Politecnico di Torino

ScuDo
Scuola di Dottorato ~ Doctoral School
WHAT YOU ARE, TAKES YOU FAR

Doctoral Dissertation
Doctoral Program in
Computer and Control Engineering (XXXVI Cycle)

Analysis and optimization of disease monitoring via computational solutions

Sofia Ostellino

Supervisor(s):
Prof. Alfredo Benso, Supervisor
Prof. Gianfranco Politano, Co-Supervisor

Politecnico di Torino
July 24, 2024

I hereby declare that, the contents and organisation of this dissertation constitute my own original work and does not compromise in any way the rights of third parties, including those relating to the security of personal data.

Sofia Ostellino

Turin, 24 July 2024

Summary

The diagnose of neurological diseases presents a challenge, but disease monitoring does even more so: the evolution of the symptoms and the characteristics of a disease continues to evolve over the course of the patient's life. Here is where the development of computational methods can have a real impact on the quality of clinical care, as the heterogeneity of neurological conditions requires the integration of different types of clinical data. The aim of computational solutions is to solve real clinical needs and answer to questions that are part of the everyday life of doctors: this is why, during the Research here presented, there has always been a special consideration for the opinion of the experts, an attention to the state-of-the-art literature, and a scrupulousness towards solutions that are also flexible and scalable. The aim of the Research was therefore answer to a variety of questions, addressing practical and clinical needs:

- What are the main needs when developing computational solutions for the monitoring of neurological diseases such as MS and Alzheimer's disease? How straightforward is it to use public datasets and publicly available tools?
- Is it useful to use different types of clinical data for disease monitoring?
- Can the solutions developed for the other inquiries be applied to cover the needs of another field, such as the gene therapy with AAVs?

The first step was an in-depth analysis of the state-of-the-art to gain knowledge in the field. Afterwards, various sources of data were used to identify how the raw data could be processed, modified, or combined, to extract information via Deep Learning. Challenges and questions related to different topics were addressed adapting, according to the needs, the processing pipeline for the images, the methods of data collection, analysis and storage, and the deep learning frameworks trained.

The developed processing pipeline bridges the gap between heterogeneous datasets and their use for training deep learning networks. Its

performances are comparable to the state-of-the-art, but it does not rely on external plug-ins, as many of the existing tools, avoiding inconsistencies and versioning issues. Furthermore, it guarantees direct access to functions' parameters for customizations.

Moreover, the analysis and practical use of the ADNI database for exploring the feasibility of computational solutions for the monitoring of Alzheimer's Disease, allowed to explore the use of different clinical data for the prediction of disease conversion in time through of a Deep Learning framework. The use of tabular and imaging data, and the choice of specific types of imaging modalities was evaluated in different settings. The approach used to tackle the aforementioned topics was then transferred to the field of gene therapy, on the possibility of improving the analysis of the vectors that carry the genome of interest. The ITR structures, characteristic of every vector, are the focus of the analysis, given their importance and attention in the state-of-the-art: the use of the graphic method of the dotplots gave interesting results.

The methodology followed can be, in retrospect, defined effective: the research questions posed at the beginning lead to a in depth-analysis of the state-of-the-art both from a technical and a medical point of view. The approach required – since the beginning – to acquire hands-on experience on the tools proposed by the literature, and on the sources of data essential to the developing of the proposed solutions. The variety of topics that were subject of research constituted both a resource and a limitation: on one hand, they directed every developing step to be scalable and flexible and, on the other hand, they limited the available time to focus only on one research question/topic.

This Thesis is dedicated, with love and solidarity,

to all the students of Gaza

that will never see the day of their graduation.

Table of contents

List of Figures

List of Tables

List of Abbreviations

[A] AAV	Adeno Associated Virus
[B] AD	Alzheimer's Disease
[C] CI	Cognitive Impairment
[D] ICD	Invasive Ductal Carcinoma
[E] DL	Deep Learning
[F] DMT	Disease Modifying Therapy
[G] DTI	Diffusion Tensor Imaging
[H] MCI	Mild Cognitive Impairment
[I] MRI	Magnetic Resonance Imaging
[J] MS	Multiple Sclerosis
[K] NP	Neuropsychological
[L] PET	Positron Emission Tomography
[M] TP	Time-point

Chapter 1

Introduction

1.1 Computational solutions for clinical needs

Neurological diseases, such as Multiple Sclerosis (MS) and Alzheimer's Disease (AD), have different characteristics and evolution, and affect different populations, but, when it comes to disease diagnosis and monitoring, they share some aspects. The symptomatology, with its visible and subtle symptoms, presents a high degree of variability and heterogeneity between patients. Clinical protocols and clinical evaluation scales are valuable tools: they are standardized and backgrounded by years of scientific research and experience. To make a diagnose, clinicians must rely on a multiple set of tools that include, among others, magnetic resonance imaging, neuropsychological evaluations, psychomotor evaluations, and an accurate review on the patient's medical history. An early diagnose is fundamental to act promptly on the course of the disease, especially when it comes to chronic and progressive conditions that involve multiple physiological domains.

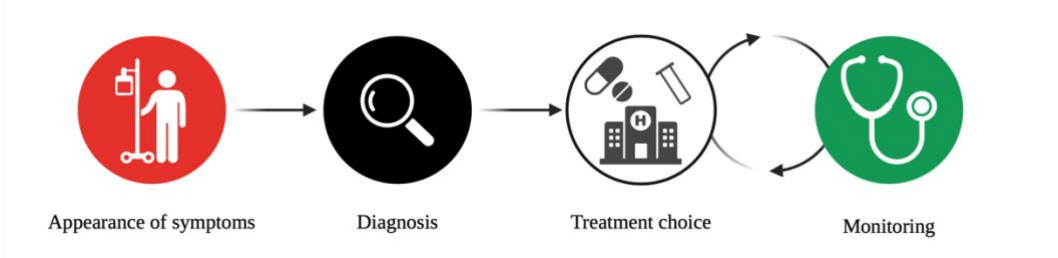


Figure 1 Graphical representation of the clinical workflow

As much as a diagnose of a disease presents a challenge, its monitoring is even more so: the evolution of the symptoms, their aggressiveness and manifestations vary inter and intra-patients, and the monitoring does not end at a well-defined moment – such as diagnosis – but continues and changes over the course of the patient's life, require a constant adjustment of the therapy. Here is where the development of computational methods can have a real impact on the quality of clinical care.

The heterogeneity of neurological conditions requires the integration of different types of clinical data: computational solutions can hence represent powerful tools when they are based on different and substantial sources of data, which is essential for the robustness of Artificial Intelligence solutions.

Many tools can be found, together with clinical scales, that already address these problems. Some support clinicians in the analysis of medical images, others allow an accurate analysis of computational neuropsychological (NP) tests, other help the clinicians during the visitation of patients, and so on. In the case of MS, for example, the Extended Disability Status Scale (EDSS) and the Multiple Sclerosis Functional Composite (MSFC) are two traditionally accepted clinical measures (COMs) of the disease severity and progression. Still, they cannot be considered complete, reliable, and exhaustive tools. Namely, EDSS is more used than MSFC but has some limitations, as EDSS scores can vary a lot due to the complex scoring rules and the subjectivity of the examiner. EDSS lacks linearity between score difference and clinical severity and it relies heavily on the evaluation of motor function and the ability to walk. Neither EDSS nor MSFC use MRI evaluation in their scores, even though MRI images are an important source of information, both for MS and AD for detecting lesions and signs of brain atrophy, for example.

The acceptance and efficacy of computational solutions that tackle many limitations of traditional methods are widely demonstrated, but we are far from saying that the technical challenges behind the implementation can be easily handled. In this context is thus crucial to adopt and implement tools that rely on multidimensional and composite prognostic biomarkers, taking into consideration the widest possible set of symptoms.
In order to efficiently monitor a disease, it is necessary to integrate different clinical data, following an approach that aims to getting closer to performing what is called personalized medicine, and that is yet an unmet clinical need.

1.2 Aim and challenges

The aim of computational solutions is, out of others, the need to solve real clinical needs and answer to questions that are part of the everyday life of doctors: this is why, during the research here presented, there has always been a special consideration for the opinion of the experts, an attention to the state-of-the-art literature, and a scrupulousness towards solutions that are also flexible and scalable.

Some of the technical challenges that are common when dealing with the mentioned applications include the need to use public and private datasets to increase the numerosity of the dataset used for algorithm training, as well as the need to handle different types of data that require different types of processing. When it comes to medical images, for example, many image formats are stored with different formats and different datasets are differently labelled and organized.

The aim of the Research was therefore answer to a variety of questions, addressing practical and clinical needs.

- What are the main needs when developing computational solutions for the monitoring of neurological diseases such as MS and Alzheimer's disease?
    - How straightforward is it to use public datasets and public available tools?
- Is it useful to use different types of clinical data for disease monitoring?
- Can the solutions developed for the other inquiries be applied to cover the needs of another field, such as the gene therapy with AAVs?

To answer the questions the following steps were followed:

- The main clinical needs related to the monitoring of neurological conditions were analysed together with the problems encountered when dealing with the processing of medical data and existing tools.
- The creation of a dataset for the integration of clinical data to predict with deep learning the conversion to Alzheimer's Disease in time.
- A 4-months internship at ProtaGene GmbH gave the possibility to transfer some of the solutions to a different and bioinformatic problem: the analysis of the structure of vectors for gene therapy.

1.3 Structure of the work and methodologies

In the following chapters, a brief overview of the neurological conditions that were the focus of this research will be given, and the literature will be reviewed as well. This has particular importance as, when approaching questions related to the medical field, a basic knowledge of the topic is needed to implement solutions that answer specific medical concerns. To better understand the characteristics of MS and Alzheimer's disease, an extensive review of the literature was conducted. Comparing existing solutions and dialoguing with clinicians took on particular importance. The clinicians (Giovanni Giulietti, Marco Bozzali and Laura Serra) from the Fondazione St. Lucia in Rome, Italy, played an important role.

Later, part of the research centred on a different topic, a bioinformatic problem, and on the possibility of applying the approach and solutions previously used on the forementioned issue to a different field. In this case, the field of analysis was related to the assessment of adeno-associated virus (AAV) vectors safety and vectors characteristics during the production of vectors for the delivery of gene therapy.

The opportunity to face the possibility of this transition was given by a 4-months visiting Ph.D. period as an internship student at ProtaGene GmbH, Heidelberg (DE).

AAV vectors are platforms for gene therapy that gained a lot of attention in the past years. They can be engineered (rAAVs) to be internalized in cells, so that the genome mounted in the vector can be integrated in the host genome to obtain the desired therapeutic effect. State-of-the-art literature points out how vector heterogeneity correlates with vector safety and functionality. AAVs are delimited by two structures called ITRs, and ITRs lay the focus of the project, as they correlate with vector heterogeneity and, if analysed with the proper tools, can reveal themselves to be very informative.

In current literature, it is possible to find many attempts to solve the urgent and complex need to provide a tool that can analyse vector characteristics during production. Most of the methods rely on NGS technologies, although these present many limitations such as long waiting times, for sequencing and data analysis, a scarce ability to quantify ITRs heterogeneity, and do not provide an informative data representation.

This is why it has been relevant to lay the foundations for a project aimed at the development of a tool for gene therapy vector analysis, based on a deep learning framework, that is independent of coverage analysis and NGS, and that can give an informative representation of results related to vector characteristics. Given the importance of ITRs, they were chosen as the focus of the analysis, together with a graphical representation method for the comparison of sequences, the dotplots. The challenges encountered in this project were in many ways like those faced when dealing with the development of deep-learning solutions for the monitoring of neurological conditions, where the use of medical images was one of the main points, as well as the need to deal with datasets not thought for these applications. Part of the processing pipeline developed for the MRI images was reused in this different scenario, as well as part of the concept for the creation of a dataset for the neural network.

More details and insights on the research summarized in this introduction will be found later on in Chapter 2 *Literature review*, and in Chapter 3 *Theoretical framework*.

Chapter 2

Literature review

2.1 Background knowledge

Computational methods can improve disease diagnose and monitoring, and many other medical science aspects. Before diving in the methodologies, it is important to contextualize the background on which this research lays.

2.1.1 Multiple sclerosis (MS)

Multiple sclerosis (MS) is a complex chronic autoimmune disease characterized by intra-patient and inter-patient variability in terms of disease course, progression, and efficacy of treatments. MS, like other neurodegenerative conditions, is a wide spectrum disease which multi-symptomatic characteristics require a comprehensive health evaluation at both diagnosis and follow-up visits. Disease monitoring is crucial and challenging as MS involves multiple physiological domains: there is not a single and widely accepted bio-marker informative enough to be used for efficiently planning a personalized treatment (Ostellino, 2022).
Some of the most common MS symptoms are listed in Figure 2, and they can arise at different degrees of gravity. Such symptoms can result in permanent disability or can manifest for short periods of time when there is disease activation (episodes called "poussè"), to later regress.

MS affects the central nervous system, causing myelinated axons to be attacked by the immune system, resulting in lesions that cause motor and cognitive impairment at different degrees of gravity. Its mechanism is yet not fully understood, and generally, given the progressive nature of the disease, an early diagnose in fundamental. As showed in Figure 2, when the myelin of the axons is attacked, lesions can appear both in brain and in spinal cord, being visible in magnetic resonance images.

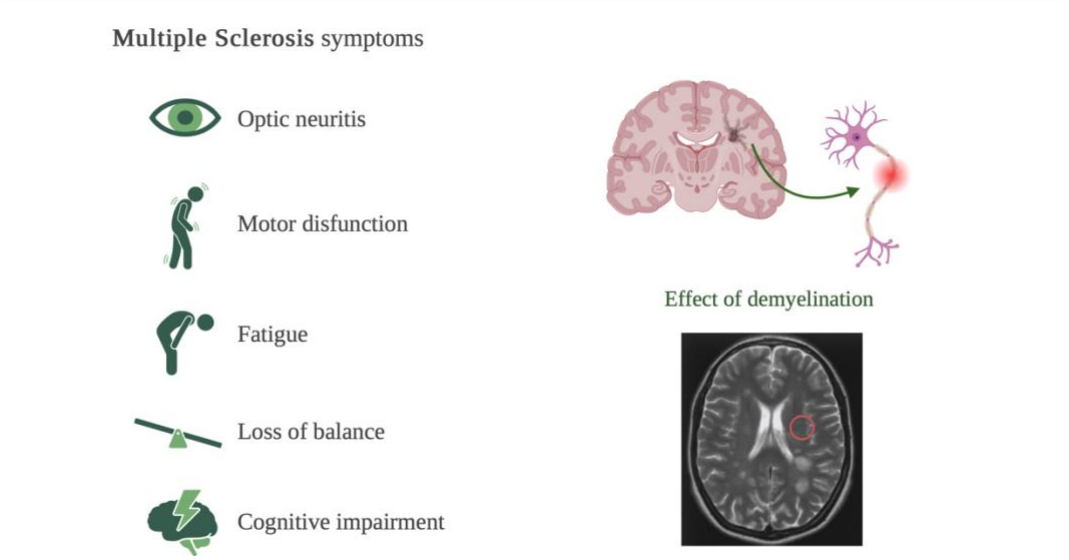


Figure 2 Most common MS symptoms and characteristics

MS is diagnosed combining evidence from several exams:

- MRI data: MRI imaging (of brain and medulla) helps assessing lesion load and brain atrophy: many MRI sequences are performed (such as T1-weighted images, T2-weighted etc...)
- Visual, motor and sensory components are examined with evoked potentials
- The cerebrospinal fluid (CSF) is analysed looking for signs of inflammation
- A general neurological examination is useful to assess reflexes and balance

Three MS clinical phenotypes are well-defined according to the time interval between relapses, the recovery capacity, and the accumulation of disability: Relapsing-remitting MS (RRMS) and Primary-progressive MS (PPMS) are the two main clinical sub-types that characterize the onset in the 85% and 10–15% of the cases, respectively; over 2/3 of RRMS cases will convert to Secondary-progressive MS (SPMS) within 10/15 years (I. Grossman, 2010).

Figure 3 shows an example of progression for different MS phenotypes: the onset is represented with the yellow symbol, and the green line traces the accumulation of disability in time. The lightning symbol stands for the poussès.

Several DMTs are available. DMTs aim at delaying the progression of the disease and the consequent accumulation of disability, primarily avoiding the formation of new lesions. It is crucial to choose the most suitable treatment for each patient, as well as to change it when it loses efficacy, minimizing the risk of administering the wrong combination of therapies, and losing precious time and economical resources. In current medical practice, therapies are often selected and modified by trial-and-error (Ostellino, 2022).

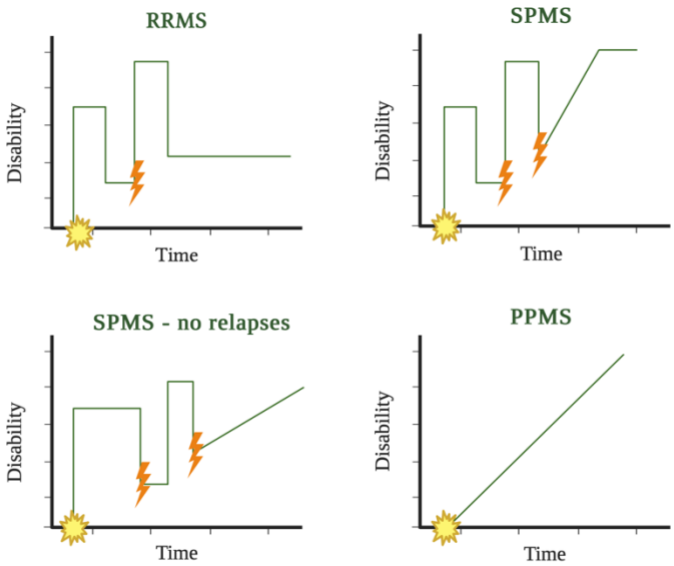


Figure 3 Graphical representation of MS phenotypes progression.

