will be used to train a classifier to identify Alzheimer's disease, to assess whether the performance of it is impacted by the method used to process images or not, and if yes in which ways.

The text starts by giving an overview of the anatomy of the brain and the regions affected by neurodegenerative diseases, it then shortly describes the technologies used to acquire images and the terminology and technology involved in their processing for the application described. After this an overview of the state of the art in the software commonly used for this task, the description of the recent trends and experimental tools are given.

In chapter 2 the methods used to perform the bibliography research and the description of the software that will be used in this work is then explained. Furthermore, the methods used to analyse the results and to train the classifier as well as the description of the data used is given.

Chapter 3 describes the results obtained in the study. Portraying the comparison of the output data of the established software and the experimental alternative, as well as the performances obtained by the classifier in both cases. The results are then commented in the conclusion.

1.1 Anatomy

1.1.1 General description

The nervous system is divided into two major sections: the central nervous system, consisting of the brain and the spinal cord, and the peripheral nervous system, which includes the nerve tracts connecting the rest of the body to the central nervous system. The brain itself can be divided into three main divisions: the cerebrum, the brainstem, and the cerebellum, all of which are housed within the cranium.

The cerebellum is a structure in the brain that consists of two hemispheres separated by the longitudinal fissure. These hemispheres are interconnected primarily by the corpus callosum. Each hemisphere of the cerebellum has three surfaces and three poles. By identifying the major sulci on the surface of the cerebellum, it becomes possible to distinguish the cerebral lobes: frontal, parietal, temporal, occipital, and the insular lobe, which is hidden within the depths of the Sylvian fissure. The cerebral lobes derive their names from the overlying bones of the skull.

The frontal lobe of the cerebelum, which corresponds to the frontal bone, is located beneath its vertical portion on the orbital roof. Above the tentorium is situated the occipital pole, which is associated with the occipital bone. In the middle cranial fossa is located the temporal lobe, and it is related to the posterior wall of the orbit. The left and right hemispheres are connected by the corpus callosum.

Based on morpho-functional data, the brain can be further categorized into specific regions:

- Telencephalon: This region includes the cerebral hemispheres, also known as the cerebrum. The cerebrum is the largest and most evolutionarily recent part of the brain and contains the cortex, which plays a crucial role in higher cognitive functions.

- Diencephalon: The diencephalon refers to the area "between" the

brain. It consists of structures such as the thalamus, metathalamus, and hypothalamus. Together with the telencephalon, the diencephalon forms the prosencephalon.

- Mesencephalon: The mesencephalon is formed by structures like the tectal plate, tegmentum, and cerebral peduncles.

- Rhombencephalon: This division includes the metencephalon (pons and cerebellum) and the myelencephalon (medulla oblongata).

The surface of the brain exhibits a high variability, as several distinct patterns of the sulci have been described. The variability of the sulci and gyri can be found not only in different individuals but also in the same brain, between the two hemispheres, which show some anatomical and physiological peculiarities. [3]

The brain is covered by meninges. Which are three membrane layers that cover and protect the brain and the nervous system. They also provide a support system for blood vessels, nerves, lymphatics and the cerebrospinal fluid that surrounds the nervous system.

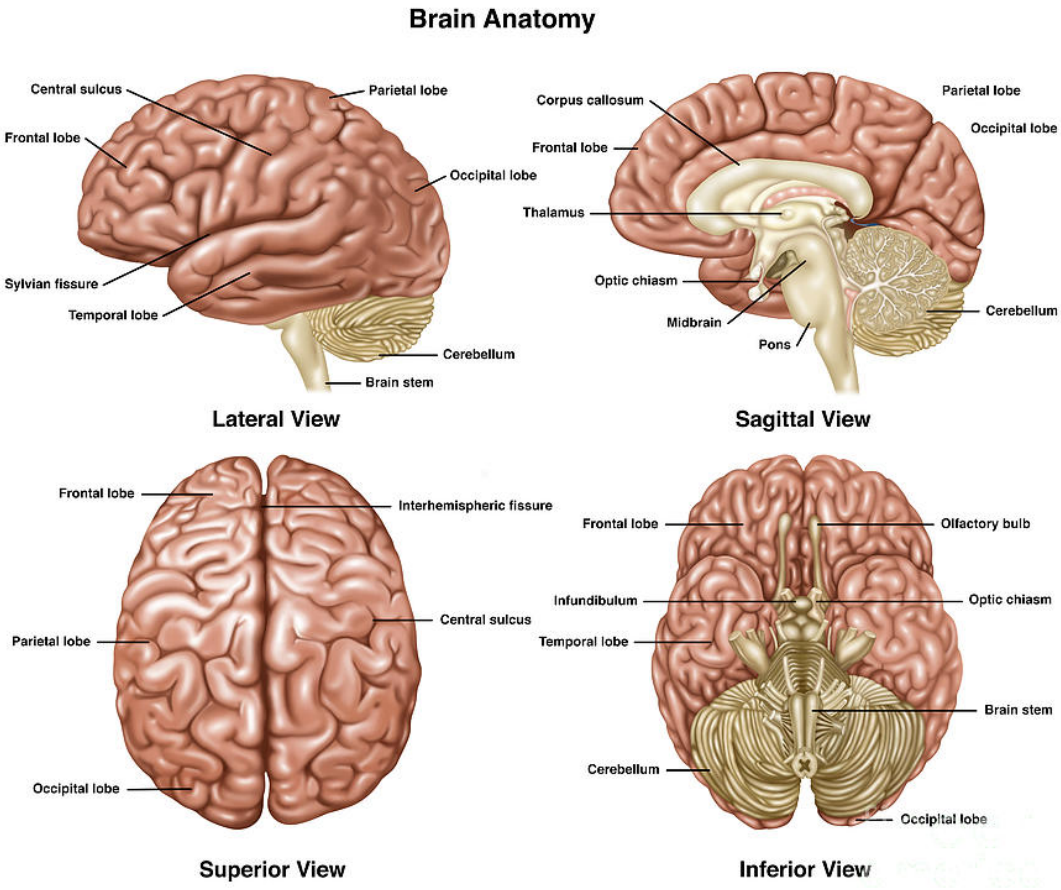


Figure 1.1: Human brain anatomy chart from [3]

1.1.2 Temporal lobe

Some regions are more affected than others by the ageing process and neurodegenerative diseases such as Alzheimer, specifically, the most affected are the temporal regions, which are responsible for memory loss and reduced brain functions. The temporal region is divided as follows.

Ventricles

Ventricles are fluid-filled cavities in the brain that contain cerebrospinal fluid (CSF). There are two lateral ventricles located within the cerebral hemispheres, and they communicate with the third ventricle through the interventricular foramen.

The lateral ventricles have a C-shaped structure within the hemispheres, consisting of a frontal horn, a body, or cell media, an occipital horn, and a temporal horn that extends into the temporal lobe. The posterior part of the third ventricle opens into the aqueduct of Sylvius, a single midline canal located in the mesencephalon (midbrain). Through the aqueduct, the CSF flows into the fourth ventricle, which is located in the rhombencephalon (hindbrain).

At the level of the fourth ventricle, the CSF exits the brain and enters the cisterns of the posterior cranial fossa as well as the spinal cord through three openings. The production of CSF occurs in the ventricles by the choroid plexus, which is located within them.

Hippocampus

The belt of the limbic lobe is situated beneath the corpus callosum. The limbic system is formed by the hippocampal formation and amygdala, septum pellucidum, hypothalamus, as well as the central olfactory system. The hippocampus is a large C-shaped structure located in the medial wall of the temporal lobe. It is a vital component of the medial temporal lobe memory system, and plays a key role in memory mechanisms. The parahippocampal gyrus surrounds the hippocampus.

Amygdala

The Amygdala is an almond-shaped structure formed by a group of different nuclei. It is located in the dorsomedial portion of the temporal lobe. The cortical and medial nuclei are olfactory centres, and the basal, lateral, and central nuclei have limbic functions. [4]

1.2 Clinical context

1.2.1 Dementia

Dementia typically develops over a span of several years, during which it is believed that individuals may be asymptomatic while pathological changes are accumulating in the brain. Initially, individuals and their relatives may notice subtle impairments in recent memory, and as time progresses, other cognitive domains become affected. Difficulties in planning and executing complex tasks gradually become more apparent. It is important to note that the specific trajectory and progression of dementia can vary between individuals and depend on the underlying cause of the condition.

The standard assessment of dementia involves multiple components to ensure a comprehensive evaluation. These typically include:

- History and Clinical Examination: The healthcare professional will gather information about the individual's medical history, including any symptoms or changes in cognition. A thorough clinical examination will be conducted to assess neurological function and overall health.

- Laboratory Tests: Laboratory tests, such as thyroid stimulating hormone (TSH), serum folic acid, serum vitamin B12, and blood count, may be performed to identify any underlying medical conditions that could contribute to cognitive impairment.

- Informant Interview: Speaking with a relative or informant who knows the individual well can provide valuable insights into changes in behaviour, cognition, and daily functioning.

- Neuroradiological Evaluation: Neuroimaging techniques such as magnetic resonance imaging (MRI) or computed tomography (CT) scans may be used to assess the structure and function of the brain, ruling out other potential causes of cognitive decline.

Before a dementia diagnosis is made, it is crucial to exclude or address any other physical or mental disorders that may be contributing to the cognitive impairment. These could include conditions such as depression, medication side effects, or metabolic abnormalities.

During the neurological examination, a comprehensive assessment of major cognitive domains is conducted. This typically includes evaluating memory function, executive functions (such as planning and problem-solving), language abilities, attention, and visuospatial skills. By assessing these cognitive domains, healthcare professionals can gain a better understanding of the specific cognitive deficits present and their impact on daily functioning.

A neuroradiological examination is commonly recommended in recent consensus guidelines for the assessment of dementia. While many tests are typically conducted after a cognitive deficit has been identified, individuals who have abnormalities detected through brain imaging performed for other reasons may subsequently be evaluated for cognitive deficits.

It is important to note that currently, there is no known cure for dementia. However, certain treatments can help to slow the cognitive and functional decline or alleviate associated behavioural and psychiatric symptoms of dementia. Early diagnosis of dementia carries numerous benefits, including the ability to plan and prepare for the future, avoiding inappropriate hospitalizations, and making use of emerging interventions aimed at delaying or preventing the progression of more severe stages of the disease.

1.2.2 AD

Alzheimer was first described at the beginning of the 20th century by Alois Alzheimer, a German psychiatrist who presented the case of a woman of relatively young age (51 years) who presented a rapidly deteriorating memory and other psychiatric disturbances. A new pathological finding, the neurofibrillary tangle, made this condition unique. At this point, it was split into two clinical conditions depending on the age of onset, and Alzheimer's disease (AD) was a term reserved for presenile dementia affecting individuals younger than 65 years of age. AD is now generally recognized as a single entity with a prevalence that increases sharply after age 65. [2]

Several risk factors have been identified in epidemiologic studies. The most potent one identified being the presence of the APOE $\epsilon 4$ allele. The lifetime risk for individuals with the allele is at least three times higher compared to the ones without. Multiple other risk factors have been identified such as head injury, low serum levels of folate and of vitamin B12, elevated plasma and total homocysteine levels, and family history of AD or dementia, as well as fewer years of formal education and in general less cognitively stimulant lives. [5]

Available evidence suggests that mild and even severe dementia is underdiagnosed in clinical practice. An example of a method used for this end is the Mini-Mental-State examination, which, despite having a good specify, has a low sensitivity, indicating that the test itself will leave a substantial portion of the cases of dementia undetected.

AD is assumed to be a slow process and discernible in older people. The symptoms are not visible for years and are hard to detect. But detection of AD in the early stages is essential before starting any clinical procedures. MCI is an initial stage of AD and might convert to AD. So identification of MCI is of great significance.[6]

1.2.3 Brain maps

In regard to the human brain, no reference standards currently exist to quantify individual differences in neuroimaging over time, but several studies were conducted. One example is a study which analyzed over 100'000 MRI scans, and used centile scores and trajectories to quantify brain structural changes. The approach used was based on GLAMMS modelling and exploited the great scale of data available to optimize model selection and estimate the non-liear trends. Centile scores across a range of conditions were computed to benchmark each individual scan in the context of normative age-related trends.

The studies identified previously unknown neurodevelopmental milestones and provides open science resources for standardized assessment of MRI data. This work demonstrates the feasibility of building brain charts to benchmark individual differences on a global scale throughout life.

These discovered highly significant differences in central scores across large groups of cases diagnosed with multiple disorders, with effect sizes ranging from medium to large. The greatest overall difference was found in Alzheimers's disease, with a maximum difference localized in the grey matter volume in biologically female patients. Clinical case-control differences generally followed the same trend in cortical thickness and surface area. Schizophrenia ranked third behind AD and MCI [7].

1.2.4 Alzheimer's modelling

Recently, similar normative models have been developed for specific diseases such as Alzheimer's as well. Notable is an attempt to do so using more than 5000 images processed with the trained deep learning model AssemblyNet. This longitudinal study uses numerous MRI scans for multiple Datasets such as C-MIND, NDAR, ABIDE and ADNI.

The implemented process is the following. After collecting the images, the first step was denoising the images and correcting inhomogeneity.

Affine registration into the Montreal Neurological Institute (MNI) space was performed using ANTS. The volumes were then processed using the 250 U-Nets implemented in AssemblyNet. The quality control procedure included a before and after visual inspection to determine if the subject was classified as an outlier and to see if there was a segmentation failure. In this case, it was deleted.

To compensate for the variability introduced by head size difference models were estimated on normalized volumes, different models were tested for the trajectory, and a brain region was considered statistically different for Alzheimer's patients if the 95% confidence intervals no longer overlapped.

This led to the identification of 19 brain structures that significantly diverged during the lifespan between Alzheimer's and healthy ageing models. In the image below there are the regions which show the most differences between healthy and not healthy subjects [8].

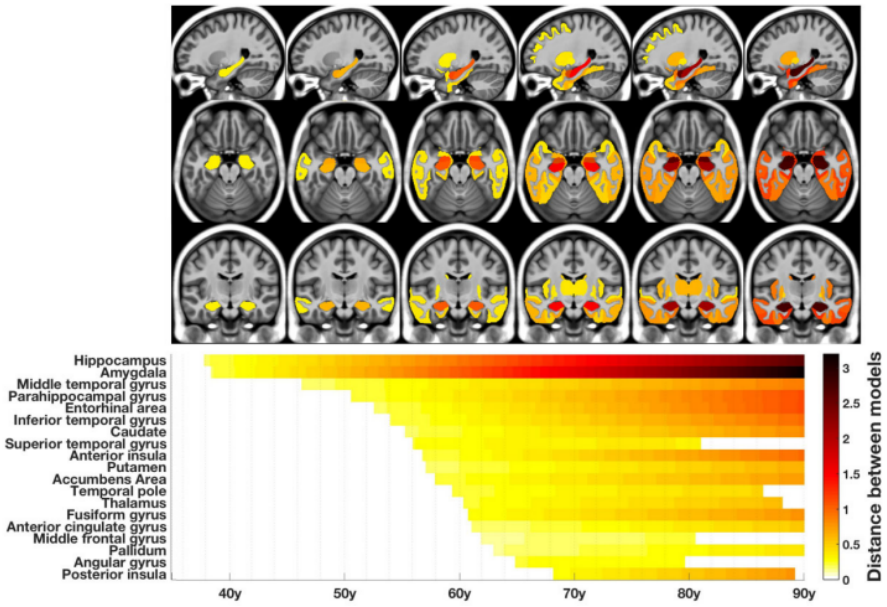


Figure 1.2: Results of the study [8]

1.2.5 Reserve

The cognitive reserve construct seeks to explain the brain's ability to compensate for degeneration caused by age or neuropathology. It has been proposed to explain the observed discrepancy between the degree of brain injury or pathology and its clinical manifestation. Individual differences in cognitive processes or neural networks are assumed to exist, which allow some people to better compensate for age-related degeneration or neurological disease

Two models of reserve have been developed to characterize this ability, the passive and the active models. In the passive model, reserve is mediated through anatomical substrate characteristics such as brain size and number of neurons and synapses, This has been proved to not be directly correlated to symptoms manifestation[5]. Active models imply that when there is damage to the brain tissue, an active and efficient effort from the brain to compensate for the injury using pre-existing compensatory processes is going on. The ability to do this is different in every individual.

1.2.6 Reserve models

It follows the definition of two models belonging to the categories of active and passive models.

1.2.7 Brain reserve

Brain reserve [9] is an example of a passive model, where reserve derives from brain size or neuronal count. The hypothesis is that larger brains can sustain more damage before clinical deficit emerges because sufficient neural substrate remains to support the normal function. This approach is a threshold model. The model recognizes that there are individual differences in brain reserve capacity. Once brain reserve capacity is depleted past some fixed critical threshold, specific clinical or functional deficits emerge. [10] This measure was found to be not related with the prediction of the cognitive functions and the onset of dementia in healthy subjects. [11]

1.2.8 Cognitive reserve

Cognitive reserve is the definition of the brain's ability to compensate for the degeneration caused by age or neuropathology. It is defined in the article by Stern in 2009 [12] as a method to explain the discrepancy between the degree of brain injury and its clinical manifestation. Cognitive reserve is not fixed throughout all stages of life but continues to evolve through experience [13]. Studies suggest that this is important in the development of Alzheimer's symptoms, as individuals with high CR manifest them when the damage is already more consistent[14].

There is no standard for assessing it, but there are different methods, most of them are fairly simple and not reliable enough. The evaluation has some issues in patients with a high and low CR. For example, if CR is high, impairment is not detected even if it is present, similar issues are present for low CR.

A review performed on the methods to measure CR identified some of the most reliable proposed criteria, they vary between each other

but have some similarities. The studies were performed in Europe and Australia. In all the frameworks, variables such as education and occupation with cognitive stimulant activities were also considered. In all studies, the participation in each variable is evaluated in each stage of life but uses different methods. The conclusion is that right now still no consensus on the measurements of cognitive reserve has been reached, and some methods, which use different approaches, can be identified. [15]

1.2.9 Ageing

Population ageing is rapidly accelerating worldwide, and this has profound implications for the planning and delivery of health and social care.

The ageing process is by itself a decrease in physiological reserves which can still support acceptable functioning in the steady state, but cannot adapt to any additional or physiological stress. Successful ageing depends on the homeostatic reserve of different physiological systems. The clinical condition of frailty is the pathologic expression of the normal ageing process. [16]

Frailty is a state of vulnerability to poor resolution of homoeostasis after a stressor event. This condition is the consequence of an age related decline of multiple physiological systems, which results in a very high sensibility to external stressor events and the vulnerability to sudden health status changes as a consequence. It is estimated that a quarter to half of people over 85 years are frail. [17]

The term cognitive frailty was first used in 2006 to indicate a state of cognitive vulnerability in MCI and other similar entities exposed to vascular risk, with a subsequently increased progression to dementia.[18]

Frail brain

The effects of ageing can be seen on the structural as well as physiological characteristics of the brain. The overall loss of the majority of regions is minimal, but the altered synaptic functions affect disproportionally neurons with a very high metabolic demand. Microglial cells also characterize the structural and functional changes in the ageing brains.

There is accumulating evidence from observational studies to support a temporal association between frailty, cognitive impairment, and dementia. [17]

Other frail systems

There are other systems affected by the frailty status of course, but they are of limited interest in this work. Some examples of frail body characteristics can be seen in the immune system, where the ageing is characterized by a decline in stem cells, alteration in T-lymphocyte production, blunting of the B-cell led antibody response and reduced phagocytic activity of neutrophils, macrophages, and natural killer cells.

The effects are clear in the skeletal muscle as well, where sarcopenia is considered a key component of frailty. [17]

1.2.10 Cognitive scores

MMSE

The Mini-Mental State Examination (MMSE) is the best-known and the most often used short screening tool for providing an overall measure of cognitive impairment in clinical, research and community settings.

This is a 30-question cognitive assessment that measures various aspects of cognitive functioning, such as attention, orientation, memory, registration, recall, calculation, language, and the ability to draw a complex polygon. Originally, it was not designed to detect early-stage dementia, differentiate between different types of dementia, or predict long-term dementia development. However, the MMSE offers several advantages, including its quick administration, availability of translations in multiple languages, and wide acceptance among healthcare professionals and researchers as a diagnostic tool. In this test, the presence of cognitive decline is determined based on the total score. Typically, a cut-off score of 23/24 is used to identify individuals with suspected cognitive impairment or dementia.

This threshold is not universally applicable as it has been proved that sociocultural variables, age, and education, among other factors can influence individual scores. Therefore, local standards must be developed for each population. [1]

Frailty indexes

Frailty is a multidimensional geriatric syndrome characterized by a decline in physical and cognitive reserve, it has received increased attention in the last years and multiple indexes have been developed. In detail, two major approaches are common[19]:

- **Frailty phenotype:** assesses the individual frailty level based on the presence of five criteria: an individual is considered frail if they present at least three of these five and pre-frail if they present at least one. Robust people do not present any of these characteristics. [20]

- **Frailty index:** is also called cumulative deficit model, and it assumes that the more deficits a person has, the more likely it is that the person is frail. Defects are comprised of symptoms, signs, disability disease, and lab experiments, ranging in severity and carrying equal weights. The index is often expressed as a ratio between the number of deficits present and the total number of deficits assessed.[21]

Scores of reserve

After the concept of brain reserve proved to be a non-efficient mean of discovering how clinical damage results in a clinical manifestation, cognitive reserve was the focus of many research endeavours. The development of an efficient index for its evaluation became thus an investigation topic of interest. Multiple methods have been developed, specifically the few that can be identified as being viable options. They vary in length and complexity and in the included variables. Most of them consider the relationship with external factors as well. There is no standard, since CR is a relatively new concept in literature. [15]

1.3 Technical context

1.3.1 MRI

MRI is a medical imaging technique used in radiology to form pictures of the anatomy and the physiological processes of the body. MRI scanners use strong magnetic fields. It is a medical application of nuclear magnetic resonance (MNR) which is also used in other fields, an example being the NMR spectroscopy.

MRI works by exploiting the magnetization properties of atomic nuclei, a powerful external magnetic field is employed to align the protons that are normally randomly oriented and subsequently, it is disturbed and perturbed, using an external RF signal. The nuclei then return to their original alignment through various processes, and this emits an RF signal that can be measured and processed using the Fourier transform. [22]

The frequency information, contained in the signal from each location in the image plane, is mapped to corresponding intensity levels, which are then displayed as shades of gray in a matrix arrangement of pixels. By varying the sequence of applied and collected RF pulses, different types of images are created.

There are multiple MRI sequences that can be used, the most common ones belong to the category of spin echo, which are T1 and T2 weighted images. Some other categories are: gradient echo, inversion recovery, diffusion-weighted sequences.[23]

T1

T1-weighted imaging (T1WI) is a basic MRI pulse sequence that reveals differences in tissue T1 relaxation times. It relies on the longitudinal relaxation of a tissue's net magnetization vector. Spins aligned in an external magnetic field (B0) are shifted to the transverse plane with a radio frequency (RF) pulse, and they gradually return to equilibrium. Tissues have varying rates of realignment, indicated by their T1 values.

Fat quickly realigns, appearing bright on T1WI, while water realigns slowly, resulting in a low signal and dark appearance.[24]

Other sequences

The T1 sequence images will be the ones used in this work, some other sequences and technologies are:

- T2 weighted image (T2WI): is one of the basic pulse sequences on MRI. The sequence weighting highlights differences on the T2 relaxation time of tissues.

- Diffusion-weighted: measures the diffusion of water molecules in biological tissues.

- Gradient echo: it does not use a 180 degrees RF pulse to make the spins of particles coherent. Instead, it uses magnetic gradients to manipulate the spins, allowing the spins to dephase and rephase when required.

- Inversion recovery: it is an MRI sequence that provides high contrast between tissue and lesion.

- Functional MRI (fMRI): measures signal changes in the brain that are due to changing neural activity.

