Czech Technical University in Prague
Faculty of Electrical Engineering
Department of Cybernetics

Classification and Prediction of Interindividual Differences from Multimodal Neuroimaging Data

Doctoral thesis

Mgr. Barbora Rehák Bučková

Ph.D. programme: Electrical Engineering and Information Technology
Branch of study: Artificial Intelligence and Biocybernetics
Supervisor: Ing. Mgr. Jaroslav Hlinka, Ph.D.

Prague, August 2023

Thesis Supervisor:
Ing. Mgr. Jaroslav Hlinka, Ph.D.
Department of Complex Systems
Institute of Computer Science of the Czech Academy of Sciences
Pod Vodárenskou věží 271
182 00 Prague 8
Czech Republic

Copyright © August 2023 Mgr. Barbora Rehák Bučková

Declaration

I hereby declare I have written this doctoral thesis independently and quoted all the sources of information used in accordance with methodological instructions on ethical principles for writing an academic thesis. Moreover, I state that this thesis has neither been submitted nor accepted for any other degree.

In Prague, August 2023

.....
Mgr. Barbora Rehák Bučková

Abstract

Abstract: In the past years, the field of neuroimaging has undergone significant changes, not only with respect to advances in technology but also in analytical approaches. Classical statistics has given way to machine learning techniques, prompting a shift towards subject-level classification and prediction, offering insights beyond traditional group-level differences. However, this transition is not without its challenges, as it necessitates addressing the lack of appropriate methodologies, issues with model interpretability, and a scarcity of independent replication studies.

This thesis aims to explore, design, and critically evaluate such novel approaches in neuroimaging data analysis. The main focus is on developing and applying innovative techniques to achieve individual-level classification and prediction using neuroimaging data of different (neuro)imaging modalities. It dissects the key stages of neuroimaging data analysis, addressing data acquisition, preprocessing, feature design, and analytical methods. Special attention is given to dimensionality reduction and confounder control, as these factors significantly impact the reliability and interpretability of the results.

Across five research works, the thesis addresses the lack of independent replication studies, the problems of longitudinal prediction, and the potential of modality integration in neuroimaging data analysis. It emphasises dimensionality reduction as a vital tool to handle natural dimensionalities among modalities, offering promising solutions for further research. The thesis concludes with an exploration of pre-trained normative models and their significance in enhancing individualised prediction in longitudinal neuroimaging data analysis.

In summary, this thesis contributes to the advancement of neuroimaging data analysis by offering novel and robust methodologies for subject-level classification and prediction. Through critical evaluations and original works, it provides insights into multimodal fusion, individualised classification and prediction, and the potential of normative models.

Keywords: Neuroimaging, Classification, Prediction, Machine Learning, Multimodality, Normative Models.

Abstrakt

V posledních letech prošel obor neurozobrazování významnými změnami, a to nejen s ohledem na technologický pokrok, ale také na používané analytické přístupy. Klasická statistika ustupuje technikám strojového učení, což podnítilo posun směrem ke klasifikaci a predikci na individuální úrovni, které nabízejí poznatky přesahující tradiční skupinové rozdíly. Tento přechod však není bez problémů, neboť s sebou nese nutnost řešení nedostatku vhodných metodologických postupů, problémy s interpretovatelností modelů a absenci nezávislých replikačních studií.

Cílem této práce je prozkoumat, navrhnout a kriticky zhodnotit metodiky analýzy neurozobrazovacích dat. Zaměřuje se především na vývoj a použití inovativních technik k dosažení klasifikace a predikce na individuální úrovni s využitím neurozobrazovacích dat různých (neuro)zobrazovacích modalit. Rozebírá klíčové fáze analýzy neurozobrazovacích dat, a sice jejich měření, předzpracování, návrh příznaků a analytické metody. Zvláštní pozornost je věnována redukci dimenzionality a kontrole zavádějících faktorů, jelikož významně ovlivňují spolehlivost a interpretovatelnost výsledků.

V pěti výzkumných kapitolách se práce zabývá nedostatkem nezávislých replikačních studií, problémy longitudinální predikce a možnostmi integrace modalit v analýze neurozobrazovacích dat. Zdůrazňuje redukci dimenzionality jako důležitý nástroj pro zvládání přirozených dimenzionálních rozdílů mezi modalitami a navrhuje slibné postupy pro další výzkum. Práci uzavírá zkoumání předtrénovaných normativních modelů a jejich významu pro zlepšení individualizované predikce při analýze longitudinálních neurozobrazovacích dat.

Závěrem lze říci, že tato práce přispívá k rozvoji analýzy neurozobrazovacích dat tím, že nabízí nové a robustní metodiky pro klasifikaci a predikci na úrovni jednotlivce. Prostřednictvím kritických hodnocení a originálních prací přináší nové poznatky o multimodální fúzi, individualizované klasifikaci a predikci a potenciálu normativních modelů.

Klíčová slova: Neurozobrazování, Klasifikace, Predikce, Strojové učení, Multimodalita, Normativní modelování.

Acknowledgements

It is incredibly challenging to recapitulate five years of my life and research into one page and make justice to everyone who supported and helped me along my PhD journey. Nevertheless, I will try.

First and foremost, I want to thank my supervisor, Jaroslav Hlinka. Doctor Hlinka accepted me as his PhD student based on one interview and one recommendation, trusting my naive enthusiasm. He was truly an inspiration and an incredible mentor who provided me with unique resources to focus fully on this journey. I would also like to thank my other supervisors: Andre Marquand, Christos Davatzikos and Nicola Spotorno, whom I mostly met through my research visits to their research groups, who welcomed me warmly and were always happy to offer their expertise.

Next, I would not be able to follow through with my research with the energy and dedication needed without the constant support of my amazing colleagues. They believed in me at my lowest and were the most enthusiastic fans I could wish for. Thanks to Aničk̆a and Isa for all the coffees, overnights and advice, Jakub Kopal for teaching me cold-headedness, and Jakub Kořenek for endless discussions about everything and nothing. Of course, the thanks are extended to everyone I was lucky to meet along the way in Prague: Standa, Lucka, Nikola, Pavel, Giulio, David, Alberto, Luigi, Kajari, Michal, Helmut, Arthur, Jacob, Madhu and everyone else. Next, my big gratitude goes to my friends who welcomed me during my research visits and, in the short time we knew each other, created a place that felt like home for me abroad. In Nijmegen, there were Saige, Charlotte, Hannah, Stijn, Pieter, and Frank; in Philadelphia, I have to mention Hao, Kalina, Shepher, Ioanna, Andrew, and Abid.

Finally, none of this would have happened without the love and encouragement of my family, who blindly supported me, listened to my rants and joys lacking any context, and, despite circumstances, were always there. I will be forever grateful to my husband, Rastislav Rehák, without whom I would not have the courage even to apply for the programme. Who, despite spending the entirety of COVID with me in a one-room apartment, decided it would be a good idea to marry me and continues to support my academic growth even at his own cost.