To give a number of the impact of MS on Italian population basing on the data available, 2.2. million people suffer from MS in the world, and 122.000 only in Italy, where MS affects a person every 500[1].

2.1.2 Alzheimer's Disease

Alzheimer's disease (AD) is the most common form of senile brain disorder which is caused by the beta amyloid peptide deposition (Murphy, 2010), as Figure 4 shows: the deposition of amyloid beta plaques results in AD progressive manifestation and in brain atrophy, which effects are visible with MRI images (Figure 4). Alzheimer's disease is not reversible and there is no existing cure, but the neuropathology related to Alzheimer's disease can be

[1] Italian Ministry of Healthcare, last update May 2019

detected several years before severe Alzheimer's disease clinical manifestations. The disease generally starts with a light deterioration of cognitive reserves and gradually worsens into a more severe form of dementia.

The initial phase of the disease is identified as Mild Cognitive Impairment (MCI) and has a wide spectrum that varies from cognitive difficulties that are challenging to detect to more evident cognitive deficits. Around the 10% and 15% of MCI patients per-year tend to convert to AD. Given the nature of its long, progressive, and variable prodromal phase, early discrimination of those patients that develop Alzheimer disease from those who manifest a stable MCI is fundamental. Therefore, administering a treatment prior to Alzheimer's disease conversion can efficiently decelerate its evolution.

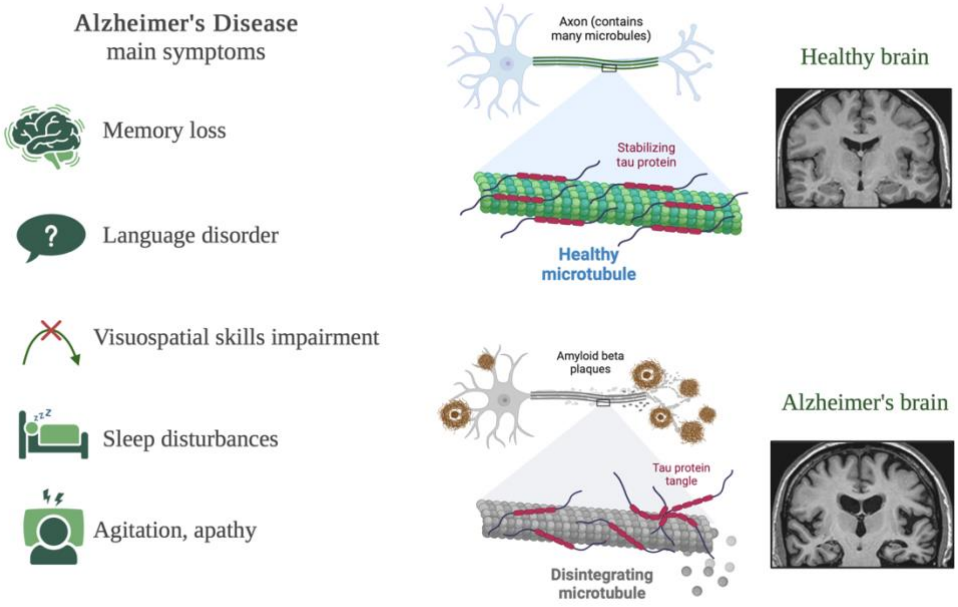


Figure 4 AD main symptom and example MRI imaging where brain atrophy is visible

Alzheimer's disease diagnosis is carried out relying on several clinical data, such as MRI structural and functional data, DTI imaging, neuropsychological tests' scores, genetic data, and others. Recent literature proposes various methods for AD detection that rely on Deep Learning principles. Some of them are focused on the classification of AD, MCI and control subjects using functional MRI or structural MRI data, others are based on multi-modality

imaging data. Not only AD detection is needed, but also the ability of algorithms to predict the probability of conversion from MCI to AD is required. Approaches that use complementary information and heterogeneous sources of data might have a decisive impact on the ability to early identify and consequently treat those subjects with a higher probability of developing Alzheimer. More than 600.000 people - only in Italy - suffer from Alzheimer Disease dementia, and 1 million suffer from other form of dementia[2]. The estimated number of people involved in the care - relatives, caregivers etc. - is more than 3 million people, only in Italy.

2.1.3 AAVs and gene therapy

Adeno-associated-virus (AAVs) vectors are the most used platforms for delivering gene therapy. They consist in a protein capsid and a single-stranded DNA of circa 4k bases. The genome is encapsulated by two T-shaped structures called ITRs (inverted repeat sequences), as exemplified in Figure **5**. When AAVs are engineered, the genome of interest is inserted between the ITRs, replacing the viral genome, following a procedure which details are beyond from the purposes of this introduction, but that allows the vector to be internalized in the cell, together with the genome of therapeutic interest.

Gene therapy is based on the introduction of the desired genome into patients to alter gene expression or the expression of proteins (Au, 2022), to correct disease mutations in three ways: replacing the defective gene with a functional copy, silencing the mutated version of the gene, and adding a therapeutic gene or synthetic construct (Au, 2022). ITRs can form two configurations, named *Flip* and *Flop*, depending on the position of the B-B' and C-C', as illustrated in Figure **6**.

[2] Italian Ministry of Healthcare, last updated September 2023

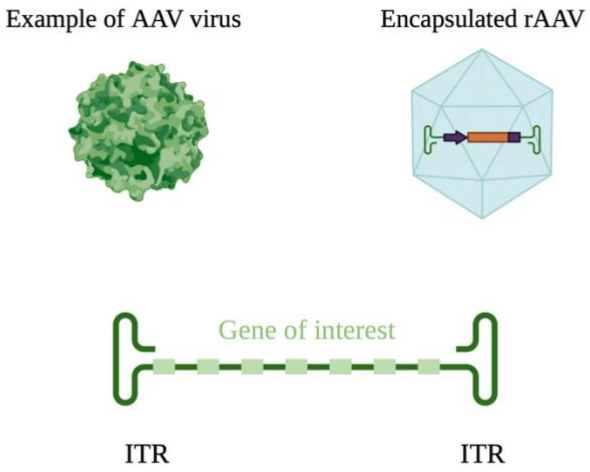


Figure 5 Graphical representation of wild-type AAVs, engineered rAAVs, and rAAV sequences

Of the existing wild-types AAVs, AAV2 is the most frequently used for these applications, alongside with AAV1, AAV8, and AAV9. There are many needs related to the production of AAVs, especially related to the assessment of vector quality and safety during vector production. As ITRs have an impact on genome packaging and on viral replication, they have been of particular interest in recent literature (later discussed in Section 2.2.2 AAVs vectors safety assessment: state-of-the-art.

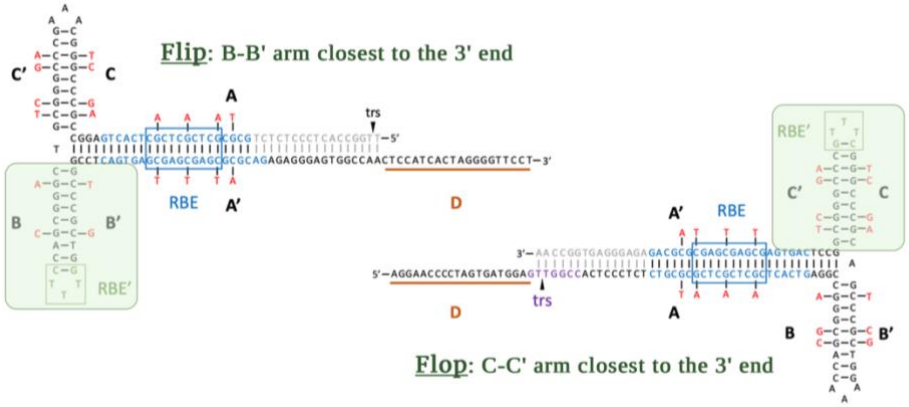


Figure 6 Description of the ITRs sequences (Flip and Flop configuration)

Vector safety is associated to the heterogeneity of the prepared vectors, and vector heterogeneity is strongly bound to ITRs characteristics.

The assessment of the quality and safety during production is one of the main challenges of gene therapy. Up-to-date methods base on the use of next-generation sequencing (NGS) technologies to quantify and investigate the characteristics of preparations, for example to identify mutations or the heterogeneity of the entire vector. Recent studies (S. Namkung, 2022) have demonstrated the feasibility of obtaining full-length resolution with NGS, covering ITR-to-ITR.

The problem of using NGS, however, is the long waiting time, and the difficulty in characterizing ITRs heterogeneity, that is very informative for assessing vector safety. Current literature points out the importance of targeting ITRs during the analysis, and of providing an informative result representation.

2.2 Literature review

The background context provided in Section 2.1 Background exposes the many needs belonging to several aspects of medical science. On one hand, there are challenges related to the diagnosis and the monitoring of neurological conditions. The two diseases hereby discussed are both common and heterogeneous conditions, united by the lack of a single clinical measure informative enough to assure easy diagnosis or monitoring and tailored therapeutic interventions. On the other hand, some issues related to the gene therapy field are discussed, regarding the necessity of assessing vector quality and safety. These issues are all due to real needs that physicians and researchers face every day.

2.2.1 Diagnose and monitoring of MS and AD: state-of-the-art

Personalized medicine is one of the focuses of current literature: nowadays, healthcare facilities and providers have to face an increasing number of patients that suffer from conditions such as Alzheimer's Disease or Multiple Sclerosis. Chronic conditions that can interests patients from a relatively young age, present a challenge in terms of money, time, and quality of provided assistance. To comprehend all the domains impaired by such diseases, it is common practice to put side by side different clinical examinations. When it comes to the diseases that are object of this dissertation, the assessment of cognitive impairment (CI) assessed via NP tests, for example, plays a fundamental role for both and can provide reliable insights on disease progression. Yet, personalized medicine for such conditions is an unmet clinical need.

Any computational solution that wants to tackle these issues, integrating the knowledge extracted from many clinical evaluations, must be trained on a huge amount of data. We can therefore here introduce another important point of this research, that is the use of private and public collection of medical

data for these purposes. Many public datasets exist, the most relevant ones are: MSBase[3], the Italian Multiple Sclerosis Registry[4], and ADNI[5].

State-of-the-art-literature was analysed targeting those works that couple CI assessment with NP tests with the evaluation of other clinical features and measures such as MRI imaging, motor and gait functions, and others. Some of the neuropsychological tests usually administered to MS patients are the SDMT, the PASAT, the CANTAB and CAB, the MSPT, and the MS Suite Test.

These tests can be computerized, to overcome the limitations of the paper-and-pencil administered tests, with proved accuracy and reliability. The administration of computerized tests brings many advantages, allowing a reliable as well as systematic collection of data, and a standardization of tests administration. D'Amico et al. (E. D'Amico, 2020), for example, showed how the performances of the CAB test battery correlate with measures of brain volume, and suggest the possibility of investigating brain pathology via cognitive assessment. A faster CI exam is the SDMT, a 15-minute-long computerized test developed starting from the MS suite Test: Pham et al. (L. Pham, 2020) showed how this test correlates with the T2-w lesion load.

These are just two examples of interesting studies that show how it is possible to correlate different clinical examinations, and more example can be found here (Ostellino, 2022).What is relevant to point out is that computerized NP assessment paired with imaging evaluation, stands as an informative and reliable procedure against the use of a single clinical measure.

Another matter that emerges when diving into the literature, is that it is not always simple using different dataset, especially when it comes to medical images. This specific problem was addressed, as described in Chapter 4 *Methods*.

Many tools and software exist to bridge the gap between the creation of deep learning applications and the sources of data needed for model training. It is typically not possible to feed a neural network with raw medical images, for several reasons. The raw images require a variable number of pre-processing, and the raw images are stored in formats such as NIfTI (.NIfTI) or DICOM (.dcm), and not in standard formats such as PNG or JPEG.

[3] https://www.msbase.org
[4] https://registroitalianosm.it/en/
[5] https://adni.loni.usc.edu

Here is a list of the most used tools for image processing:

- CBS Tools
- JIST
- TOADS-CRUISE
- BrainSuite
- Volbrain

The evaluation of these tools allowed to identify the following limitations: they lack in easy customization, and have often strict requirements in terms of settings, making it difficult to simply obtaining structured and uniformed data-sets (Ostellino, 2022).

What is in this sense needed is, firstly, a solution that allows an easier integration of imaging data coming from different sources. Chapter 4 *Methods* describes the pipeline that was developed, to be later used as a backbone of a deep learning architecture that is aimed at a second need, being providing clinicians with a tool for disease monitoring, able to integrate various types of clinical evaluations (also described in Chapter 4 *Methods*).

2.2.2 AAVs vectors safety assessment: state-of-the-art

Medical images, as described above, represent a crucial part of clinical works. As introduced in Section 2.1.3 AAVs and gene therapy, in the gene therapy with AAVs vectors field there are many needs, one of these being having a tool that can be used during vector production for vector analysis. Recent literature explores the possibilities of targeting ITRs to extract information about the properties of the manufactured vectors. ITRs are indeed interesting, and their functioning is yet not fully understood (K.I.Berns, 2020).This lack of a discussion regarding the applied knowledge about ITRs in vector production and quality assessment (P. Wilmott, 2019), opens questions about how to improve the analysis of vectors to estimate their safety.

To fill this gap, the idea of *medical image* was extended to comprehend the dotplots, as illustrated more in detail in Chapter 3 *Theoretical framework*. In this sense, dotplots are not only a graphical method to confront sequences, buy they are also a depiction of the structure of the vectors and, in particular for the sake of this analysis, of the ITRs.

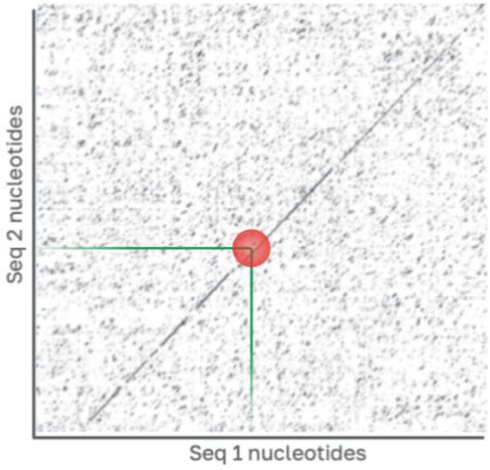


Figure 7 Example of a dotplot where a matching between sequences is highlighted in red

Figure 7 shows an example of a dotplot: on the axis of the dotplot there are the two sequences of nucleotides. The graph is seen as a matrix filled basing on nucleotide matching: the gaps and the adjacent lines in the graph show matches, missing bases and other characteristics of the two sequences. Moreover, the two sequences can be the same, or two different ones: this allows to extract information basing on the design of the plot (see Section 3.3.4 The dotplots and the FlexiDot tool for details).

The method proposed by Tran et al. (Tran, 2020), for example, analyses a population of vectors relying on coverage analysis - like most state-of-the-art works - managing to get several information about the preparation. One of the main findings is that the presence and type of ITRs is informative of the preparation quality: a 1-to-1 ratio of ITRs in flip and flop configuration, for example, characterizes an ideal preparation. The results visualization, as showed in Figure 8, raises several questions. The problem with this type of representation is that is not very informative per-se.

Another noteworthy work is the one by Zhang et al. (J. Zhang, 2022), where the authors explore the importance of assessing the quality of administered rAAVs to identify the presence of contaminants. The need of creating a tool that can be used in practice during vector production with a focus on ITRs is therefore not only urgent, but also relevant. The examples that were brought hereby, rely on coverage analysis: the possibility of relying on another paradigm is later presented in Chapter 4 *Methods* where the approach targets the analysis of AAVs, providing an informative result representation.

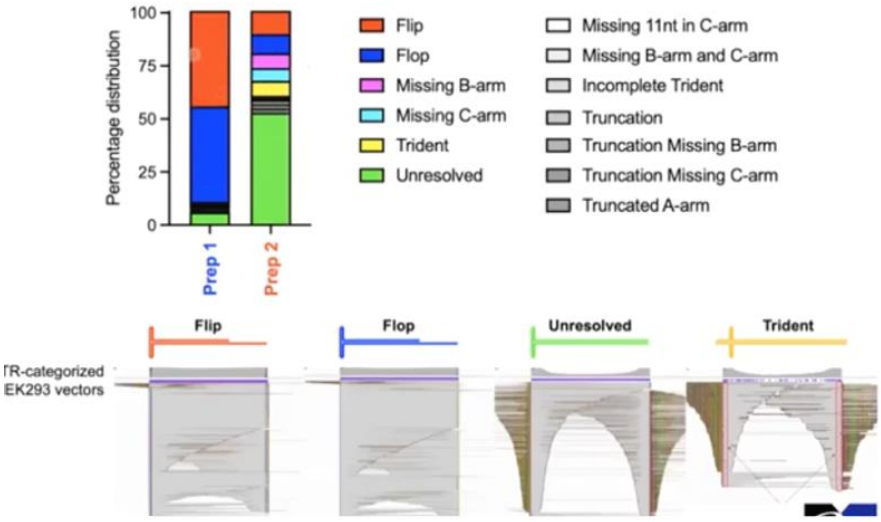


Figure 8 Result representation provided by the Tran et al. paper

Chapter 3

Theoretical framework

The concepts introduced and explained in detail in Chapter 1 *Introduction* and Chapter 2 *Literature review* serve as base for the following Chapter, which aims to present the assumptions, concepts and tools related to the development of this Research. Given the variegate number of topics that is covered, it is worth highlighting once more the common thread that ran throughout the work. Right from the start, some assumptions regarding the use of Artificial Intelligence in healthcare were grounded: the application of AI in this field must be treated consciously, given the common concerns related to privacy, biased results and the trust granted by clinicians. Moreover, the efficacy of the use of AI in clinical settings its not only determined by its performances in terms of accuracy, but also by how and at what extent these applications are accepted by the medical field (Asan, 2020). A well-trained CAD system does not divest doctors of any responsibilities, as they remain the ultimate decision-makers.