1.3.2 Registration

MRI scans from multiple individuals will vary greatly due to differences in slice orientation and brain features (i.e. brain size and shape varies across individuals). Therefore, it is generally useful to normalize scans to a standard template. Normalization is the process of translating, rotating, scaling, and maybe warping a brain to roughly match a standard template image. After normalization, it may be informative to report locations using stereotaxic ("Talairach") coordinates. This format uses three numbers (X,Y,Z) to describe the distance from the Anterior Commissure (the 'origin' of Talairach space). The X,Y,Z dimensions refer to left-right, posterior-anterior, and ventral-dorsal respectively. [25]

- **Linear registration (Affine registration):** Meaning that it will translate, rotate, zoom and shear one image to match it with another. Affine transformations have twelve degrees of freedom. These are also called linear transformations because a transformation applied in one direction along an axis is accompanied by a transformation of equal magnitude in the opposite direction. Sometimes the differences between subjects are such that the linear transformation is not sufficient to achieve good registration. The local deformations permitted by a non-linear method may then do a better job.

- **Non-linear registration:** They are not subject to the constraints mentioned above. For example, a nonlinear transformation can enlarge the image in one direction while shrinking it in the other direction.[26]

Talairach coordinates

Talairach coordinates, also known as Talairach space, were developed starting from 1967 by Talairach as a system for identifying small regions of the brain during epilepsy surgery. In the Talairach coordinate system, brain regions are labelled by their Brodmann numbers. [28]

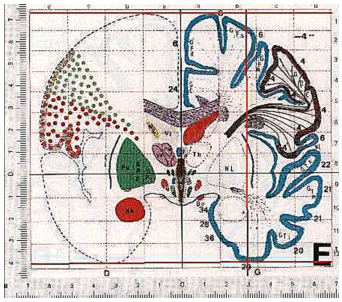


Figure 1.3: Talairach Atlas from [27]

It is a 3-dimensional coordinate system (known as an 'atlas') of the human brain, which is used to map the location of brain structures independent of individual differences in the size and overall shape of the brain, a sketch can be seen in image 1.3. It is still common to use Talairach coordinates in functional brain imaging studies and to target transcranial stimulation of brain regions. [29]

A major caveat of using the Talairach coordinate system is that the coordinates were based on a single, post-mortem case study rather than an average of multiple brains. [30]

MNI

The MNI defined a new standard brain by using a large series of MRI scans on normal controls. The goal was to define a brain that is more representative of the population. A new template that was approximately matched to the Talairach brain in a two-stage procedure was created. First, 250 normal MRI scans were taken, and various landmarks were manually defined, in order to identify a line very similar to the AC-PC line and the edges of the brain. Each brain was scaled to match the landmarks to equivalent positions on the Talairach atlas. This resulted in the 250 Atlas. Subsequently, an extra 55 images were then taken and registered to the 250 Atlas using an automatic linear registration method. The images were manually registered to the 250 Atlas to create the MNI 305 Atlas. The MNI305 was the first MNI

template. The current standard MNI template is the ICBM152, which is the average of 152 normal MRI scans that have been matched to the MNI305 using a 9-parameter affine transform. The International Consortium for Brain Mapping adopted this as their standard template. [31]

1.3.3 Segmentation

Image segmentation is one of the most important tasks in medical image analysis, and is often the first and the most critical step in many clinical applications. In brain MRI analysis, image segmentation is commonly used for measuring and visualizing the brain's anatomical structures, analysing brain changes, for delineating pathological regions, and for surgical planning and image-guided interventions. Clinical imaging studies encompassing large cohorts of patients and control commonly apply automatic image segmentation tools. It involves the identification of different brain tissue without giving any information on its functions, as it can be seen in figure 1.4. [32]

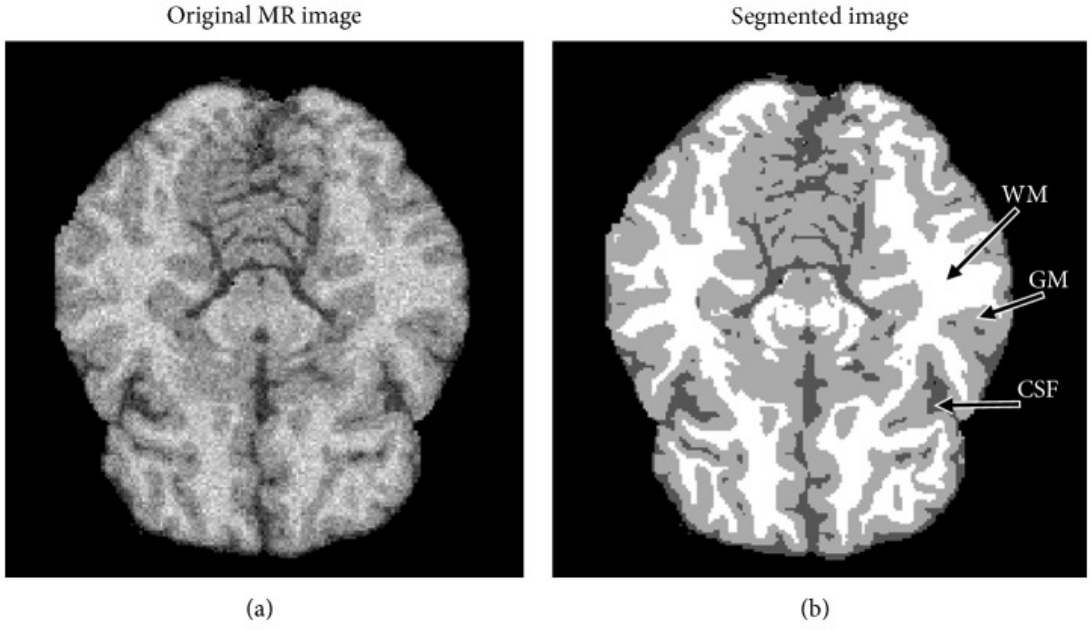


Figure 1.4: Example of the result of a segmentation process

1.3.4 Atlases

An atlas of the brain defines the shape and location of brain regions in a common coordinate space. They provide spatial reference systems for neuroscience that allow navigation, characterization, and analysis of information based on anatomical location. Some examples are:

- Desikan-Killiany Atlas (Cortical) 1.5: 34 ROIs in each hemisphere, used by Freesurfer [33]

- ASEG (subcortical): 37 regions, used by FreeSurfer. [34]

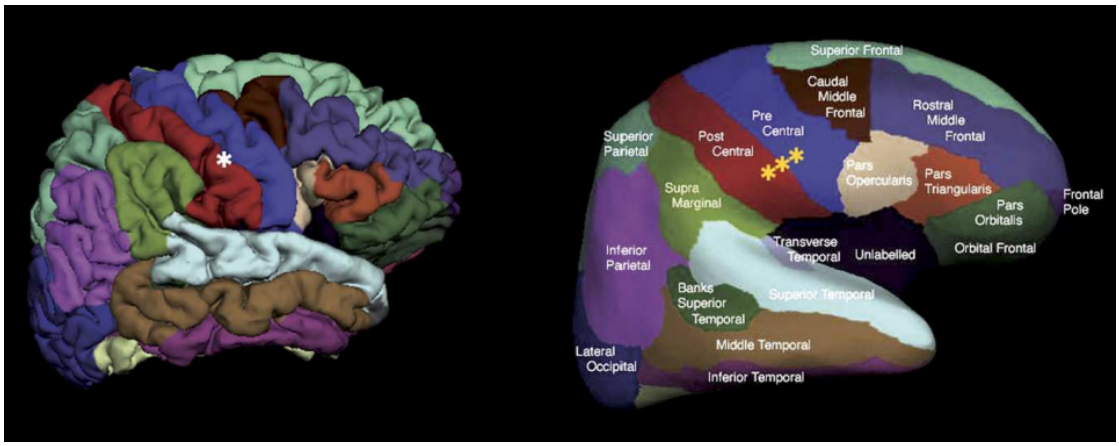


Figure 1.5: DKT Atlas regions of interest (Deflated left, Inflated right)

1.3.5 Parcellation

The difference between the segmentation and parcellation process is that the former only recognizes the difference between type of tissues, assigning to each voxel a class label to identify it[35]. Parcellation is a more advanced process. To date, several sophisticated software packages have been developed to achieve this goal. One example is the Integrated Registration and Segmentation Tool (FIRST) of the FSL software library, and FreeSurfer, the software used in this work, which will be described in detail afterwards.

Parcellation means defining distinct partitions in the brain, be they areas or networks, that comprise multiple discontinuous but closely interacting regions. The brain is parcellated according to atlases, and classified according to their function. [36]

The exploration of the complex structures and functional networks in which the brain is segregated into has been conducted using many techniques including clustering methods such as Gaussian mixtures models, meta-analytic connectivity methods and edge detection methods, among may others. The primary goal of the parcellation process is to reveal brain organization. But cognitive, developmental and clinical research have frequently used parcellation as a means to identify differences in the functional organization of the brain based on condition, age, and presence of psychopathology. Brain parcellations are used to extract data from a set of parcels, or brain regions, that comprise the pre-identified functional network. The widespread adoption of these parcellations in cognitive neuroscience has allowed extensive exploration of individual differences in functional brain organization.

1.4 State of the art

In this chapter the state of the art in brain parcellation will be presented, starting from well established softwares and moving to research and experimental methods.

1.4.1 FreeSurfer

Introduction and description

FreeSurfer is an open source package for the analysis and visualization of structural, functional, and diffusion neuroimaging data from cross-sectional and longitudinal studies. It is developed by the Laboratory for Computational Neuroimaging at the Athinoula A. Martinos Center for Biomedical Imaging.

It implements a full processing stream for MR imaging data that involves skull-stripping, bias field correction, registration, and anatomical segmentation as well as cortical surface reconstruction, registration, and parcellation. FreeSurfer also includes fMRI and diffusion tractography toolboxes, a robust visualization interface, utilities for statistical group analysis. FreeSurfer is the structural MRI analysis software of choice for the Human Connectome Project.

Over the years many functionalities have been added, recently a liberal open source licence that allows great freedom in the use of the source code was adopted. Other updates included tools for accurate cross-modal intra-subject registration, combined volume and surface cross-subject registration, probabilistic estimation of cytoarchitectonic boundaries, automated tractography, and longitudinal analysis.

It is widely adopted, and it has been used to improve the understanding of an array of neurological disorders. It has a great interoperability with the software FSL, which will be shortly described later. [37]

Output

The output folder of FreeSurfer is standardized and contains many files created during the processing phase, it has a dimension in the order of the hundreds of megabytes for a single subject's MRI. It contains the following folders:

- **label**: this directory stores various label files that represent different anatomical regions or structures in the brain. These labels can be used for region of interest (ROI) analysis or to define specific areas for further processing.
- **mri**: this directory contains the preprocessed MRI data and intermediate results generated during the processing steps. It may include files such as the original DICOM images, the brain-extracted volume, skull-stripped volume, and various transformation matrices.
  - **orig**: located inside the "mri" folder, contains the original file as it was given as input.
- **scripts**: FreeSurfer generates a "scripts" directory that contains log files and scripts used during the processing steps. These files can be helpful for reviewing the processing history and for troubleshooting any issues.
- **stats**: this directory contains statistical information and summary measures derived from the segmentation and parcellation results. It may include files like cortical thickness measurements, volume statistics for different brain regions, and other statistical summaries.
- **surf**: this directory contains the surface-based data, including the reconstructed cortical surface models (such as *lh.pial* and *rh.pial*) and other surface-related files.
- **tmp**: the temporary directory stores temporary files generated during the processing and can be safely deleted once the analysis is complete.

Freeview

Each neuroimaging software package has a data-viewer, or an application that allows the user to look at the data. AFNI, SPM and FSL all have data viewers which are very similar between each other and have the task of loading imaging data and viewing it in three dimensions.

The FreeSurfer viewer is called Freeview and can be easily launched by terminal, the interface can be seen in the image 1.6, it can read NIFTI images and FreeSurfer specific formats such as *.mgz* and *.inflated*.

In the listing below an example of the bash commands to visualize the results of a FreeSurfer pipeline on a subject's MRI it is reported.

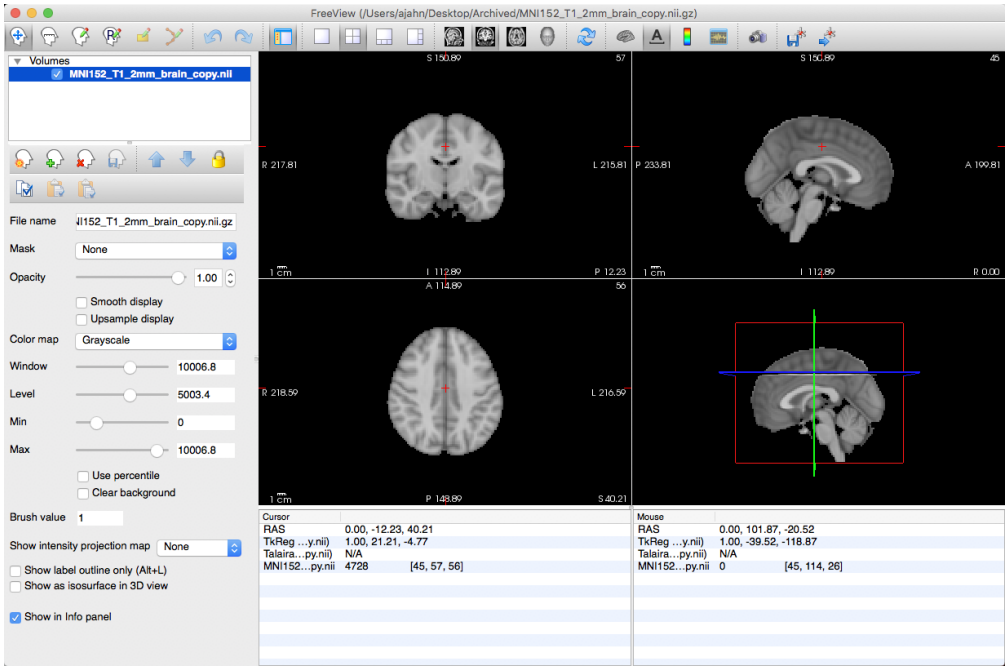


Figure 1.6: FreeView Interface

```
source \$FREESURFER\_HOME/SetUpFreeSurfer.sh

#example/mri/orig.mgz \
#example/mri/mask.mgz \
#example/mri/aparc.DKTatlas+aseg.deep.mgz

freeview -v \
example/mri/T1.mgz \
example/mri/wm.mgz \
example/mri/brainmask.mgz \
example/mri/aparc+aseg.mgz:colormap=lut:opacity=0.2 \
example/mri/aseg.mgz:colormap=lut:opacity=0.2 \
-f example/surf/lh.white:edgecolor=blue \
example/surf/lh.pial:edgecolor=red \
example/surf/rh.white:edgecolor=blue \
example/surf/rh.pial:edgecolor=red
```

Listing 1.1: example of bash code to setup freeview

1.4.2 Other softwares

SPM12

SPM (Statistical Parametric Mapping) is an fMRI analysis software package that is run in Matlab. In addition to fMRI analysis, SPM contains toolboxes for performing volume-based morphometry and effective connectivity.[38]

CAT12

CAT12 is a powerful suite of tools for morphometric analyses with an intuitive graphical user interface, but also usable as a shell script.[39]

FSL

FSL is a comprehensive library of analysis tools for FMRI, MRI and DTI brain imaging data. It runs on Apple and PCs (both Linux, and Windows via a Virtual Machine), and is very easy to install. Most of the tools can be run both from the command line and as GUIs. [40]

ANTs

ANTs is a software package for normalizing data to a template. Most of the templates provided on the ANTs website are in MNI space. [41]

1.4.3 Machine learning

Machine learning is a subset of artificial intelligence, and it is the process of training a computer to solve a problem given to it. In recent years, the application of ML in different fields to solve problems faster than before has gained significant interest due to the current availability of cheaper computing power and inexpensive memory.

Machine learning differentiates itself from deep learning by the input that it is given to it. In fact, while machine learning works with features extracted and chosen beforehand, and constructs an output based on that processed input, deep learning accepts raw data as input, this translates to it being more complex and less interpretable, but with a higher ability to generalize and less dependent on human decisions.

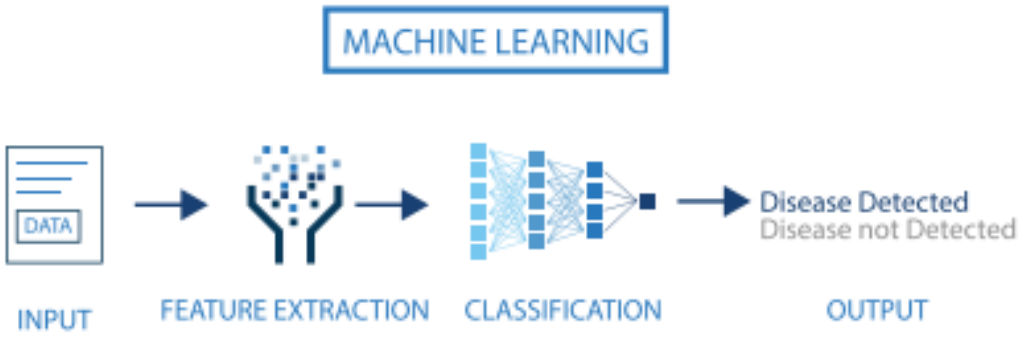


Figure 1.7: Description of machine learning

Machine learning has many branches that deal with different problems and types of data, for example it may deal with labelled data or unlabelled data and the goal may be to solve a classification problem, to predict a value through a regression, or to find unknown patterns in data. The main categories of machine learning are:

- **Unsupervised**: unsupervised learning works with unlabelled data, which means that one does not know the relationship between data, machine learning is used to find patterns in them. Some example of algorithms are k-means, DBSCAN or dimensionality reduction methods such as PCA.

- **Supervised**: it is the type used in this study, and it implies having data of which are know the relationship and the classification to obtain, this is a necessity and the algorithm will be trained on them before being applied to unknown data. The problem that a supervised algorithm deals with may be of classification, which means each element to a class, or of regression, which means predicting a value.

- **Reinforcement learning**: Reinforcement learning differs between the previous paradigms because in this case the goal is optimizing a "fitness" function. In the training phase the algorithm does not know the correct answer but knows if the prediction it made is good or not.

In this case machine learning will be used to deal with a classification problem. Some models that may be used in dealing with these type of problems are:

- **Logistic regression**: Despite being a regression, it belongs to the class of classification because it does not assign a value or a probability to an element but directly assigns a class to it. It is a binary classification algorithm because it can only work with two classes. The name logistic derives from the log operation, on which the regression is based.

- **Decision trees**: Decision trees are one of the conceptually simplest machine learning algorithm. They have many nodes in which a decision is taken on a feature of the input. The last node in which a class is assigned is called leaf. The upside of this method are its interpretability and the fact that it works with little data. Some downsides are that it overfits very easily and it is very unstable.

- **SVM**: SVM maps training examples to points in space to maximize the width of the gap between the two categories. It is one of the most robust prediction methods. They can perform linear and nonlinear classification using what is known as "kernel trick", implicitly mapping their inputs into high-dimensional feature spaces.

- **Neural Networks**: Neural networks are the most known method of machine learning and deep learning. They are based on perceptrons which theoretically mimic the behaviour of a biological neuron, so have many inputs and many outputs. Each neuron has a bias and an activation function, the combination of many in different architecture can approximate any function.

An important concept in machine learning is **Feature selection**, which is the process by which we can select the most relevant variables for the model. It differs from another method that can be used to reduce the number of inputs which is **Feature extraction** because the former selects the most informative features while the latter creates new ones by combining them. So new ones are created, and they are not a subset of the original.

1.4.4 Deep learning

Deep learning is a branch of artificial intelligence, and it is a more advanced approach which enables a computer to automatically extract, analyse and understand the useful information from the raw data by imitating how humans think and learn. The automatic learning of features makes this set of techniques accurate and of excellent performance. DL is replacing standard ML algorithms in many fields, medicine being one of them.[6]

Introduction

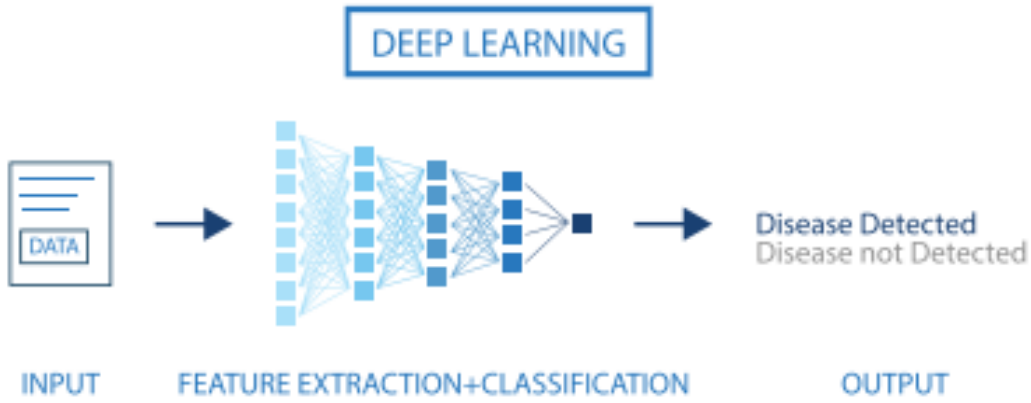


Figure 1.8: Definition of deep learning

The difference between machine learning and deep learning is the type of data they work with, machine learning algorithms use structured labelled data to make predictions, meaning that specific features are organized into tables and defined previously [42]. Deep learning eliminates the pre-processing that is typically involved with machine learning. These algorithms can ingest and process unstructured data, like text and images, and automatizes feature extraction, removing the dependency on human expertise. Machine learning and deep learning models are usually characterized by different kinds of learning as well. Deep learning is based on artificial neural networks, which aim to mimic the

human brain, the basic block of a Neural Network is the Perceptron. [43]

Perceptron

The perceptron is a simple algorithm for supervised learning of binary classifiers. It can be thought as a function that maps its input x (a real-valued vector) to an output value f(x), to which then a threshold can be applied.

The basic components of a perceptron are:

- Input Layer: the input layer consists of one or more input neurons, which receive input signals from the external world or from other layers of the neural network.
- Weights: each input neuron is associated with a weight, which represents the strength of the connection between the input neuron and the output neuron.
- Bias: a bias term is added to the input layer to provide the perceptron with additional flexibility in modeling complex patterns in the input data.
- Activation Function: The activation function determines the output of the perceptron based on the weighted sum of the inputs and the bias term. Common activation functions used in perceptrons include the step function, sigmoid function, and ReLU function.
- Output: the output of the perceptron can be a single binary value, either 0 or 1, which indicates the class or category to which the input data belongs, or in neural networks, the value of the activation function.

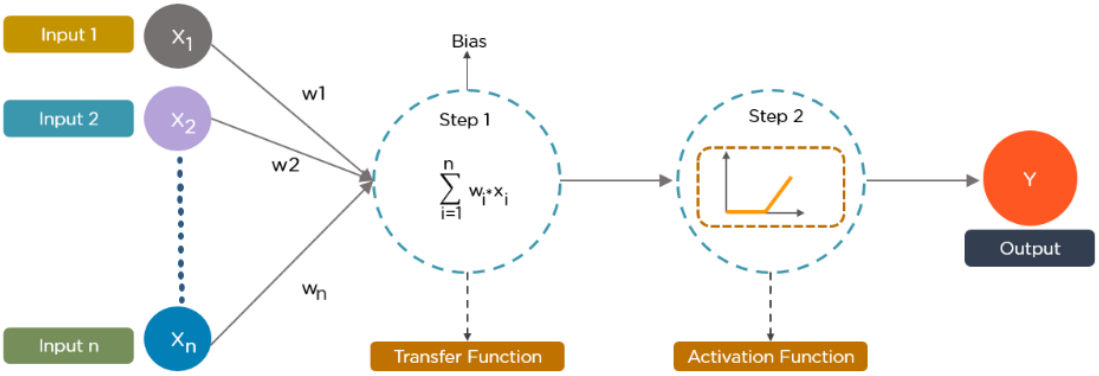


Figure 1.9: Building block of a neural network

Image processing

Deep Neural networks have found applications in many fields and many specialized architectures have been created, some examples include RNNs and CNNs.