List of Tables

List of Figures

List of Acronyms

AAL Automated Anatomical Labelling.
AD Axial Diffusivity.
AI Artificial Intelligence.
ALFF Amplitude of Low Frequency Fluctuations.
ANOVA Analysis of Variance.
AUC Area Under the Curve.

BBS Berg Balance Scale.
BOLD Blood Oxygenation Level Dependent.

CNN Convolutional Neural Network.
CSF Cerebrospinal Fluid.

DWI Diffusion Weighted Imaging.

EDSS Expanded Disability Status Scale.
EEG Electroenecephalography.
EPI Echo-planar Imaging.

FA Fractional Anisotropy.
FC Functional Connectivity.
FDR False Discovery Rate.
fMRI functional Magnetic Resonance Imaging.
FOV Field of View.

GAF Global Assessment of Functioning.
GLM General Linear Models.

ICA Independent Component Analysis.
IDP Image-Derived Phenotype.

LASSO Least Absolute Shrinkage and Selection Operator.

MD Mean Diffusivity.
ML Machine Learning.
MNI Montreal Neurological Institute.
MRI Magnetic Resonance Imaging.
MSIS Multiple Sclerosis Impact Scale.
MSWS Multiple Sclerosis Walking Scale.

PANSS Positive and Negative Syndrome Scale.
PCA Principal Component Analysis.

RD Radial Diffusivity.
RF Radiofrequency.
ROC Receiver Operator Characteristic.
ROI Region of Interest.
rs-fMRI resting-state fMRI.

SC Structural Connectivity.
SPM Statistical Parametric Mapping.
SVM Support Vector Machines.
SVR Support Vector Regression.

T Tesla.
TBSS Tract-based Spatial Statistics.
TE Echo Time.
TFCE Threshold-free Cluster Enhancement.
TR Repetition Time.
TUG Timed Up and Go.

VBM Voxel-Based Morphometry.

Contents

Preface

Motivation

Historically, the field of neuroscience sought to understand the structure of the nervous system via the study of *post mortem* samples [1], whereas the allocation of functional processes was conducted through behavioural studies of disabled patients [2], [3]. This process slowly progressed until the second half of the 20th century, which witnessed considerable changes as neuroscience found its place as a distinct scientific discipline, rapidly evolving due to technological advancements. The development of non-invasive neuroimaging techniques such as electroencephalography, magnetic resonance, and its derivatives enabled us to study the brain more thoroughly than ever before.

Yet, despite these breakthroughs, the field faced scepticism and challenges. In 1988, Terrence Sejnowski, a pioneer in computational neuroscience and large-scale simulations, questioned the potential of neuroimaging data analysis, arguing that the brain acquisition methods were "severely limited in spatial and temporal resolution," undermining the potential of data analysis [4]. However, by 2014, in a surprising turnaround, Sejnowski advocated for *"Putting big data to good use in neuroscience"* in an article published in *Nature Neuroscience*[5]. This apparent contradiction can be attributed to the tremendous progress in neuroimaging methods, as illustrated in Figure 1, which demonstrates the technological advancements made over the intervening years. These advances have not only expanded our data acquisition capabilities but also called for the development of analytical approaches used to process acquired data in order to understand and interpret them correctly.

A decade after Sejnowski's initial criticism, the term *Neuroinformatics* was defined as "Combining neuroscience and informatics research to develop and apply advanced tools and approaches essential for a major advancement in understanding the structure and function of the brain" [6]. Indeed, the involvement of rigorous computational methods in the field was a great promise for neuroscience in the 21st century. Soon, the anatomical and functional organisation of a healthy brain was described as a topologically complex interconnected network – connectome, which balances regional and functional speciali-

sation and strong integration [7]–[9]. Connectome consists of two complementary parts – functional and structural. Each is constructed using distinct neuroimaging techniques and carries either information about the (dis)similarity of activity across the brain in the former case or about the existence (and intensity) of anatomical white-matter axonal connections in the latter case. Although both sub-connectomes provide valuable insights, it's becoming evident that combining data from different brain imaging methods is necessary. Doing so prevents us from studying two separate types of human brains: one with structure but limited function and the other with function but unclear structure.

Both sub-connectomes offer valuable insights to the research community, but it became increasingly clear that integrating information from multiple neuroimaging modalities was essential to avoid studying two distinct human brains–one with a structure but limited functional capacity and the other with meaningful functions but no clear structure.

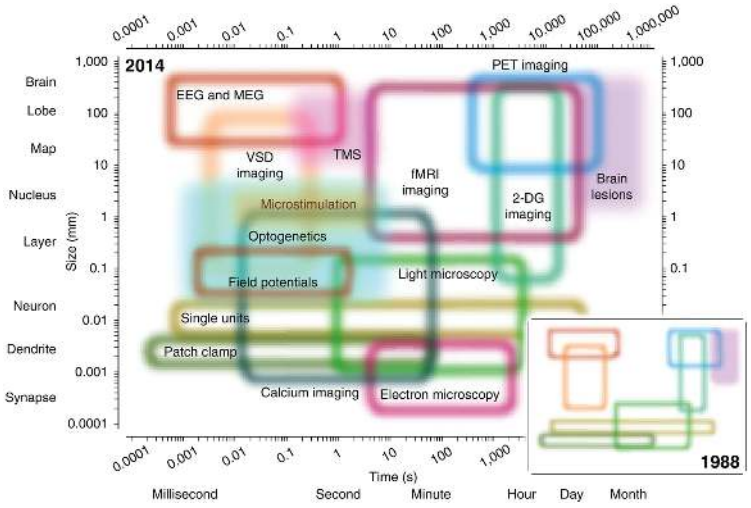

Figure 1: **The comparison of acquisition methods of the brain in 1988 and 2014 in spatiotemporal domain.** Open regions represent measurement techniques; filled regions, perturbation techniques [5].

Further progress in the field is documented in bibliometric studies. In 2017, Yeung *et al.* constructed a landscape of neuroscientific research mapping the most frequent terms used in neuroscientific literature within a period of 2006-2015 [10]. Whereas in 2006, the terms *performance*, *task*, and *processing* were among the most used, they were absent in the 2015 ranking, when *scale*, *diagnosis*, and *classification* were more in focus. Especially the term *classification* indicated the formation of significant movement in the whole-brain data analysis. During this period, new research objectives emerged, extending beyond the exploration of general principles of cognition. Instead, the focus shifted towards unravelling inter-individual differences through extensive datasets, laying the foundation for the future of individualised medicine. Methodologically, machine learning techniques

became natural candidates to aid in this endeavour.

Initially used for brain image processing, machine learning dominates a new niche within neuroimaging data analysis: brain biomarker identification. Indeed, the prospect of uncovering objective neuroimaging patterns that signify the onset and progression of neurological and mental disorders holds significant promise [11]. However, as the scope of problems solved with machine learning changes, it is necessary to re-evaluate the way the methods are used and the interpretational power assumed by their use. While a neural network applied for precise segmentation may not be fully understood, a neural network used to develop unique brain signatures characterising certain brain states should be comprehensible to health practitioners [12].