Medical imaging is an indispensable tool for the diagnose and monitoring of many diseases. However, different imaging techniques require diverse steps for the processing of the images, as they present different artifacts. Many solutions are proposed to automatise the pre-processing of images, and despite the high performances of many of these solutions, it is a task still - often - preferably performed by hand (Dhar, 2023).

From these findings we can draw two ideas:

- many restrains to the use of AI in clinical practice is often related to a lack of trust
- many existing tools are not easy to use or exchangeable between applications.

During the Research here presented, these two points were always kept into consideration. Based on the recent literature that examines the challenges of the application of AI in healthcare, one way to improve the trust in AI is to favour explainable CADs, that provide a visual interpretation of the results (Dhar, 2023). Moreover, deep learning methods require inputs in a structured manner, and it is not easy to have data already organized in such a way to be straightforwardly used for AI (Asan, 2020). This line of thought was followed when identifying the dotplots as a valid tool for the analysis of AAVs vectors, and when developing solutions related to the use of MRI images for deep learning, as well as the access to well-known public datasets.

Innovating the medical field with the help of AI is a common goal in the scientific community, keeping in mind that real necessities are those that need to be solved, and it is often minor inconveniences related to the implementation and the collection of data to represent the biggest challenges. Combining clinicians' knowledge and experience with the analytical power of AI have a disrupting effect on the quality of provided healthcare.

3.1 The dialogue with clinicians

During this Research, especially at its earliest stages, a close dialogue with clinicians had a huge role in implementation choices, as much as state-of-the-art literature. In the first place, about the use and processing of MRI images, the analysis of the literature helped understanding the innovations in this field do exist, but few of them where actually possible to use for the scope of the research. Such scope was clarified by discussing with the physicians about what they usually look for during the monitoring of a neurological disease, or about what they find more challenging.

Taking the importance of medical images in a clinical work-flow as read , in order to work on a deep learning architecture able to deal with clinical data of different kind, it was important to consolidate a backbone for the use of medical images as input to a network, and to understand how to efficiently extract information from a public dataset, as described in detail in the next chapter.

3.2 The collaboration with ProtaGene GmbH

The introduction of AI in the bioinformatic workflow, in the context of a company and of the needs of clients, can follow a similar approach to the introduction of AI in the clinical workflow. In this specific case, during the collaboration with ProtaGene GmbH, the task was understanding how a specific type of image could be use with the aim of improving the analysis of AAV vectors quality for gene therapy. The foundations of this project were laid together with Raffaele Fronza.

The main question was targeting the problem using as few resources as possible: methods for vector quality assessment already exist, and their limitations had to be addressed. Moreover, this needed to be done without necessary relying on experimental data related to previous projects, not to incur into data confidentiality issues. Another important point was changing the approach towards this problem, being alignment free and providing an informative representation of results, to be scalable to many of the requirements related to the analysis of vectors for gene therapy.

ITRs were, for this reason, perfect candidates for these tasks, and for the construction of the backbone theory for this project, opening a series of questions regarding the use of the NCBI public dataset, and the translation of the information presented in form of a sequence of bases, into an image that could be considered informative.

3.3 Tools and theories: state-of-the-art

The following section will present the tools and Python[6] libraries used for this Research, to able the reader to dive more easily in Chapter 4 *Methods*. Python was used as a programming language for every implementation, given its popularity, richness of libraries, and possible applications.

3.3.1 MRI images processing

Being MRI images an essential part of this Research, it was fundamental to rely on a stable and customizable pre-processing pipeline. To do so, state-of-the-art literature was used to identify the most necessary steps for the

[6] https://www.python.org/

processing of images acquired with magnetic resonance technology, therefore identifying the critical aspects of already existing tools. Once this phase was completed, the building blocks of a custom-made pipeline were clear, and needed to be realised with the optimal Python library.

Figure 9 summarizes the main blocks that constitute the processing pipeline. It includes the basic steps for an optimal processing. For the detailed description, see Section 4.5.1 Preprocessing pipeline.

Figure 9 Main steps of the processing pipeline

Most of these steps refers to well-established procedures and theories. Therefore, the reference literature used for this part of investigation includes both recent and less recent papers.

A scanning protocol is normally used during a standard MRI scan for MS monitoring, and such protocol can vary between clinical facilities. Standard norms are accepted as reference: according to these, T2 weighted sequences are the preferred one for brain scans as, also without the injection of contrast liquid, they generally allow an easier identification of new lesions, or the assessment of the status of old lesions.

In Figure 10 Example of T1-w and T2-w two images are shown: the one on the left shows a T1-weighted (T1w) example of image, the one on the right shows a T2w image. The difference is evident, especially when considering the contrast of the image (Guizard, 2015).

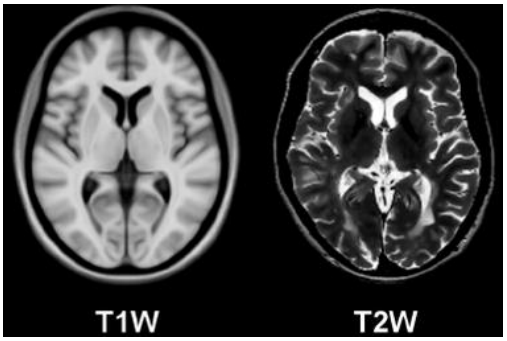


Figure 10 Example of T1-w and T2-w MRI images

T1w images are informative to assess the status of the so-called black-holes lesions, areas of permanent nerve damage, and to assess the degree of brain atrophy, when necessary. Therefore, the pipeline is thought in such a way that focus lays on T2w images, and on the relevance of the assessment of the lesion load.

Image registration and conversion to NIfTI format

The image registration step allows matching the space of the image to a reference space, referring the image to a standard atlas, being a prerequisite for any analysis that aims at a comparison across datasets or in time (Toga, 2019) as different MRI scans can present different slice spacing and slice resolution due to the scan settings and protocol (Alam, 2016). For a better understanding, see Figure **11** where an example of image registration is given.

On the left there is a T2-w scan that consists in 70 different images (or "slices"), and on the right there is the same scan after the registration on a reference atlas. The reference atlas here considered and later used is the ICBM Average Brain (Mazziotta, 1995). Registration algorithms determine the transformation needed to match the source image with the target atlas (ICBM) to optimize the similarity index between them. The registered image is obtained linearly interpolating the initial image domain into the new domain, to match the anatomical references.

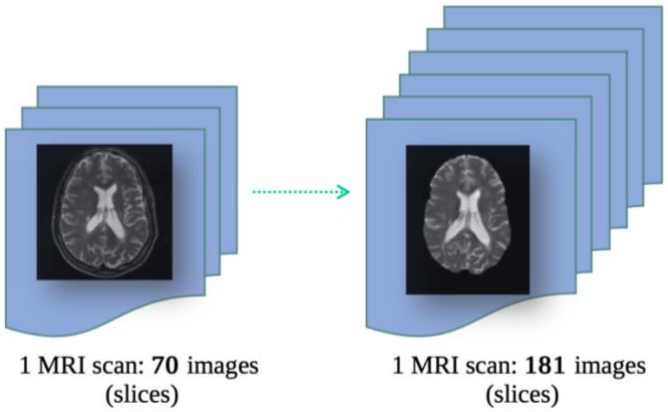


Figure 11 Effect of the atlas registration on an MRI scan

The different number of slices between the initial and the resulting scan is given by the fact that the resolution of the initial image is matched with the atlas resolution (in the case of the ICBM atlas, $1mm^3$) via interpolation. In this way, congruous datasets can be created, where every slice corresponds to a known anatomical reference despite the scan having been acquired with different machines and with different settings.

The different number of slices between the initial and the resulting scan is given by the fact that the resolution of the initial image is matched with the atlas resolution (in the case of the ICBM atlas, $1mm^3$) via interpolation. In this way, congruous datasets can be created, where every slice corresponds to a known anatomical reference despite the scan having been acquired with different machines and with different settings.

Bias field correction and noise reduction

MRI images are typically affected by an artifact called bias field, that appears in the image as an intensity inhomogeneity. It is due to spatial inhomogeneity in the magnetic field and is therefore unavoidable. The bias field correction step is crucial to adjust intensity discrepancies in intensity within tissues with the same physical properties. The effect of the bias field becomes visible in images acquire with magnetic fields stronger than 1.5T, that is the standard field for the monitoring of MS.

There are several methods for performing bias field correction, such as minimizing the image entropy or fitting the histogram of the local neighbourhood to global histogram of the image (Despotović, 2015).

The N4 method for bias field correction was selected for the purpose of this Research (Tustison, 2010). The N4 method is an improved version of the well-known N3 nonparametric intensity normalization method, being fully automatic and applicable to almost every MR image.

Brain extraction

Brain extraction, also known as skull stripping, is an important step that allows to isolate the relevant information from the image leaving out interests that are not of interest for the scope of the analysis. In this context, tissues like the dura mater and the skull are not of interest and might only add more sources of noise and errors. Like the bias field correction, the brain extraction is a necessary step for most pre-processing pipelines, and many methods are proposed and adopted in the state-of-the-art algorithms: some of these rely on deep learning networks, others on more traditional morphological operations.

In order to obtain good performances, the pipeline that was created for the image processing useful in this Research uses a combination of morphological operations to reach two goals: First, to keep guarantee good performances in terms of processing times and, secondly, to easy the access to the internal functioning of the skull stripping procedure, to personalise it in case of necessity, and to allow the method to work independently from the MRI modality. More details about the approach inspired by the work of (Gambino, 2011) are given in the Chapter 4 *Methods*.

Image processing toolkits

Many toolkits and corresponding Python libraries for image processing do exist and respond to different needs. Therefore, it is essential to choose one among these that best adapt to the requirements. In this case, several libraries were considered and tested in terms of performances. In the Table **1** a summary of the most known toolkit is given.

Table 1 Toolkits used to develop the image processing pipeline

Toolkit	*Main features*	*Python library*
Insight toolkit (ITK)[7]	• Cross-platform library that provides a suite of software tools for image analysis (Avants, 2014) • Provides many I/O image formats and several image processing algorithms	ITK Python package[8]
NiBabel[9]	• Read and write access to common neuroimaging file formats • Supports geometrical and morphometry files • Limited support for DICOM	NiBabel lib
Advanced Normalization Tools (ANTs)	• C++ library available through the command line • Available in Python as medical image library	ANTsPy [10,11]

After having evaluated the tools, the ANTsPy library was selected to be included in the processing pipeline. More details are discussed in Chapter 4 *Methods*.

3.3.2 DL applications with fastai

The fastai[12] library was used for all the implementations in this Research that include DL. As (Howard, 2020) highlight, this library provides

[7] https://itk.org
[8] https://itkpythonpackage.readthedocs.io/en/master/Quick_start_guide.html
[9] https://nipy.org/nibabel/index.html
[10] https://antspy.readthedocs.io/en/latest/
[11] http://stnava.github.io/ANTs/
[12] https://docs.fast.ai/

compromises between flexibility, ease of use and performances, distributing open-source material and high-quality courses.

Many recent application use fastai as a backbone, ranging in a series of topics. Among these, it is also possible to find many examples in the medical field. In the work of Paciorek et al. (Paciorek, 2024), for example, fastai is used to create a deep learning model for cardiac pathologies detection in MRI T1 mapping, obtaining results that justify the use of MRI T1 cardiac images without contrast for the detection of several cardiac conditions. Praveen et al. (Praveen, 2022) rely on fastai for the realization of a model able to detect malignant tissue in histological 2D breast tissue images. They also show, among the rest, the benefits of using fastai, that provides libraries and packages, offering precise outcomes in deep learning. Fastai allowed also Chaudhury et al. (Chaudhury, 2023) to address the problem of ICD breast cancer detection in histological images, combining a Gradient Color Activation Mapping and image colouring mechanism with a discriminative fine-tuning methodology employing a one-cycle strategy using fastai techniques.

These are just some of the most recent state-of-the-art publications where the advantages of using fastai concretely benefits the developers.

3.3.4 The dotplots and the FlexiDot tool

This Paragraph introduces the theory behind the dotplots used in the bioinformatic field, how they were used in this Research as medical images for the analysis of AAV vectors. How this topic relates to the rest of the Research is described in Chapter 4 *Methods*.

Dotplots are a visualization method used to compare two sequences of nucleotides, looking for similarities between the two, providing a global representation through the graph that is itself an output matrix. It can be used to compare a sequence to itself (obtained a self-dotplot) or to compare two different sequences. The sequences are placed along the axis (x-axis and y-axis), and the graph is filled based on nucleotide matching: regions of the sequence on the x-axis are compared to the entire query sequence (y-axis), and the size of the region of interest depends on the defined window size. A mismatch limit must also considered, and when 0 mismatches are found in the region, then a dot is placed at the appropriate (x,y) coordinates.

In Figure 12 A) a sequence is compared with itself with a standard window size of 10 nucleotides: this results in a straight diagonal line. In Figure 12 B) the window size is reduced to 5, resulting creating a higher likelihood of matching.

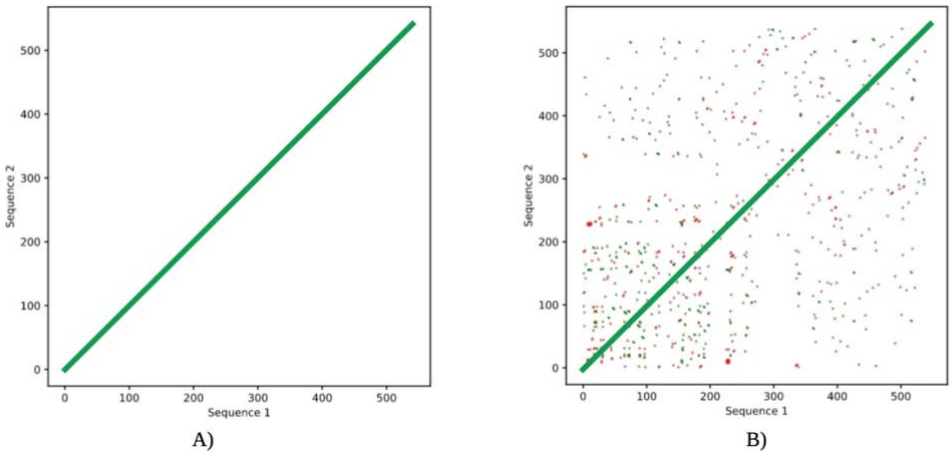


Figure 12 Comparison of dotplots obtained with different window sizes

Figure 13 suggests why dotplots can be very useful to providing an immediate representation of the characteristics of a sequence. Hypothesizing that the reference sequence lays on the x-axis, and that on the y-axis there is the same sequence, but with mutations (e.g. insertions, deletions), the localization and type of such changes in the sequence is clearly visible in the graph. In a) there is a mismatch, b) a deletion and c) represents an insertion.

Hence, in the context of evaluating the characteristics of AAV vectors during vector production, dotplots are a valuable tool which benefit was explored in this Research.

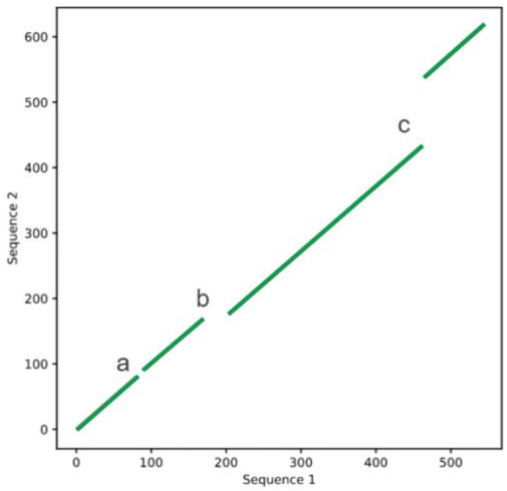


Figure 13 Effect of mutations on the structure of a dotplot

Many applications and Python libraries exist[13], the majority of which are only usable online or locally[14], or are standalone WEB applications[15] that are difficult to personalize, or which parameters and inner functioning are not accessible to the user. Considering the necessities of this Research, the FlexiDot tool has been chosen. It was introduced by Seibt et al. (Seibt, 2018) as a cross-platform dotplot suite for the generation of high-quality self, pairwise and all-against-all visualizations, implemented in Python 2.7 and relying on Biopython[16]. Its performances were also compared to other tools such as Dotter, Gepard, and PolyDot, and the benefits of FlexiDot are presented in the literature (Seibt, 2018).

FlexiDot documentation is available17 (Kathrin M. Seibt, 2018) and described from top to bottom, and specifically thought for routinely analyzing sequences. Its flexibility allows the creation of different types of dotplots in different configurations, and it grants access to many of the parameters that control the inner functioning of the algorithm, such as ambiguity handling mismatch toleration, and color shading as an indicator of similarity.