CNNs mimic to some degree the way humans classify images. They recognize specific patterns or features anywhere on the image that distinguish between particular object classes. Typically, they work by first identifying low level features on the input image, these are then combined to form high level features. Eventually, the presence or absence of these higher level features contributes to the probability of any given output class. This hierarchy is built using a combination of specialized hidden layers. In general CNNs include convolutional layers pooling, activation and full connected layers. The pooling layer is mainly used to reduce the resolution of feature maps. Activation layers introduce nonlinear factors and improve expression ability, the fully connected layer acts as a classifier [44]

Network structure

In a classification network, which is the first type to be developed for image processing, the input is sent through a convolution filter and then through a max pooling layer. Repeating this process reduces the

size of the channels, but increases its number (a new channel for each convolutional operation). After repeating this operation a number of times and after each channel is reduced to just a few pixel, the feature maps are flattened. The pixels are treated as separate units and fed into one or more fully connected layers before reaching the classification layer.

There are many tuning parameters to be selected in constructing such a network, apart from the number, nature, and sizes of each layer. Dropout learning can be used at each layer, as well as lasso or ridge regularization.

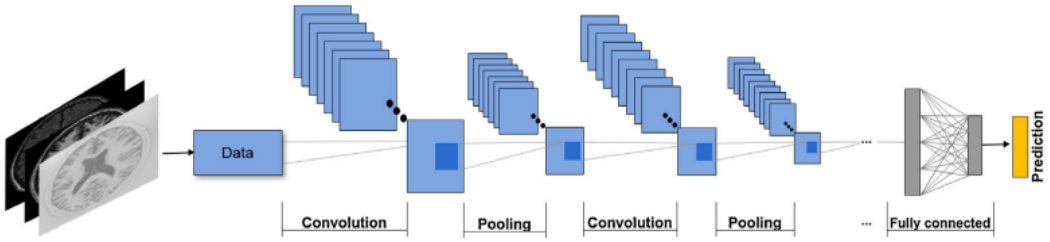


Figure 1.10: Generic CNN architecture

Convolution layers

A convolution layer is made up of many convolution filters, which rely on a simple operation called convolution, which basically amounts to repeatedly multiplying matrix elements and then adding the results. Each of the layers described uses a number of filters to pick up a variety of differently oriented edges and shapes on the image.

These filters are not new in image processing. The distinguishing characteristic of convolutional neural networks is that the filters are learned to perform well on the specific classification task. They operate on localized patches in the input images, and the weights are constrained.

Pooling layers

A pooling layer is a way to condense a large image into a smaller summary image. While there are a number of possible ways to perform this operation, a very common approach is the max pooling operation, which summarizes each non-overlapping 2×2 block of pixels in an image using the maximum value in the block.

Activation layers

The activation function is to introduce nonlinearity into the CNN. In a practical problem, the data is often not linearly separable and without activation it is difficult for CNNs to achieve a good effect on linearly indivisible data. *Sigmoid* and *tanh* are two of the earliest proposed activation functions.

$$sigmoid(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad (1.1) \qquad \mathrm{ReLU}(x) = \begin{cases} 0, & x \leq 0 \\ x, & x > 0 \end{cases} \quad (1.2)$$

Fully connected layer

The fully connected layer plays the role of a "classifier" in the whole convolutional neural network. If operations such as convolutions and pooling map the original data to the hidden feature space, fully connected layers and activation functions map the learned distributed feature representation to the sample space. In traditional CNN, usually more than one fully connected layer are used to construct a fully connected network.

Batch normalization

Deep neural network tuning is very difficult and often causes internal covariate shift, which is a phenomenon by which when the parameters change in the network, the data distribution of internal nodes ranges as well. A solution is to reduce the network convergence using batch normalization, which normalizes the output signal in an ideal range by

correcting the parameters only after calculating the average correction for a batch of images.

Dropout

Overfitting is one of the most encountered problems in deep learning. To solve it dropout can be used. It consists in randomly setting to 0 some neurons in the hidden layer, effectively ignoring them, which alleviates overfitting. [45]

Transformers and self attention models

Nowadays, transformers are the de facto standard in NLP and are being applied successfully to other problems such as computer vision image processing. What is remarkable about this model is its ability to deal with sequential data and complex relationships. The key feature is the attention mechanism. The attention mechanisms allow the model to focus on different parts of the input sequence when making predictions, rather than having to process the entire sequence at once. [46]

1.4.5 U-Net for biomedical segmentation

CNNs immediately after their conception outperformed the state of the art in many visual recognition tasks, specifically in classification tasks. In biomedical, and many other applications, the desired output should also include segmentation (a class label is supposed to be assigned to each pixel). The first approach to solve this problem used a sliding window setup to predict the class label of each pixel by providing the region, this solution has many drawbacks and has been replaced in a short time.

The first paper describing a U-Net (U stands for the shape of the network in diagrams, which resembles a U, as it can be seen in the image 1.11) was published in 2015, and it features a FCN (fully convolutional network) modified in such a way to open the possibility of being trained with very little images and yield precise segmentation. The main idea

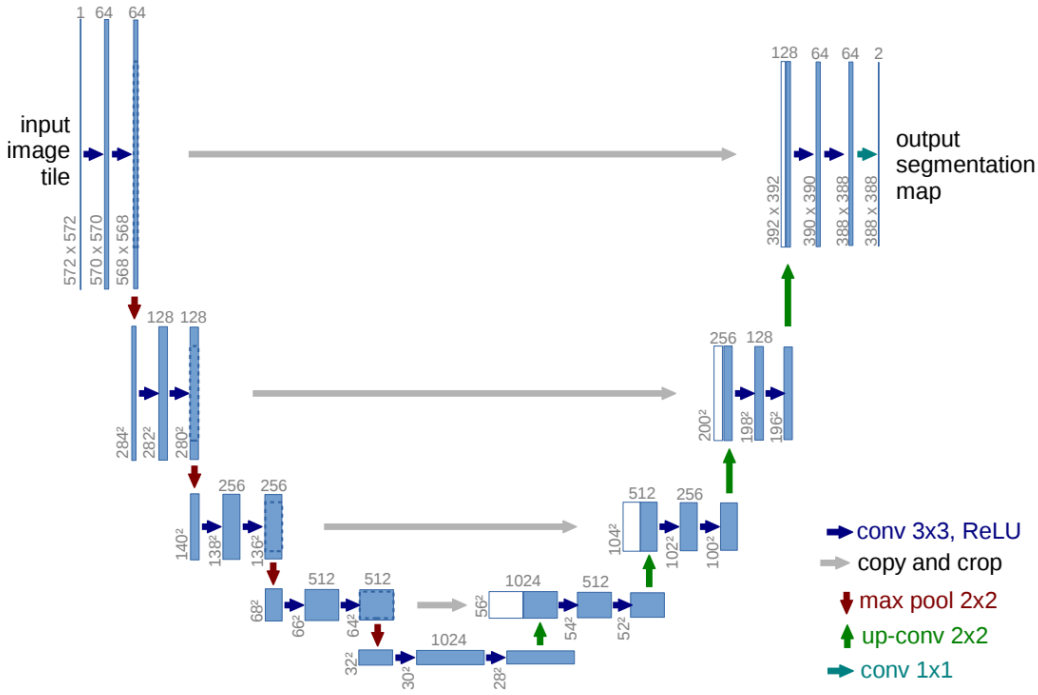


Figure 1.11: Ronnenberg's Original U-Net Architecture

is to supplement a usual contractive network with successive layers that feature upsampling operation, in order to increase the resolution of the output to the original shape. For generalization, high resolution features from the contracting path are combined with the upsampled output, and a successive convolution layer can then learn to assemble a more precise output based on this information. The study was validated using very little subjects and extensive data augmentation with elastic deformations and obtained remarkable results. Since then, most of the works in biomedical image segmentation are based on this architecture. [47]

1.4.6 State of the art in deep learning

Follows a brief description of some models that represent the state of the art in this field of research, created using the findings of the bibliographic research described in the following chapter.

QuickNat

This paper achieves fast segmentation results in just 20 seconds and introduces auxiliary label training to address the limited access to medical imaging data. This approach improves accuracy and reliability compared to existing methods, and notably outperforms FreeSurfer when trained solely on its output. The architecture consists of three networks operating on different axes, followed by view aggregation for final segmentation as it can be seen in the image 1.12. Each F-CNN has an encoder/decoder structure with skip connections, unpooling layers, and dense connections. The network is optimized using a joint loss function combining multi-class DICE loss and weighted logistic loss. The key contributions of this project are the auxiliary label training strategy and the F-CNN architecture, which have undergone extensive testing. [48]

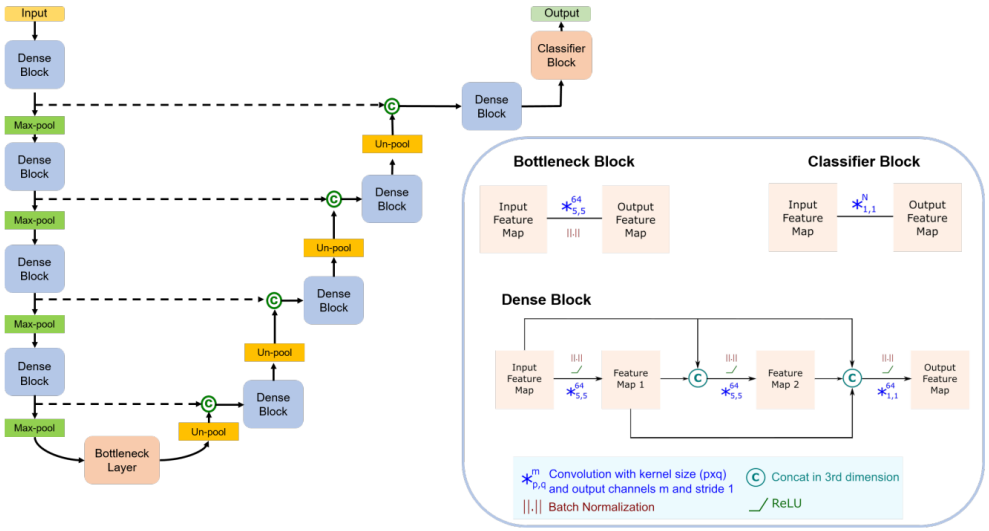


Figure 1.12: QuickNAT structure

AssemblyNet

This system is based on an ensemble method based on numerous CNNs processing different overlapping brain regions. Inspired by parliamentary decision-making systems, it is made of two “assemblies" of U-Nets, as can be seen from the image **??**. AssemblyNet introduces sharing of knowledge among neighbouring U-Nets, an “amendment” procedure made by the second assembly at higher-resolution to refine the decision taken by the first one, and a final decision obtained by majority voting. It is inspired by state-of-the art label fusion methods such as SLANT, and achieves competitive performances.

Each U-Net processes a sub-volume of the global volume, the results are then aggregated. A defining characteristic is the nearest neighbour transfer learning strategy proposed. Two chambers also interact with each other using an expected final decision that represents the prior knowledge and is passed between the two chambers. According to the researchers, it is trained on very little data and still is able to obtain very up to standards results. [49]

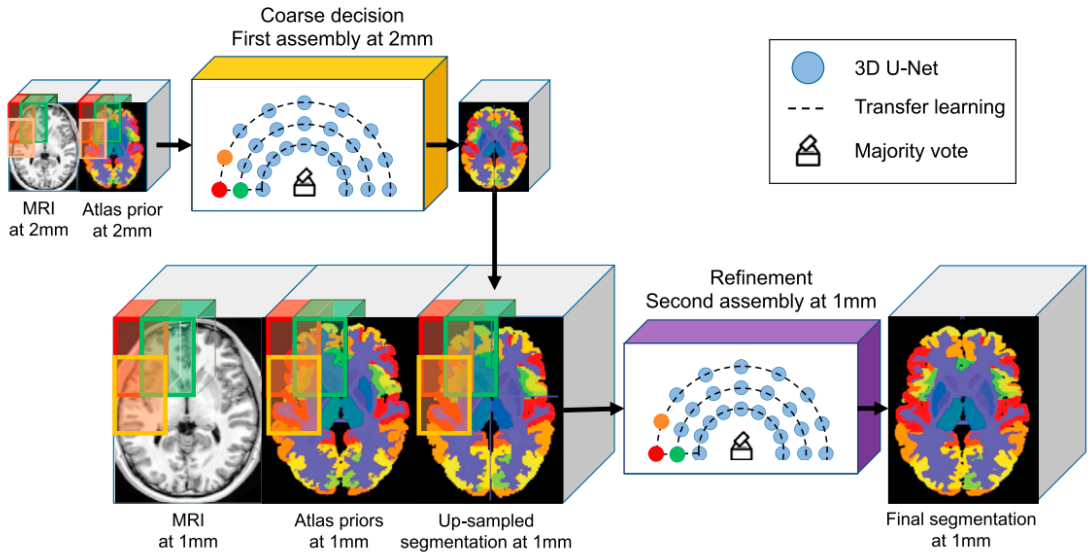


Figure 1.13: AssemblyNET structure

SLANT

This method uses a number of independent 3D fully convolutional neural networks (FCN) for high resolution whole brain segmentation, the structure can be seen below 1.14. For the training it takes inspiration from QuickNAT, using at first unlabelled data and consequently manually annotated volumes. In this work each network is only dealing with a particular spatial location as multiple networks are used, the task of each network is simplified to focus on the patches of similar parts of the brain with lower variation. The input is registered according to the MNI305 standard. The decoder is compatible with 33 labels outputs. 3D output channels have been employed in the deconvolutional layers in each 3d U-Net. Since the entire MNI space is divided into k overlapping subspaces, a majority voting label fusion method was employed to get the final segmentation results. [50]

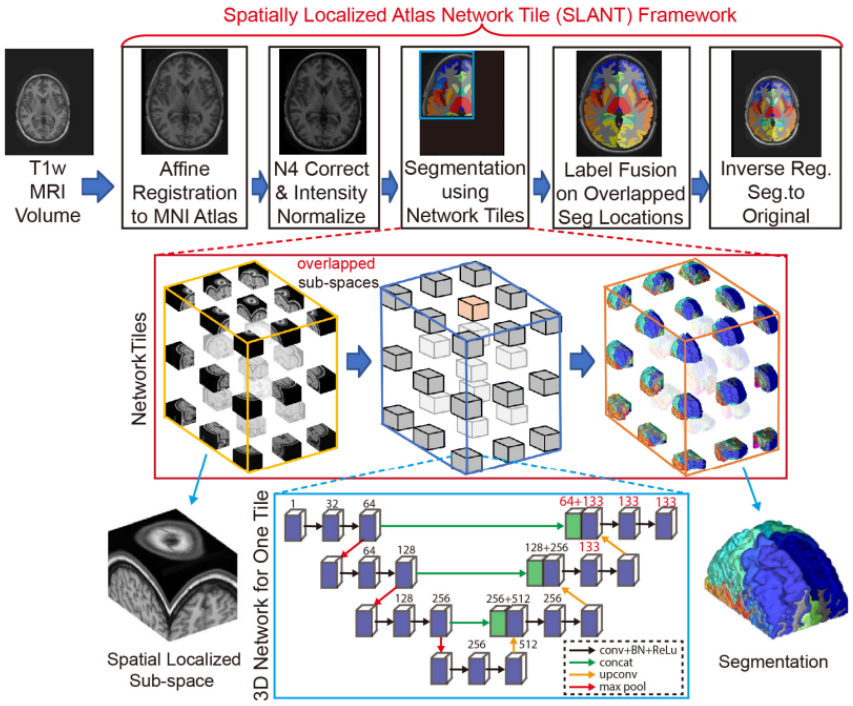


Figure 1.14: SLANT structure

Whole brain segmentation with full volume neural network

The paper aims to perform whole brain segmentation of numerous brain structures using a full volume framework and an advanced FCN architecture capable of accurate segmentation of small neuroanatomical regions, the structure can be seen in the image below 1.15. Unlike existing methods, this framework provides holistic predictions for each full volume, simplifying the process without requiring fusion strategies or multiple passes. The model is trained using complete ground truth labels through empirical risk minimization and backpropagation. The backbone consists of multiple stages and parallel branches, incorporating bottleneck modules, transitional layers, and fusion layers for feature integration. Mixed precision training is employed to reduce resource requirements. This is still an experimental approach due to the complexity of full volume CNNs. [51]

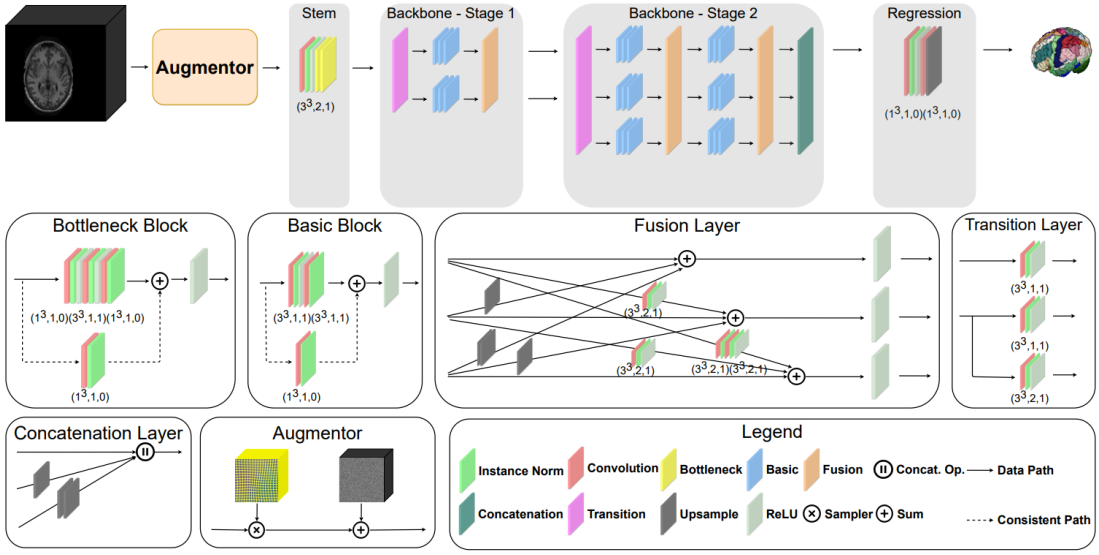


Figure 1.15: 3D FCN structure

FAST-AID BRAIN

It is an efficient 2.5D based deep learning method for automatic segmentation of the human brain into 132 cortical and non-cortical regions. It is a U-Net-like fully convolutional network to the three principal views, and it fuses them based on the intersection points and the hierarchical relations in end-to-end fashion. It has a rather new architecture that can fuse 2D information with spatial context. The class imbalance is managed with label hierarchies, and supervision is used to learn from partially labelled data to segment the whole brain and estimate the inter-cranial volume. A very extensive data augmentation process was performed and experiments on different atlases as well to evaluate accuracy and robustness of the trained model. As stated before it is composed by a fully convolutional encoder-decoder network on three orthogonal planes, only one backbone is used to keep the complexity of the network low while incorporating 3D information. The main features of this model are the hierarchical softmax, the use of only one network for the tree planes and the CE loss. [52]

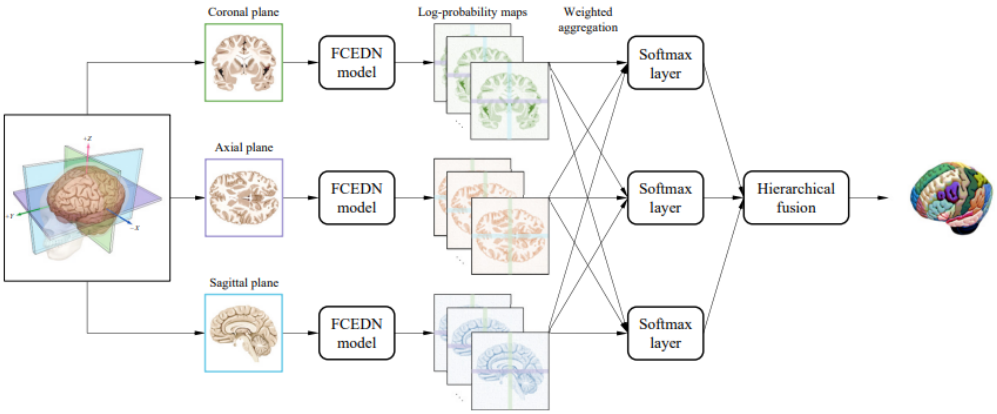


Figure 1.16: FAST-AID structure

UNesT

It is a Transformer based model. This kind of models have recently demonstrated exceptional representation learning capabilities in computer vision and medical analysis. They reformat the image into separate patches and realizes global communication via the self attention mechanism. It has some downsides, such as the difficulty of preserving positional information between patches, which can lead to suboptimal performance. This paper proposes a solution inspired by the nested hierarchical structures in vision transformer. which achieves local communication among spatially adjacent patches sequences by aggregating them hierarchically. It is tested against SLANT27, and it is shown to outperform it. [53]

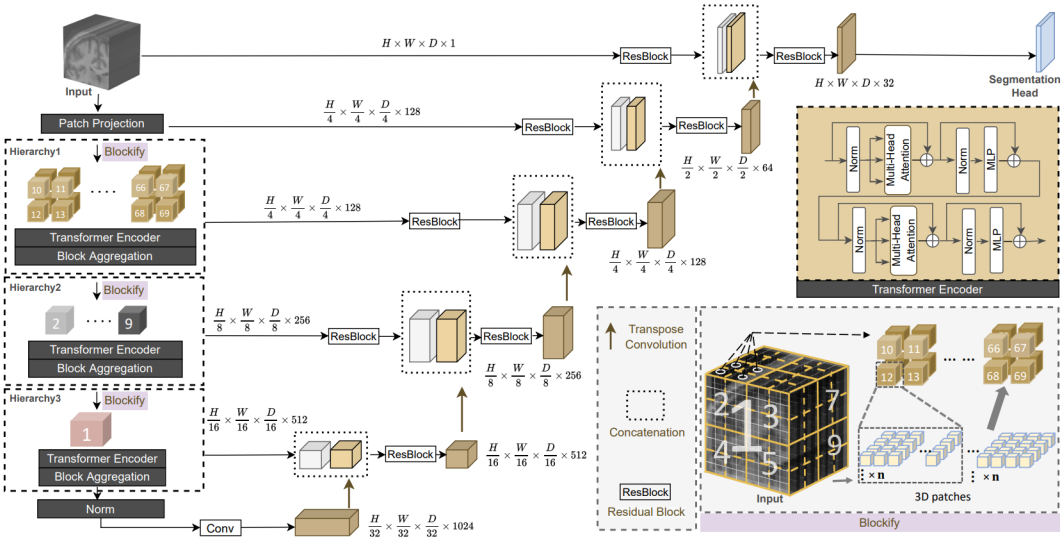


Figure 1.17: UNesT structure

Chapter 2

Material and Methods

2.1 Bibliographic research

Before the implementation of a model, a review of the state of the art was performed in the following manner.

2.1.1 Introduction

The research was performed using different databases of research literature, mainly from Google Scholar and PubMed as well as IEEExplore and using keywords such as Whole brain segmentation, brain parcellation, Alzheimer's detection using deep learning and imaging, deep learning parcellation of brain tissue.

Research was performed as well using the search engine paper with code which has a repository of papers which are supposed to have the code linked and available. This was very useful in the research as most of the papers found through other means did not provide access to the source code. The research focused on recent articles after 2018 since deep learning and computational problems in general is a field in rapid evolution.

Using this method 119 articles were identified, of which, based on the

abstract, 17 articles proposed a method of interest and are representative of the state of the art on the subject as it can be seen if the table 2.1, of these 6 were selected as potential candidates because they implemented the whole pipeline and made the code available to download for research purposes or had a linked repository in abstract or article, the process is shown in a simplified version in the image 2.1.