The context of my dissertation thesis is set in this challenging interplay of trade-offs between simplicity and complexity, interpretability and accuracy, and individuality and generalisability. Throughout my thesis, I will introduce and explore analytical techniques for analysing neuroimaging data, considering the aforementioned trade-offs and trying to strike a balance. I will examine how to account for the interindividual differences and disentangle variability of interest from confounders. Finally, I will connect all of these topics by exploring options for multimodal analysis of neuroimaging modalities.

Aims of the thesis

The thesis has two overarching aims. Firstly, to review and explore advanced methodologies in neuroimaging data analysis, including its essential components such as preprocessing techniques, dimensionality reduction, and machine learning algorithms.

Secondly, to improve and critically apply these methodologies to address classification and prediction challenges in neuroimaging.

This will be achieved by:

1. Conducting a replication study to assess the outcomes of deep learning algorithms in neuroimaging data analysis, focusing on influential features and applicability.

2. Investigating the integration of structural and diffusion neuroimaging data for longitudinal prediction.

3. Integrating multiple neuroimaging modalities and investigating their combined potential to capture complex patterns and relationships within the brain.

4. Contributing to the neuroimaging community by sharing neuroimaging data.

5. Developing modality-free normative modelling method for longitudinal analysis.

Outline of the thesis

The thesis is organised as follows:

Chapter 1

In the first chapter, I introduce fundamental anatomical and physiological concepts related to the brain, providing essential background knowledge for the subsequent chapters. I also introduce the radiomic framework of an analytical pipeline, which will be further expanded in the following chapters.

Chapter 2

The second chapter offers an overview of the historical development and advancements in non-invasive brain imaging techniques. It explains the underlying physical principles and technical intricacies involved in data acquisition across modalities relevant to this thesis.

Chapter 3

Each neuroimaging modality captures distinct information about the brain and also presents unique challenges and artifacts that must be addressed through tailored preprocessing procedures. In this chapter, I focus on preprocessing neuroimaging data, discussing specific techniques and strategies for artifact correction and quality enhancement. Additionally, I explain the extraction of neuroimaging features from individual modalities and provide insights into multimodal fusion approaches.

Chapter 4

Although neuroimaging data analysis is a relatively young field, it already underwent dynamic changes with respect to the methods we use for their analysis. This chapter reviews the key concepts and ideas that have shaped the current landscape of machine learning in neuroimaging analysis, emphasising their appropriate application context. The chapter also highlights the significant challenge of dimensionality in neuroimaging data and examines various dimensionality reduction methods that can be employed prior to model fitting. Finally, it explores analytical approaches suitable for neuroimaging data analysis.

Chapter 5

The final theoretical chapter sheds light on the often-overlooked issue of confounders in neuroimaging data analysis. Confounders are typically clinical variables that can influence

both the outcome of interest and explanatory variables, posing a significant risk of bias in neuroimaging studies. This chapter discusses potential strategies and methods for mitigating the impact of confounders, aiming to enhance the validity and reliability of neuroimaging data analysis.

Chapter 6

The use of deep learning algorithms in neuroimaging data analysis has gained significant traction. However, it is essential to also critically examine the outcomes of these algorithms in terms of influential features and replicability. In the first results chapter of this thesis, I will replicate the findings of a deep neural network that identified differences in EEG time series between men and women. I extract the key features identified by the authors as crucial for the classifier and replicate the findings using a simpler method on the longitudinal dataset.

Chapter 7

In this chapter, I integrate structural and diffusion neuroimaging data on subjects who have experienced a stroke to predict cognitive deficits one year after the neuroimaging data were collected. The analysis not only combines the two modalities but also explores different levels of spatial resolution to investigate the impact of the grain of the analysis on the results.

Chapter 8

Building upon the multimodal analysis, I extend my investigation to incorporate three modalities. The goal of this chapter is twofold: first, to classify patients with multiple sclerosis and healthy controls, and second, to identify neuroimaging correlates that indicate motor deficit. I evaluate multiple approaches across different experimental setups to determine the most effective approach.

Chapter 9

Using the methodology designed in the previous chapter, this section strives to enhance its efficacy in classifying individuals with schizophrenia and healthy controls. I revisit feature design and data integration techniques with the aim of achieving improved performance.

Chapter 10

The scarcity of training data poses a significant challenge when using deep learning approaches in neuroimaging data analysis. Data-sharing initiatives are crucial to address

this issue and promote progress in the field. In this chapter, I describe my efforts to contribute to the community by sharing neuroimaging data.

Chapter 11

In this chapter, I revisit the discussion on two major challenges in neuroimaging data analysis: confounders and limited sample size. Using normative modelling, I developed a method that enables the analysis of longitudinal neuroimaging data based on pre-trained models derived from large-scale datasets, eliminating the need for the original data. This method can be applied to any neuroimaging data for which pre-trained models exist. I demonstrate the effectiveness of this method in the longitudinal analysis of first-episode schizophrenia subjects, showcasing superior results compared to traditional approaches of controlling for confounders.

Chapter 12

The final chapter of this thesis encompasses the discussion of my original findings and their alignment with the aims of the thesis. Additionally, I outline future extensions and ongoing projects that are relevant to the scope of this thesis.

While the findings presented in this thesis are the results of my independent research, I use the plural form throughout to discuss general concepts and findings, emphasising the collective knowledge and experiences within the field of neuroimaging data analysis. Additionally, at the outset of each research chapter, due acknowledgement is given to my colleagues who have contributed their valuable insights to specific aspects of the research.

Chapter 1

Navigating brain

In this introductory chapter, we will provide the essential knowledge related to the brain that is relevant to this thesis. It is crucial to acknowledge that neuroscience is a vast and multidisciplinary field, relying on the collective expertise of professionals with diverse backgrounds and levels of knowledge in various domains. This thesis is written from the perspective of a data scientist. Therefore, the primary objective of this chapter is not to delve deeply into the intricacies of action potentials' physiology or the detailed anatomy of the brain. Instead, we aim to provide a conceptual framework that encompasses the key aspects influencing neuroimaging data acquisition to the extent necessary for its appropriate analysis. For a comprehensive, in-depth course on neuroscience, we recommend [13].

1.1 Zooming in

Although not the most abundant cells of the brain, neural cells or neurons are the foundational components of its architecture. Neurons are essential in transmitting and processing information within the nervous system, enabling communication and coordination across body functions. Their unique structure is finely tuned to efficiently transmit electrical signals throughout the body. It consists of the cell body, known as the soma, and the axon (Fig. 1.1).

The soma contains the nucleus of the cell. It serves as a foundation for the dendrites, which are abundant tree-like structures extending from the neuronal body and receiving incoming signals from other neurons.