[13] https://pypi.org/project/dgenies/
[14] https://www.bioinformatics.babraham.ac.uk/projects/redotable/
[15] https://en.vectorbuilder.com/tool/sequence-dot-plot.html
[16] https://biopython.org

[17] https://github.com/molbio-dresden/flexidot

The functionalities that were most useful for the scope of the application in this Research were:

- FlexiDot plotting modes
  - Self-dotplots: to compare a sequence with itself
  - Pairwise dotplots: to compare different sequences
  - Collage output option when more than one sequence is used
- Variable k-value

Here is an example – taken from the documentation – of the main commands for the execution: a fasta file (test.fas) containing more sequences is given as input to create a single image (-c y) as output (Figure **14 Example of a FlexiDot output image**) and not six different graphs, containing a collage (-p 0) of dotplots in different configurations, disposed on 3 columns (-m 3), with no color shading (-x n).

```
>> python flexidot.py -i test.fas -p 0 -k 10 -c y -x n -m 3
```

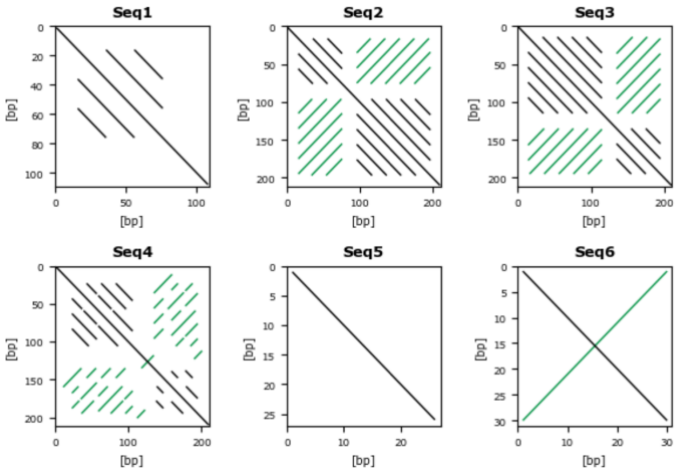


Figure 14 Example of a FlexiDot output image

This example shows some of the many parameters that can be used to obtain a visualization that respond to the users' necessities. However, in order to use this tool, it was necessary to automatize it, avoiding having to run it manually: details on this and on the use of FlexiDot for the analysis of AAVs are in Chapter 4 *Methods*).

Chapter 4

Methods

4.1 Aim of the Research

As partially already discussed in Section 1.2 Aim and challenges this Research dealt with several topics and aimed to explore different medical field-related questions. Figure **15** Map of the research topics and tools describes, graphically, the interconnections between the topics, the tools and developed solutions, and with which tools the questions were addressed, to introduce the reader to the contents of this Chapter.

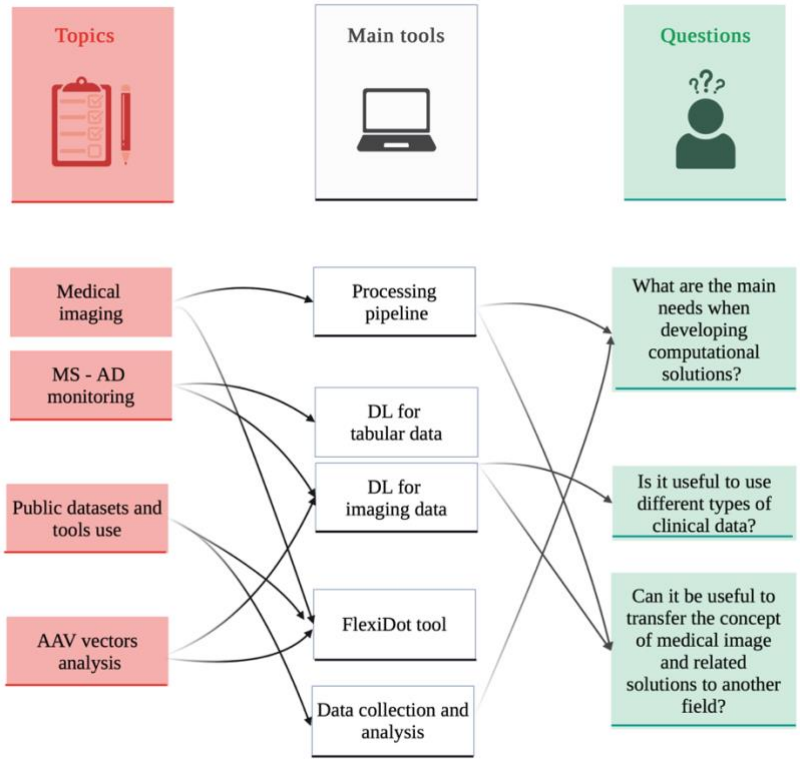


Figure 15 Map of the research topics and tools

Firstly, there are three main topics (listed in the first column), and each tool (second column) that was used, addressed one or more topics. This allowed to tackle the different questions from different perspectives. Moreover, this interdisciplinary approach did not declinate in a pedantic fashion, hyper focusing on a single aspect or disease. It aimed to embrace common challenges and questions related to different subjects: it started from neurological diseases monitoring and landed in the gene therapy field, adapting according to the needs the processing pipeline for the images, the methods of data collection, analysis and storage, and the deep learning frameworks. Lastly, the third column in Figure **15** shows which solution connected to the corresponding research question.

4.2 Data collection and analysis

The Research method that was followed was quantitative: data, being imaging data, clinical data, or genomic sequences, was used to identify how the raw data could be processed, modified, or combined, to extract information.

Data were differently extracted and handled, depending on its nature and purpose of analysis. In this Section the method used for data collection and analysis will be described, exploring more in detail this fundamental aspect, exposing the challenges as well met, and how they impact the work itself.

This Section will be organized as follows: basing on the topics summarized in Figure **15**, the methods used for data collection and analysis will be described in order and divided by topic: Section 4.2.1 Medical imaging and neurological diseases will describe the collection, analysis, and manipulation of imaging data related to MS and AD, while Section 4.2.2 AAV sequences will expose how the wild-type AAV sequences were collected and stored for later purposes.

This part of the work made it possible to face the challenge of using public datasets, and the main difficulties will be also disclosed. Later, Section 4.4 Materials and equipment will briefly detail the equipment used.

4.2.1 Medical imaging and neurological diseases

Two different datasets were used, one for each disease targeted by the Research. For MS, the ISBI 2015[18] dataset was utilized, and for AD the ADNI [19] dataset was made use of. The rationale behind the choice of these two datasets was driven by an analysis of the state-of-the-art literature, and later identified ISBI2015 and ADNI as most adequate to the scopes of the Research.

In fact, there are not many datasets containing medical data (imaging data, or other clinically relevant information) that have all the characteristics that make them suitable:

- It is relatively easy to get access to them
- They are well organized with good documentation and description of how the data was collected
- They have a good numerosity, and heterogeneously represent the target population.

4.2.1.2 The ISBI 2015 dataset

The ISBI 2015 dataset was used during the 12th International Symposium on Biomedical Imaging: it consists of MRI images and corresponding masks of 5 subjects, for a total of 20 different scans, as 4 different scans – corresponding to different time-points (TP) – are available for each subject. T2-w, MPRAGE T1-w, PD, and FLAIR modalities are available for each subject.

Accessing and navigating this database did not represent an issue, given the small number of subjects (unlike the ADNI dataset). In total, the dataset contains 1475 images, distributed like Figure **16** ISBI 2015 describes.

[18] https://biomedicalimaging.org/2015/program/isbi-challenges/
[19] https://adni.loni.usc.edu/

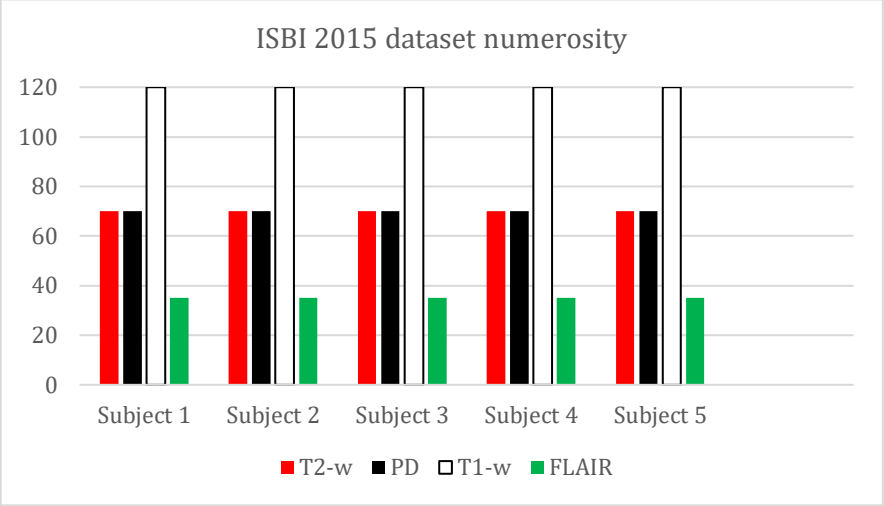


Figure 16 ISBI 2015 dataset numerosity per subject

As the T2-w modality was the one of interest in this Research (Abderrahim, 2020), and as 70 images are available for each scan, only the 350 images in this modality were later used to implement the pre-processing pipeline (see Section4.5.1 Preprocessing pipeline).

The main reason that led to the selection of this dataset, it that it provides, for every scan of every subject, the lesion masks obtained – manually – after image preprocessing by two independent readers: this was useful to verify the functioning of the processing pipeline later developed. Moreover, the processed images are included as well in the dataset, and they were obtained with the MIPAV software: the type of pre-processing proposed for the ISBI images was used to define the main requirements of the processing pipeline, object of further discussion.

The raw T2-w images for each subject were downloaded together with the processed images and the lesion masks. The software ImageJ[20] was used for the first visual examination of the images, as it was necessary to take note of the resolution of the scans, of the number of images (from now on, also addressed as slices) per scan. Figure… shows, as an example. the T2-w second scan, of the Subject 01, visualized with ImageJ.

[20] https://imagej.net/ij/

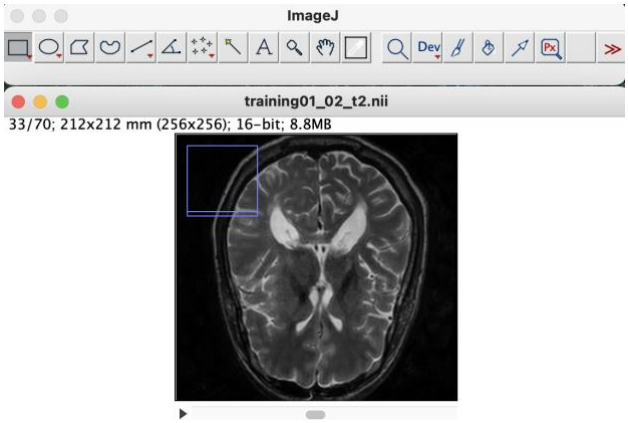


Figure 17 ImageJ example of visualization of a .nii image (NIfTI)

4.2.1.3 The ADNI dataset

The ADNI dataset belongs to an on-going initiative and is – at present – the most used dataset for the development of computational solutions regarding the Alzheimer's disease. ADNI began in 2004 as ADNI-1, an initial 5-years study that later was followed by ADNI-GO, ADNI-2, and ADNI-3, respectively in 2009, 2011, and 2016. Table 2 describes the 4 phases of ADNI, with their goals and the patient cohorts. Thanks to how ADNI was built, patients that were initially included in ADNI-1, are still part of ADNI-3, and this allows to have, for such patients, data relative to many different time-point as they were followed and reassessed over time. ADNI results are shared through the USC Laboratory USC Laboratory of Neuro Imaging's Image and Data Archive (IDA), and available for research purposes upon request.

Table 2 Description of different ADNI phases

	ADNI-1	ADNI-GO	ADNI-2	ADNI-3
Primary goal	Develop biomarkers as outcome measures for clinical trials	Examine biomarkers in earlier stages of disease	Develop biomarkers as predictors of cognitive decline	Study the use of tau PET and functional imaging techniques
Duration/start	5 yrs/2004	2 yrs/2009	5 yrs/2011	5 y/2016
Cohort	200 elderly controls	Existing ADNI-1	Existing ADNI-1 and	

400 MCI 200 AD	200 early MCI	ADNI-GO 150 elderly controls100 early MCI 150 late MCI 150 AD	Existing ADNI-1, ADNI-GO, ADNI-2 133 elderly controls 151 MCI 87 AD

At present, 2380 patients are part of ADNI, and more than 15000 time-points are avaiable in total. Figure **18** ADNI infographic is the infographic avaiable on the ADNI website that gives the first insights on the content of the dataset in terms of the distribution of age and gender, and of the research group: this information is of particular importance, as not only patients with diagnosed Alzheimer's disease are included in the study, but also patients with diverse forms of cognitive impairment, that often happen to be prodroms of the disease.

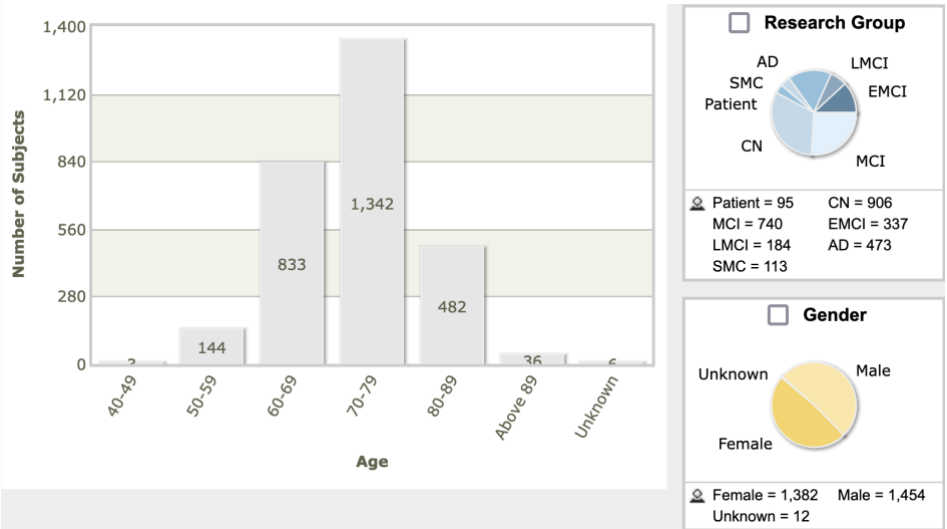


Figure 18 ADNI infographic [21]

Table **3** ADNI patient cohorts summarized the abbreviations used to classify the ADNI patients, and that will be used from now on in this

[21] https://ida.loni.usc.edu/login.jsp

dissertation and indicates from which phase of the ADNI project was the patient cohort included.

Table 3 ADNI patient cohorts with the indication of the enrolment

Research Group	***Abbreviation***	***ADNI phase***
Alzeimer's Disease	AD	Since ADNI-1
Mild Cognitive Impairment	MCI	Since ADNI-1
Control	CN	Since ADNI-1
Early Mild Cognitive Impairment	EMCI	Since ANDI-GO
Late Mild Cognitive Impairment	LMCI	Since ADNI-2
Significant Memory Concern	SMC	Since ADNI-2

The types of data included in ADNI are the following:

- Clinical data
- Imaging MRI data
- Imaging PET data
- Genetic data
- Biospecimen data

To give and idea of the level of detail and organization of ADNI, see Figure **19** The acquisition of different data types in different ADNI phase (in time) : for each ADNI phase and at each time-point of screening (from the initial screening, to those conducted after several months), the overmentioned data types were collected and organised inside the dataset. Figure **3** Graphical representation of MS phenotypes progression.

Figure 19 The acquisition of different data types in different ADNI phase (in time) [22]

Despite the good documentation and organization of the various phases was well documented, an extensive and detailled analysis of ADNI is lacking in the state-of-the-art literature. This can represent a problem when it it necessary to search for data entries of different nature and distributed over time Figure **22**. For the scope of this Research, it was necessary to conduct such analysis: many inconsistencies were found and needed to be tackled to retrieve and organise data avoiding errors and misinterpretations (see Section 4.2.1.3 a) ADNI extensive analysis).

Given that only the MRI imaging data, together with the clinical data, were interesting for the Research, the study info material was selected as in Figure **20** "Study Info" material selection, after a careful evaluation of all the avaiable files.

[22] https://adni.loni.usc.edu/data-samples/data-types/

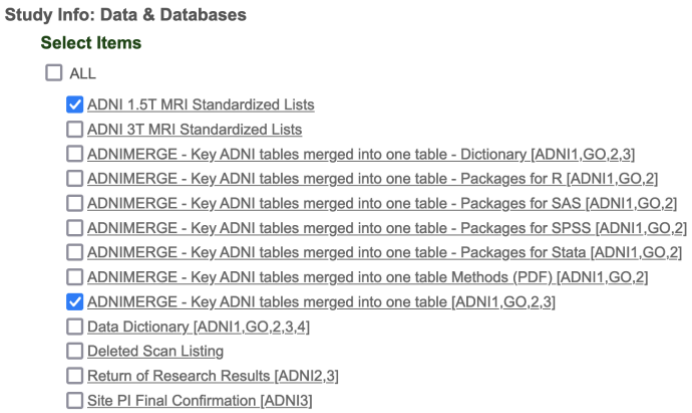


Figure 20 "Study Info" material selection on the ADNI website

The *ADNI MERGE* file was of particular importance, and it was used as basis to select patients, to keep track of the different time-points per subject, to identify issues related to discrepancies in the keywords and labels used thorought different the 3 ADNI phases, and to access the clinical data relative to cognitive impairment and others, as discussed later on.