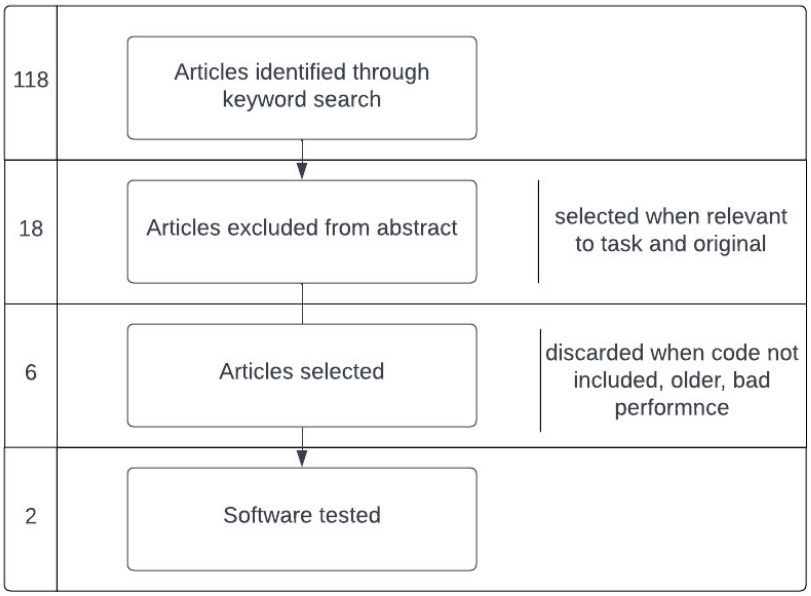


Figure 2.1: Segmentation model selection

The first objective was understanding the state of the art on the subjects. Numerous reviews were investigated and the state of the art was assessed, large studies with already very complex networks are being tested with very successful results. The networks focus only on the segmentation or parcellation process, to come close to a similar efficiency to FreeSurfer only the last years' publications may be considered. A few base models have achieved a greater success and are often referenced when talking about the topic, such as QuickNat. These papers were taken into account as reference models, only few of them were adequately tested and could be used for this research.

Some of the requirements were:

- Open-source code available on GitHub or through request and trained model provided.

- Atlas on which the network based the segmentation on, giving priority to networks that were able to output a result similar to FastSurfer.

- Completeness of the work.

An important parameter that was taken into account is the simplicity of use and preprocessing and post-processing steps that are required to be implemented for the correct use of the method. This is a necessary step since the solution end goal is to be implemented in a pipeline to be used with many patients and ideally in a clinical or research purpose. These requirements drastically reduced the number of studies published that could fit the goal. After filtering redundant studies and studies excluded for different reasons such as incomplete data or unavailable code, a few were left.

MONAI

Project MONAI was originally started by NVIDIA and King's College London to establish an inclusive community of AI researchers for the development and exchange of best practices for AI in healthcare imaging across academia and enterprise researchers. MONAI Core is the flagship framework created by Project MONAI, and it would not have been possible to accelerate this development without the development of existing toolkits such as Nvidia Clara Train, NiftyNet, DLTK, and DeepNeuro. [54]

1	QuickNAT: A Fully Convolutional Network for Quick and Accurate Segmentation of Neuroanatomy
2	HyperDense-Net: A hyper-densely connected CNN for multi-modal image segmentation
3	A Learning Strategy for Contrast-agnostic MRI Segmentation
4	FAST-AID Brain: Fast and Accurate Segmentation Tool using Artificial Intelligence Developed for Brain
5	Whole Brain Segmentation with Full Volume Neural Network
6	FastSurfer - A fast and accurate deep learning based neuroimaging pipeline
7	Concurrent Spatial and Channel Squeeze and Excitation in Fully Convolutional Networks
8	VoxResNet: Deep Voxelwise Residual Networks for Volumetric Brain Segmentation
9	Deep Learning Framework for Real-time Fetal Brain Segmentation in MRI
10	AssemblyNet: A large ensemble of CNNs for 3D whole brain MRI segmentation
11	Spatially Localized Atlas Network Tiles Enables 3D Whole Brain Segmentation from Limited Data
12	3D fully convolutional networks for subcortical segmentation in MRI: A large-scale study
13	3D Whole Brain Segmentation using Spatially Localized Atlas Network Tiles
14	Early diagnosis of Alzheimer's disease using machine learning: a multi-diagnostic, generalizable approach
15	RP-Net: A 3D Convolutional Neural Network for Brain Segmentation From Magnetic Resonance Imaging
16	An Open-Source Tool for Longitudinal Whole-Brain and White Matter Lesion Segmentation
17	Nested Hierarchical Transformer: Towards Accurate, Data-Efficient and Interpretable Visual Understanding

Table 2.1: Table of the selected articles

2.1.2 Implementation process

The process of selecting the model to use for the experiments had some conditions, the first is that the code needed to be available to the public. This already filtered out some options. Furthermore, it needed to be pretrained and needed to have an output codified with a usable Atlas. The output needed to have clinical validity and to be comparable with a gold standard, i.e. FreeSurfer. This implied a limited number of options were available. The first article that was looked into was FAS-AID BRAIN.

The choice was made because of the following reasons:

- Framework: it is implemented using a popular framework: MONAI
- New: it is published in 2022, this means that most of the features included are up-to-date
- Many brain regions identified: it uses a very complex atlas with more than 100 brain regions identified.

Downsides:

- Different atlas compared to FastSurfer.
- Memory intensive.
- No postprocessing.

It proved to be problematic because difficulties in its implementation slowed down the work more than an acceptable amount of time, the work was not complete and the workstation kept having memory problems in running it. Furthermore, the output needed further processing to be comparable to the one of FreeSurfer, which, as implied before, was the gold standard for this work.

This led to the change of the focus of this study. The model that was thus selected is FastSurfer, which, as the name implies, seeks to be a substitute for FreeSurfer, thus being optimal for this task. Fastsurfer was already included in a pipeline that includes both the registration steps and the preprocessing steps, which include the parcellation of the brain according to the ASEG and APARC atlases, which will be the focus of this work. The upsides are:

- Post-processing included: the relevant statistics are extracted automatically by a tool provided, and are comparable to FreeSurfer.
- Same atlas as FastSurfer: Since it was proposed as a substitute of FreeSurfer it uses the same Atlas, The DKT Atlas, which makes the comparison between the two methods easy.
- Automatic pipeline implemented in Docker: the implementation makes it easy to use for large number of data.

The downside is:

- Developed in 2020, so not very recent.

2.2 FastSurfer

2.2.1 Introduction

The Fastsurfer CNN is implemented in a pipeline inspired by FreeSurfer, it includes cortical surfaces analysis and thickness analysis using a surface processing pipeline that integrates the neural network architecture at its core to provide the same volume and surface results. this includes cortical surfaces, thickness maps and summary statistics in cortical regions following the DTK protocol atlas.

The pipeline which processes the output of the CNN has a number of innovations. Traditionally, surfaces are generated through a pipeline consisting of several time-consuming steps:

1. Meshes are smoothed and mapped to a sphere to localize topological defects
2. Surface placement along the white matter is fine-tuned and a second expanded surface is placed at the outer grey matter (GM) boundary, also providing thickness estimates on every point on the cortex.
3. Surfaces are then carefully mapped to the sphere a second time (minimizing metric distorsions), registered to a spherical atlas and segmented into cortical parcellations (DTKatlas).

In the FastSurfer pipeline, the above FreeSurfer pipeline is modified to yield surface results of FreeSurfer. A significant speed-up compared to is obtained by omitting several steps that have become obsolete such as skull stripping and non-linear atlas registration, given that the high quality segmentation can be achieved easily.

The steps taken are the following:

1. A brainmask is created by closure, dilation, and erosion of the labels, including the ventricle labels.

2. A quick bias field corrected brain image and linear Talairach registration are retrospectively contracted, as these results are needed later for some relevant statistics.

3. The mesh is created using a marching cube algorithm rather than the traditional approach aiming at higher mesh quality and reduced number of vertices.

4. A fast mapping to the sphere is obtained using the Eigenfunctions of the Laplace operator to perform a spectral embedding of the original white matter surfaces quickly, solving the Laplace-Beltrame Eigenvalue problem.

5. After topology fixing and GM surface creation, the DTK GM segmentations from the image are registered onto the surface and the surface ROI statistics are computed. These statistics, such as mean thickness and curvature averages per region, mimick FreeSurfer surface segmentation.

2.2.2 Architecture

Description of the architecture

FastSurferCNN is a network architecture that allows to segment the brain into 95 classes excluding the background in less than 1 minute on the GPU. It is composed of three FCNNs, similar to the ones implemented by Roy et al. in QuickNat[48] which consist of a sequence of 4 dense encoder and decoder blocks separated by a bottleneck layer as shown in the image 2.2. Fastsurfer has some improvements, which are competitive dense blocks and spatial information aggregation.

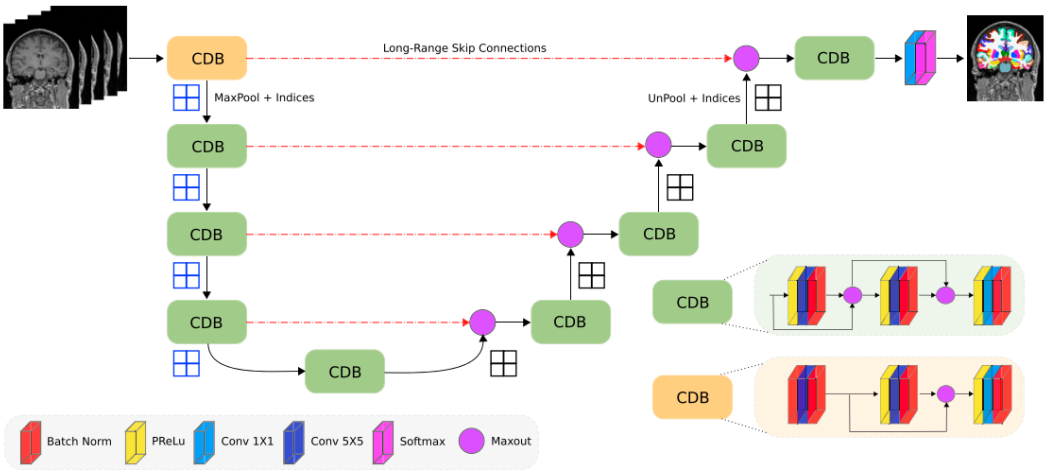


Figure 2.2: FastSurfer architecture

Competitive dense blocks

Competitive dense blocks employ a distinct concept where the conventional concatenation operation in dense connections is replaced with max-out activations, shown in the image 2.3. This substitution instigates competition among feature maps and effectively reduces the parameter count compared to traditional dense blocks. Consequently, a lightweight model is formed by solely retaining the maximum value at a particular position, instead of stacking the output of the previous layer from the previous model. This approach keeps the number of input channels and parameters constant in every convolutional layer. Furthermore, the competition extends to long-range skip connections, ensuring consistent competitive behaviour throughout the network.

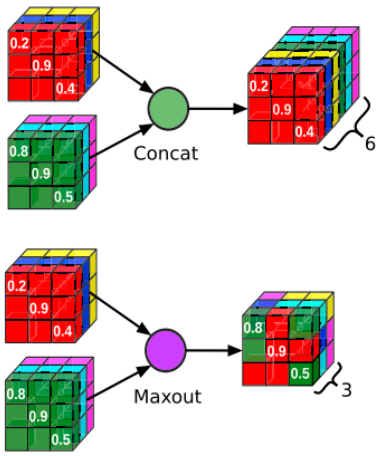


Figure 2.3: Competitive dense block

Spatial information aggregation

3D deep neural networks are not feasible for numerous classes, however 2D networks with single slice inputs lose information on the 3D spatial dependency between the inputs, which can be crucial for correct segmentation of neuroanatomical structures. For this purpose, a multi-slice input is passed to the network [55]. The spatial information aggregation approach involves the processing of a 7-channel image by stacking three preceding, the current, and three succeeding slices. This technique aims to segment only the middle slice. Essentially, this approach combines the benefits of 3D patches, which capture local neighborhood information, with 2D slices, which provide a global view. The obtained results are then compared to those obtained using 2D inputs.

View Aggregation

The brain is a 3D structure and this needs to be taken into account, since in this case there are 3 F-CNNs, one for each plane, Their output needs to be combined to create only one probability map. The three views (coronal, axial and sagittal) are thus aggregated through a weighted average, as shown in image 2.4. This boosts accuracy for cortical folds and subcortical structures. In the sagittal view, since it is not possible to differentiate left from right hemisphere the lateral labels are merged, reducing the number of classes from 78 to 50.

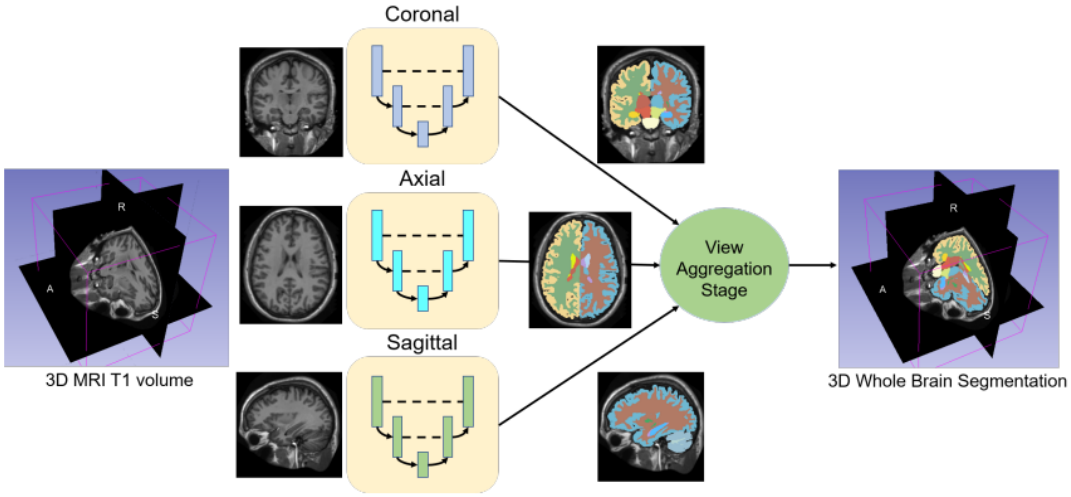


Figure 2.4: View Aggregation example from QuickNat [48]

2.2.3 Training process and datasets

For training purposes, a total of 140 representative subjects were selected from datasets including ABIDE-II, ADI, LA5C, and OASIS. To validate the model, 20 subjects from the MIRIAD dataset were used. Empty slices were filtered out, resulting in an average of 145 single view planes per subject and a total of 20,000 images per network. Data augmentation techniques were applied to enhance the training set. Moreover, the training set was balanced based on various parameters such as age, gender, and others.

2.2.4 Testing documented

Metrics used and statistics

The FastSurfer Pipeline was validated in terms of accuracy, generalizability, reliability, and sensitivity using DICE, ICC, and group analyses on volume and thickness of the regions of interest (ROIs) as well as thickness maps.

- Dice Coefficient and HD: The evaluation of segmentation performance for different network architectures involves the use of Dice Coefficient (DSC) and Hausdorff Distance (HD), with the average HD serving as the metric for comparison. DSC is utilized in two ways: firstly, to directly compare the performance of different network architectures against each other, and secondly, to estimate the similarity between the predictions achieved with FastSurferCNN and FreeSurfer v6.0 on previously unseen datasets, thus assessing the generalizability of the models.

- The agreement between cortical thicknesses and subcortical volumes in consecutive scans using the OASIS1 test-retest set were calculated. After averaging across hemisphere, the ICC as well as the upper and lower bound with $\alpha = 0.5$ level of significance are calculated for each region.

- Identical linear models were fitted to FastSurfer's and FreeSurfer's results, explaining thickness or volume by diagnosis controlling for age, sex and head size. The p-values of the diagnostic effect are monotonically connected to the absolute value of the t-statistic which in turn is a scaled version of Cohen's d, where the scaling factor depends on sample size, a direct comparison of p-values is possible given that both methods operate on the same input.

Results from Paper

FastSurfer Obtained a High Generalizability and high performance compared to FreeSurfer:

Comparison to manual reference: The segmentation performance of the networks was compared to a manual standard. The DSC and average HD were computed on the Mindboggle101 dataset (78 subjects) with manually labeled cortical as well as manual subcortical segmentations on 20 of these subjects.

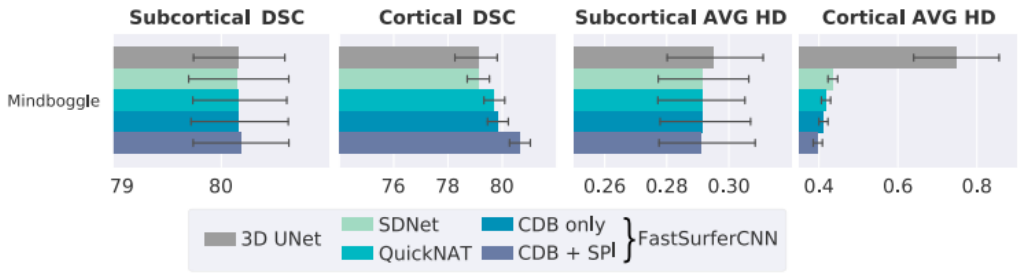


Figure 2.5: Results of Fastsurfer

Comparison to FreeSurfer: To evaluate segmentation accuracy, the DSC, and AVG HD was computed with FreeSurfer Labels on five Datasets (ADNI, OASIS1, HCP, MIRIAD, and THP). It was benchmarked against SDNet and QuickNAT and a 3D FCN Network.

Reliability: Test-retest reliability is assessed as the agreement between the evaluations of two scans in a short time frame. The ICC is calculated on the OASIS1 test-retest dataset with 20 participants. It is shown to be higher for Fastsurfer than FreeSurfer.

Generalizability: It was tested on machines from different manufacturer, as well as different age groups, gender, and diseases. A small decrease in segmentation DSC can be observed with disease progression. Event though Fastsurfer was primarly trained with images coming from Philips machines, good results were observed with any vendor.

2.3 Datasets

Many public datasets are available that contain MRI volumes. Two were partially used in this project.

2.3.1 ADNI

ADNI is an association of medical centres and universities located in the USA and Canada. Its main aim is to provide open-source datasets to discover biomarkers and to identify and track AD accurately. It developed to become an ideal source of longitudinal multisite MRI and PET images of patients with AD, MCI and NC and elderly controls. The data sets were formed to make the detection system powerful by providing baseline information regarding brain structure and metabolism and also through clinical cognitive and biochemical data. This study has been taking place since 2004 in multiple phases, i.e. ADNI1, ADNI2 and ADNI3. [56]

2.3.2 OASIS

It is an open source data set of MRI images that can be used freely. It consisted of 416 subjects initially, all being right-handed and aged 18-96 years. Both female and male patients were present. One hundred of them with an age above 60 were diagnosed with very mild to moderate AD. For each MRI, three to four weighted scans with high contrast to noise ration are present. The total volume of the brain and the estimation of the intracranial volume is used for analysing normal ageing and Alzheimer's disease, 20 dementia patients are included as well. [57]

2.4 Methods

In this section, the instruments used in this project are described.

2.4.1 Docker gpu

Docker is an open source software platform to create, deploy and manage virtualized application containers on a common operating system, with as ecosystem of allied tools. Docker container technology was created in 2013. The technology is available through the operating system. A container packages the application service or function with all the libraries, configuration files, dependencies, and other necessary parts and parameters to operate. Each container shares the service of one underlying operating system. Docker images contain all the dependencies needed to execute the code inside a container, so containers that move between docker environments with the same OS work with no changes. Docker uses resource isolation in the OS kernel to run multiple containers on the same operating system (OS). This is different from virtual machines, that encapsulate the entire OS with executable code on top of an abstract layer of physical hardware resources. Docker was created to work on Linux platforms but was extended to offer support for non-Linux operating systems including Microsoft Windows and Apple OS X, AWS and Microsoft Azure. [58]

For our goal Docker was used on Linux to run an image of the trained network using an additional toolkit to make it compatible with NVIDIA GPUs which is the The NVIDIA Container Toolkit and provides different options for enumerating GPUs and the capabilities that are supported for CUDA containers. [59]

2.4.2 Workstation

Managing a large amount of data and processing MRI images require significant time and resources.

To expedite the process, the study utilized a similar workstation as

described in the paper. It was reported that the total processing time for one image was indeed around one hour, which was empirically confirmed. The network implementation employed Docker with NVIDIA GPU support and was automated, enabling relatively fast processing of hundreds of images within a matter of days or weeks.

Furthermore, the workstation housed all the necessary data, including the data required for processing with FastSurfer and FreeSurfer. Consequently, the memory requirements were substantial, amounting to several terabytes of memory at disposal.

2.4.3 Database

The used data was stored on the workstation, as well as some other data from different projects. The access was possible to a subset of ADNI and OASIS that had already been processed with FreeSurfer and to the tables with the characteristic of the subjects for both datasets. The data needed preparation because it was not very well organized, this meant that some time needed to be spent organizing tables and finding all the respective images on the computer. Writing code to do so meant learning how to use pandas, which revealed itself to be a very important instrument for carrying out this project.

The table that included information about patients such as age and pathology as well as the patient ID was also used to save the folders names with the subjects' information, Its description is in the image 2.2. The table was loaded using pandas and cleaned, keeping only the age, the pathology and the subject ID. Using an appropriate script, the location of the file in the computer was found and added to the table.

Regarding ADNI the matter was a little more complex since the data was stored in two tables that had information with some overlaps. They needed to be filtered and then matched to the data available on the workstation which was a little more fragmented as well. Once sorted out, the dataset to work with was organized as follows:

A better balance was needed between subjects who are healthy and

n	int
ID	strig
path	string
age	int
sex	bool
main_condition	categorical
mmse_score	int
processed	bool
processed_path	string

Table 2.2: Details of the table with the dataset information

people who are not, so, while for the testing phase and for the developing of the algorithm for the data preprocessing the OASIS dataset was used, for the actual testing it was used ADNI, which is more balanced.

2.4.4 Python

Python is an interpreted, object-oriented, high-level programming language with dynamic semantics. Its high-level built in data structures, combined with dynamic typing and dynamic binding, make it very attractive for Rapid Application Development, as well as for use as a scripting or glue language to connect existing components together. Python's simple, easy to learn syntax emphasizes readability and therefore reduces the cost of program maintenance. Python supports modules and packages, which encourages program modularity and code reuse. The Python interpreter and the extensive standard library are available in source or binary form without charge for all major platforms, and can be freely distributed. [60]

2.4.5 pyCharm

PyCharm is a dedicated Python Integrated Development Environment (IDE) providing a wide range of essential tools for Python developers, tightly integrated to create a convenient environment for productive

Python, web, and data science development. It is available in a community edition and a professional edition, with more features available. [61]

2.4.6 Pandas

Pandas is a remarkably popular open-source Python library that is widely used in the data science field. It has become a go-to tool for data manipulation and analysis, thanks to its many data structures and functions that simplify the handling and manipulation of tabular data. This tabular data can be anything from spreadsheets to SQL tables. One of the great advantages of pandas is that it offers a variety of functions for reading data from different file formats. It also offers a wide range of functions for data cleaning, filtering, aggregation, and transformation. [62]

2.4.7 Matplotlib

Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. It has many functionalities such as:

- Creating publication quality plots.
- Making interactive figures that can zoom, pan, update.
- Customizing visual style and layout.
- Exporting to many file formats.
- Embedding in JupyterLab and Graphical User Interfaces.
- Using a rich array of third-party packages built on Matplotlib. [63]

2.4.8 Scikit-learn

Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modelling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python. This library is built upon NumPy, SciPy and Matplotlib. [64]

2.4.9 Other libraries

Many python libraries were used for many tasks such as:

- seaborn - for improved visualization.
- numpy - to deal with matrices and mathematical operations.
- imlearn - for machine learning with imbalanced datasets.
- scipy - for statistics.
- penoguin - for statistical tests.