On the other hand, the axon is typically a single elongation of the neuron, varying in length from micrometres to over a meter, depending on the type and location. The primary role of the axon is to transmit electrical signals to other neurons. To facilitate efficient transmission, certain parts of the axon are insulated by a lipid substance called

myelin, which is produced by either oligodendrocytes or Schwann cells, depending on the neuron's location within the body. Along the length of the axon, there are specific regions where myelin is absent, exposing the axonal membrane. These exposed regions are referred to as the nodes of Ranvier, and they play a crucial role in facilitating the conduction of nerve impulses. During the propagation of an action potential, the electrical impulse "leaps" along the myelin sheath from one node of Ranvier to the next, in a process known as saltatory conduction. This conduction mode is considerably faster and more energy-efficient than the continuous conduction observed in unmyelinated axons.

At the end of the axon, there are specialised structures called terminal boutons. These terminals form connections, known as synapses, with other neurons or target cells. When an action potential reaches the axon terminals, it triggers the release of chemical messengers called neurotransmitters into the synapse. These neurotransmitters then bind to receptors on the target cells, transmitting the signal from one neuron to the next or to the effector cells, such as muscles or glands.

Overall, neurons employ two mechanisms to transmit signals. The first mechanism is the action potential, a rapid electrical charge that travels along the neuron's axon, starting from the soma. Typically, action potentials are contained within a single neuron. However, in certain cases, when neurons are in close proximity, the electric signal can "jump" between the axon of one neuron onto the dendrite of another through specialised connections called *gap junctions*.

The second mechanism involves the neurotransmitters released from the terminal endings of one neuron's axon into the synapse. They then bind to receptors on the dendrites of the target neuron. This interaction can either depolarise the target neuron, producing an excitatory postsynaptic potential or hyperpolarise it, producing an inhibitory postsynaptic potential. These changes in the target neuron's membrane potential influence its likelihood of generating an action potential. When a sufficient number of neurons in close proximity release neurotransmitters onto the dendrites of a target neuron, it can trigger an action potential, initiating further communication within the nervous system.

1.2 Zooming out

The human central nervous system has evolved into four distinct parts: the spinal cord, brainstem, cerebellum, and cerebrum. The spinal cord extends from the base of the brain and serves as a channel for transmitting signals between the brain and the rest of the body. The brainstem is the evolutionarily oldest part of the brain and consequently houses vital centres responsible for essential bodily processes such as breathing.

Cerebellum, which translates to "little brain", exhibits numerous fissures resembling

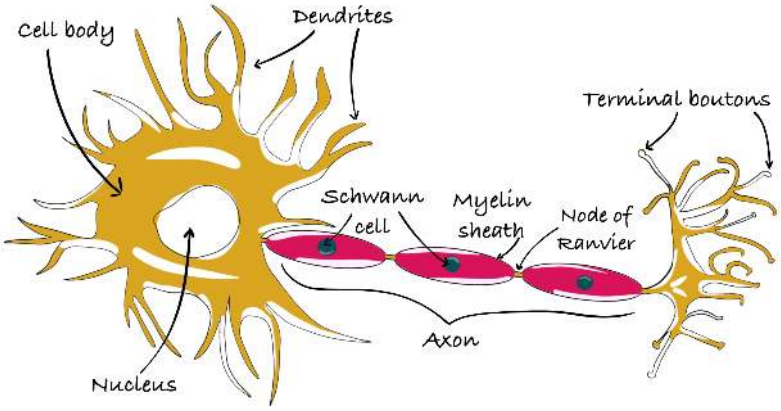

Figure 1.1: **A scheme of a neuron.**

the cerebrum. It is situated behind the brainstem at the back of the skull. The cerebellum is highly specialised in coordinating and regulating voluntary movements, maintaining balance, posture, and facilitating motor learning. However, it has also been associated with cognitive functions like attention, language, and emotional regulation.

The cerebrum, the brain's largest and most prominent part, is characterised by its smooth outer surface with fissures and ridges, known as sulci and gyri. The longitudinal fissure divides it into two hemispheres, yet they remain interconnected and communicate. Each hemisphere contains four lobes: frontal, occipital, temporal, and parietal (Figure 1.2). These lobes exhibit varying degrees of functional specialisation. For instance, the occipital lobe houses the primary visual cortex responsible for processing visual stimuli. In contrast, the prefrontal cortex in the frontal lobe is associated with higher cognitive functions such as planning and problem-solving.

The outer layer of the cerebrum is known as the cerebral cortex. It is predominantly composed of grey matter, which consists of (mostly) six layers of neuronal cell bodies organised into cortical columns, forming a thin layer measuring just a few millimetres. Despite its thinness, grey matter plays a crucial role in information processing and integration in the brain. Underneath, brain regions are interconnected by white matter, which consists of axons packed tightly together. The distinction between grey and white matter is visually notable, with white matter appearing white due to myelin sheaths around the axons. Another important component to consider is Cerebrospinal Fluid (CSF), a clear and colourless fluid that fills the cavities within the brain and spinal cord. Although not technically a tissue, CSF is often referred to as such for generalisation purposes. It serves as a protective medium, absorbing mechanical shocks, regulating the chemical environment of the brain, aiding waste removal, and maintaining stable intracranial pressure. Differentiating between these three "tissue," types is crucial for neuroimaging data analysis.

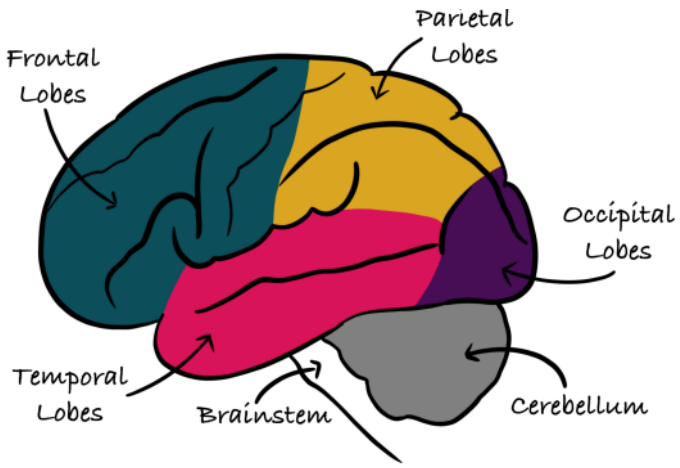

Figure 1.2: **A sketch of the brain.**

1.3 Radiomics and neuroimaging data analysis

As we will learn from later chapters, imaging techniques revolutionised our knowledge of the human body and significantly widened our options for its study. However, most of these techniques are universal and don't have to be used solely for studying the brain.

Radiomics focuses on the high-throughput extraction of numerous image features from radiographic scans [14]. Originally mainly used for analysing oncological scans, it is rooted in mining hospital databases, which house standard radiological images ensuring a large and diverse pool of scans [15]. The essence of radiomics lies in the power of big data, naturally intertwining itself with advanced Machine Learning (ML) methods and Artificial Intelligence (AI). It has a well-defined framework for conducting the analysis [16].

While there is a significant overlap between neuroimaging data analysis and radiomics in extracting and analysing imaging features, they are not entirely synonymous. Neuroimaging data analysis only focuses on brain images. It often operates within the context of predesigned studies that seek to address specific questions about brain function, which require data beyond what is available in hospital records alone. Additionally, although advancing, the data-sharing initiatives within and between hospitals are not yet a standard practice making a small number of subjects an analytical challenge, limiting the application of AI methods without risking overfitting.