Basing on the content of *ADNI MERGE*, other important files were selected from ADNI, here listed and described. As a note for the reader, it is highlighted that this initial work esclusively dedicated to the analysis of ADNI was necessary to orientate among the moltitude of documents avaiable with no clear description nor clear indications: this represented a gap that, initially, made it difficult to use ADNI flowlessy. Figure **21** ADNI MERGE file example gives an example of an *ADNI MERGE* section with the keywords useful for the identification of different subjects, and different evaluations.

RID	COLPROT	ORIGPROT	PTID	SITE	VISCODE	EXAMDATE	DX_bl	AGE	PTGENDER
2	ADNI1	ADNI1	011_S_0002	11	bl	08/09/05	CN	74.3	Male
2	ADNI1	ADNI1	011_S_0002	11	m06	06/03/06	CN	74.3	Male
2	ADNI1	ADNI1	011_S_0002	11	m36	27/08/08	CN	74.3	Male
2	ADNIGO	ADNI1	011_S_0002	11	m60	22/09/10	CN	74.3	Male
2	ADNIGO	ADNI1	011_S_0002	11	m66	04/03/11	CN	74.3	Male
2	ADNI2	ADNI1	011_S_0002	11	m126	23/06/16	CN	74.3	Male
2	ADNI2	ADNI1	011_S_0002	11	m120	22/09/15	CN	74.3	Male
2	ADNI2	ADNI1	011_S_0002	11	m108	13/10/14	CN	74.3	Male
2	ADNI2	ADNI1	011_S_0002	11	m102	17/03/14	CN	74.3	Male
2	ADNI2	ADNI1	011_S_0002	11	m96	09/09/13	CN	74.3	Male
2	ADNI2	ADNI1	011_S_0002	11	m90	25/03/13	CN	74.3	Male
2	ADNI2	ADNI1	011_S_0002	11	m84	26/09/12	CN	74.3	Male
2	ADNI2	ADNI1	011_S_0002	11	m78	22/03/12	CN	74.3	Male
2	ADNI2	ADNI1	011_S_0002	11	m72	19/09/11	CN	74.3	Male
2	ADNI2	ADNI1	011_S_0002	11	m132	27/09/16	CN	74.3	Male
2	ADNI3	ADNI1	011_S_0002	11	m144	18/10/17	CN	74.3	Male
3	ADNI1	ADNI1	011_S_0003	11	bl	12/09/05	AD	81.3	Male
3	ADNI1	ADNI1	011_S_0003	11	m06	13/03/06	AD	81.3	Male
3	ADNI1	ADNI1	011_S_0003	11	m12	12/09/06	AD	81.3	Male
3	ADNI1	ADNI1	011_S_0003	11	m18	12/03/07	AD	81.3	Male
3	ADNI1	ADNI1	011_S_0003	11	m24	12/09/07	AD	81.3	Male
4	ADNI1	ADNI1	022_S_0004	22	bl	08/11/05	LMCI	67.5	Male

Figure 21 ADNI MERGE file example

In total, ADNI MERGE consists (at the time of the analysis, 2022) of 15754 lines and 115 columns: each line corresponds to a different time-point, and each column refers to clinical variables (as described below in Section 4.2.1.3 b) Creation of a database based on ADNI), and some IDs such as:

- Two different variable identify the subject: RID and PTID (often indicated also as Subject)
- COLPROT indicates the ADNI phase in which the assessment was performed, ORIGPROT tells in which ADNI phase the subject was recruited
- VISCODE stands for the type of visit conducted at a specific time-point bl is the baseline visit, m06 is the visit performed 6 months after the baseline, m12 is the visit performed 12 months after the baseline, etc…
- The EXAMDATE column is misleading and not well documented, therefore it was not considered (many inconsistencies where found when looking for the image files at a certain time-point). The EXAMDATE, for example, often does not correspond to the date of the MRI scan, altough the scan refers to the same time-point
- The DX_bl column indicates the diagnosis carried out at the baseline visit, and the DX column indicates the diagnosis carried out at the time-point
    - A patient can be stable, for example being reassessed as MCI, or can be assessed as AD if the disease progressed. In some cases, the diagnose reverse: some patients assessed, for

example, as AD at a time-point, are diagnosed as MCI in the following time-point. This is clinically impossible and clearly represents a mistake: these cases had to be identified and removed from the considered time-points.

A clear understanding of how to use these labels is fundamental to download the right image data from the web portal: Figure **22** ADNI - Advanced Sear shows the advanced search window that allows to retrieve the images selecting one or more subjects, referring to one or more visit/time-points, and selecting one or more of the avaiable image modalities. The selection of the image scan corresponding to the correct time-point is possible, selecting the visit description (Figure **22** ADNI - Advanced Sear in black) that correspond to the time-point. Table **4** Description of the relevant ADNI documents describes shortly why each document was selected and how it was used.

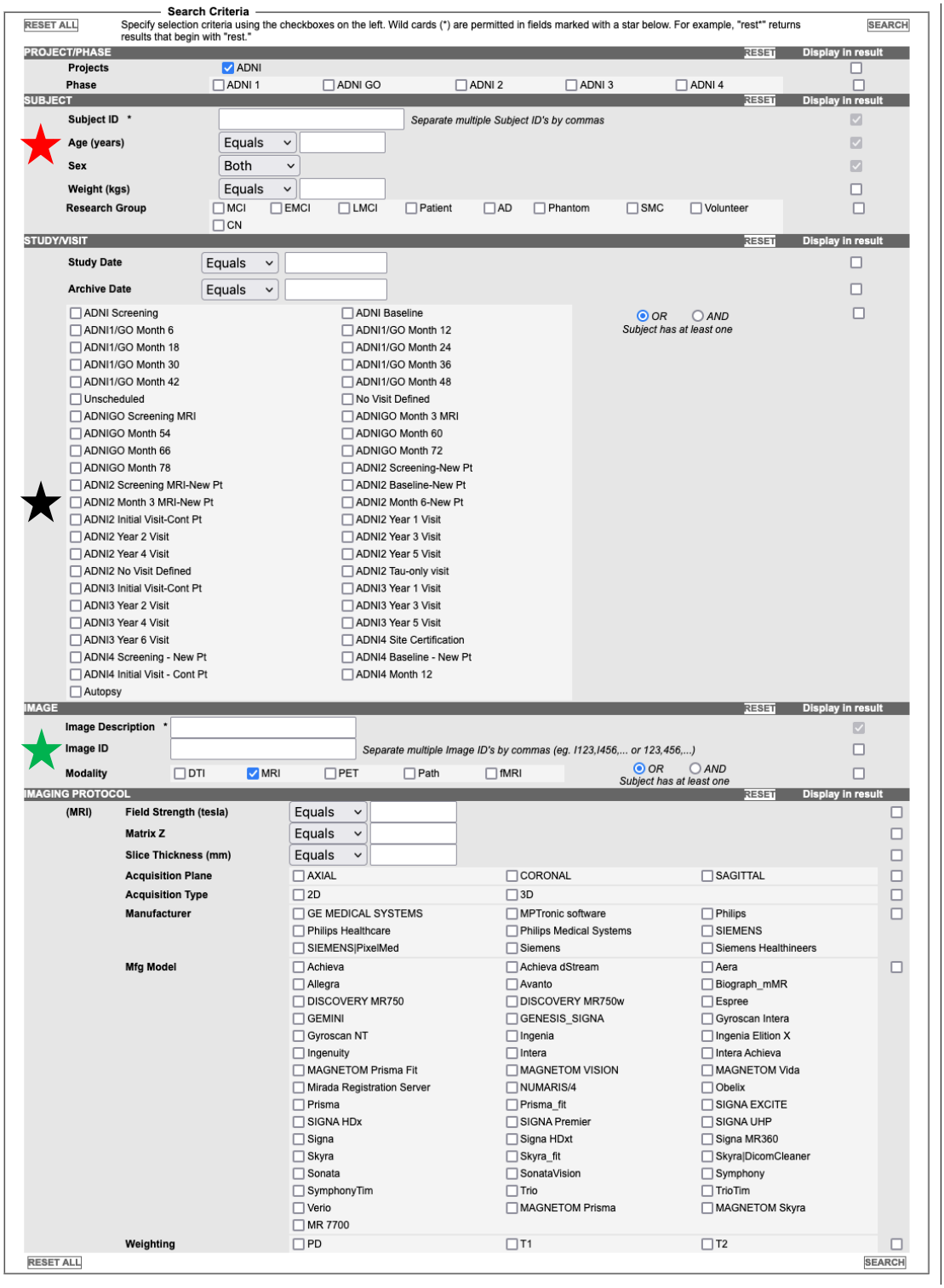


Figure 22 ADNI - Advanced Search interface

Table 4 Description of the relevant ADNI documents that were used

Document (name and extension[23])	*Use*
ADNIMERGE.csv	Summary of ADNI, patient info and tabular data
REGISTRY.csv	Useful for the download of the images related of a subject at a specific time-point. The match between the ADNI MERGE entry (*VISCODE*) and visit description label is needed. Unlike ADNI-1 and ADNI-GO, the ADNI-2 database assigns participant visit data a generic visit code that does not clearly indicate the longitudinal progression of the participant. The label VISCODE in ADNI MERGE does not correspond to the VISCODE label in REGISTRY.csv, but to VISCODE2 (Davis).
ADNI_VISCODE.csv	Deacribes the correspondande between the VISCODE and the visit description that allows the download of the images.
search.csv	File generated from the seach interface including all the images avaiable. Used to verify if the association beween time-point and image scan was correct.
MRILIST.csv	Allows to retrieve the IMAGEID and the SERIESID, as different image scans can correspond to the same visit.

[23] Avaiable here – after registration

Before introducing, in Section 4.2.1.3 b) Creation of a database based on ADNI, the detailled data types of interest in ADNI and the selection of the subjects for the creation of a dataset based on ADNI, **Figure 23** Example of navigation between the documents for subject 023_S_4115 and **Figure 24** Example of navigation between the documents for subject 011_S_0 simulates the process of selecting a time-point for a sebject, identify the images of interest and download these. This should help the reader to understand how ADNI can be navigated to correctly select the data of interest.

- **Figure 23** Example of navigation between the documents for subject 023_S_4115 shows the process of retrieving the information to download the Axial T2-FLAIR images of the subject 023_S_4115 (which is possible through the SERIESID and IMAGEUID numbers), for the time-point identified as m12. This examples shows that, as this subject was enrolled during ADNI-2, the VISCODE identifiers, correctly chosen in the ADNI VISCODE table, allow to identify the correct visit description
- **Figure 24** Example of navigation between the documents for subject 011_S_0shows the example of a subject enrolled in ADNI-1, and for which is simpler to retrieve the images corresponding to the correct time-point

It is once more worth mentioning, that the literature lacks a detailled description of ADNI, and the analysis here performed was essential to the proceeding of the Research.

ADNI MERGE

RID	DX_in	DX_fin	COLPROT	ORIGPROT	PTID	VISCODE
4115	MCI	MCI	ADNI2	ADNI2	023_S_4115	m12

REGISTRY

Phase	ID	RID	SITEID	VISCODE	VISCODE2
ADNI2	7624	4115	17	v11	m12
ADNI2	10778	4115	17	v12	m18
ADNI2	13250	4115	17	v21	m24
ADNI2	15894	4115	17	v22	m30

ADNI VISCODE

Visit Code	Visit Description	Number of Subjects
v06	ADNI2 Initial Visit-Cont Pt	344
v11	ADNI2 Year 1 Visit	895
v21	ADNI2 Year 2 Visit	744

MRILIST

SUBJECT	VISIT	SEQUENCE	SCANDATE	STUDYID	SERIESID	IMAGEUID
023_S_4115	ADNI2 Year 1 Visit	MoCoSeries	02/10/12	52934	169431	337815
023_S_4115	ADNI2 Year 1 Visit	ASL PERFUSION	02/10/12	52934	169432	337816
023_S_4115	ADNI2 Year 1 Visit	MPRAGE GRAPPA2	02/10/12	52934	169433	337817
023_S_4115	ADNI2 Year 1 Visit	Axial T2-FLAIR	02/10/12	52934	169434	337818
023_S_4115	ADNI2 Year 1 Visit	Field_mapping	02/10/12	52934	169435	337819
023_S_4115	ADNI2 Year 1 Visit	MPRAGE	02/10/12	52934	169437	337821
023_S_4115	ADNI2 Year 1 Visit	AXIAL_T2_STAR	02/10/12	52934	169439	337823

Figure 23 Example of navigation between the documents for subject 023_S_4115

ADNI MERGE

RID	DX_in	DX_fin	COLPROT	ORIGPROT	PTID	VISCODE
3	AD	AD	ADNI1	ADNI1	011_S_0003	m12

REGISTRY

Phase	ID	RID	SITEID	VISCODE	VISCODE2
ADNI1	34	3	107	sc	sc
ADNI1	56	3	107	bl	bl
ADNI1	836	3	107	m06	m06
ADNI1	2834	3	107	m12	m12
ADNI1	5284	3	107	m18	m18
ADNI1	7138	3	107	m24	m24

ADNI VISCODE

Visit Code	Visit Description	Number of Subjects
sc	ADNI Screening	1669
bl	ADNI Baseline	1350
m06	ADNI1/GO Month 6	895
m12	ADNI1/GO Month 12	726

MRILIST

SUBJECT	VISIT	SEQUENCE	SCANDATE	STUDYID	SERIESID	IMAGEUID
011_S_0003	ADNI1/GO Month 12	B1-calibration Body	12/09/06	5186	19093	24693
011_S_0003	ADNI1/GO Month 12	B1-calibration Head	12/09/06	5186	19094	24694
011_S_0003	ADNI1/GO Month 12	Axial PD-T2 TSE	12/09/06	5186	19095	24695
011_S_0003	ADNI1/GO Month 12	Axial PD-T2 TSE	12/09/06	5186	19095	24701
011_S_0003	ADNI1/GO Month 12	MPRAGE	12/09/06	5186	19096	24696
011_S_0003	ADNI1/GO Month 12	MPRAGE Repeat	12/09/06	5186	19097	24697

Figure 24 Example of navigation between the documents for subject 011_S_0003

4.2.1.3 a) ADNI extensive analysis

As previously mentioned, ADNI consist of more than 15000 time-points, and for each time-point different types of data acquired during the evaluation are available. The number of evaluations per subject is not constant, as Figure **25 Time-points distribution for 43 subjects** shows for a sample of subjects (identified by their RID number, that is an alternative to the SUBJID seen before). Another aspect worth mentioning, is that not for every evaluation a diagnose is noted (i.e. a subject was diagnosed as AD, or if was stable as MCI): for the subject with RID 31, for example, more than 20 evaluations are available, but for less than 10 the diagnose carried out by the examinators is known.

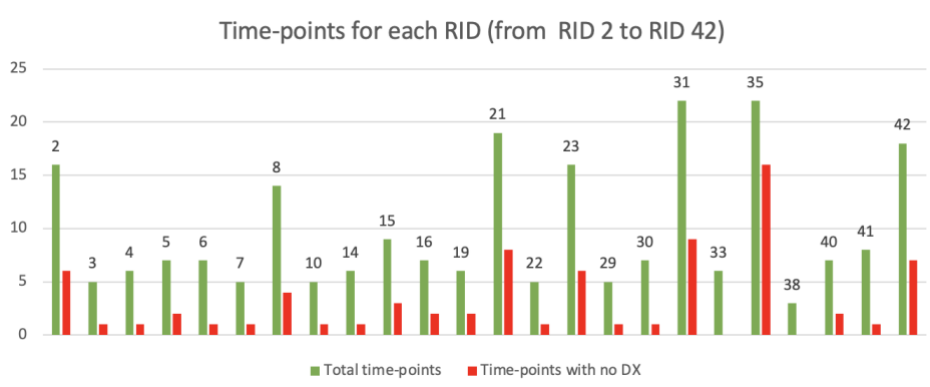


Figure 25 Time-points distribution for 43 subjects

Table **5** ADNI time-points numerosity gives an idea of ADNI numerosity: of the total TP of interest, only 10751 were used to proceed with the analysis and selection of data: given the interest in the clinical evidence related to the progression of the disease in time, the single TP with no follow-up, and those with no indication of the diagnose where immediately excluded [24].

Table 5 ADNI time-points numerosity

Total TP	Single TP	TP with no diagnose	Subjects
15754	253	4750	2379
		TP of interest	10751

To have an idea of the distribution of the type of diagnose, considering the number of subjects and not the TP after having excluded the subjects with a single evaluation, see in Table **6** Diagnose type that:

- 403 subjects, presented a diagnose of CN that remained stable between the baseline (BL) visit, and the last visit
- For 69 subjects, the disease progressed converting in MCI, and for 30 it converted to AD
- 362 subjects, diagnosed as MCI when enrolled, converted to AD in the last evaluation
- In red, the subjects that shows a reversion in the diagnose (these were later excluded)

[24] The numerosity refers to the year (2022) during which the analysis was performed

Table 6 Diagnose type (CN, SMC, MCI, AD) numerosity

	CN	SMC	MCI	AD
CN_BL	403	0	69	30
SMC_BL	250	0	22	3
MCI_BL	67	0	560	362
AD_BL	0	0	3	357
			Tot subjects	2126

Table **7** Number of time-points that remain stable or convert considers the number of TP that show conversion, reversion, or stability, considering the diagnose that was made at the first and last visit.