2.5 Statistics

2.5.1 Statistical tests

Sometimes a study may just describe the characteristics of a sample, such as a prevalence study. In this case the statistical analysis involves only descriptive statistics. Another option is that studies are conducted to test a hypothesis and derive inferences from the sample results to the population, this is known as inferential statistics. The goal of inferential statistics may be to assess difference between groups, establish an association between two variables, to predict a variable from another, or to look for agreement between measurements. One of the most important instruments in inferential statistics are the statistical tests. [65]

An important definition is the difference between paired and unpaired observations. Paired observations are made on the same individual but in different conditions, for example before or after or in different parts of the body. Comparison between individuals are usually not paired.

The type of data must be assessed, the data may be categorical or numerical. Normally distributed data can use parametric tests, which are more statistically powerful.

It is important to define how many measures will be compared. The choice of tests in fact differs whether two or more than two measurements are being compared. This includes more than two groups (unmatched data) or more than two measurement in a group.[66] [67].

In this case the data is paired since they come from the same subjects but from different processing methods, the distribution is not parametric, thus the appropriate test chose according to this table are:

- **Wilcoxon Signed-rank test**: The Wilcoxon test, also known as the Wilcoxon signed-rank test, is a non-parametric statistical test used to compare paired data or dependent samples. It ranks the absolute differences between paired observations and calculates a

test statistic based on the sum of the ranks for positive or negative differences. It is robust against outliers and does not require the data to follow a specific distribution, making it a suitable alternative to the paired t-test when assumptions are violated. [68]

- **Mann-Whitney U-test**: The Mann-Whitney U test, also known as the Wilcoxon rank-sum test, is a non-parametric statistical test used to compare two independent groups or samples. It assesses whether there is a significant difference between the distributions of the two groups. The test involves ranking all the observations from both groups together, calculating the sum of ranks for each group, and comparing the sums of ranks to obtain a U statistic. The test is robust against non-normality and does not assume equal variances. It provides a reliable method for analysing data when the assumptions of parametric tests are not met, allowing researchers to determine if there is a statistically significant difference between two independent groups. [69]

2.5.2 Effect size

As described before, significance is the magnitude of the evidence which the scientific observation produces regarding a postulated hypothesis. It relies on the hypothesis that the observation is intimately affected by some degree of randomness, and that it is always possible to figure out the way the observation would look like when the phenomenon is completely absent.

The result of hypothesis testing is the probability for which it is likely to consider the observation was shaped by chance rather than by the phenomenon. If the result is not significant, there are two possibilities: the result is actually not significant, or a phenomenon does exist, but its small effect is overwhelmed by the effect of chance.

The second option poses the question of whether the experimental setting actually makes it possible to show a phenomenon when there is really one. In order to achieve it, there is the need to quantify how

large, or small, the expected effect produced by the phenomenon is in respect to the observation through which we aim to detect it.

Hypothesis testing assumes that the null hypothesis is always determinable, and usually it is zero. This means that under a practical standpoint achieving such precision is impossible for large datasets. The testing procedure would make it too sensible to trivial differences, making them look like insignificant even when they are not. With respect to the experimental designs, we can assume that each observation taken on a case of the study of a population corresponds to a single trial. Therefore, enlarging the sample would increase the probability of getting a p-value even with a small effect.

This creates the necessity of having a dimensionless measure to estimate the size of the effect. It needs to be dimensionless, as it should return the same information regardless of the system used to take the observations [70]. There are many methods for computing the effect size. For both the tests that were used the metric chosen is the following:

$$r = \frac{|z|}{sqrt(n)} \tag{2.1}$$

2.5.3 ICC

For any instrument and method that needs to be validated, reliability must be evaluated, it is defined as the extent to which measurements can be replicated. It not only reflects the degree of correlation but also the agreement between measurements.

The reliability of a measure is typically quantified using a value between 0 and 1, where values closer to 1 indicate greater reliability. In the past, the Pearson correlation coefficient, paired t-test, and Bland-Altman plot have been commonly used as measures for reliability. However, it is important to note that the paired t-test and Bland-Altman plot are primarily used to establish agreement rather than reliability, and the Pearson correlation coefficient only measures the degree of correlation. Therefore, they are not ideal measures for assessing reliability.

In contrast, the Intraclass Correlation Coefficient (ICC) captures both the degree of correlation and the degree of agreement between measures. ICC was initially introduced by Fisher in 1954 as a modification of the Pearson correlation coefficient. However, the modern ICC is calculated using mean squares obtained from the analysis of variance. It is worth noting that there are various forms of ICC available, and the appropriate one should be chosen based on the specific application or context in question.

Based on the Model, Type, and the Definition of the relationship considered to be important, 10 different forms of ICC coefficient have been theorized [71].

The selection of the correct ICC form for a reliability study can be guided by three steps.

Model selection:

1. **One way random effect**: each subject was selected by a different set of raters who were randomly chosen from a random population of different raters.
2. **Two way random effect**: if raters are randomly selected from a larger population of raters, this is chosen if we intend to generalize to any rater who possess the same characteristic of the selected rater.
3. **Two way mixed effect**: if the selected raters are the only raters of interest.

Type selection: depends on how the measurement protocol will be conducted in actual application.

1. **Mean of k raters**: if it is planned to use the mean value of k raters as an assessment basis.
2. **Single rater**: if it is planned to use the measurement from a single rater as the basis of the actual measurement.

Definition Selection:

1. **Absolute Agreement**: if different raters assign the same score to the same subject.
2. **Consistency**: if raters' scores of the same group of subjects are correlated in an additive manner.

For this application the appropriate ICC measure is the **two way mixed effect, single rater definition for Consistency**.

2.5.4 Bonferroni correction

The Bonferroni correction is a multiple-comparison correction used when several dependent or independent statistical tests are being performed simultaneously (while a given α value may be appropriate for each individual comparison, it is not for the set of all comparisons). In order to avoid a lot of spurious positives, the α value needs to be lowered to account for the number of comparisons being performed[72]. The simplest expression is the following:

$$p = 1 - \frac{\alpha}{m} \tag{2.2}$$

2.5.5 Normalization

All the features were normalized according to the total intracranial volume for every subject before every subsequent processing. The column age was also added. The APARC table went through some processing as well, in fact redundant columns were deleted, and the age column was added.

2.5.6 Visualization methods

Violin plot

The violin plot synergistically combines the box plot and the density trace into a single display that reveals structure found within the data. It still shows the same information as a box plot, which are the centre, the spread, the symmetry, and outliers. The violin plot includes a

box plot with some modifications. First a circle replaces the medial line which facilitates quick comparisons when viewing multiple groups. Second, outside points which are traditionally classified as mild and severe outliers are not defined by individual symbols.

The density trace supplements traditional summary statistics by graphically showing the distributional characteristics of batches of data. A histogram, which is a simple density estimator, shows the distribution of data values along the real number line. To solve the shortcomings of the histogram, it can be substituted by the density trace described by Chambers.

Bland-Altmann plot

It is an important and established method for measuring the agreement between methods in clinical practice. When a new method is proposed, the common practice is to assess its value by comparison with another established technique. It is not certain that either method given an unequivocally correct measure when the degree of agreement between them is assessed. The standard method is often referred to as "gold standard", but this does not and should not imply that it is measured without error.

What is important to assess is the agreement between different methods of measurement. It is of interest to know by how much the new method is likely to differ from the old, so that if this is not enough to cause problems in clinical interpretation the old method can be replaced by the new. A visual approach to this problem is the Bland-Altmann plot. It represents on the x-axis the mean between the gold standard measure and tested measure, and on the y-axis the difference between the two. [73]

2.6 Testing process

The initial goal of the analysis was to validate the accuracy of FastSurfer and compare its results with those obtained from FreeSurfer. To achieve this, considerable effort was put in developing an efficient data processing framework, which involved constructing an extensive codebase. A statistical analysis was performed on the processed data. This allowed to assess the comparability of the processed brain volumes generated by FastSurfer with the results from FreeSurfer, while also highlighting any differences. Additionally, a classification task was implemented, aiming to evaluate the performance of FastSurfer as a novel software tool in automating Alzheimer's detection.

2.6.1 Code

The workflow included different sections. It was implemented starting from 5 main classes, as shown in the picture 2.6 3 of which are a composition of the base statistics class, as it is shown below.

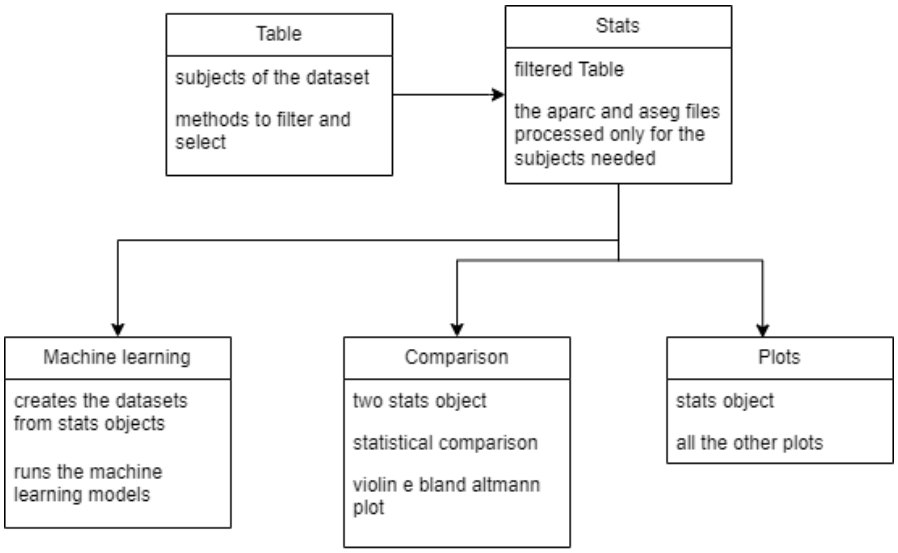


Figure 2.6: Class diagram

Filtering the table

The first step imvolved in the process included creating a summary table that combined information from the original table and data in the folders. This summary table contained the paths to the original images, which were obtained using a file search script, as well as other relevant information from the original table, depending on the datasets. Manual consultation of this table allowed copying the paths to a text file, which served as the input for running the network. This manual action was necessary to maintain control over the processed data and is the only step that requires manual intervention each time.

Furthermore, a method was implemented to update the table. This method enables easy tracking of the subjects that have already been processed by performing a search in the destination folder. Additionally, it adds the path to the folder where the processing results are stored to the table. This automated process is highly useful for maintaining dataset integrity.

Processing

Fastsurfer was run using Docker, which offers several conveniences. To automate the container creation for each subject, a Bash script was written for Linux. This script selects the output folder and creates a separate folder for each subject. It then creates a container and repeats the process for all the files listed in a selected text file, which contains the paths copied from a table.

Postprocessing

The next step involved automating the extraction of features and reorganizing them for post-processing. To achieve this, certain text files needed to be read and interpreted. It is important to note that the output files from FastSurfer differed from those generated by FreeSurfer, so separate scripts were developed for each.

Building upon the previously created table, only the subjects with

processed data were selected. Further filtering conditions could be applied to create datasets containing only the desired data. Once the selection and filtering of the relevant data were complete, the statistics extracted by the neural network were located on the computer and loaded. Regular expressions were used to interpret the data, extracting and adding everything to a dictionary. Subsequently, the dictionary was converted into a dataframe and saved. The resulting dataframe contained all the extracted features organized by subject, specifically for the subjects of interest. This simplified the interpretation of the results and made it much easier to analyse the data.

Data analysis

In this section, a class was created that inherits from the Table class. This new class includes several methods to facilitate all the necessary operations and stores all the datasets. By instantiating an object of this class, users can easily perform comparisons and statistical analyses between results. The class is designed to handle both the statistics extracted from FreeSurfer and FastSurfer, and the resulting objects from either software are interchangeable and compatible with the class.

This approach allows for a seamless integration and analysis of data from both software tools, providing flexibility and convenience in conducting comparisons and statistical evaluations.

Statistical analysis

Once all the data was correctly identified and organized in tables and folders, some information could be extracted from it. To do this, some objects that are compositions of the previous ones were created. In detail, one allowed the comparison between two data classes, the other between an indefinite number of data classes, even though many data series plotted at the same time make the result impossible to interpret.

Comparison

It has a number of methods for both visualization and statistical analysis. All the plots are saved in the images' folder, which is a subfolder of the output folder, and a distinctive name that contains the name of the comparison object, the atlas that was compared and the number of plots contained in the image. The number of subplots can be set when the function is called. Furthermore, the columns can be selected or excluded with two arguments to allow easier data interpretation and to plot only the data that are of interest.

A single iteration method has been written for both the violin plot and the bland Altmann, for the latter an additional section does the matching between the datasets to have paired data.

Violin plot

To visualize the violin plot, the library Seaborn was used, and the series had to be constructed in the correct way before being able to plot it. Substantially the only thing that was done was selecting one series at a time, if it met the conditions, which are that the series is either not in the argument column to exclude or the argument column to keep. It was then checked that the series contains numerical values, and then it was passed to the plotting function.

Bland-Altmann

Creating a Bland-Altman plot requires additional effort to ensure accurate results. The function takes two series as input, but an additional step is needed to match the data to check if it is paired. This step makes sure that every element of each of the two series has a corresponding one. Not matching elements are deleted. Thereafter, the visualization function is called. Similarly to before, which columns to keep and which to discard can be chosen.

Statistical tests and other indexes

Statistical tests are performed calling the function on each column depending on whether the data is paired or not the statistical test to perform is chosen: either the Wilcoxon signed-rank test or the Mann–Whitney U test. In this section, the ICC is computed as well if dataset is paired. The same function used for the Bland-Altmann plot is used to make sure the data series are paired if needed. The results of the statistical tests are then saved in *.csv* files which contain all the information of every structure identified.

Other plots

This class accepts an indefinite number of stat class objects, even though too many make the visualization impossible. It has one main method and the principle is the same as before: the images are saved in the same folder and have unique names that are identified by the name of the object as well as the data represented, the number of subplots can be also decided before plotting. It can be used to plot a combination of data with models fitted to it, mostly used with linear regression.

Linear Regression

To facilitate the visual analysis of the plots generated by the output of the two software programs, a simple linear regression was performed on the data. This regression model captures the relationship between the variables and aids in comparing the software outputs.

All the characteristics of the fitted models, including the R-squared (R2) values for each region, are saved in an Excel table in the common output area. This table provides a comprehensive overview of the performance of the regression models, allowing a further analysis and comparison between the software outputs.

2.6.2 Output

The application is structured with several classes that perform different functions, all the output folders are created in a directory specified when the object of each class is created. Each object creates a subfolder with parameters passed to it. If every module is run with the same data folder the result will have the following structure:

Figure 2.7: Example of an output of the folder structure

2.7 Machine learning

For the machine learning experiments, multiple trials were conducted. During the development stage, all features were initially used, resulting in low accuracy results as expected. Below the pipeline of the machine learning algorithm training process is shown.

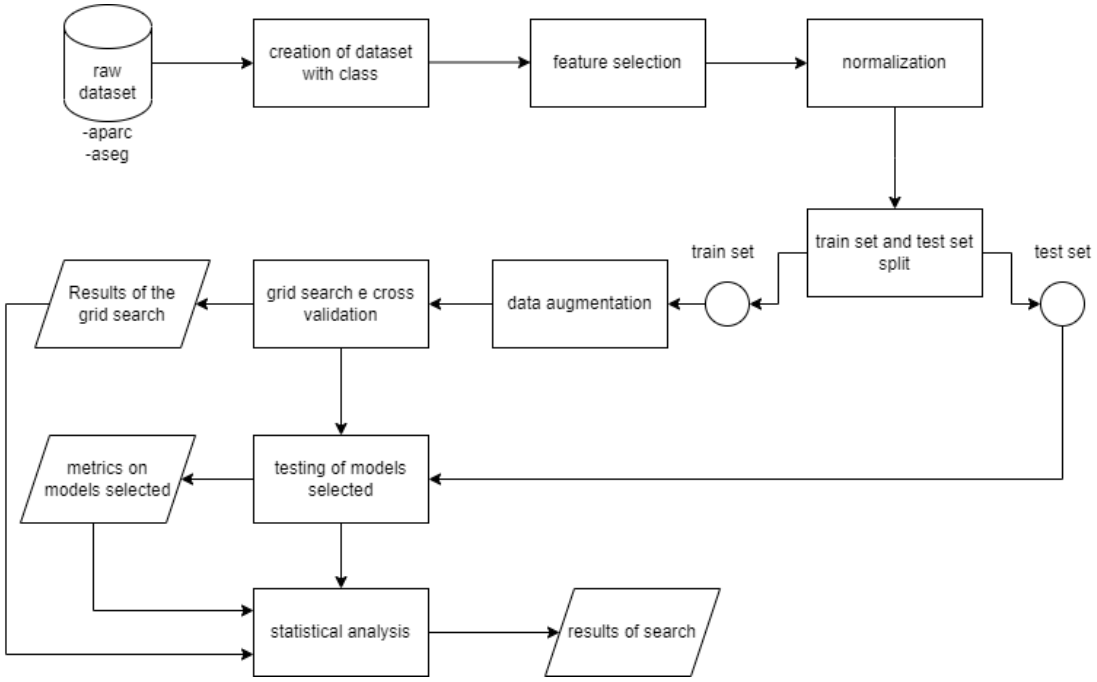


Figure 2.8: Pipeline of the machine learning process

2.7.1 Models tested

- **Logistic Regression:** Logistic regression is a widely used algorithm for binary classification. It models the relationship between the independent variables and the binary outcome using the logistic function. It is often interpretable and efficient for large datasets.

- **Random Forest:** Random forest is an ensemble learning method that combines multiple decision trees to make predictions. Each tree is constructed using a random subset of features and aggregated

to obtain the final prediction. Random forest can handle complex relationships and is robust against overfitting.

- **Support Vector Machine (SVM):** SVM is a powerful algorithm for both classification and regression tasks. It aims to find an optimal hyperplane that maximally separates different classes. SVM can handle high-dimensional data and is effective in cases where the data is not linearly separable by transforming it into a higher-dimensional feature space.

2.7.2 Dataset

The dataset at this point is composed by mostly not healthy subjects with a number of 175, the healthy subjects were 103. These were the two classes considered, and the AD subjects are considered in the same class as MCI. This is expected to result in a lower model performance.

2.7.3 Train and test sets construction

Different strategies to construct the training and test set were used:

- Down-sampling: the first strategy tested was simply creating a dataset with a balanced number of healthy and pathologic subject and to use that for training the classifiers. This dataset was used to construct the balanced training and test set. The advantages of this technique are that the dataset is balanced by definition and the elements of it are all original and unique. The disadvantage is that the numerosity of the dataset is reduced, and before that it was not very numerous.

- Up-sampling: this method uses the opposite approach compared to the one discussed above, in this case what it is done is balancing the classes by upsampling the less numerous one, there will then be duplicates, this solves the numerosity problem, but the elements are not unique and this could introduce a bias depending on the

construction of the training and test set.

- Training balanced: this last method uses a balanced dataset for training and all the other subjects are used for testing, this has the advantage of having many elements in the test set without wasting the data available, the test set is highly unbalanced and metrics for balanced datasets are not reliable, the training set is still not of the optimal dimension.

- SMOTE: the balancing was done using the SMOTE method implemented by the *imbalanced-learn* library. Which is a method that creates synthetic samples for the minority classes by performing a k-nearest neighbour interpolation to find elements close to each other in the feature space. [74]

2.7.4 Feature selection

It is the process of selecting which variables to use to train the classifier, it was done in two ways.

Automatic: at first, the feature selection method based the F-test for classification algorithm implemented with sklearn, that assigns a correlation score of the variable with the output.

From literature: features were selected manually from the combination of the information coming from papers and the statistical information we had available in the data, from the analysis done beforehand.

2.7.5 Normalization

Normalization is needed for this dataset since all the features have different ranges. Not doing any normalization procedure would return biased results.

Min-Max scaling: the Min-Max scaling strategy is the first that was

tested, and it is one of the most simple and used methods.

$$\text{Normalized Value} = \frac{\text{Original Value} - \text{Min Value}}{\text{Max Value} - \text{Min Value}}$$

2.7.6 Model selection

Grid search

To find the most suitable algorithm, the taken approach is often grid search. Grid search is a tuning technique commonly implemented in libraries such as sklearn. It aims to determine the optimal values of hyperparameters through an exhaustive search. This search is performed on specific parameter values of a model, also known as an estimator. In sklearn, this functionality is provided by the *GridSearchCV* function.

Repeated k-fold

During the training process, the models were trained using the k-fold technique, which involves splitting the data into k subsets or folds. The training is then performed k times, where each time a different fold is used as the validation set and the remaining folds are used for training, as seen in the image 2.9. This helps to average the results and ensures that the model's performance is not biased towards a specific subset of the data.

In order to improve the reliability of the results, this process is repeated multiple times. By repeating the k-fold technique for a number of times, we can obtain a more robust characterization of the model's performance. This approach helps to account for any potential variations in the data or randomness in the training process, providing a more accurate assessment of the model's capabilities.

2.7.7 Used metrics

The process began with an unbalanced dataset, which was then balanced using various techniques. Multiple metrics were employed to evaluate

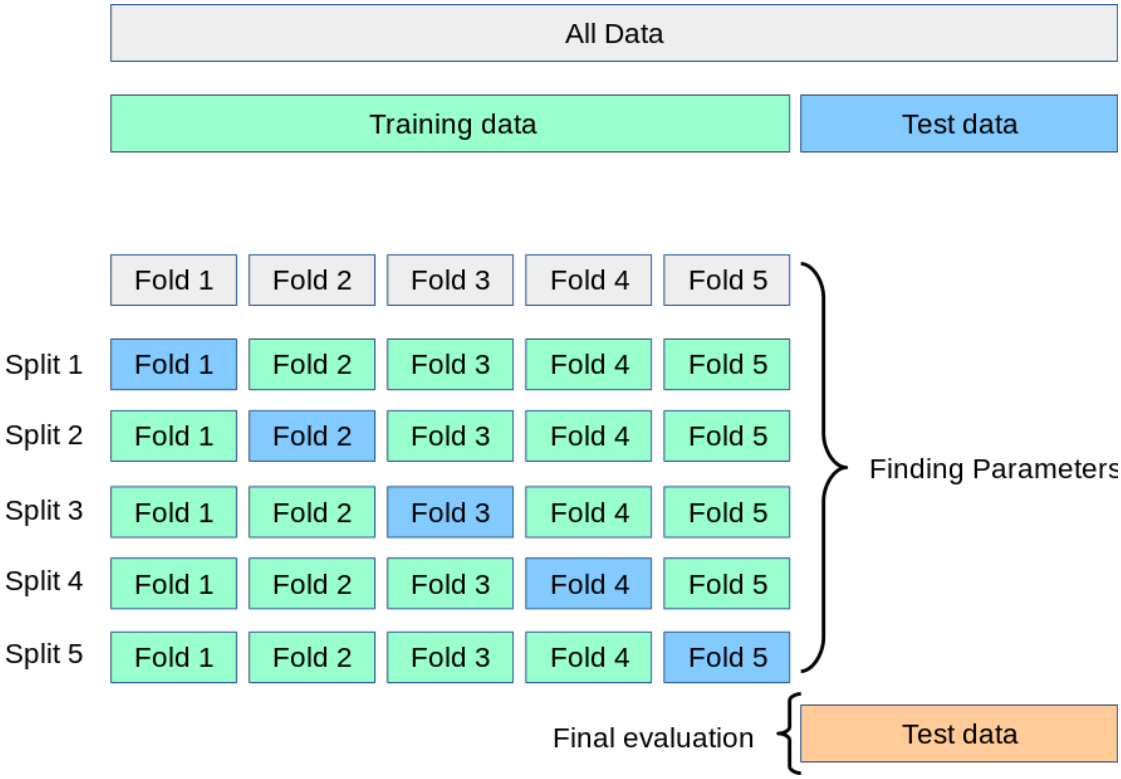


Figure 2.9: k-fold process example

model performance, and the Matthews Correlation Coefficient (MCC) was selected as the criterion for choosing the best model through grid search. This approach aimed to ensure a reliable classification model by addressing dataset imbalance and optimizing performance using the MCC.