However, despite these differences, the well-established framework and methodologies of radiomics offer valuable insights and tools that can be effectively applied to neuroimaging data analysis.

The radiomics analytical pipeline consists of the following steps, which will be more thoroughly discussed in later chapters:

1. **Planning**: This crucial stage involves thorough research planning, including setting clear goals, designing realistic means to achieve those goals, anticipating potential

problems and devising appropriate solutions. Although not directly related to this thesis, the importance of planning cannot be undermined.

2. **Data curation**: The quality of study results heavily relies on effective data acquisition and curation. As the saying goes, "Garbage in, garbage out." Therefore, investing time, energy, and resources into ensuring proper acquisition practices is essential [17]. Curation becomes even more critical in radiomics since, unlike in smaller neuroimaging studies, large neuroimaging databases are not constructed with a single well-defined question in mind. If a database is primarily designed to be large, real-life biases may permeate the AI algorithms used for data mining. Therefore, the curation stage becomes crucial for ensuring unbiased results. Given that knowledge of the data acquisition process is vital for proper preprocessing and data analysis, we will discuss it in the following chapter.

3. **Preprocessing and segmentation**: Once the data is acquired, it is essential to address noise and artifacts accompanying any data acquisition. Understanding the acquisition process allows us to anticipate the presence and form of artifacts, enabling the design of appropriate solutions to enhance the signal-to-noise ratio as much as possible.

4. **Feature extraction**: Feature extraction requires a deep understanding of the problem at hand and the ability to extract meaningful information from neuroimaging data. Features can be generated at different temporal and spatial resolution levels, depending on the dataset size and the task at hand. Striking the right balance between generating too many or too few features is crucial. This challenge is further amplified when performing multimodal analysis, where features should reflect dependencies and correlations across different modalities. The issue of proper feature generation will be discussed in a separate chapter, together with the preprocessing stage.

5. **Model building**: Selecting an appropriate model for addressing the research question (typically classification or prediction) requires trade-offs between simplicity and complexity, avoiding underfitting or overfitting and achieving both individuality and generalisability. At this stage, the limited number of subjects, a common bottleneck in neuroimaging data analysis, becomes particularly evident. With a small number of subjects (up to a hundred), the significant disproportionality between the available number of features is a considerable impediment and requires caution and more conservative approaches. However, while intuitively, more samples allow for the use of more sophisticated algorithms, it does not guarantee their superiority

over simpler approaches. It is, therefore, advisable to have a simple classifier as a baseline for comparing more sophisticated methods. The factors to consider when choosing a model will be discussed in a dedicated chapter.

6. **Internal test performance**: Building a model goes hand in hand with appropriate validation. The performance of the model should be assessed using internal tests to ensure its effectiveness within the given dataset.

7. **External validation**: Once the model achieves satisfactory performance, it is good practice to also evaluate its performance on an independent dataset. In radiomics, where population-level conclusions are often drawn, the model should ideally be validated on an independent dataset from a different scanning site. However, this can be challenging due to confounding factors such as scanner effects, especially in neuroimaging data analysis. In such cases, transfer learning methods may prove useful. The discussion of confounders is covered in the final theoretical chapter of this thesis.

Chapter 2

The principles of neuroimaging

In this chapter, we will briefly cover the selected neuroimaging techniques that are commonly used in contemporary neuroscience research. While the chapter does not aim to be comprehensive, it will provide sufficient information for the reader to follow the thesis. For those who wish to delve deeper, we recommend [18].

2.1 Setting up the stage

Although neuroscience, the study of the nervous system (including the brain, spinal cord, and peripheral nervous system), has roots dating back to ancient Egypt [19], the non-invasive study of brain structure and function is much more recent. In 1895, Wilhelm Roentgen's discovery of X-rays revolutionised non-invasive visualisation of the human body, opening new avenues for research, especially on hard tissues [20]. However, X-rays were unsuitable for brain imaging for several reasons, including the harmful effects of ionising radiation. Moreover, the hard tissue surrounding the brain absorbs most of the X-rays, limiting the details of the brain's structure. Nevertheless, the technology was the cornerstone of innovation, giving rise to methods such as computed tomography in 1961. Unlike traditional X-rays, computed tomography takes a series of X-ray images from various body angles, producing 3D images. Additional improvements in post-processing diminished the effect of hard tissue interference, enabling computed tomography to produce brain images of significantly higher quality. Nevertheless, the core problem of harmful radiation persevered, and an alternative need to be found to improve brain imaging.

Shortly after the introduction of computed tomography, Paul Lauterbur published the first images obtained through nuclear magnetic resonance [21] (Figure2.1). Nuclear magnetic resonance describes the behaviour of atomic nuclei in a strong magnetic field when exposed to radio waves. Originally, this technique was developed for studying atomic nu-

clei and their interactions in molecular structures. However, as Lauterbur notes in his 1974 article:"*When the magnetic field gradients imposed upon objects are large compared with the peak widths produced by intrinsic effects, the inhomogeneously broadened signals contain information on the spatial distributions of the nuclei or electrons at resonance. Under these circumstances, the magnetic resonance becomes a technique for studying structure above the molecular level.*" [21].

In summary, nuclear magnetic resonance uses strong magnets and sensitive techniques to detect the weak signals emitted by hydrogen atom nuclei in tissue. This non-invasive imaging method produces three-dimensional images of the target tissue and is considered safe, with no harmful effects from the time spent in the scanner. Because of its favourable properties, it became an essential tool in most clinical facilities by the 1980s. However, due to negative connotations associated with the term "nuclear", the original name was changed to Magnetic Resonance Imaging (MRI). MRI was a significant milestone leading to another imaging revolution, for which Paul Lauterbur and Peter Mansfield were awarded a Nobel Prize in 2003.

Back in the early 20th century, amidst the frenzy around X-rays, another method emerged for studying the brain. In 1924, Hans Berger recorded the first human electroencephalogram [22]. Unlike X-rays, Electroenecephalography (EEG) was not designed to be an imaging technique but rather a method to study the activity of the brain by measuring excitatory postsynaptic potentials. The device comprises two main parts: a cap and an amplifier. The cap consists of electrodes that attach to the scalp and measure the electrical activity of the brain. The amplifier records and amplifies the signals from individual electrodes filters out electrical noise, and then digitises and records the signals with a computer.

Compared to imaging devices, the construction and deployment of the EEG device were considerably easier, and as a result, by the 1940s, it had become widely used by cognitive scientists. Over time, EEG research has provided valuable insights into various brain disorders, including epilepsy, stroke, encephalitis, and others. Despite not being an imaging technique by design, modern high-density EEG tools, when combined with advanced computational techniques, can be regarded as a form of neuroimaging, albeit with limited spatial resolution.