Table 7 Number of time-points that remain stable or convert

	CN	MCI	AD
CN_BL	2822	175	6
MCI_BL	121	3531	411
AD_BL	0	0	1567

When assigning to the TPs with no diagnose, the diagnose of the previous TP, the only numerosity that increases it the one of the so-called “stable” TP, not of interest for the Research that focuses on the progression in time. This step was therefore not performed, and these TP directly excluded from the pool (see Table **8** Numerosity of time-points after assigning the diagnose of the previous time-points to the missing values).

Table 8 Numerosity of time-points after assigning the diagnose of the previous time-points to the missing values

	CN	**MCI**	**AD**
CN_BL	*4853*	175	6
MCI_BL	121	*5173*	411
AD_BL	0	0	*2606*

The next step was defining the concept of *disease conversion*, and to construct two datasets based on ADNI: one consisting of tabular data, and one consisting of imaging data. There are different cases that can be considered conversion of the disease:

- A subject, initially classified as CN, is classified as MCI (EMCI and LMCI were considered as MCI)
- A subject, initially classified as MCI, is classified as AD
- A subject, initially classified as CN, is classified as AD

As **Table 7** Number of time-points that remain stable or convert shows in color, the numerosity of the classes of interest, meaning those that show a conversion in the diagnose, is not consistent: it was important to introduce permutations, expanding the size of the classes of switch considering not only the TP that precedes the switch of the diagnosis, but also all the previous TPs. Table **9** Example of conversion class expansion exemplifies how this is done: hypothesizing that a subject converts to MCI from CN after 12 months from the first assessment, instead of classifying as stable all the TP that precede the conversion, these can be considered as TP of conversion 12, 6 and 0 month before the conversion to MCI. A patient, however, can present multiple conversions in case of progressing disease.

This step is relevant also from the clinical point of view, as there is no biomarker for predicting the conversion from a clinically stable profile to the diagnose of AD.

Table 9 Example of conversion class expansion

Visit (months)	**DIAGNOSE**	**CONVERSION**	**MODIFIED CONVERSION**
0	CN	CN-CN	CN-MCI / 12m
06	CN	CN-CN	CN-MCI / 6m
12	MCI	CN-MCI	CN-MCI / 0m
18	MCI	MCI-MCI	-

Table **10** Distribution of the classes of switch for each time interval shows the distribution of the classes of switch in each interval of interest, expressed in months: most switches are condensed within 36 months.

Table 10 Distribution of the classes of switch for each time interval

Interval (months)	**CN-MCI**	**CN-AD**	**MCI-AD**
3-12	106	2	406
12-24	120	12	459
24-36	407	141	588
36-100	55	14	28

The ADNI MERGE file was therefore modified, adding the time information about the conversion, and considering the permutations described above (Figure **26** ADNI MERGE modified, as categorical and numerical variables (see Delta_cont and Delta_cat).

RID	DX_in	DX_fin	Delta_cont	Delta_cat	COLPROT	ORIGPROT	PTID	SITE	VISCODE	*VISCODE-new*	Visit Description
2	CN	MCI	12	Under12	ADNI2	ADNI1	**011_S_0002**	11	m120	*v41*	ADNI2 Year 4 Visit
2	CN	MCI	36	Under36	ADNI2	ADNI1	**011_S_0002**	11	m96	*v21*	ADNI2 Year 2 Visit
2	CN	MCI	60	Under36	ADNI2	ADNI1	**011_S_0002**	11	m72	*v06*	ADNI2 Initial Visit-Cont Pt
2	CN	MCI	72	Under36	ADNIGO	ADNI1	**011_S_0002**	11	m60	*m60*	ADNIGO Month 60
2	CN	MCI	97	Under36	ADNI1	ADNI1	**011_S_0002**	11	m36	*m36*	ADNI1/GO Month 36
2	CN	MCI	126	Under36	ADNI1	ADNI1	**011_S_0002**	11	m06	*m06*	ADNI1/GO Month 6
2	CN	MCI	132	Under36	ADNI1	ADNI1	**011_S_0002**	11	bl	*bl*	ADNI Baseline
2	CN	MCI	12	Under12	ADNI2	ADNI1	**011_S_0002**	11	m72	*v06*	ADNI2 Initial Visit-Cont Pt
2	CN	MCI	24	Under24	ADNIGO	ADNI1	**011_S_0002**	11	m60	*m60*	ADNIGO Month 60
2	CN	MCI	48	Under36	ADNI1	ADNI1	**011_S_0002**	11	m36	*m36*	ADNI1/GO Month 36
2	CN	MCI	78	Under36	ADNI1	ADNI1	**011_S_0002**	11	m06	*m06*	ADNI1/GO Month 6

Figure 26 ADNI MERGE modified after the expansion of the conversion classes

This information was put together with the imaging data available for the subjects selected at this point, preliminary selecting 46 subjects, listed here:

Table 11 List of the IDs of the selected subjects

037_S_4071	016_S_1326
100_S_0035	029_S_5135
067_S_0098	073_S_0518
100_S_0015	018_S_0057
023_S_0388	100_S_0892
011_S_4105	094_S_1015
016_S_4902	020_S_0883
037_S_1078	009_S_4741
126_S_0680	021_S_4857
014_S_0548	130_S_0285
116_S_0361	037_S_0539
033_S_0741	057_S_0779
062_S_1299	014_S_4577
068_S_0442	012_S_4188
137_S_4331	126_S_4712
029_S_5166	009_S_4530
014_S_4668	128_S_1406
114_S_5234	012_S_5121
033_S_4179	052_S_0952
021_S_0141	031_S_4947
023_S_4035	011_S_0326
941_S_4100	127_S_6241
	130_S_2391

	sc	bl	m06	m12	m18	m24	m30	m36
	ADNI Screenin	ADNI Baseline	ADNI1/GO Mo	ADNI1/GO Mo	ADNI1/GO Mo	ADNI1/GO Mo	ADNI1/GO Mo	ADNI1/GO Mo
Description	0	0	0	0	0	0	0	0
3D BRAIN FDG FORE	0	1	0	0	0	0	0	0
3D BRAIN AV-45 FORE AV45	0	0	36	0	0	0	0	0
3 Plane Localizer	0	0	4	0	0	0	0	0
MPRAGE SENSE2	0	0	6	0	0	0	0	0
Field Mapping	0	62	200	134	49	70	0	46
MPRAGE	0	0	0	0	0	0	0	0
Axial T2-TSE with Fat Sat	0	0	3	0	0	0	0	0
Axial T2-Star	0	0	12	1	0	0	0	0
Axial T2-FLAIR	0	0	4	0	0	0	0	0
Resting State fMRI	0	7	100	77	29	51	0	43
B1-Calibration PA	0	42	247	186	71	113	0	77
B1-Calibration Body	0	21	335	219	90	153	0	207
3-plane localizer	0	12	228	180	80	136	0	115
Axial PD/T2 FSE	0	9	141	111	53	86	0	68
MP-RAGE	0	2	114	98	39	62	0	56
MP-RAGE REPEAT	0	0	24	18	12	9	0	0
3-pl T2* FGRE S	0	35	28	28	16	23	0	12
ADNI Brain PET: Raw	0	0	0	0	0	0	0	0
Accelerated Sag IR-FSPGR	0	0	8	0	0	0	0	0
Axial T2 Star	0	0	6	0	0	0	0	0
Axial FLAIR	0	0	0	0	0	0	0	0
Sag IR-FSPGR	0	0	0	0	0	0	0	0
Ax T2 FSE with Fat Sat	0	0	5	0	0	0	0	0
Calibration Scan	0	2	0	0	0	0	0	0

Figure 27 ADNI images numerosity for different time-points

Moreover, it was necessary to analyze the numerosity of the dataset in terms of images: ADNI was therefore scanned to assess the modalities and the number of scans for each time label (see Figure **27** ADNI images numerosity that shows an example of the numerosity analysis). This was done to have more insights on ADNI content, and to identify the image modality that is most common among the subjects and the TPs. The T1-3D (MPRAGE and SPGR sequences) images were selected as images of interest, being the most numerous one in the interval of interest (between 0 and 36 months). These were later used, as described in Section Deep Learning for tabular and imaging data.

Other important information contained in ADNI MERGE, is clinical scores and clinical data of different sort, such as NP tests scores, for a total of 86 variables. 37 variables were selected out of these, depending on their clinical relevance assessed by the literature and on their numerosity, as described later.

To summarize the study on the ADNI dataset, the following challenges were faced:

- There is a lack of detailed documentation and description

- Data is differently labelled, and many inconsistencies can be found
- Data is not entirely collected systematically
- There is lack of a reference if the literature about relevant biomarkers for AD progression.

4.2.1.3 b) Creation of a database based on ADNI

One of the points of this Research was understanding the impact and role of different sources of data on a neural network for the prediction of the evolution of AD in time. The analysis of the ADNI dataset to quantify the amount and the typology of data that could be used for this purpose was successful and fundamental, also thanks to the active support of physicians of the Fondazione St. Lucia to identify the real clinical needs.

The subjects were selected basing on agreed criterions (exclusion of the subjects with more than one TP of reversion, exclusion of the subjects with only one TP), expanding the dataset via permutations, considering only the subjects with more than one TP, and excluding the TP with no diagnosis label.

A detailed description of the dataset and its use can be found in Section Deep Learning for tabular and imaging data.

4.2.2 AAV sequences

For the analysis of AAV sequences, with the aim of focusing on the ITR structures, it was not possible to rely on private data, but on public data. This was another opportunity to engage with another type of dataset, differently organized and labelled than strictly clinical ones.

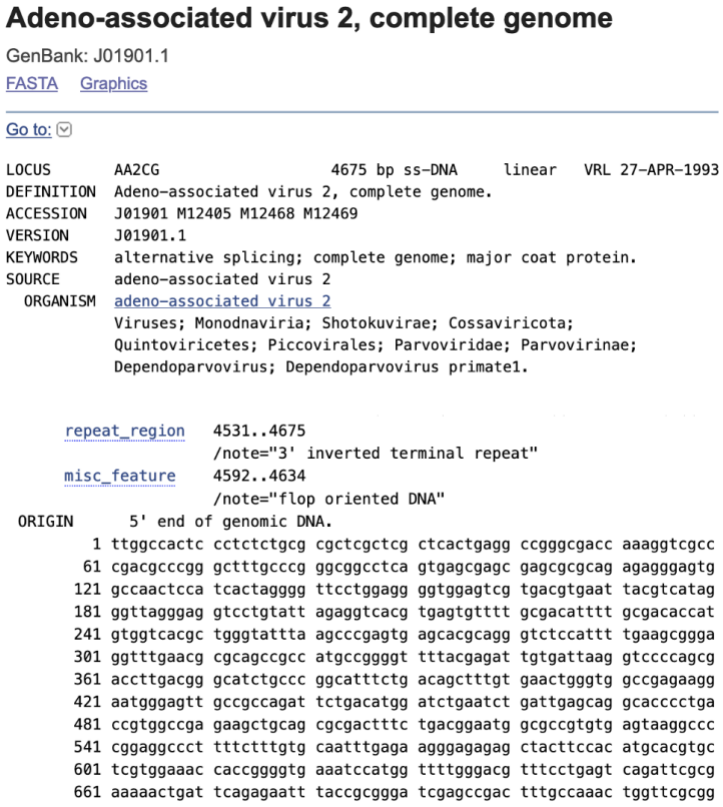
The NCBI dataset [25] was used to download the sequences of the following wild-type AAV: AAV1, AAV2, AAV3, AAV4, AAV5, AAV6, AAV7, AAV8. The following are the NCBI Reference Sequence IDs:

AAV virus	**NCBI Reference Sequence**
AAV1	NC_002077
AAV2	NC_001401
AAV3	NC_001729.1
AAV4	NC_001829

[25] https://www.ncbi.nlm.nih.gov/datasets/genome/

AAV5	NC_006152
AAV6	AF_028704
AAV7	NC_006260
AAV8	NC_006261

Figure **28** Example taken from the NCBI dataset shows an example of the information provided by the dataset: some general information about the genome is given, together with an annotation about the ITR sequences, such as their position in terms of bases, and their nature (if flip or flop). This, however, is not true for every AAV.



Adeno-associated virus 2, complete genome

GenBank: J01901.1

FASTA Graphics

Go to:

```
LOCUS       AA2CG                   4675 bp ss-DNA     linear   VRL 27-APR-1993
DEFINITION  Adeno-associated virus 2, complete genome.
ACCESSION   J01901 M12405 M12468 M12469
VERSION     J01901.1
KEYWORDS    alternative splicing; complete genome; major coat protein.
SOURCE      adeno-associated virus 2
  ORGANISM  adeno-associated virus 2
            Viruses; Monodnaviria; Shotokuvirae; Cossaviricota;
            Quintoviricetes; Piccovirales; Parvoviridae; Parvovirinae;
            Dependoparvovirus; Dependoparvovirus primate1.

     repeat_region   4531..4675
                     /note="3' inverted terminal repeat"
     misc_feature    4592..4634
                     /note="flop oriented DNA"
ORIGIN      5' end of genomic DNA.
        1 ttggccactc cctctctgcg cgctcgctcg ctcactgagg ccgggcgacc aaaggtcgcc
       61 cgacgcccgg gctttgcccg ggcggcctca gtgagcgagc gagcgcgcag agagggagtg
      121 gccaactcca tcactagggg ttcctggagg ggtggagtcg tgacgtgaat tacgtcatag
      181 ggttagggag gtcctgtatt agaggtcacg tgagtgtttt gcgacatttt gcgacaccat
      241 gtggtcacgc tgggtattta agcccgagtg agcacgcagg gtctccattt tgaagcggga
      301 ggtttgaacg cgcagccgcc atgccggggt tttacgagat tgtgattaag gtccccagcg
      361 accttgacgg gcatctgccc ggcatttctg acagctttgt gaactgggtg gccgagaagg
      421 aatgggagtt gccgccagat tctgacatgg atctgaatct gattgagcag gcacccctga
      481 ccgtggccga gaagctgcag cgcgactttc tgacggaatg gcgccgtgtg agtaaggccc
      541 cggaggccct tttctttgtg caatttgaga agggagagag ctacttccac atgcacgtgc
      601 tcgtggaaac caccggggtg aaatccatgg ttttgggacg tttcctgagt cagattcgcg
      661 aaaaactgat tcagagaatt taccgcggga tcgagccgac tttgccaaac tggttcgcgg
```

Figure 28 Example taken from the NCBI dataset for the AAV2

4.2.2.1 Dotplot images dataset creation

Merely retrieving the AAV sequences was not the goal: the extraction – manual or with the indications provided by the database – of the ITR sequences was the first scope. Secondly, to create a dataset to train a network able to infer the ITR origin via dotplots, a first dataset needed to be generated (see Figure **29** AAV analysis).

Figure 29 AAV analysis steps

Further details on the use of Flexidot to create dotplots will be given in Section 4.5.3 Flexidot tool: creating self-doplots for each ITR of the AAVs using different k-values (see Section 3.3.4 The dotplots and the FlexiDot tool), did not allow to obtain a numerous database.

K-values of 2, 3, 5, and 7 were used, and 65 dotplot images generate. Figure **30** Dotplots images simple processing shows the simple processing steps for make the images created by Flexidot later usable by a neural network, without including irrelevant information such as axis and labels (Figure **31** Effect of the processing on a Dotplot image).

Figure 30 Dotplots images simple processing

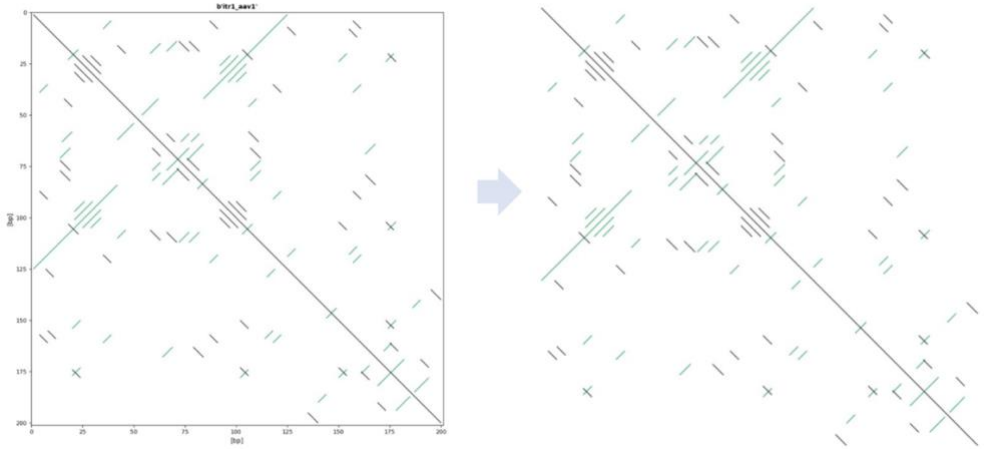


Figure 31 Effect of the processing on a Dotplot image

Therefore, a maximum of 10 small mutations (< 5 bases) where randomly inserted in the ITR sequences to:

- Introduce variability

- Increase the robustness of the classifier
- Simulate biological noise

The mutations that were considered were insertions, deletions and substitutions of basis, and selected on a uniformly distributed mutation rate. The position on the sequence at which applying the mutation was randomly chosen as well, depending on the length of the sequence. The details of each mutation (position, type of mutation, mutation lenght) were separately saved to keep track of the mutations applied on each sequence.

Here is an example of how this information saved:

MUTATIONS:
Number of mutations on the seq: 5
Mutations: ['I', 'I', 'D', 'D', 'S']
Insert @109 len3
Insert @31 len5
Delete @81 len1
Delete @30 len3
Substitute @24 len1

Figure **32** Effect of mutations on the ITR dotplots shows the effect of mutations on the dotplots: by observing this, expanding the dataset introducing these mutations was considered informative and effective.