Matthews correlation coefficient (MCC)

The best model was selected according to the maximization of this coefficient, the metric is not natively present as an option for the sklearn library, but it can be easily made available. It is a more reliable statistical rate which produces a high score only if the prediction obtained good results in all of the four confusion matrix categories (true positives, false negatives, true negatives, and false positives), proportionally both to the size of positive elements and the size of

negative elements in the dataset.

2.7.8 Validation

Once the best model has been selected, the next step is to validate the chosen model. During this validation process, multiple metrics are computed for each selected model. Since this is a classic binary classification problem, the metrics commonly used in such cases are derived from the confusion matrix.

Confusion matrix

The confusion matrix is a specific table layout that allows visualization of the performance of an algorithm, typically a supervised learning one. It provides a tabulation of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). From this matrix, various metrics can be derived.

Receiver operating characteristic (ROC) - Area under the curve (AUC)

A receiver operating characteristic (ROC), or simply ROC curve, is a graphical plot which illustrates the performance of a binary classifier system as its discrimination threshold is varied. It is created by plotting the fraction of true positives out of the positives (TPR = true positive rate) vs. the fraction of false positives out of the negatives (FPR = false positive rate), at various threshold settings. TPR is the sensitivity, and FPR is one minus the specificity or true negative rate.

Other metrics computed

- **Precision (P)**: The proportion of correctly identified positive instances out of the total instances predicted as positive.

  Formula: $P = \frac{TP}{TP+FP}$

- **Recall (R)**: The proportion of correctly identified positive instances out of the total actual positive instances.

  Formula: $R = \frac{TP}{TP+FN}$

- **Negative Predictive Value (NPV)**: The proportion of correctly identified negative instances out of the total instances predicted as negative.

  Formula: $NPV = \frac{TN}{TN+FN}$

- **Positive Predictive Value (PPV)**: The proportion of correctly identified positive instances out of the total instances predicted as positive.

  Formula: $PPV = \frac{TP}{TP+FP}$

Repetition

The process was repeated 100 times to be able to perform a statistical analysis. The mean and standard deviation of the accuracy were computed to demonstrate the reliability of the process. The mean represents the average accuracy, while the standard deviation indicates the variability of the accuracy values. This analysis helps assess the consistency and stability of the model's performance.

2.7.9 Analysis of results

The results were checked for normality with the Tukey test [75] and then either the independent t-test or the Mann Whitney U-Test was performed on the MCC and ROC-AUC values of the test. This was done to check whether the results are statistically different using the features extracted by FastSurfer and FreeSurfer.

Chapter 3

Results

3.1 Introduction

For a more accurate comparison and measurement of the features of the regions the parcellated areas from the ASEG Atlas were normalized according to the total intracranial volume, which can be more informative since it is relative to the subjects. The APARC features were not normalized.

3.2 OASIS

The OASIS dataset, despite being the one used for the building of the software was not used since the majority of the subjects are healthy. For our goal it was a necessity to have a balance between the healthy subjects and the not healthy subjects that were considered. Despite this, All the subjects present in the part of the OASIS dataset present on the workstation were processed and were used to prove that FastSurfer is actually empirically viable as a solution for brain segmentation. The decision to switch to ADNI was made in a relatively advanced phase of the project because it is more balanced.

3.3 Literature review

The normative model which was built using the deep learning model AssemblyNet gave some guidelines on how to model accurately the progression of Alzheimer. Using the model the progression of Alzheimer was characterized. The regions that indicate the presence of an illness in the most evident way were identified as being the following. [8]

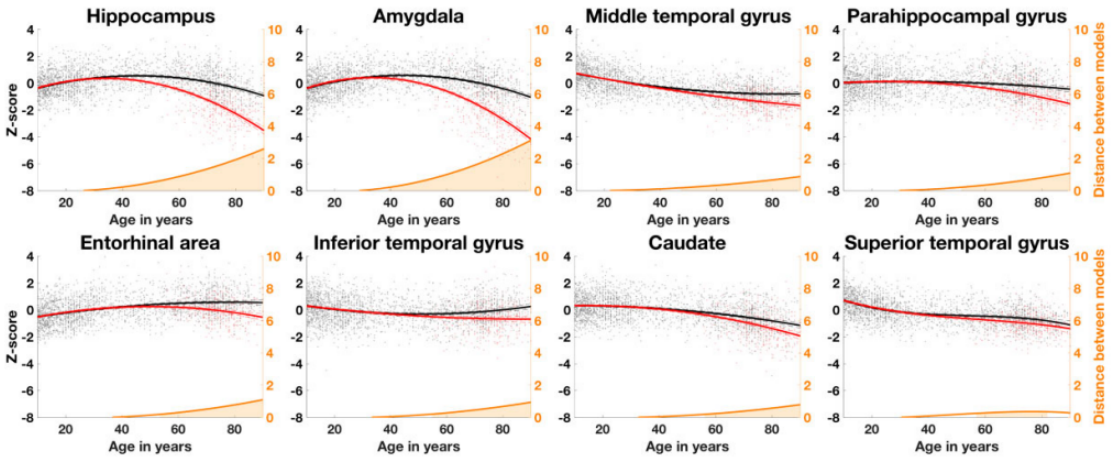


Figure 3.1: Regions identified

3.4 ADNI

Following is the age distribution and the pathology information about the dataset. MCI and AD were considered as one class.

Class	n
NC	103
MCI	130
AD	45
Total	278

Figure 3.2: Age distribution table

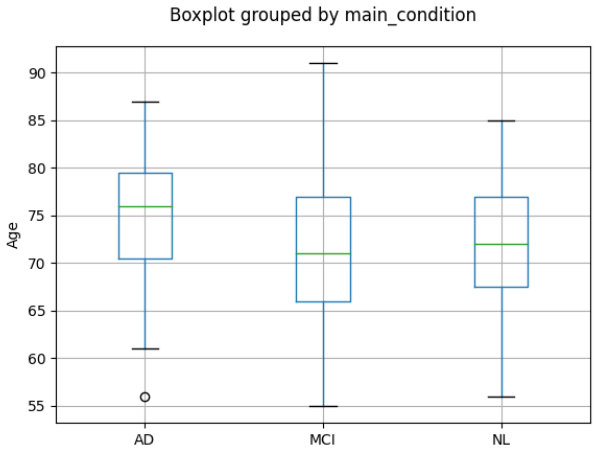
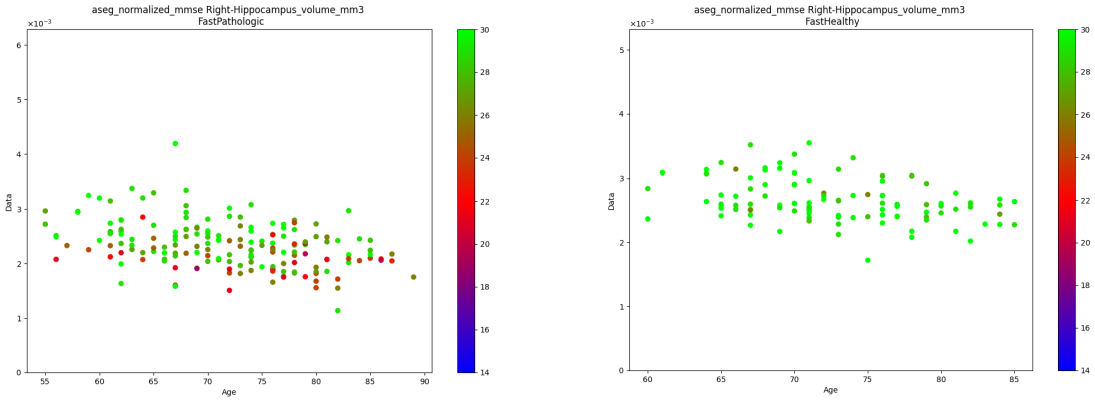


Figure 3.3: Box-plot of age distribution

3.4.1 Relationship with MMSE

In a plot that was implemented with the use of a MMSE colour coding, it can be seen how some regions show a trend, like the hippocampus, which as expected shows atrophy in correlation with a lower MMSE score. The MMSE score is not directly correlated to the pathology, and there is no threshold set to classify a subject as healthy based on this score as seen in the image below.



(a) Fastsurfer Pathologic **(b)** FastSurfer Healthy

Figure 3.4: MMSE score, pathology and volume

3.4.2 Comparisons between conditions

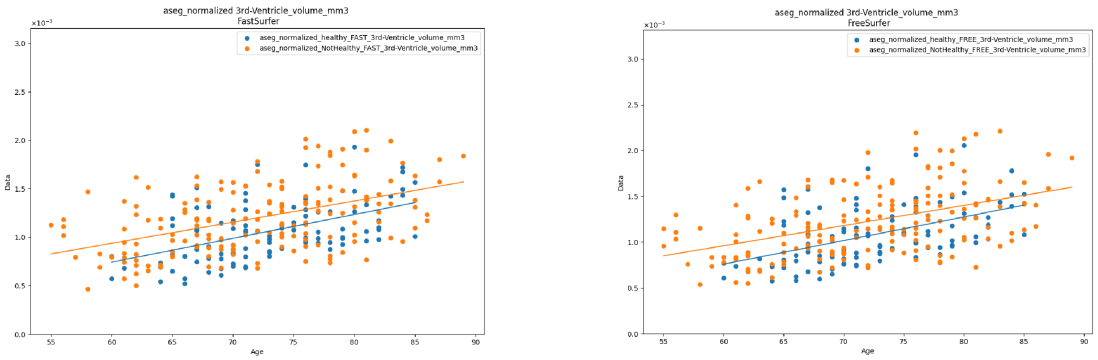
ASEG

In the ASEG atlas, certain brain volumes show significant differences between healthy and unhealthy subjects. These differences are consistent across both the traditional FreeSurfer method and the FastSurfer method.

In addition to the amygdala and hippocampus, the accumbens also tends to be larger in healthy subjects. In addition, certain regions exhibit more pronounced differences in unhealthy subjects. These include increased cerebrospinal fluid (CSF) volume and wider ventricles.

These informative features, the size differences in the amygdala, hippocampus, and accumbens, along with the increased CSF volume and wider ventricles, can be utilized to accurately estimate the health status of a patient.

It is worth noting that some regions, such as the hypointensities and the 5th ventricle, were not computed in this analysis. It is not surprising since these areas indicate defects in the image where it is not possible to see a brightness level, but the image is saturated. Usually they can be present for many reasons which include lesions, but not in every subject in a general population.



(a) FastSurfer **(b)** FreeSurfer

Figure 3.5: Comparison between conditions 3rd Ventricle ASEG

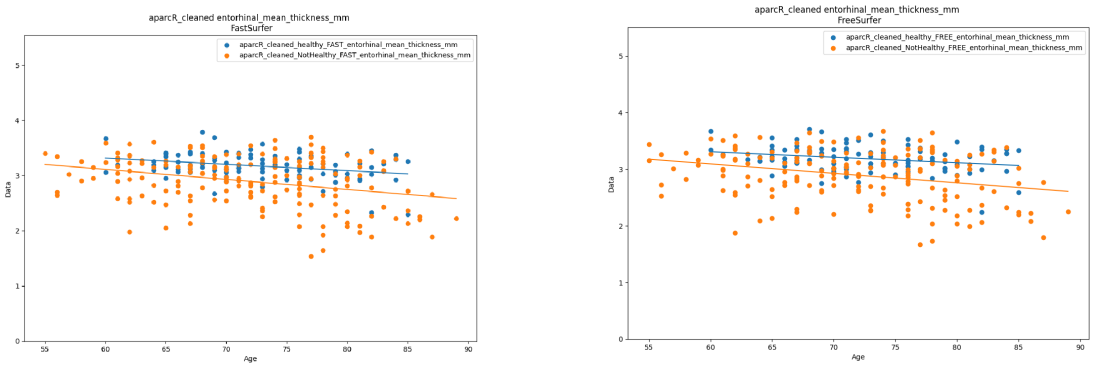
APARC

The analysis reveals that the thickness measurements, in general, do not provide significant information for distinguishing between healthy and unhealthy subjects. However, there are noticeable differences observed in the mean thickness of the entorhinal region and the mean area of the inferiotemporal region between these two groups.

It is important to mention that certain features, specifically the inferior-parietal mean area and thickness, were not computed for the unhealthy subjects in FastSurfer, but only by FreeSurfer.

Regarding the remaining brain areas, no significant differences were observed among the pathological subjects, and there were no anomalies in the behaviour of the software that are worth highlighting.

The statistical analysis supports the visual observations, confirming that the differences noticed are indeed significant.



(a) FastSurfer **(b)** FreeSurfer

Figure 3.6: APARCL enthorinal mean thickness estimation

3.4.3 Comparisons between methods

This section gives an overview of the macroscopic differences between the patients with a healthy brain compared to the pathologic patients.

ASEG

When comparing the two methods, it is observed that FastSurfer estimates higher volumes for the Caudate and Amygdala, particularly on the right side. Discrepancies are also noted in the Right Caudate Volume, Right Accumbens area volume, and Right Ventral DC volume.

Significant differences are found in the estimates of the Choroid Plexus volume, CSF volume, and Mask volume, all of which are overestimated in FastSurfer. The CC Anterior volume exhibits similar behaviour, as does the Right Choroid Plexus (indicating a symmetric issue), and the Left Ventral DC volume. Only FreeSurfer identified the Optic Chiasm and 5th Ventricle, while neither method detected the Left Vessel. The Left Accumbens area is overestimated in FastSurfer. The most significant difference is observed in the CC Central volume.

The opposite trend is found in the normalized mask, which also shows a little variance and a noticeable difference.

APARC

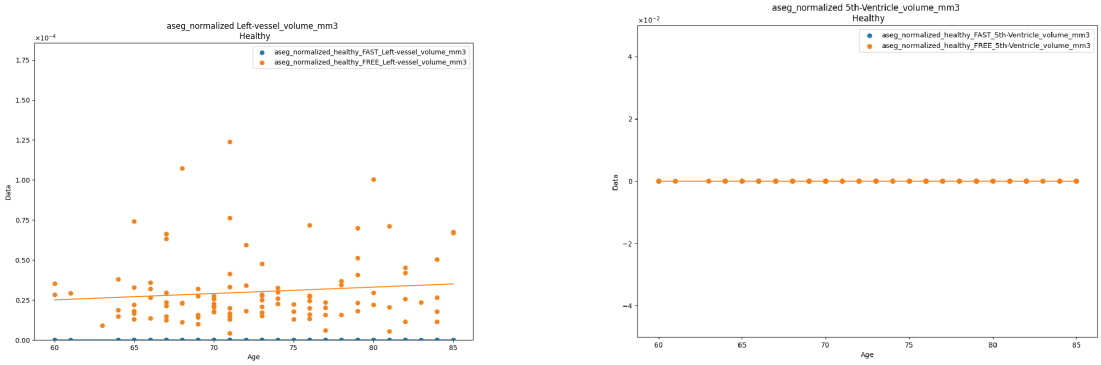
Considering the APARC atlas regions features, comparing FastSurfer and FreeSurfer a number of differences can be noticed, starting from healthy subjects. A number of regions are consistently overestimated by FastSurfer, such as the paracentral mean area, caudalmiddlefrontal mean area, transversetemporal mean area, isthmuscingulate mean area and fusiform mean area. Some regions were underestimated such as the cortex volume and the inferiortemporal mean thickness, enthorinal mean area. Of these, some more than others, for example the isthmucingulate mean area, and some less, such as the inferiortempoeral mean thickness. In general, a statistical difference can be noticed in many of the structures.

Some other areas, were found to be overestimated by FreeSurfer compared to FastSurfer. This is evident from the statistical tests performed. The areas that were overestimated by FreeSurfer include the parahippocampal mean area, superior frontal mean thickness, supramarginal mean area, lateral orbitofrontal mean thickness, pars opercularis, rostral middle frontal mean thickness, posterior triangularis mean thickness, and fusiform mean thickness as well as others. Some regions show a bigger difference than others on average, and in general there is no pattern of over or underestimation in the comparison between the two tools. It is important to note that for a specific goal it is not necessary that the performance on every region, but some of them may be more important than the rest in a specific application. In general, APARC atlas features are more alike between the two tools than the region volumes estimated.

Notable differences

The results show that there are some regions which are computed only by one of the two methods, in general these are computed by FreeSurfer and not FastSurfer.

- Hypointensities - only FreeSurfer
- 5th Ventricle - not any
- vessel Volume - only FreeSurfer
- Optic Chiasm - only FreeSurfer



(a) Hypointensities Healthy **(b)** 5th Ventricle Healthy

Figure 3.7: Hypointensities and vessel volume example

Some regions that show a very pronounced difference and constant, as for example the whole brain mask. It can be noted by the plot of the mask value. 3.8

Some regions show a difference which is less consistent between the processing with FastSurfer and FreeSurfer. It is reported the p-value again for the regions showed below as an example of what visually the result corresponds to.

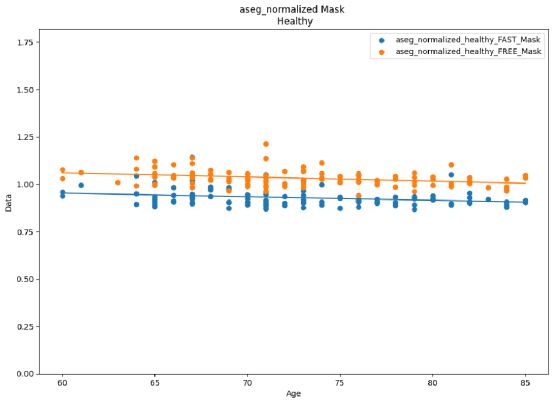
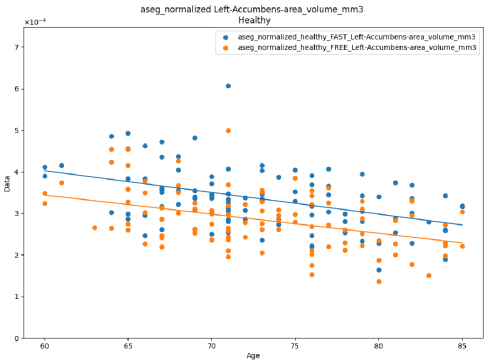
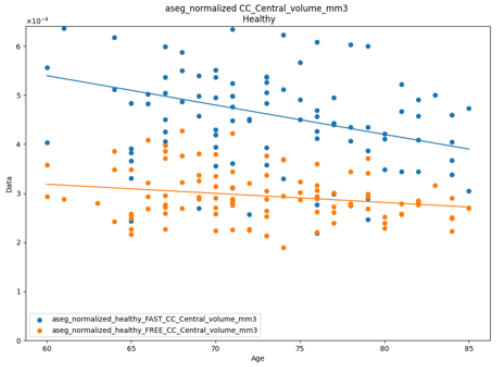


Figure 3.8: Comparison of the FastSurfer and FreeSurfer Output for the mask volume



(a) Left Accumbens Healthy



(b) CC Central Volume Healthy

Figure 3.9: Comparison between methods for the left Accumbens and the CC central volume

3.4.4 Bland-Altmann plot

The Bland-Altmann plots were plotted for every region, shows the agreement between measures. It is used to visually assess the reliability of a method usually compared to a gold standard, it can be used between methods to show the agreement of their measures. In this case it was computed only for the paired tests, so to compare the two methods and not to compare the two pathologies. Follow two examples of plots, of two regions with different ICC values shown in the table 3.1. The plot have the same scale of the y-axis proportionally to the mean values of the x-axis.

Region	Healthy	Pathologic
APARCR_entorhinal_mean_thickness_mm	0,856	0,943
Right-Lateral-Ventricle_volume_mm3	0,998	0,998

Table 3.1: ICC of the regions displayed in the plots

3.4.5 Violin plots

A violin plot depicts distributions of numeric data for one or more groups using density curves. The width of each curve corresponds with the approximate frequency of data points in each region. It was plotted for every region to compare both FastSurfer and FreeSurfer and the pathologic and healthy subjects.

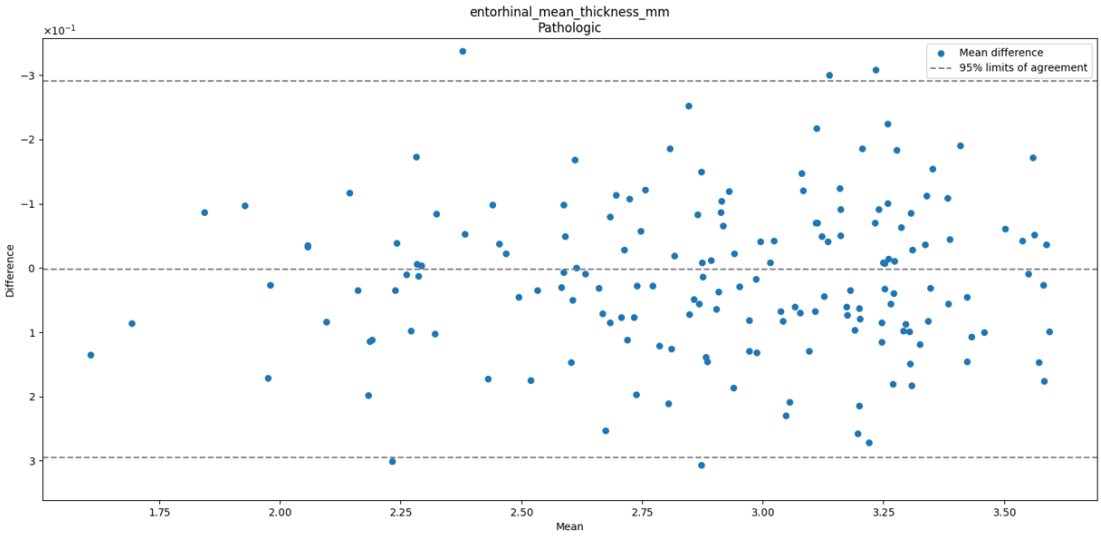


Figure 3.10: Bland-Altmann for enthorinal mean thickness, APARCR, Healthy

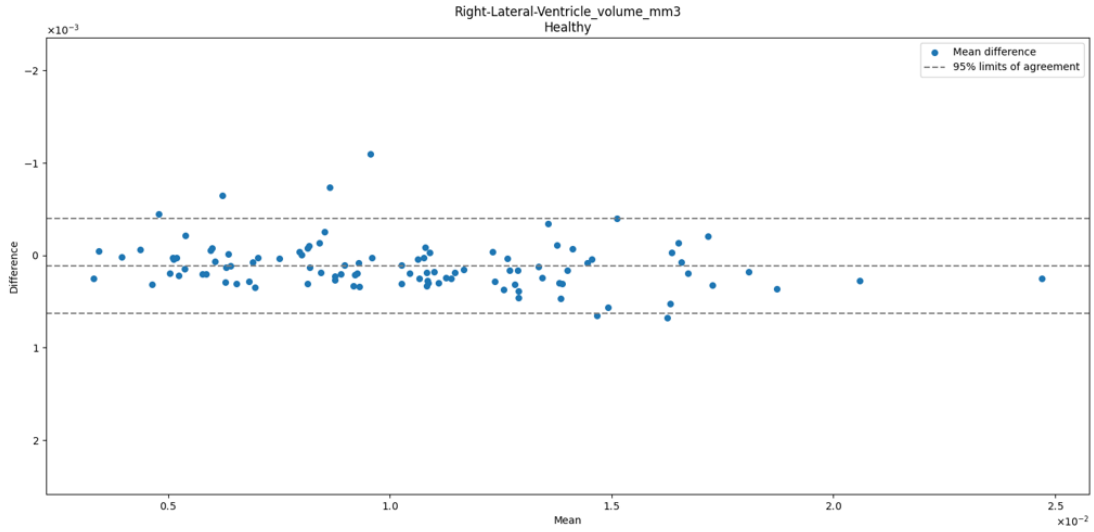


Figure 3.11: Bland-Altmann for Right-Lateral-ventricle volume, ASEG, Pathologic

3.5 Statistical analysis

Follow the results of the statistical analysis performed.