2.2 MRI principles

Although it is not essential for this thesis to deeply understand the physical principles of MRI, it is necessary to grasp the basic laws and specifics tied to individual modalities. In this part, we will describe the basic principles of MRI and its three modalities – structural

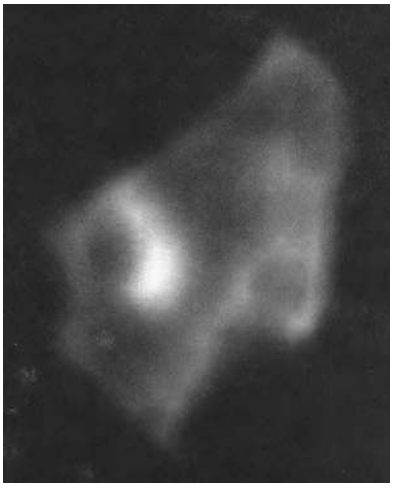

Figure 2.1: **Lauterbur's MRI (at that point called NMR) image of a cherrystone clam.** Adopted from [21].

MRI, functional Magnetic Resonance Imaging (fMRI) and Diffusion Weighted Imaging (DWI) [23]–[25].

The human brain is predominantly composed of water, with hydrogen atoms being the most abundant. Each hydrogen atom has a single positively charged proton that spins on its axis, generating a small magnetic field that can be manipulated using an MRI scanner. The principle of MRI relies on the alignment of these hydrogen protons in the main magnetic field before being disrupted by a specific Radiofrequency (RF) pulse. After the RF pulse is applied and discontinued, the protons return to their original alignment and emit an RF pulse that is detected, amplified, and converted into a digital signal. In short, MRI detects and measures the signal re-emitted by the hydrogen atoms in the body's tissues when exposed to strong magnetic fields. The device consists of three essential parts: a superconducting magnet, a gradient coil, and an RF coil.

Magnet

The superconducting magnet is the most significant part of the device. This component determines the strength of the magnetic field produced by the machine, known as B_0, and is measured in the units of Tesla (T). While the commonly used devices have a strength of 1.5T and 3T, advanced 7T scanners are increasingly becoming popular. To highlight the sheer magnitude of the generated magnetic field, the Earth's magnetic field has a strength of only approximately $3.2*10^{-5}$T. The strength of the field is crucial, as it directly impacts the resolution and quality of the final image. Once activated, the protons in the tissue align in the direction of the B_0 magnetic field and precess around it in phase coherence at a characteristic frequency called Larmor frequency, which is directly proportional to the strength of the field.

Gradient coil

Gradient coils superimpose a varying magnetic field on the B_0 magnetic field, enabling spatial encoding of the MR signal along a specific direction and, consequently, the reconstruction of 3D images. Switching on and off the gradient coil at specific times allows the MR signal to be selectively excited and detected from a specific volume element. The setup of the gradient coil can also produce spatially varying magnetic fields to suppress unwanted signals or enhance specific features in the MR image. Two important parameters of gradient coils are the strength of the gradient field they can produce, and the switching times, both of which affect the thickness of the tissue we can study.

Radiofrequency coil

The RF coils generate the B_1 field, causing part of the protons originally aligned with the B_0 field to rotate away into a transverse plane. More specifically, the RF coils transmit radiofrequency pulses at Larmor frequency, causing the protons to absorb energy and enter a higher-energy state, that is, not aligned with the B_0 field. When the pulse is turned off, the energy is released, and protons return to their original state, emitting a radiofrequency signal detected by the RF coil. MRI devices contain a variety of RF coils specific to imaging different parts of the body. These coils can be separate sets for emitting and receiving the signal or one coil serving both purposes.

As with any advanced technology, MRI is controlled by a "program" called a *pulse sequence* which coordinates the switching of gradient and RF coils during acquisition. Pulse sequences depend on several parameters that affect the acquisition. These parameters are routinely reported in every neuroimaging study and significantly affect the contrast and quality of the resulting image:

- **Repetition Time (TR)**: The time interval between the beginning of one RF pulse and the beginning of the next RF pulse (Figure 2.2). The shorter the TR, the less time the hydrogen protons have to recover and align with the B_0 field. It is measured in milliseconds.

- **Echo Time (TE)**: Describes the timing between the application of an RF pulse and the collection of the resulting signal (Figure 2.2). In other words, it is the time between the centre of the RF pulse and the peak of the echo signal received by the

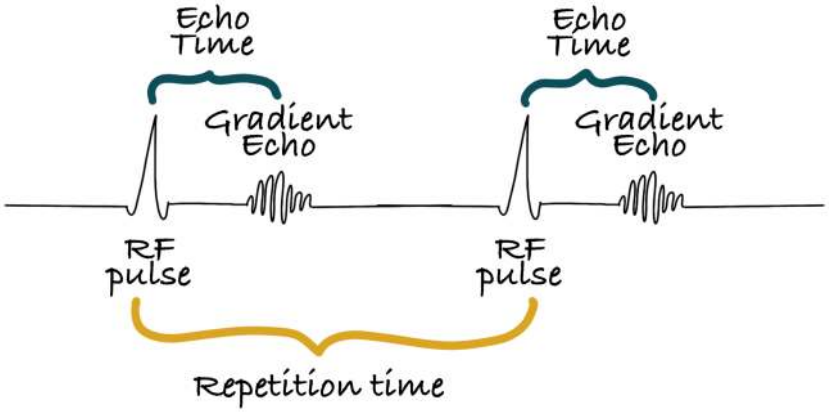

Figure 2.2: **Repetition and echo time in pulse sequence.**

receiver coil. It is typically measured in milliseconds and can be adjusted during the MRI scan to optimise image contrast and resolution.

- **Flip angle**: The angle by which the hydrogen atom is rotated away from the original B_0 field after an RF pulse. It is measured in degrees or radians.
- **Field of View (FOV)**: The size of the area that's being visualised (in mm).
- **Matrix size**: The number of pixels in the FOV.
- **Slice thickness**: the thickness of the "slice of brain" being scanned (mm).

Different types of images are created using different pulse sequences, depending on the tissue properties we want to highlight. The final signal is then represented as either a pseudo three-dimensional or four-dimensional image series, depending on the sequence used. Similar to a pixel, the smallest unit of information in a 2D image, the equivalent unit in 3D structures is referred to as a voxel. Voxel can be visualised as a cube, with two of its dimensions determined by the FOV and matrix size, while the third dimension is the slice thickness. The voxel size varies depending on the pulse sequence, ranging from, for example, 0.5 mm $\times$ 0.5 mm $\times$ 0.5 mm for detailed structural images to, for example, 2 mm $\times$ 2 mm $\times$ 2 mm for fMRI series. In the following sections, we will review the relevant tissue contrasts and link them to specific neuroimaging modalities.