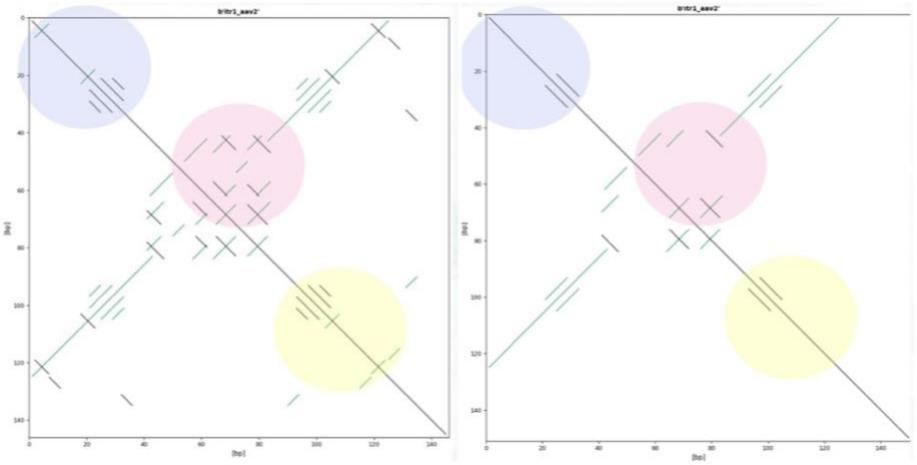


Figure 32 Effect of mutations on the ITR dotplots

By doing so, running Flexidot on these sequences as well, *2113* self-dotplot images were generated: despite not being a substantial dataset, it was

the first basis to test the possibility of using dotplots to classify a dotplot image, recognizing the type of ITR and to which AAV it correspond.

More details are given in Section 4.5.3 Flexidot tooland in the Results chapter.

4.4 Materials and equipment

The analysis of data, and the implementation of the computational solutions, were all conducted running script and programs on a MacBook Pro with the following specifications:

- Processor: 2,6 GHz 6-Core Intel Core i7
- Memory: 16 GB 2667 MHz DDR4

In addition, the Kaggle platform[26] and the Google Colab[27] notebooks have been used. The GPU (NVIDIA T4(x2)[28]) acceleration option available on Kaggle was fundamental for running the image processing steps and the training of the networks.

[26] https://www.kaggle.com/
[27] https://colab.research.google.com/
[28] https://www.kaggle.com/discussions/product-feedback/361104

4.5 Description of the main tools

The following section contains details on the main tools and proposed solutions: this will allow the reader to dive into the Results chapter, understanding the rationale for choosing materials, methods and procedures, and the obtained results.

4.5.1 Preprocessing pipeline

The development of the image processing pipeline wanted to tackle practical issues that are often encountered when using this type of application. This led to the idea of creating a flexible image processing pipeline that:

- Does not rely on many external plug-ins, to decrease the risk of encountering versioning problems
- Can handle different image formats, to make it feasible to be integrated in the analysis of differently organized datasets.

The state-of-the-art and the theory framework were previously described as a background of this part of the Research in Chapter 3.3.1 MRI images processing

After the analysis of existing processing tools (CBS Tools, JIST, TOADS-CRUISE, and BrainSuite), the processing pipeline was developed to be customizable and optimized.

Figure **33** Detailed processing pipeline describes the main blocks that constitute the pipeline, and details about each block.

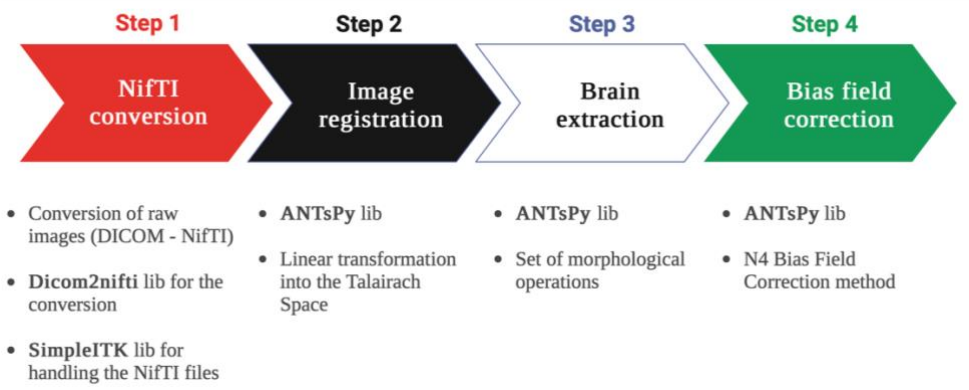


Figure 33 Detailed processing pipeline (steps and tools)

The Python libraries chosen for this implementation where selected basing on their general performances, documentation, and feasibility to be integrated together in the process. As the images needed to be manipulated in the Python environment, after the conversion to the NIfTI format, they had to be converted to Numpy[29] arrays: Table **12** Numpy conversion processing times shows the conversion times of three common libraries for the conversion from NIfTI (previously converted from DICOM, if necessary) to Numpy.

Table 12 Numpy conversion processing times

ITK TIME	4.4 sec
NIBABEL TIME	2.9 sec
ANTsPy TIME	1.7 sec

The **image conversion** step is fundamental to generalize and work with different types of images or datasets: as previously mentioned, this pipeline supports two different (and standard) raw images format, the DICOM and the NIfTI.

The **image registration** developed for this pipeline took inspiration from the suggested steps for processing of the ISBI 2015 dataset. The images are rigidly registered to the corresponding reference standard atlas in the MNI

[29] https://numpy.org

space [30]. The pipeline allows the selection of different atlases, as it is sufficient to download the desired model and import it. To register the image, the similarity index between the reference image and the image of interest is computed, and the algorithm optimizes it: after this, the images is linearly interpolated into the reference domain. As different scans can have been acquired with different scan setting and can therefore present differences in resolution (i.e. the number of slices – images – per scan), the registration step uniforms the number of slices to the reference atlas, making sure to have an anatomical reference.

The **brain extraction** phase consists of many steps, summarized in Figure **34** Brain extraction steps.

Figure 34 Brain extraction steps

These steps were inspired by the work of (Gambino, 2011). The following operations are performed:

- N (default: N=3) erosions with a cross kernel
- N - 1 dilatations with a cross kernel
- Extraction of the brain mask
- Final erosion (N cross kernel)
- The mask that is obtained is multiplied to the original image to obtain the brain with no surrounding tissues that are not of interest (the result of these steps is shown in Figure **35** Result of the brain extraction).

[30] https://nist.mni.mcgill.ca/icbm-152lin/

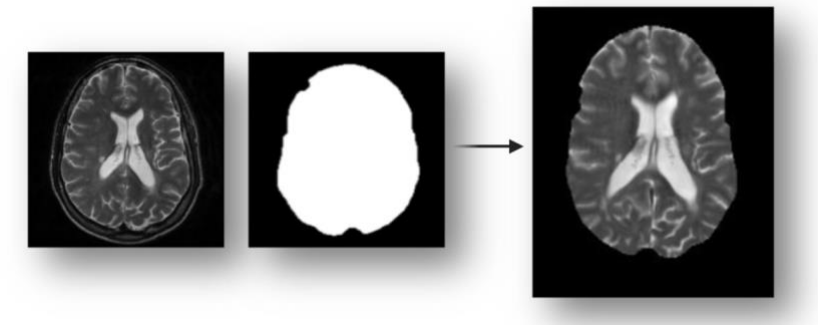


Figure 35 Result of the brain extraction on a MRI image

Before running the pipeline, it is possible to access some parameters that influence each of the core blocks:

- For the registration step, it is possible to select the type of transform and the type of reference atlas to be used
- For the brain extraction step, it is possible to select the size of the kernels and the number of morphological operations
- For the bias field correction step, it is possible to set the correction parameters.

As detailed in Section Deep Learning for tabular and imaging data the workflow of the pipeline was also integrated with the volbrain tool[31]: this tool was used to extract the cortical thickness images. The volbrain tool provides brain parcellation, cortical thickness, intracranial cavity, brain tissues using a 3D T1w MR scan.

To do so, the vol2Brain pipeline requires a 3D T1w MRI as NIfTI file, giving as output the desired images plus a report. The possibility of inserting the volbrain step into the workflow of the pipeline showed its flexibility in handling different image types.

4.5.2 Deep Learning for tabular and imaging data

The analysis of the ADNI dataset was useful to quantify the amount and the typology of data that could be collected to create a dataset to be used for the training of deep learning networks for the prediction of the evolution of AD in

[31] https://volbrain.net/

time, that considers heterogeneous sources of clinical data (i.e. tabular and imaging data), and to assess the impact on the performances of the network.

The tabular data included the variables (continuous or categorical) contained in ADNI MERGE. Out of the 86 available, 37 interesting ones were selecting, to be down reduced to 21, excluding the variables that referred to brain volume metrics. Moreover, the data included corresponded to the selected patients (see Section 4.2.1.3 a) ADNI extensive analysis).

The continuous variables that were selected to be included in the tabular DL model are:

- General variables: AGE, PTEDUCAT, APOE4
- NP tests scores: CDRSB, ADAS11, ADAS13, ADASQ4, MMSE, RAVLT_immediate, RAVLT_learning, RAVLT_forgetting, RAVLT_perc_forgetting, LDELTOTAL, DIGITSCOR, TRABSCOR, FAQ, MOCA, mPACCdigit, mPACCtrailsB

The categorical variables that were selected are: DX_in, PTGENDER, PTETHCAT.

Two different tabular datasets were created: one with 2 categorical classes (related to the time information, meaning labelling the TP with conv36 if the conversion happens between 0 and 36 months, and convover36 if the conversion happens after 36 months). The dataset with 3 classes differentiates between a conversion between 0 and 24 months, between 24 and 36 months, and over 36 months. The information about the switch of the diagnose is included as well in the label.

Such classes were the target of the neural networks that were tested. Three different networks models (fastai) were trained, giving as input the tabular datasets:

- Resnet 18
- Resnet 34
- Resnet 50

These models were chosen as the use of ResNet models is promoted in present literature (Xu, 2023). Moreover, the ResNet models come with a

different number of layers (18, 34, 50, 101, 152): the first three configurations were selected, to reduce the chances of promoting overfitting.

The same networks were also trained giving as input the T1-w MRI images corresponding to the TP of interest of the selected subjects.

The models were trained on 10 epochs, with a learning rate of 10^{-2} to obtain comparable results. Such results are presented and discussed in the Results chapter.

4.5.3 Flexidot tool and the neural network

Figure 36 Comparison shows the interesting parallel between the structure of a dotplot and the structure of an ITR: specific ITR sections (such as A, B, B', C, C', A' and D), that characterize the ITR itself, can be recognized in the dotplot.

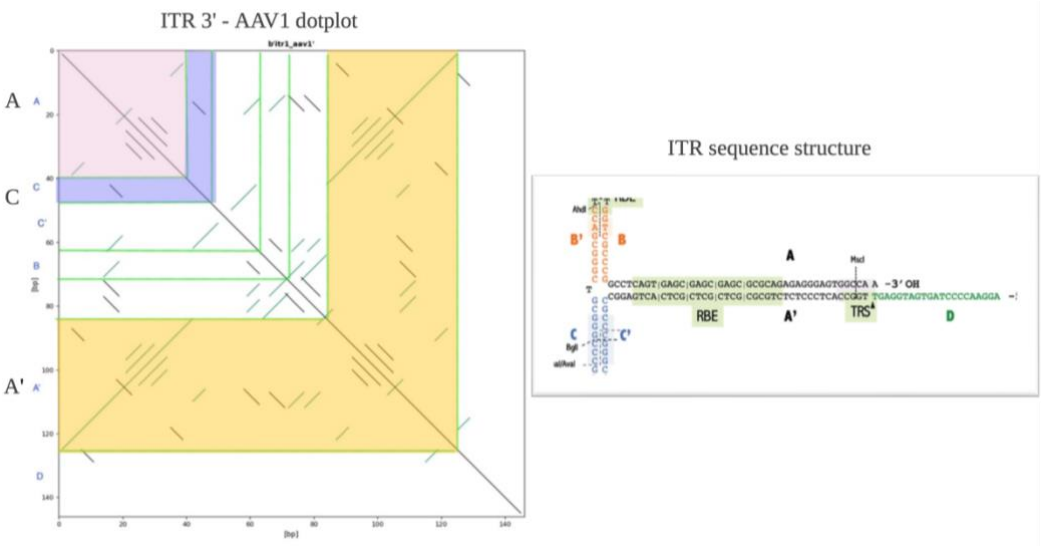


Figure 36 Comparison a dotplot of a ITR and the structure of the same ITR

The execution of Flexidot was automatized: as it normally runs using the command line, needing to manually insert the command every run, the flexidot.py script available on the GitHub repository[32] was modified to make it possible to integrate it with the Colab Shell on a Google Colab notbook: the

[32] https://github.com/molbio-dresden/flexidot

creation of the dotplots for the non-mutated and mutated sequences was therefore easier and faster, not subject to typo errors.

A dataset of *2113* self-dotplots, as described in 4.2.2.1 Dotplot images dataset creation, was created to train a network to be able to classify a dotplot giving as output the wild-type of AAV to which the ITR belongs, and which of the two ITRs it represents. Figure **37** Training images for dotplots classification shows a batch of the images that are used for the training of the network. As it is visible, the dotplots present different degrees of detail, as different K-values were used to create them. The processing pipeline presented before was modified to take as input the PNG images, crop them, resize them, and store them accordingly.

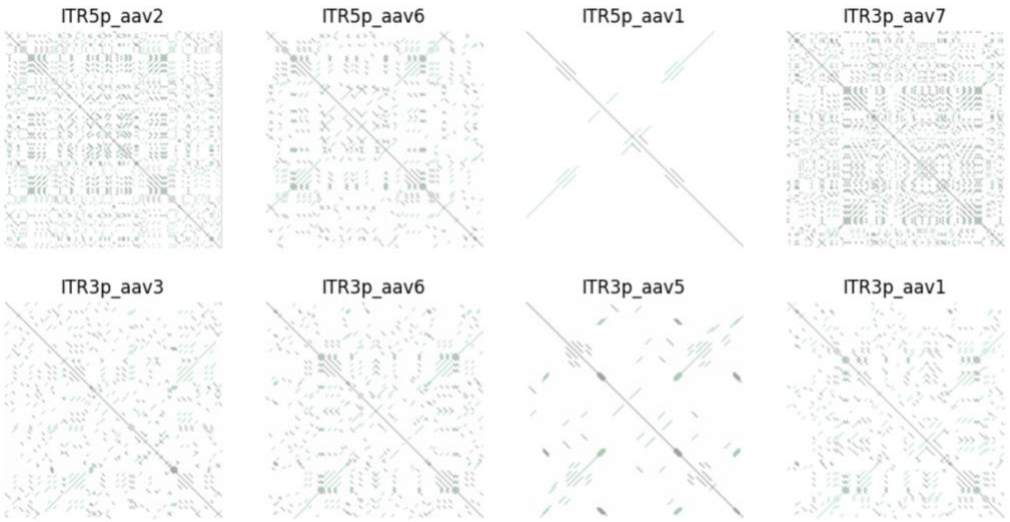


Figure 37 Training images for dotplots classification

The model was trained to recognize patterns and features associated with different ITR sequences and AAV types. A convolutional Resnet18 model, pre-trained network given the size of the dataset, was fine-tuned for single-label classification with a learning rate of 10^{-3}. The results will be discussed in the next chapter.

Chapter 5

Results

5.1 Answering the research questions

In Section 1.2 Aim and challenges the aim and challenges of the Research were presented. This Chapter directly refers to the questions there posed, and here mentioned again:

1) What are the main needs when developing computational solutions for the monitoring of neurological diseases such as MS and Alzheimer's disease?
    a. How straightforward is it to use public datasets and publicly available tools?
2) Is it useful to use different types of clinical data for disease monitoring?
3) Can the solutions developed for the other inquiries be applied to cover the needs of another field, such as the gene therapy with AAVs?

The following paragraphs take each of the three questions, presents the obtained results, and answers the question.

5.1.1 Research question 1

Question 1: *What are the main needs when developing computational solutions for the monitoring of neurological diseases such as MS and Alzheimer's disease? (How straightforward is it to use public datasets and publicly available tools?)*

The development of the processing pipeline, and the analysis of different datasets allowed to address this question. The pipeline aimed to fill the gap between heterogeneous datasets and the use of these as datasets for the training of Deep Learning networks: available image pre-processing tools have specific requirements and might not be adequate for extensive usage with heterogeneous datasets.

The processing times of the pipeline are here summarized:

Registration	4.5 sec
Other processing time	143.52 sec

The processing times are comparable to the state-of-the-art: after the registration is performed, the rest of the pipeline runs in 143.52 seconds, and this includes the steps of brain extraction, the bias field correction, and the storage of the images as .png or .jpg files to be used for Deep Learning applications.

- These performances, as already said, are comparable to the state-of-the-art, but it does not rely on external plug-ins, as many of the existing tools, avoiding inconsistencies and versioning issues.
- Being implemented in Python, it guarantees the possibility to directly access each functions' parameters to allow further customizations/optimizations.
- It differs from other existing software as it incorporates all the steps that are needed when preparing heterogeneous datasets for deep learning application, from the raw MRI scan (in different image formats) to the image that will be the input of a deep learning algorithm.
- It does not only work with a single image type (for example, only with T1-w or with T2-w MRI images), and it can handle as input DICOM and NIfTI formats, that are the most used ones in the field.