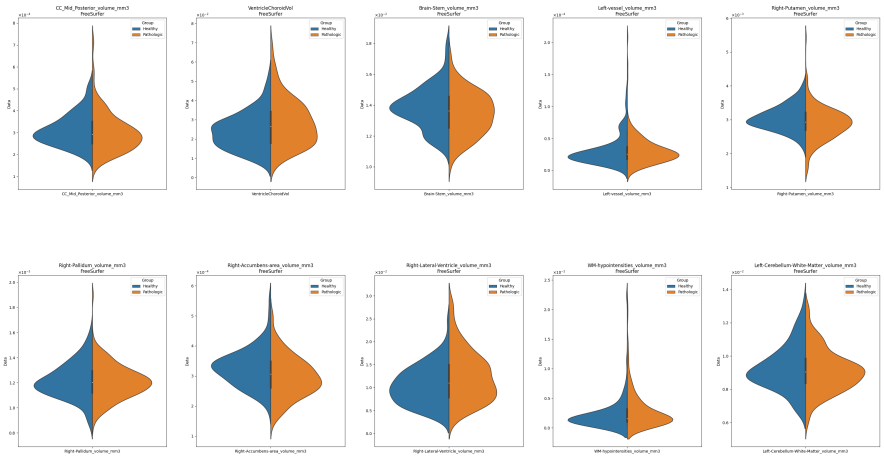


Figure 3.12: Example of Violin Plot for subjects processed with FreeSurfer

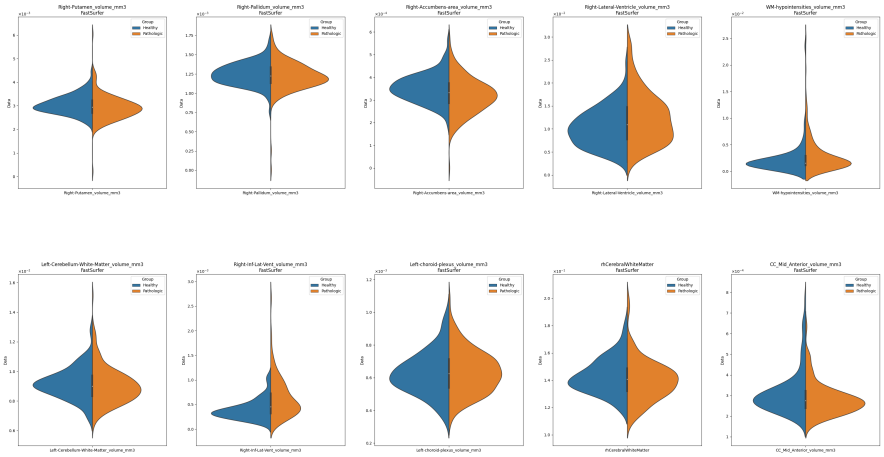


Figure 3.13: Example of Violin Plot for subjects processed with FastSurfer

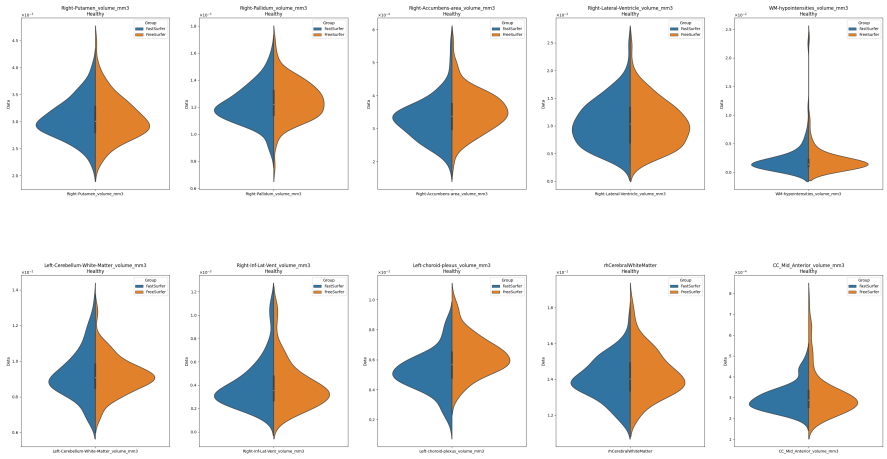


Figure 3.14: Example of Violin Plot for Healthy Subjects.

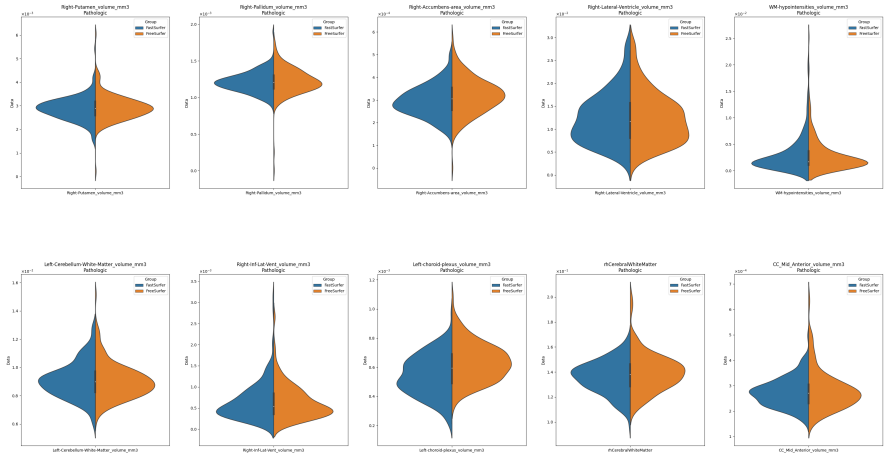


Figure 3.15: Example of Violin Plot for not Healthy Subjects

3.5.1 Statistical tests results

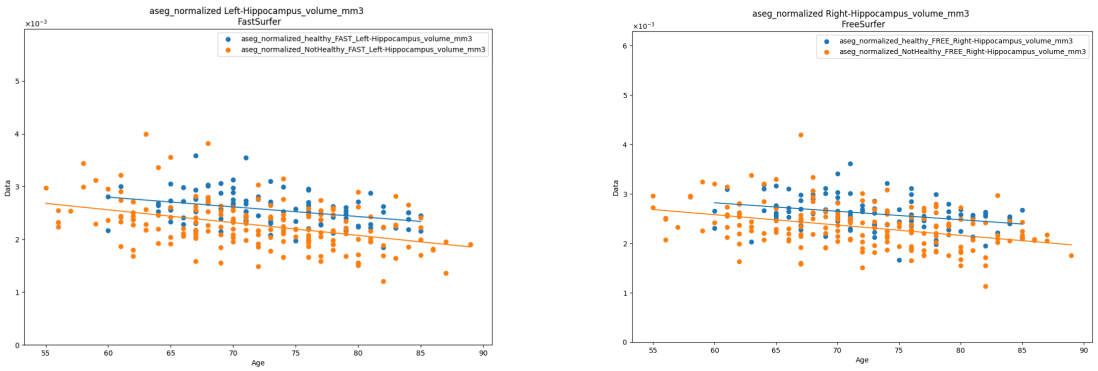
What can be observed from the results of the statistical tests is that both FastSurfer and FreeSurfer exhibit a significant difference in numerous regions. Moreover, the differences between the software packages are generally greater than the discrepancies between pathologies for subjects processed using the same software. These findings are based on the utilization of specific statistical tests, with the Wilcoxon test employed for paired comparisons of the software packages and the Mann-Whitney test used for unpaired comparisons between healthy and non-healthy subjects. Notably, the lowest p-value observed in the comparison between pathologies is e-11, whereas multiple negligible p-values ranging as low as e-30 are obtained when comparing the software tools. This may be accepted since what is more interesting in this analysis is the reliability between methods which is assessed by the ICC values, not always a low p-value relates to a low reliability of the measure.

According to the statistical tests results the regions which are the most informative when trying to discriminate between pathologic and healthy subjects are the same between the processing results of the two methods, the results for amygdala and the hippocampus, which are both in the literature and in the results shown to be the most indicative of a pathology are listed below in table 3.2. Other regions that showed little correlation are the inf-lat-vent volume and the enthorinal mean thickness, as well as the superiortemporal mean thickness.

Region	FreeSurfer	FastSurfer
Left-Amygdala_volume_mm3	5.92E-09	5.96E-11
Left-Hippocampus_volume_mm3	6.59E-10	1.22E-10
Right-Amygdala_volume_mm3	1.68E-08	2.39E-10
Right-Hippocampus_volume_mm3	9.25E-10	4.96E-10

Table 3.2: Results for Amygdala and Hippocampus

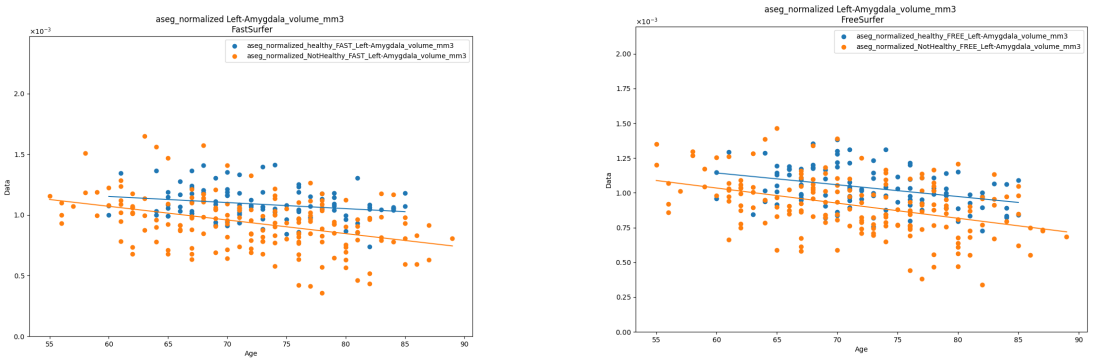
Below, the plots of the values of two areas chosen from the table above can be seen. It can be observed what those statistical differences mean visually below in image 3.16b and 3.17b.



(a) FastSurfer **(b)** FreeSurfer

Figure 3.16: Comparison between methods for Hippocampus

In general, it is important to note that a low p-value does not necessarily imply poor estimation of a specific area. In this context, the ICC holds greater relevance. Some areas have a poor p-value and a high ICC. The opposite is also true. For example, despite having a decent ICC, regions like the Accumbens may exhibit a very low p-value. Interestingly, this trend persists regardless of whether the average estimation is higher or lower. The table 3.3 shows the relationship between the lowest p-values and the corresponding ICC.



(a) FastSurfer **(b)** FreeSurfer

Figure 3.17: Comparison between methods for Amygdala

Region	Healthy stat test p value	Healthy ICC value
aparcR cleaned rostralmiddlefrontal mean thickness mm	1.26E-18	0.883
aparcL cleaned superiorfrontal mean thickness mm	1.26E-18	0.917
aparcL cleaned Cortex,MeanThickness	1.27E-18	0.952
aparcR cleaned lateralorbitofrontal mean thickness mm	1.30E-18	0.845
aparcL cleaned rostralmiddlefrontal mean thickness mm	1.30E-18	0.862

Table 3.3: Comparison between p-values and ICC

3.5.2 ICC

As stated before, the ICC was only computed for the agreement between methods. It can be noticed that in general the ICC values are not low, in fact Fastsurfer has a high reliability if the gold standard is considered to be Freesurfer. Some observations can be made. Firstly, the ICC is generally lower in pathologic subjects, which could be expected since the higher variability of brains in pathologies. In this case a significant disparity cannot be noticed, but it is still evident. There are some outliers that show a very low ICC. The lowest is in the CC central volume and it is 0,4. This region was already described before to show significant differences between the two methods in its estimation and this is a confirmation of that claim. For more than 100 out of 197 regions considered, the ICC is higher than 0,9 for healthy subjects, which indicates a good reliability. The lowest is in the CC Central volume, with a value of less than 0,5 in ICC.

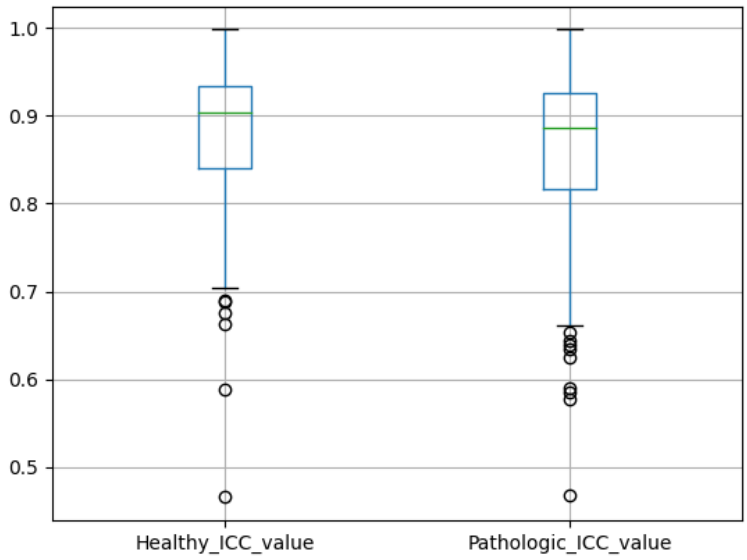
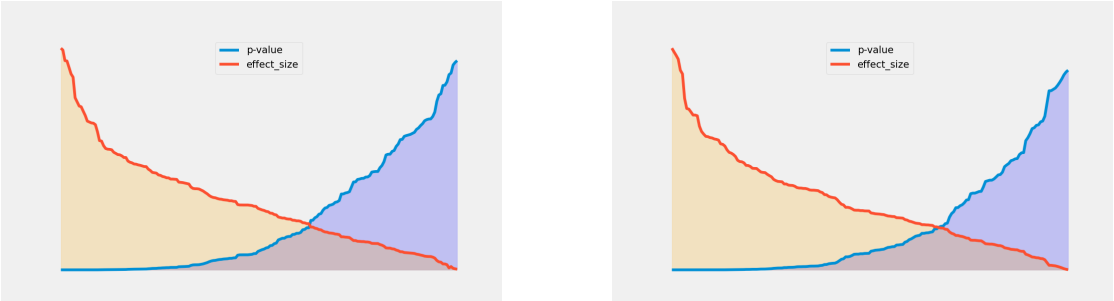


Figure 3.18: ICC values distribution

3.5.3 Effect size

The effect size was utilized to assess the magnitude of differences between the processed areas based on each method. Notably, the trend observed for effect sizes is opposite to that of p-values, which is expected and desirable as shown in figure 3.19. When comparing the areas plotted in order of p-values for FreeSurfer and FastSurfer, it becomes apparent that the p-values exhibit an inverse trend. In contrast, the effect sizes exhibit a pattern that aligns with the expectations.



(a) FastSurfer **(b)** FreeSurfer

Figure 3.19: Trend of variation for effect size and p-value

The effect size of the regions considered can then be thought of being inversely proportional to the p-value mentioned above, so we can expect it to be very high for these regions, as shown below in table 3.4. Furthermore it shows the effect direction as well, an insight that cannot be seen with FastSurfer.

Region	FreeSurfer	FastSurfer
Left-Amygdala_volume_mm3	-0.418	-0.470
Left-Hippocampus_volume_mm3	-0.443	-0.462
Right-Amygdala_volume_mm3	-0.405	-0.455
Right-Hippocampus_volume_mm3	-0.439	-0.446

Table 3.4: Effect sizes of some of the most informative regions

3.6 Combination of methods

In this section an example of how the methods mentioned above one can see the overview of how the information can be extrapolated. Below it is shown the Bland-Altmann, the violin plot and a scatter plot with the linear regression trend line.

Region	Right-Inf-Lat-Vent_volume_mm3
FreeSurfer_stat_test_p_value	5.55E-09
FreeSurfer_effect_size_value	0.418
FastSurfer_stat_test_p_value	3.76E-09
FastSurfer_effect_size_value	0.423
Healthy_stat_test_p_value	0.0614
Healthy_ICC_value	0.981
Healthy_effect_size_value	0.212
Pathologic_stat_test_p_value	0.0128
Pathologic_ICC_value	0.986
Pathologic_effect_size_value	0.216

Table 3.5: Values for the analyzed region

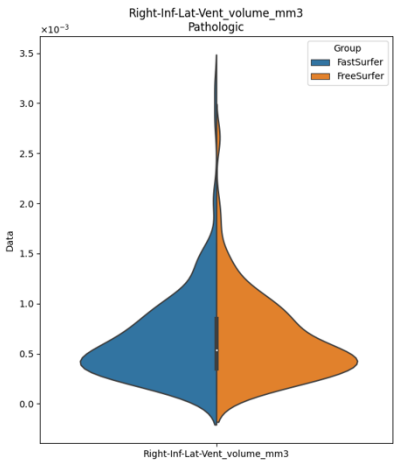


Figure 3.20: Violin plot

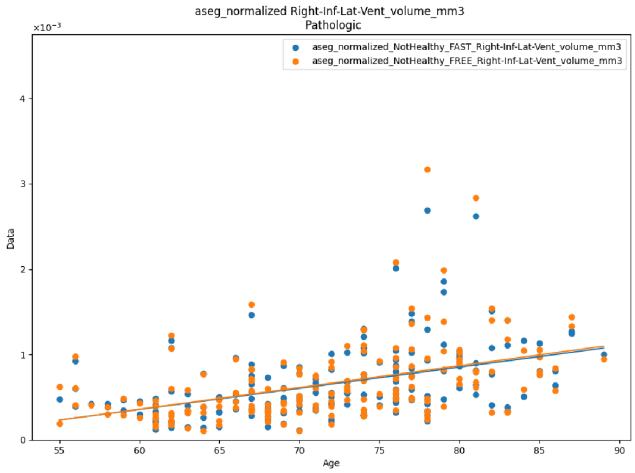


Figure 3.21: Linear regression

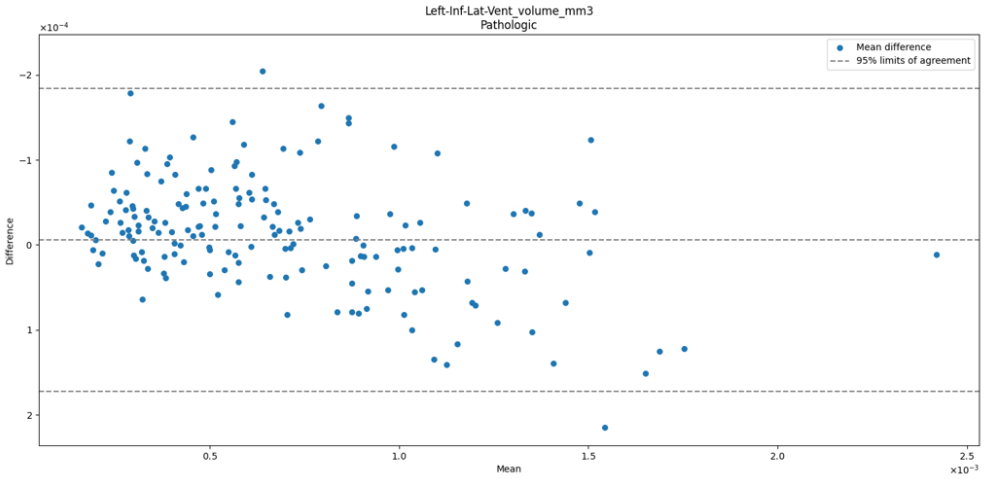


Figure 3.22: Bland-Altmann plot

Figure 3.23: Plots related to the analysed feature

3.7 Machine learning

The initial phase of the machine learning process involved selecting specific features and using a limited number of models. The metrics chosen for evaluation were accuracy and balanced accuracy, with the latter being more suitable for unbalanced datasets. Even though the dataset was unbalanced, it was later balanced for the training process.

Following the initial testing phase, a grid search was conducted to evaluate various algorithms. The grid search systematically tested each algorithm using different parameter combinations. To ensure the reliability of the process, a k-fold cross-validation with three repetitions was employed, allowing to assess the performance of the models on multiple subsets of the data.

3.7.1 Feature selection

The feature selection process was conducted manually, focusing on identifying the most informative features. This selection was based on a careful examination of the data, considering the agreement between FastSurfer and FreeSurfer results. It was observed that the regions showing the greatest differences were consistent between the two methods, further supporting the reliability of the selected features.

- Features used:
  - aseg_normalized_Left-Hippocampus_volume_mm3
  - aseg_normalized_Right-Hippocampus_volume_mm3
  - aseg_normalized_Right-Inf-Lat-Vent_volume_mm3
  - aseg_normalized_Left-Amygdala_volume_mm3
  - aseg_normalized_Right-Amygdala_volume_mm3
  - aparcR_cleaned_entorhinal_mean_thickness_mm
  - aseg_normalized_Left-Inf-Lat-Vent_volume_mm3
  - aparcR_cleaned_superiortemporal_mean_thickness_mm
  - aparcL_cleaned_superiortemporal_mean_thickness_mm

3.7.2 Grid search

The grid search was performed using the *GridSearchCV* function provided by Scikit-learn. This function enables training multiple models by testing all possible combinations of the specified parameters. To conduct the grid search, dictionaries of parameters were created for each model, and the function was called for each model separately. The results from each model were then aggregated into a single dataset. As mentioned earlier, the grid search was performed using k-fold cross-validation to ensure reliable and robust model evaluation.

Model	**C**	**gamma**	**Kernel**
SVM	0.1	0.1	linear
	0.1	0.1	poly
	0.1	0.1	rbf
	0.1	1	linear
	0.1	1	poly
	0.1	1	rbf
	0.1	10	linear
	0.1	10	poly
	0.1	10	rbf
	1	0.1	linear
	1	0.1	poly
	1	0.1	rbf
	1	1	linear
	1	1	poly
	1	1	rbf
	1	10	linear
	1	10	poly
	1	10	rbf
	10	0.1	linear
	10	0.1	poly
	10	0.1	rbf
	10	1	linear
	10	1	poly
	10	1	rbf
	10	10	linear
	10	10	poly
	10	10	rbf

Table 3.6: Grid Search combinations part 1

Model	C	max_depth	n_estimators
logistic	0.1	-	-
	1	-	-
	10	-	-
RF	-	-	50
	-	-	100
	-	-	200
	-	None	-
	-	5	-
	-	10	-

Table 3.7: Grid Search combinations part 2

The results of each test, including the standard deviation and average accuracy, were saved and recorded in an Excel table.

In some cases, the volumes processed with FastSurfer proved to be even more reliable for the goal of this problem.

3.7.3 Testing of best models

The best models for each category, FreeSurfer and FastSurfer, were tested using multiple metrics. This allowed a thorough evaluation of their performance in accurately classifying the data. Repeating the process numerous times, the reliability of the process could be tested.

3.7.4 Model selection

To select the model, a grid search was conducted to optimize the selection of features and parameters. This grid search utilized cross-validation, specifically a repeated k-fold validation with 5 folds, which was repeated 3 times. This configuration was chosen as it provides a good tradeoff between model performance and computational efficiency.

By performing the grid search with cross-validation, the model's hyper-parameters and feature selection were evaluated, taking into account the model's performance across different subsets of the data. This approach helps in selecting the most optimal combination of features and parameters, leading to a robust and reliable model for the given classification problem.

3.7.5 Results

The models selected for FreeSurfer and FastSurfer have the following performances, on average, on the training set, the measures are reported with the standard deviation across all the tests.