2.3 Taking a picture

While we often visualise the brain as a three-dimensional structure, neuroimaging data is acquired as a sequence of two-dimensional slices, which are then reconstructed to form a comprehensive 3D representation. These slices are typically examined along three main

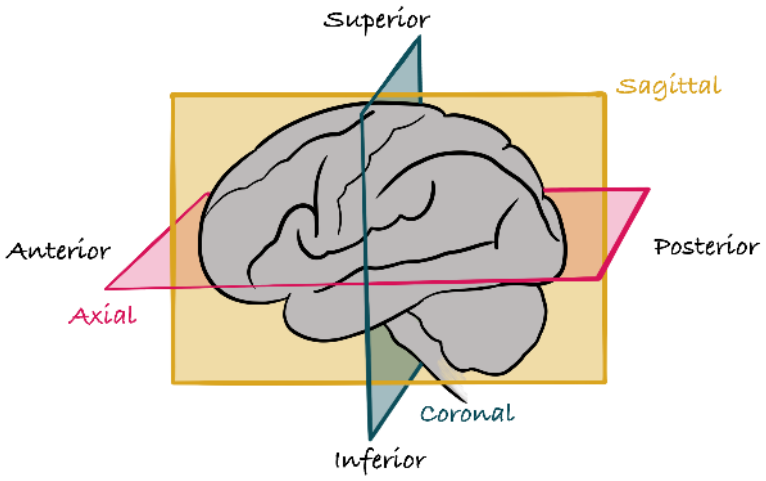

Figure 2.3: **Planes of view.**

axes: axial (horizontal), sagittal (vertical), and coronal (frontal), providing different perspectives of the brain's anatomy and functional regions (Fig 2.3).

In this context, specific directional terms are used to describe the brain's spatial orientation. The anterior direction refers to the front of the brain, while the posterior direction corresponds to the back. The superior direction signifies upward, and the inferior direction indicates downward.

2.4 Neuroimaging techniques

2.4.1 Structural MRI

Structural MRI images are used to provide detailed anatomical information about the brain. These images are available in various contrasts, depending on the required tissue-related information (Figure 2.4). They can also be used to compute a number of features, such as the volume of various anatomical parts, cortical thickness, and others. The T1 contrast is the most commonly used, albeit T2-weighted images are preferable to study brain lesions and abnormalities.

T1-weighted image

T1 refers to the longitudinal relaxation time, which represents the time it takes for hydrogen protons to return to their original state of alignment with B_0 field after being excited by the RF field. The T1 time changes based on the tissue the hydrogen atoms are a part of. For example, the T1 time of CSF is ~4,000 ms, while it is only ~900 ms for grey matter. Thus, when setting short TE and TR, the protons of CSF do not have enough time to

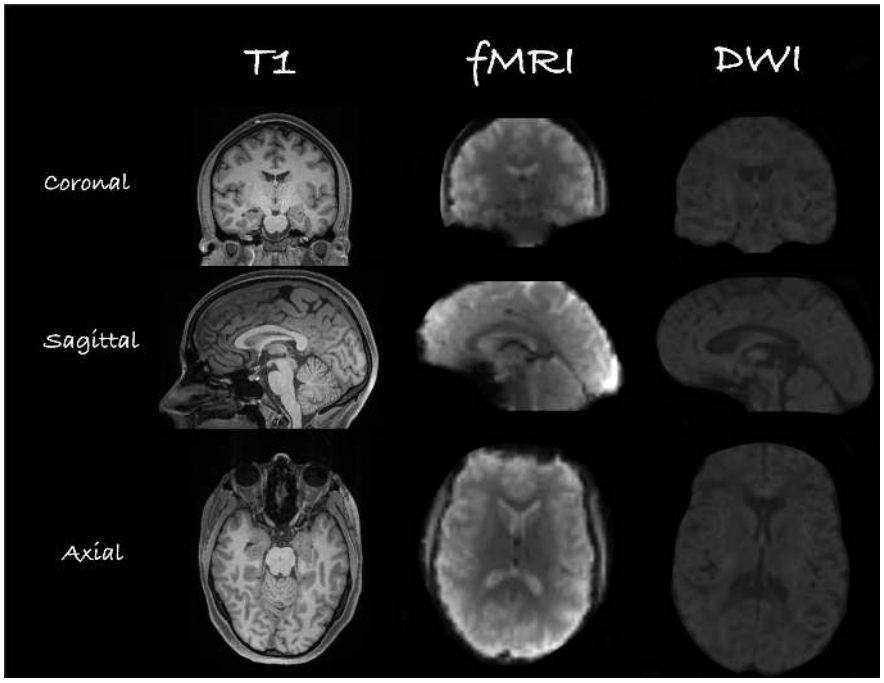

Figure 2.4: **Comparison of structural, functional and diffusion images.**

recover, making CSF appear almost black on the images, while grey matter, the protons of which recover faster, is a darker grey. This allows for a clear distinction between grey matter, white matter, and CSF.

T2-weighted image

T2 refers to the transverse relaxation time. It describes the gradual loose of phase coherence between the hydrogen protons.[1] It uses a long TR and a long TE, allowing the protons in tissues to have enough time to relax and emit their own signals before the next RF pulse is applied. During this time, the protons' magnetic moments precess at different rates, resulting in a decay of the transverse magnetisation. This sequence is mostly used to visualise fluid-filled structures such as the brain's ventricles, but also abnormalities like white matter lesions in multiple sclerosis.

T2*-weighted image

The principle of T2* is analogous with the T2 discussed above; however, on top of the decay of transverse magnetisation caused by the spins of hydrogen protons falling out of phase, the T2* also takes into account magnetic field inhomogeneities due to tissue structure. This sequence is particularly useful for fMRI and DWI.

[1]Note that before the RF pulse, all protons are aligned with the B_0 magnetic field and are in phase. After the RF pulse tips protons off the B_0 plane, the phase between the protons gradually loses sync.

2.4.2 Diffusion-weighted imaging

DWI enables the visualisation of water molecules' movement within a tissue. The T2* signal is attenuated based on the water molecules' diffusion ability, with higher diffusion resulting in a lower signal (Figure 2.4). For example, white matter in the brain mainly comprises densely packed axons, significantly restricting the directions in which water molecules can move, effectively capturing the structure of white matter.

To create a DWI image, a single T2* volume is initially acquired without diffusion attenuation. Following this, a set of images is taken from different angles, each measuring the amount of diffusion from a specific direction. A strong gradient is symmetrically applied on either side of the RF pulse, and a set of parameters must be tuned, mainly related to the amplitude and duration of the gradient and the intervals between them. These parameters are combined to produce a single **b** value. A b value of zero results in a T2* image, and as the b value increases, the attenuation becomes more profound.

Unlike structural images, DWI imaging does not produce a single 3D volume but rather a set of 3D volumes, each representing diffusion in a specific direction. This imaging technique is particularly useful in detecting acute strokes and tumours by visualising water movement in tissues.