The reasoning behind the development of the pipeline, was realizing which are the main needs behind the development of computational solutions in this field: the lack of systematically collected data makes it hard to Researcher to access reliable sources, especially for diseases like MS and AD, that are heterogeneous and require an approach that avoids focusing on a single data or examination to assess well the patient's status. As this is a limitation on which little can be done, at least at the developer end, providing the right tools to use the data that is available, is crucial.

The task of data preparation and image pre-processing cannot be ignored or underestimated when constructing datasets for Deep Learning. The

pipeline here proposed, aimed to solve this crucial need. Its flexibility and performances were assessed also practically, using it in situations that tested its characteristics:

- Using it to process the image data retrieved from ADNI, testing its ability to work with different types of MRI images
- Integrating in its workflow the volbrain tool for the extraction of the cortical thickness from the MRI images, as this image modality is not provided by ADNI
- Using it as a backbone to a Deep Learning architecture, to process the images and to store

The steps related to the development of the pipeline, and to the analysis of the used datasets (ISBI 2015, ADNI, and NCBI for the AAVs), made it possible to deal with the question regarding the usability of public datasets.

The three datasets that were used, refer to the state-of-the-art, and are used in many applications: they are considered well documented, and good organised. For ISBI 2015, and NCBI this was the case but, for ADNI, the general assumption that a well-known dataset is immediately usable and trusful was not always true.

This might due to the fact that big dataset like ADNI are not built for a specific application, nor for not be used only for computational solutions and the systematicity of the collection of data is not always well organized between different clinical centers.

5.1.2 Research question 2

Question 2: *Is it useful to use different types of clinical data for disease monitoring?*

Alzheimer's disease was the disease on which this question was directed: the initial goal was the integration and analysis of ADNI clinical data with computational methods for the prediction of AD disease conversion in time through of a Deep Learning framework.

The focus laid both on the conversion from CN-MCI, and MCI-AD, in a variable time window. Moreover, the initial idea was to integrate different

sources of data (tabular data and MRI imaging data) and evaluate its impact of the on the performances of the network in the prediction of the evolution of AD in time.

Two different types of network were tested, and their performances are here reported:

Table 13 Networks performances

Tabular network	*Variables*	*Classes*	*Accuracy*
Resnet 18	21	3	0.5
Resnet 34	37	2	0.73
Resnet 50	21	2	0.76
Image network			
Resnet 18	21	3	0.5
Resnet 34	37	2	0.6
Resnet 50	21	2	0.7

Different variables were used, dividing the dataset with different classes as described in Section 4.5.2. The *Restnet 50* showed the best performances in all the configurations. The accuracy reached in the network that uses only tabular data, and the one in the network that uses only images, are comparable, considering that the use of images is more computationally expensive, in terms of pre-processing and network traing.

It was however not possible – and this is a limitation of the study – to develop a network that could take as input both tabular and imaging data. A constrain to this part of the Research was put by the impossibility of finding an optimal implementation in a reasonable time. Answering the question that opens this paragraph is not easy: the similar performances obtained by the

networks (Table **13** Networks performances) might be due to the fact that the cortical thickness image that were used (no provided by ADNI, but obtained during the pre-processing steps), correlate with the neuropsychological clinical scores used for training the tabular network, as the cortical thickness reflects the status of the gray matter, which deteroration indicates cognitive impairment. In retrospect, the use of cortical thickness images might have been a limitation to the development of this solution, introducing redundancy. The cortical thickness images are a valuable tool for clinicians, and their use was also encouraged by the clinicians of the Fondazione St.Lucia, but might not be informative enough to be used in this type of application.

Using different type of clinical data is fundamental for a personalized approach to disease monitoring, but translating this into practice it not immediate.

5.1.3 Research question 3

Question 3: *Can the solutions developed for the other inquiries be applied to cover the needs of another field, such as the gene therapy with AAVs?*

Answering this question was made possible by having to deal with a problem related to a field that has nothing to do with neurological diseases, medical imaging, or the monitoring of a disease. The use of AAVs vectors for gene therapy has requirements that fall outside the scope of those related to the rest of the Research. This represented a challenge and allowed to verify if the solutions proposed for the other topics here presented could, with some adjustment, be re-used. The concept of medical image was revisited, translating – thanks to the dotplots – the information contained in a sequence of bases into an informative graph that can be both visually and automatically analysed. The performances of the Resnet 18 architecture in classifying the dotplots were good, reaching the 80% of accuracy: although the numerosity of the dataset on which the network was fine-tuned was not huge, this allows to draw interesting considerations: Firstly, it is possible to overcome NGS limitations when it comes to the analysis of ITR heterogeneity, suggesting an approach not yet followed in the state-of-the-art literature. Secondly, dotplots are an informative tool both for analysis and result representation that can help gain insights on structures like the ITRs, which structure, variations, and characteristics have a direct impact on the safety of vectors for gene therapy.

Besides, the processing pipeline and the concepts behind it was re-used in this context for the processing of the dotplot images, and as a backbone for the deep learning classifier, and lent itself to its personalization for this task, using a different image format, removing the unnecessary blocks (i.e. skull-stripping), adding the cropping and resizing blocks, and storing the images as was necessary.

The adoption of the proposed solutions in different clinical settings or to include larger datasets would represents although a challenge: different clinical environments pose different questions and, therefore, have different requirements, which in turn give importance to different types of data. In this context, aiming at scalability means being ready to rethink the structure of the whole framework. The flexibility of each building block is therefore a key point: including, for example, data related to other biomedical signals (such as ECG, EMG or EEG) would require including different pre-processing steps in between those already existing, or it would be necessary to exclude some of these. Let us suppose of wanting to include the analysis of EMG signals graphs instead of on MRI images, to focus the attention on the walking impairment of patients with neurological diseases: this would still require the processing of images, but the processing needs would be completely different.

5.2 Challenges and limitations

Many were the difficulties encountered during the collection and the analysis of data, as well as during the development of the tools that were described in this chapter.

As the topic covered were different, for each one of them the approach to overcome the difficulties had to be recalibrated and adjusted, although some basic assumptions were always at the base of the process: the solutions had to be as simple as possible, and scalable. For each point, a brief indication on how the challenge was tackled is given in blue.

1) During the developing of **the image processing pipeline**, two were the main challenges:

    - Existing software propose an approach that relies on many external plug-ins, creating compatibility and versioning problems

- As the aim of the pipeline was being used to handle different images formats and to navigate through differently organized datasets, it was necessary to keep into consideration which parts of the pipeline needed to allow personalization and – if required – were it is possible to add other blocks to perform other functions.

→ A lot of time and effort was put in evaluating the best Python library to be used for building the blocks, and each block was kept well separated from the preceding and the following blocks to easy the process of including additional blocks.

2) Working with **the ADNI dataset**, the following challenges had to be faced:

- There is a lack of detailed ADNI documentation, and no precise insights are given on how the dataset is organized, or on how to effortlessly retrieve a big quantity of data
    - Data is differently labelled across ADNI phases, and some inconsistencies can be also found in each phase, as data is not systematically collected
- There is no biomarker for the prediction of the conversion from CN to MCI, nor for MCI-AD conversion, so there were no indications in the literature on the most informative data, therefore the support of clinicians was fundamental

→ Before starting with the implementation of any deep learning framework for the monitoring of AD, an extensive analysis of ADNI was performed, selecting the important documentation, identifying what configured as critical in terms of patients selection, of the download of the data from the web portal, and of making sure to correctly assign the correct image scans to the correct time reference of a specific patient.

3) The challenges encountered working **with AAVs** and their ITRs were:

- The lack of a literature reference that follows the same approach, targeting ITRs with the graphical method of the dotplots
- It was not possible to use a real dataset made of sequences of produced vectors

→ Initial evaluations on the ITR sequences and the correspondance in the dotplots were fundamental to assess if the approach could be of interest in practice. A relyable dataset (NCBI) was used to retrieve the AAV sequences, and a – small – dataset for this application was crafted introducing mutations in the sequences.

Chapter 6

Discussion, conclusions and future directions

This final chapter aims to reflect on the research process, knowing that measuring the success of a Research is trivial: it can be based on the number or quality of published articles, or on the attended conferences, or it can be based on the comparison with the work of peers in the field. It is also true, however, that every research path is different in its being shaped by the bumps in the road.

The methodology followed in the work presented in this Thesis can be, in retrospect, defined effective: the research questions posed at the beginning lead to a in depth-analysis of the state-of-the-art both from a technical and a medical point of view. The approach required – since the beginning – to acquire hands-on experience on the tools proposed by the literature, and on the sources of data essential to the developing of the proposed solutions. The variety of topics that were subject of research constituted both a resource and a limitation: on one hand, they directed every developing step to be scalable and flexible and, on the other hand, they limited the available time to focus only on one research question/topic.

There are many existing and very advanced applications that address the topics object of this discussion. However, it was noticed that, often, the more a solution is complex, the more difficult it is to use it the clinical practice or for research purposes. This research began, for each topic, defining the real needs for:

- The monitoring of neurological diseases, in particular Multiple Sclerosis and Alzheimer's Disease, via computational solution that focus on more than one clinical
- The use of several well-known sources of data

- The optimization of the analysis of the structure of AAVs vectors for vector safety.

For many neurological diseases, to ensure a good monitoring of their evolution, relying on a single medical evaluation is not advisable: for the diagnose of a neurological disease, typically, a protocol is followed, including MRI imaging and neuropsychological evaluations. During the follow-up, it is more common to rely on MRI imaging – for a question of time and resources. This is where computational solutions can find their space of applicability: to develop trustworthy and accurate algorithms, it is necessary to have access to heterogeneous sources of data, that need to be properly organized as training sets.

This – despite the availability of tools – is not always easy, but can be overcome with, on one hand, the support of the clinicians to keep an eye on the clinical application and, on the other, providing a tool able to take raw MRI images to organize them in a dataset to be directly fed into DL networks. Moreover, a crucial aspect often not included in algorithms for the prediction of disease progression is the time, as for various subjects the evolution of a disease follows different patterns.

The processing pipeline proposed in this research, allows customization and optimization, and can be used in various settings and with different type of datasets, allowing as well to be incorporated as a backbone of deep learning application, to flawlessly connect the source dataset with the training of neural networks.

Furthermore, the extensive analysis of ADNI and its features constitutes another contribution to the field: the method proposed and based on the data contained in ADNI, proposes a deep learning framework that focuses on the conversion from a healthy condition to cognitive impairment and to Alzheimer's, in a variable period.

Despite not all the things initially planned were done, such as the integration of the two networks into one, and the extension of this approach to Multiple Sclerosis, the approach is not yet adopted in recent literature: usually, the solutions focus to monitor the disease on periods of 1 year or 2 years, and not on months, and the prodromes of Alzheimer's disease are often

neglected, and that is considered is the change from an healthy condition to the clinical form of Alzheimer's, without considering the MCI condition.

When coming to the contributions related to the gene therapy via AAVs field, a similar approach was followed: the first step was identifying a practical need – in this case, the analysis of vectors – and to find a practical solution – that was found in the use of dotplots as valuable tools to quantify ITRs heterogeneity and, in turn, gain information about the AAVs structure. This posed the foundations of a novel and useful paradigm aimed to avoid using NGS technologies to assess vectors' characteristics.

Many are the limitations of this research, and some of these were already mentioned in previous chapters. Each of these limitations, opens possible future implementations such as:

- The inclusion of more MCI subtypes for the prediction of Alzheimer's prodroms, and the integration of other types of data (such as genetic data, contained in ADNI as well) in a single network for the optimization of the monitoring of Alzheimer's disease
- The inclusion of more clinical data from hospitals, to test the performances of the processing pipeline in the integration of these in the database for training the DL architecture
- Testing this approach on data of patients with Multiple Sclerosis (where MRI imaging and neuropsychological aspects play an important role as well) to evaluate similarities and differences in terms of applicability and performances
- The expansion of the evaluation with the dotplots, considering pair-dotplots (built comparing ITRs that belong to different AAVs) to detect mutations in the sequence that can compromise the packaging yield. This could lead to evaluations for looking not only at the ITRs in the vector, but to a whole sequence of a produced vector to detect the viral origin via ITR recognition, to create a pair-dotplot between the ITR sequences of the produced vector and the reference wild-type sequences, and to analyze the structure of the ITRs – and retrieve information about the vector itself, as the ITRs determine its stability and functionality.

Bibliography

Guizard, N. e. (2015). Rotation-Invariant Multi-Contrast Non-Local Means for MS Lesion Segmentation. *NeuroImage: Clinical*, 376–89.

Toga, A. W. (2019). The role of image regis- tration in brain mapping. *Journal of Big Data*, 3-24.

Alam, F. e. (2016). Evaluation of medical image registration techniques based on nature and domain of the transformation. *Journal of Medical Imaging and Radiation Sciences*, 178–93.

Mazziotta, J. C. (1995). A Probabilistic Atlas of the Human Brain: Theory and Rationale for Its Development: The International Consortium for Brain Mapping (ICBM). *NeuroImage*, 89–101.

Despotović, I. e. (2015). MRI Segmentation of the Human Brain: Challenges, Methods, and Applications. *Computational and Mathematical Methods in Medicine*.

Tustison, N. J. (2010). N4ITK: Improved N3 Bias Correction. *IEEE Transactions on Medical Imaging*, 1310–20.

Avants, B. B. (2014). The Insight ToolKit image registration framework. *Frontiers in Neuroinformatics*, 44.

Gambino, O. e. (2011). Automatic skull stripping in MRI based on morphological filters and fuzzy c-means segmentation. *nnual International Conference of the IEEE Engineering in Medicine and Biology Society*, 5040–43.

Howard, J. S. (2020). astai: A Layered API for Deep Learning. *Information*, 108.

Paciorek, A. M. (2024). Automated Assessment of Cardiac Pathologies on Cardiac MRI Using T1-Mapping and Late Gadolinium Phase Sensitive

Inversion Recovery Sequences with Deep Learning. *BMC Medical Imaging*, 43.

Praveen, S. P. (2022). ResNet-32 and FastAI for Diagnoses of Ductal Carcinoma from 2D Tissue Slides. *Scientific Reports*.

Chaudhury, S. e. (2023). Deep Transfer Learning for IDC Breast Cancer Detection Using Fast AI Technique and Sqeezenet Architecture. *Mathematical Biosciences and Engineering: MBE*, 10404–27.

Seibt, K. M. (2018). FlexiDot: Highly Customizable, Ambiguity-Aware Dotplots for Visual Sequence Analyses. *Bioinformatics (Oxford, England)*, 3575–77.

Kathrin M. Seibt, T. S. (2018). *FlexiDot documentation*. Retrieved from https://github.com/molbio-dresden/flexidot/blob/master/documentation/SupplementaryData.pdf

I. Grossman, A. M. (2010). Multiple sclerosis pharmacogenetics: personalized approach towards tailored therapeutics. *EPMA*, 317–327.

Ostellino, S. B. (2022). The Integration of Clinical Data in the Assessment of Multiple Sclerosis - A Review. *Computer Methods and Programs in Biomedicine*.

Murphy, M. P. (2010). Alzheimer's Disease and the Amyloid-β Peptide. *Journal of Alzheimer's Disease*, 311-23.

Au, H. K. (2022). Gene Therapy Advances: A Meta-Analysis of AAV Usage in Clinical Settings. *Frontiers in Medicine*.

S. Namkung, N. T. (2022). Direct ITR-to-ITR nanopore sequencing of AAV vector genomes. *Human Gene Therapy*.

E. D'Amico, e. a. (2020). The association between MRI brain volumes and computerized cognitive scores of people with multiple sclerosis. *Brain Cogn*.

L. Pham, e. a. (2020). Smartphone-based symbol-digit modalities test reliably measures cognitive function in multiple sclerosis patients. *MedRxiv*.

Ostellino, S. B. (2022). Brain MRI Images Pre-processing of Heterogeneous Data-sets for Deep Learning Applications. *Proceedings of the 15th International Joint Conference on Biomedical Engineering Systems and Technologies*, 115-20.

K.I.Berns. (2020). The unusual properties of the AAV inverted terminal repeat. *Human Gene Therapy*, 518–523.

P. Wilmott, L. L. (2019). A user's guide to the inverted terminal repeats of adeno-associated virus. *Human Gene Therapy Methods*, 206–213.

Tran, N. T. (2020). AAV-Genome Population Sequencing of Vectors Packaging CRISPR Components Reveals Design Influenced Heterogeneity. *Molecular Therapy. Methods & Clinical Development*.

J. Zhang, P. G. (2022). Subgenomic particles in raav vectors result from DNA lesion/break and non-homologous end joining of vector genomes. *Molecular Therapy - Nucleic Acids*, 852–861.

Dhar, T. e. (2023). Challenges of Deep Learning in Medical Image Analysis—Improving Explainability and Trust. *IEEE Transactions on Technology and Society*, 68–75.

Asan, O. e. (2020). Artificial Intelligence and Human Trust in Healthcare: Focus on Clinicians. *Journal of Medical Internet Research*, 852–861.

Abderrahim, M. e. (2020). Comparative Study of Relevant Methods for MRI/X Brain Image Registration. *The Impact of Digital Technologies on Public Health in Developed and Developing Countries*, 338–47.

Davis, M. (n.d.). *ADNI2 VISCODE2 Assignment.* San Diego: ADCS, Department of Neurosciences, University of California.

(2009). *Journal of English for Academic Purposes*, pages 180-191.

Xu, W. e. (2023). ResNet and its application to medical image processing: Research progress and challenges. (240), 107660.