Dataset	Model	MCC	std
FastSurfer	RF	0,62	0,01
FreeSurfer	RF	0,65	0,01

The metrics computed on the test set, on average, are the following:

score	mean	std
best_score_training	0.65	0.01
accuracy	0.83	0.02
sensitivity	0.76	0.03
specificity	0.79	0.02
PPV	0.87	0.03
NPV	0.24	0.03
roc_auc	0.83	0.02
MCCscore	0.66	0.05

Table 3.8: FreeSurfer results

score	mean	std
best_score_training	0.62	0.01
accuracy	0.78	0.03
sensitivity	0.80	0.06
specificity	0.79	0.5
PPV	0.76	0.02
NPV	0.20	0.06
$\mathrm{roc}_a uc$	0.78	0.03
MCCscore	0.56	0.05

Table 3.9: FastSurfer results

On the MCC scores on the test set, the statistical test performed showed the following results:

Metric	**p-value**	**Null Hypothesis**
$\mathrm{best}_s core_t raining$	2.56×10^{-34}	Rejected
MCCscore	1.96×10^{-27}	Rejected
ROC-AUC	1.39×10^{-27}	Rejected

Table 3.10: Statistical tests results

3.8 Discussion

The results of the analysis indicate a slight advantage for FreeSurfer compared to FastSurfer. One possible explanation for this finding is that FreeSurfer demonstrates a higher sensitivity to inter-subject differences. This can be observed by examining the p-values and effect sizes of the 15 most informative regions generated by both software packages.

Interestingly, FreeSurfer appears to be particularly sensitive to small regions. In fact, almost all the regions that were exclusively computed by one software package were identified by FreeSurfer. This suggests that FreeSurfer may be more adapt at capturing subtle variations in brain structures, especially in localized areas.

Despite these differences, both FreeSurfer and FastSurfer demonstrate a high level of reliability overall. When the outputs of these software packages are utilized in practical applications, such as the classification of Alzheimer's disease patients, the results obtained from both methods are good. However, the they are still significantly better with FreeSurfer in this experiment, as shown by the table. 3.10

Both software packages offer valuable insights into the analysis of brain structures, and their outputs can be used reliably in various applications, including disease identification.

3.8.1 Problems and future improvements

Future improvements could dive into what pushes FastSurfer in overlooking some regions compared to FreeSurfer and if it would be important to work on it. On a more practical note, to improve the results of this work in the future, some steps that could be taken are:

- Optimizing the code or switching to a faster language than Python.
- Using a bigger dataset to obtain more reliable results.
- The machine learning process could be done more thoroughly, for example, by performing more experiments with different preprocessing method and implementing a feature selection process.

Chapter 4

Conclusion

Based on the analysis and research, it was concluded that FastSurfer represents a highly viable alternative to FreeSurfer in the task of brain parcellation. FastSurfer possesses numerous advantageous qualities, most notably its exceptional computational speed, making it a powerful tool for efficient and accurate brain analysis. Despite this, for the model analysed the conventional methods still obtained better results in the classification task and sensibility to small regions.

These results emphasize the significant advances made by deep learning in the field of brain imaging analysis. It has reached a stage where it can obtain comparable results to the conventional methods that are still regarded as the gold standard. Deep learning networks like FastSurfer can reduce processing times for brain parcellation often by several hours. This breakthrough would have important implications for clinical workflows, enabling faster and more efficient diagnosis and treatment planning.

Furthermore, the successful integration of FastSurfer into a pipeline similar to that of FreeSurfer, makes it a promising rival for contemporary software solutions. It is increasingly evident that tools like FreeSurfer may face limitations and possible obsolescence in various applications, given that similar results can now be obtained in a much faster manner.

The rapid advancements in deep learning-based methodologies mean that probably in the near future they will be powerful alternatives that will potentially match and surpass the contemporary technology.

In conclusion, this study highlights the efficacy of FastSurfer as an alternative to FreeSurfer for brain parcellation. It highlights the transformative impact of deep learning in the field of brain imaging analysis. Moving forward and further exploring the potential of these emerging technologies will undoubtedly lead to improved diagnostic accuracy, improved workflows, and enhanced patient care, and the technology has the potential to match and surpass the conventional softwares.

Bibliography

[1] Agustín Ciapponi- Erick Sanchez-Perez- Antri Giannakou- Olga L Pedraza- Xavier Bonfill Cosp- Sarah Cullum Ingrid Arevalo-Rodriguez- Nadja Smailagic- Marta Roqué-Figuls. «Mini-Mental State Examination (MMSE) for the early detection of dementia in people with mild cognitive impairment (MCI)». In: (July 2021). DOI: https://doi.org/10.1002/14651858.CD010783.pub3 (cit. on pp. 1, 17).

[2] Rudy J Castellani; Raj K Rolston; Mark A Smith. «Alzheimer Disease». In: (Sept. 2010). DOI: 10.1016/j.disamonth.2010.06.001 (cit. on pp. 1, 9).

[3] Di Ieva Antonio. *Brain anatomy - from a clinical and neurosurgical perspective: a clinically oriented manual of neuroanatomy*. English. Other. Copyright A. Di Ieva 2011 2018. Version archived for private and non-commercial use with the permission of the author. 2011 (cit. on pp. 4, 5).

[4] American Association Neurosurgeons. *Anatomy of the Brain*. 2023. URL: https://www.aans.org/en/Patients/Neurosurgical-Conditions-and-Treatments/Anatomy-of-the-Brain (visited on 2023) (cit. on p. 6).

[5] MRCP; Alex M. Rossor; et al Rhian Jenkins BSc; Nick C. Fox. «Intracranial Volume and Alzheimer Disease, Evidence Against the Cerebral Reserve Hypothesis». In: (Feb. 2000). DOI: 10.1001/archneur.57.2.220 (cit. on pp. 9, 12).

[6] Protima Khan; Md. Fazlul Kader; S. M. Riazul Islam; Aisha B. Rahman; Md. Shahriar Kamal; Masbah Uddin Toha. «Machine

Learning and Deep Learning Approaches for Brain Disease Diagnosis: Principles and Recent Advances». In: (Feb. 2021). DOI: 10.1109/ACCESS.2021.3062484 (cit. on pp. 9, 35).

[7] Richard A.I. Bethlehem; Jakob Seidlitz. «Brain charts for the human lifespan». In: (June 2021). DOI: https://doi.org/10.1038/s41586-022-04554-y (cit. on p. 10).

[8] Vincent Planche; José V. Manjon; Boris Mansencal; Enrique Lanuza; Thomas Tourdias; Gwenaëlle Catheline; Pierrick Coupé. «Structural progression of Alzheimer's disease over decades: the MRI staging scheme». In: (Apr. 2022). DOI: https://doi.org/10.1093/braincomms/fcac109 (cit. on pp. 11, 87).

[9] Robert Katzman, Robert Terry, Richard DeTeresa, Theodore Brown, Peter Davies, Paula Fuld, Xiong Renbing, and Arthur Peck. «Clinical, pathological, and neurochemical changes in dementia: A subgroup with preserved mental status and numerous neocortical plaques». In: *Annals of Neurology* 23.2 (Feb. 1988), pp. 138–144. DOI: 10.1002/ana.410230206. URL: https://doi.org/10.1002/ana.410230206 (cit. on p. 13).

[10] Yaakov Stern;a Carol A. Barnes;b Cheryl Grady;c Richard N. Jones;d and Naftali Raze. «Brain Reserve, Cognitive Reserve, Compensation, and Maintenance: Operationalization, Validity, and Mechanisms of Cognitive Resilience». In: (Nov. 2019). DOI: 10.1016/j.neurobiolaging.2019.03.022 (cit. on p. 13).

[11] Helen Christensen; Kaarin J. Anstey; Ruth A. Parslow; Jerome Maller; Andrew Mackinnon; Perminder Sachdev. «The Brain Reserve Hypothesis, Brain Atrophy and Aging». In: (Oct. 2006). DOI: https://doi.org/10.1159/000096482 (cit. on p. 13).

[12] Yaakov Stern. «Cognitive reserve: Theory and applications». In: (2013) (cit. on p. 13).

[13] Yaakov Stern. «Cognitive Reserve». In: (Mar. 2009). DOI: 10.1016/j.neuropsychologia.2009.03.004 (cit. on p. 13).

[14] Corinne Pettigrew; Anja Soldan; Yuxin Zhu; Mei-Cheng Wang; Timothy Brown; Michael Miller; Marilyn Albert. «Cognitive reserve and cortical thickness in preclinical Alzheimer's disease».

In: (Aug. 2016). DOI: https://doi.org/10.1007/s11682-016-9581-y (cit. on p. 13).

[15] Thaís Landenberger; Nicolas de Cardoso; Camila Rosa de Oliveira; Irani Iracema de L. Argimon. «Instruments for measuring cognitive reserve: a systematic review». In: (ago 2019). DOI: http://dx.doi.org/10.5935/1980-6906/psicologia.v21n2p58-74 (cit. on pp. 14, 18).

[16] Xujiao Chen; Genxiang Mao; and Sean X Leng. «Frailty Syndrome: An overview». In: (Mar. 2019). DOI: 10.2147/CIA.S45300 (cit. on p. 15).

[17] Dr Andrew Clegg MD; Prof John Young MBBS; Prof Steve Iliffe MBBS; Prof Marcel Olde Rikkert PhD; Prof Kenneth Rockwood MD. «Frailty in elderly people». In: (Mar. 2013). DOI: https://doi.org/10.1016/S0140-6736(12)62167-9 (cit. on pp. 15, 16).

[18] John E. Morley MD; Matthew T. Haren PhD; Yves Rolland MD; Moon Jong Kim MD. «Frailty». In: (Sept. 2006). DOI: https://doi.org/10.1016/j.mcna.2006.05.019 (cit. on p. 15).

[19] Matteo Cesari; Giovanni Gambassi; Gabor Abellan van Kan; Bruno Vellas. «The frailty phenotype and the frailty index: different instruments for different purposes». In: (Jan. 2014). DOI: https://doi.org/10.1093/ageing/aft160 (cit. on p. 18).

[20] Qian-Li Xue. «The Frailty Syndrome: Definition and Natural History». In: (Feb. 2011). DOI: doi:10.1016/j.cger.2010.08.009 (cit. on p. 18).

[21] Samuel D Searle; Arnold Mitnitski; Evelyne A Gahbauer; Thomas M Gill; Kenneth Rockwood. «A standard procedure for creating a frailty index». In: (2008 May). DOI: https://doi.org/10.1186/1471-2318-8-24 (cit. on p. 18).

[22] Donald B Plewes and Walter Kucharczyk. «Physics of MRI: a primer». In: *Journal of magnetic resonance imaging* 35.5 (2012), pp. 1038–1054 (cit. on p. 19).

[23] David Preston MD. *Magnetic Resonance Imaging (MRI) of the Brain and Spine: Basics*. 2006. URL: https://case.edu/med/neurology/NR/MRI%5C%20Basics.htm#:~:text=The%5C%

20most%5C%20common%5C%20MRI%5C%20sequences,longer%5C%20TE%5C%20and%5C%20TR%5C%20times (cit. on p. 19).

[24] Jeremy Jones. «T1 weighted image». In: (Mar. 2009). DOI: https://doi.org/10.53347/rID-5852 (cit. on p. 20).

[25] Peter J Kostelec and Senthil Periaswamy. «Image registration for MRI». In: *Modern signal processing* 46 (2003), pp. 161–184 (cit. on p. 21).

[26] Cheryl M Lacadie, Robert K Fulbright, Nallakkandi Rajeevan, R Todd Constable, and Xenophon Papademetris. «More accurate Talairach coordinates for neuroimaging using non-linear registration». In: *Neuroimage* 42.2 (2008), pp. 717–725 (cit. on p. 21).

[27] *Magnetism — mriquestions.com*. https://mriquestions.com/registrationnormalization.html. [Accessed 12-Jul-2023] (cit. on p. 22).

[28] Michael Strotzer. «One century of brain mapping using Brodmann areas». In: *Clinical Neuroradiology* 19.3 (2009), p. 179 (cit. on p. 22).

[29] Jean Talairach. «Co-planar stereotaxic atlas of the human brain». In: *3-D proportional system: An approach to cerebral imaging* (1988) (cit. on p. 22).

[30] Talairach Jean; Tournoux Pierre. *Co-Planar Stereotaxic Atlas of the Human Brain: 3-Dimensional Proportional System: An Approach to Cerebral Imaging - Hardcover*. Thieme, 1990. DOI: https://doi.org/10.1017/S0022215100111879 (cit. on p. 22).

[31] Matthew Brett, Ingrid S. Johnsrude, and Adrian M. Owen. «The problem of functional localization in the human brain». In: *Nature Reviews Neuroscience* 3.3 (Mar. 2002), pp. 243–249. DOI: 10.1038/nrn756. URL: https://doi.org/10.1038/nrn756 (cit. on p. 23).

[32] Ivana Despotović; Bart Goossens; Wilfried Philips; «MRI segmentation of the human brain: challenges, methods, and applications». In: (Mar. 2015). DOI: 10.1155/2015/450341 (cit. on p. 23).

[33] Bruce Fischl; David H. Salat; Evelina Busa; Marilyn Albert; Megan Dieterich; Christian Haselgrove; Andre van der Kouwe; Ron Killiany; David Kennedy; Shuna Klaveness; Albert Montillo; Nikos

Makris; Bruce Rosen; Anders M. Dale. «Whole Brain Segmentation: Neurotechnique Automated Labeling of Neuroanatomical Structures in the Human Brain». In: (Jan. 2002). DOI: https://doi.org/10.1016/S0896-6273(02)00569-X (cit. on p. 24).

[34] Rahul S. Desikan; Florent Segonne; Bruce Fischl; Brian T. Quinn; Bradford C. Dickerson; Deborah Blacker; Randy L. Buckner; Anders M. Dale; R. Paul Maguire; Bradley T. Hyman; Marilyn S. Albert; Ronald J. Killiany. «An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest». In: (Mar. 2006). DOI: 10.1016/j.neuroimage.2006.01.021 (cit. on p. 24).

[35] Xiang Feng; Andreas Deistung; Michael G. Dwyer; Jesper Hagemeier; Paul Polak; Jessica Lebenberg; Frédérique Frouin; Robert Zivadinov; Jürgen R. Reichenbach; and Ferdinand Schweser. «An Improved FSL-FIRST Pipeline for Subcortical Gray Matter Segmentation to Study Abnormal Brain Anatomy Using Quantitative Susceptibility Mapping (QSM)». In: (Feb. 2017). DOI: 10.1016/j.mri.2017.02.002 (cit. on p. 25).

[36] Simon B Eickhoff; Sarah Genon; B.T. Thomas Yeo. «Imaging-based parcellations of the human brain». In: (Oct. 2018). DOI: 10.1038/s41583-018-0071-7 (cit. on p. 25).

[37] Bruce Fischl. «FreeSurfer». In: (Aug. 2012). DOI: https://doi.org/10.1016/j.neuroimage.2012.01.021 (cit. on p. 26).

[38] John Ashburner et al. «SPM12 manual». In: *Wellcome Trust Centre for Neuroimaging, London, UK* 2464.4 (2014) (cit. on p. 30).

[39] URL: https://neuro-jena.github.io/cat12-help/ (cit. on p. 30).

[40] Mark Jenkinson, Christian F. Beckmann, Timothy E.J. Behrens, Mark W. Woolrich, and Stephen M. Smith. «FSL». In: *NeuroImage* 62.2 (Aug. 2012), pp. 782–790. DOI: 10.1016/j.neuroimage.2011.09.015. URL: https://doi.org/10.1016/j.neuroimage.2011.09.015 (cit. on p. 30).

[41] Brian B Avants, Nick Tustison, Gang Song, et al. «Advanced normalization tools (ANTS)». In: *Insight j* 2.365 (2009), pp. 1–35 (cit. on p. 30).

[42] IBM. *What is deep learning*. 2023. URL: https://www.ibm.com/topics/deep-learning (cit. on p. 35).

[43] Muhammad Imran Razzak; Saeeda Naz; Ahmad Zaib. «Deep Learning for Medical Image Processing: Overview, Challenges and Future». In: (Apr. 2017). DOI: MuhammadImranRazzak,SaeedaNaz,AhmadZaib (cit. on p. 36).

[44] Trevor Hastie; Robert Tibshirani: Gareth James; Daniela Witten; *An Introduction to Statistical Learning: with Applications in R*. Springer, 2019 (cit. on p. 37).

[45] Jian Wang; Hengde Zhu; Shui-Hua Wang; Yu-Dong Zhang. «A Review of Deep Learning on Medical Image Analysis». In: (Nov. 2020) (cit. on p. 40).

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. «Attention is All you Need». In: 30 (2017). Ed. by I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf (cit. on p. 40).

[47] Olaf Ronneberger; Philipp Fischer; Thomas Brox. «U-Net: Convolutional Networks for Biomedical Image Segmentation». In: (May 2015). DOI: https://doi.org/10.48550/arXiv.1505.04597 (cit. on p. 41).

[48] Abhijit Guha Roy, Sailesh Conjeti, Nassir Navab, and Christian Wachinger. *QuickNAT: A Fully Convolutional Network for Quick and Accurate Segmentation of Neuroanatomy*. 2018. DOI: 10.48550/ARXIV.1801.04161. URL: https://arxiv.org/abs/1801.04161 (cit. on pp. 42, 56, 58).

[49] Pierrick Coupé, Boris Mansencal, Michaël Clément, Rémi Giraud, Baudouin Denis de Senneville, Vinh-Thong Ta, Vincent Lepetit, and José V. Manjon. «AssemblyNet: A large ensemble of CNNs for 3D whole brain MRI segmentation». In: *NeuroImage* 219 (Oct.

2020), p. 117026. DOI: 10.1016/j.neuroimage.2020.117026. URL: https://doi.org/10.1016/j.neuroimage.2020.117026 (cit. on p. 43).

[50] Yuankai Huo et al. «3D whole brain segmentation using spatially localized atlas network tiles». In: *NeuroImage* 194 (July 2019), pp. 105–119. DOI: 10.1016/j.neuroimage.2019.03.041. URL: https://doi.org/10.1016/j.neuroimage.2019.03.041 (cit. on p. 44).

[51] Yeshu Li, Jonathan Cui, Yilun Sheng, Xiao Liang, Jingdong Wang, Eric I.-Chao Chang, and Yan Xu. «Whole brain segmentation with full volume neural network». In: *Computerized Medical Imaging and Graphics* 93 (Oct. 2021), p. 101991. DOI: 10.1016/j.compmedimag.2021.101991. URL: https://doi.org/10.1016/j.compmedimag.2021.101991 (cit. on p. 45).

[52] Mostafa Mehdipour Ghazi and Mads Nielsen. *FAST-AID Brain: Fast and Accurate Segmentation Tool using Artificial Intelligence Developed for Brain*. 2022. DOI: 10.48550/ARXIV.2208.14360. URL: https://arxiv.org/abs/2208.14360 (cit. on p. 46).

[53] Xin Yu et al. *UNesT: Local Spatial Representation Learning with Hierarchical Transformer for Efficient Medical Segmentation*. 2022. DOI: 10.48550/ARXIV.2209.14378. URL: https://arxiv.org/abs/2209.14378 (cit. on p. 47).

[54] M. Jorge Cardoso et al. *MONAI: An open-source framework for deep learning in healthcare*. 2022. DOI: 10.48550/ARXIV.2211.02701. URL: https://arxiv.org/abs/2211.02701 (cit. on p. 50).

[55] K. Yip and F. Zhao. «Spatial Aggregation: Theory and Applications». In: *Journal of Artificial Intelligence Research* 5 (Aug. 1996), pp. 1–26. DOI: 10.1613/jair.315. URL: https://doi.org/10.1613/jair.315 (cit. on p. 57).

[56] Michael W. Weiner et al. «The Alzheimer's Disease Neuroimaging Initiative: Progress report and future plans». In: *Alzheimer's &amp Dementia* 6.3 (May 2010), p. 202. DOI: 10.1016/j.jalz.2010.03.007. URL: https://doi.org/10.1016/j.jalz.2010.03.007 (cit. on p. 61).

[57] Daniel S. Marcus, Tracy H. Wang, Jamie Parker, John G. Csernansky, John C. Morris, and Randy L. Buckner. «Open Access Series of Imaging Studies (OASIS): Cross-sectional MRI Data in Young, Middle Aged, Nondemented, and Demented Older Adults». In: *Journal of Cognitive Neuroscience* 19.9 (Sept. 2007), pp. 1498–1507. DOI: 10.1162/jocn.2007.19.9.1498. URL: https://doi.org/10.1162/jocn.2007.19.9.1498 (cit. on p. 61).

[58] *Docker*. 2023. URL: https://www.docker.com/ (cit. on p. 62).

[59] *Nvidia$_d$ocker*. 2023. URL: https://docs.nvidia.com/datacenter/cloud-native/container-toolkit/latest/user-guide.html (cit. on p. 62).

[60] *Python*. 2023. URL: https://www.python.org/ (cit. on p. 64).

[61] *PyCharm*. 2023. URL: https://www.jetbrains.com/pycharm/ (cit. on p. 65).

[62] *Pandas*. 2023. URL: https://pandas.pydata.org/ (cit. on p. 65).

[63] *Matplotlib x2014; Visualization with Python — matplotlib.org*. https://matplotlib.org/. [Accessed 05-Jul-2023] (cit. on p. 65).

[64] *scikit-learn: machine learning in Python x2014; scikit-learn 1.3.0 documentation — scikit-learn.org*. https://scikit-learn.org/stable/. [Accessed 05-Jul-2023] (cit. on p. 65).

[65] K. A. FISHER. «Statistical Tests». In: *Nature* 136.3438 (Sept. 1935), pp. 474–474. DOI: 10.1038/136474b0. URL: https://doi.org/10.1038/136474b0 (cit. on p. 67).

[66] Priya Ranganathan. «An Introduction to Statistics: Choosing the Correct Statistical Test». In: (2021 May). DOI: 10.5005/jp-journals-10071-23815 (cit. on p. 67).

[67] Evie McCrum-Gardner. «Which is the correct statistical test to use?» In: *British Journal of Oral and Maxillofacial Surgery* 46.1 (2008), pp. 38–41. ISSN: 0266-4356. DOI: https://doi.org/10.1016/j.bjoms.2007.09.002. URL: https://www.sciencedirect.com/science/article/pii/S0266435607004378 (cit. on p. 67).

[68] R. F. Woolson. *Wilcoxon Signed-Rank Test*. Sept. 2008. DOI: 10.1002/9780471462422.eoct979. URL: https://doi.org/10.1002/9780471462422.eoct979 (cit. on p. 68).

[69] Patrick E. McKnight and Julius Najab. *Mann-Whitney scpU/scp Test*. Jan. 2010. DOI: 10.1002/9780470479216.corpsy0524. URL: https://doi.org/10.1002/9780470479216.corpsy0524 (cit. on p. 68).

[70] Cristiano Ialongo. «Understanding the effect size and its measures». In: (2016). DOI: 10.11613/BM.2016.015. (cit. on p. 69).

[71] Kenneth O Mcgraw; S.P. Wong. «Forming Inferences About Some Intraclass Correlation Coefficients». In: (Mar. 1996). DOI: 10.1037/1082-989X.1.1.30 (cit. on p. 70).

[72] Philip Sedgwick. «Multiple significance tests: the Bonferroni correction». In: (Jan. 2012). DOI: https://doi.org/10.1136/bmj.e509 (cit. on p. 71).

[73] Davide Giavarina. «Understanding Bland Altman analysis». In: *Biochemia Medica* 25.2 (2015), pp. 141–151. DOI: 10.11613/bm.2015.015. URL: https://doi.org/10.11613/bm.2015.015 (cit. on p. 72).

[74] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. «SMOTE: Synthetic Minority Over-sampling Technique». In: (2011). DOI: 10.48550/ARXIV.1106.1813. URL: https://arxiv.org/abs/1106.1813 (cit. on p. 81).

[75] Hervé Abdi and Lynne J Williams. «Newman-Keuls test and Tukey test». In: *Encyclopedia of research design* 2 (2010), pp. 897–902 (cit. on p. 85).