2.4.3 Functional magnetic resonance imaging

fMRI was developed to investigate brain activity rather than its structural characteristics. Essentially, it does not directly measure neural activity; instead, it relies on the unique magnetic properties of haemoglobin. Haemoglobin, a protein responsible for transporting oxygen to cells through erythrocytes, exists in two forms: oxyhaemoglobin, when bound with oxygen and deoxyhaemoglobin when oxygen is released. The magnetic properties of haemoglobin differ depending on its oxygenation state. Oxyhaemoglobin is diamagnetic, while deoxyhaemoglobin exhibits paramagnetic behaviour relative to the rest of the body. This paramagnetic property creates microscopic field distortions around red blood cells and vessels containing deoxyhaemoglobin, leading to a detectable decrease in T2 and T2* signals during fMRI scans.

When neural activity increases in a specific brain area, the neurons require more oxygen, which is delivered to the cells through blood, specifically oxygenated haemoglobin. Blood vessels anticipate the increased supply of blood and temporary increase in volume. Oxygen is then delivered to the area in greater supply than needed, resulting in an increase in capillary and venous blood oxygenation. The excess of oxyhaemoglobin relative to the local amount of deoxyhaemoglobin causes an increase in T2 and T2* signals, which fMRI can detect. Because the acquisition relies on the ratio of oxy and deoxyhaemoglobin

in active regions of the brain, the technique is referred to as Blood Oxygenation Level Dependent (BOLD).

Several factors affect the interpretation and acquisition of fMRI data. First and foremost, it is important to reiterate that fMRI does not directly measure neural activity but the response of blood flow to that activity. While the BOLD process is reliable, it is also spatially inconsistent across the brain.

Another crucial factor to consider is the frequency of the acquired time series. While the amount of time it takes to acquire individual slices of T1 images is not particularly relevant, when studying brain function, the whole brain must be scanned quickly and ideally simultaneously to capture comparable information (Figure 2.4). Structural sequences take minutes to acquire a whole brain image, which is too slow for capturing a single sample of brain activity. To address this issue, the Echo-planar Imaging (EPI) pulse sequence was developed by Peter Mansfield [26]. It enables the acquisition of a single brain slice in 50-100 milliseconds, allowing for a 3D image of the brain to be acquired in under two seconds [26], which should be just enough to capture the BOLD change.

2.5 Electroencephalography

The EEG is an older method than MRI, and although not a neuroimaging method *per se*, it has unique properties which make it indispensable for clinicians and scientific research.

In the first chapter, we mentioned the mechanism of communication between neurons, which could be either through action potential or neurotransmitters. The EEG measures signal that is the result of the second mechanism.

The outcome of *excitatory postsynaptic potential* (depolarisation of neuronal membrane) is a current sink (extracellular negativity) and current source (intracellular positivity), creating a current dipole. Although the current dipole of a single neuron is not strong enough to be registered by an electrode on the scalp, when multiple neurons are excited in this manner, their combined current is conducted through the brain, veins, and other tissues to the scalp. This summed current is isotropic, meaning it flows in all directions equally and can be detected as a scalp potential by electrodes.

EEG signals are obtained as time series data, with one time series per electrode. The frequency of the signal is much higher than that of fMRI, generally ranging from 250 to 1000 Hz. Despite convenient temporal properties, EEG has several limitations, particularly in terms of spatial resolution, due to the distance of the sources from the scalp, limited spatial sampling, and number of electrodes. Other factors, such as muscle activity and eye movements, can also introduce artifacts and impact the quality of the signal.

Chapter 3

Labyrinth of processing

In this chapter, we introduce the techniques used in processing neuroimaging data. Expanding on the previous chapter, we explore specific workflows tailored to each modality and situations where multiple modalities are employed. This chapter serves as a foundation for generating features essential for the classification and prediction algorithms discussed in the subsequent chapter.

Each modality offers unique benefits and generates specific features. However, with these benefits come challenges, including artifacts and distortions that must be addressed before the analysis. In this section, we revisit each modality to discuss: the common preprocessing steps and artifact correction and the types of features that can be derived from them.

It is important to note that preprocessing in neuroimaging is a vast subject that could be the focus of an entire thesis. Furthermore, acquisition artifacts often interact with one another, requiring complex algorithms for appropriate correction. For the purpose of this work, we will treat individual preprocessing steps as fully effective without discussing their specific limitations. While we acknowledge the significance of visually assessing the quality of preprocessing, we do not aim to optimise or refine them. Consequently, we will provide only a brief overview of these stages.

In the course of this chapter, we cite the influential works that impacted the preprocessing of neuroimaging data. However, for a comprehensive overview, we recommend [27], [28] for preprocessing fMRI data specifically, and [29] for processing and analysis of EEG data.

3.1 Structural imaging

3.1.1 Preprocessing

Structural imaging is a fundamental component of any neuroimaging study as it provides crucial information about the position and quality of different brain tissues. It enables the measurement of a wide range of high to low-level features. The type of features we aim to generate determines to a degree, the specific preprocessing steps applied. Two common approaches are volumetric and surface analyses. Both types are not dependent on a specific contrast and can use T1, T2, or other contrasts, although T1 is typically the most commonly used.

In the volumetric approach, structural features are primarily associated with the volume of brain regions or abnormalities. They are particularly useful for investigating volumetric changes in the brain.

The surface approach involves reconstructing voxel-based brain data as a surface composed of small triangles or vertices. As the name suggests, it is primarily used to generate features specific to the surface of the brain, such as grey matter thickness or surface area.

Skull stripping

The process of brain extraction involves removing the skull and non-brain tissues. It is not considered a mandatory step in neuroimaging data analysis, but it offers significant advantages. Firstly, it eliminates unnecessary tissues that are not the focus of the analysis, thereby avoiding potential complications in subsequent preprocessing steps. Additionally, skull stripping ensures the anonymisation of the data, enabling researchers to share information without compromising the identity of the individuals involved [30]. However, it is important to note that the process is not foolproof, and the results should be visually inspected.

Bias field correction

Bias correction is a prevalent artifact encountered in all neuroimaging modalities, resulting from imperfections in RF coils and image acquisition. It manifests as a reduction in high-frequency components of the image, leading to blurred boundaries between different types of brain tissues. Furthermore, it introduces intensity variations, causing the same tissue to exhibit different intensities. Modern techniques for bias field correction have been developed, making it relatively straightforward to identify and mitigate this bias, thereby improving the accuracy of subsequent analyses [31].

Tissue segmentation

Tissue segmentation aims to create distinct spatial masks for the three fundamental brain "tissues": grey matter, white matter, and CSF. In certain cases, such as when studying brain damage caused by conditions like glioma, stroke, or multiple sclerosis lesions, there may be a need for an additional tissue mask representing the affected area.

Registration

Registration is a crucial step of neuroimaging data analysis, which is again not limited to structural imaging data. We discriminate between inter-subject and intra-subject registration.

- **Intra-subject registration**: involves aligning different imaging modalities within a *single subject*. Each modality may have a varying resolution, distinct artifacts, and potential shifts caused by movement, so they must be aligned before analysis. This alignment is particularly critical in analysing DWI, fMRI, and any multimodal data in general. Additionally, various brain atlases can also be registered to an individual subject's image to produce higher-level features (more on this in the following section on feature generation).

- **Inter-subject registration**: is a widely used tool for making group-level inferences. In this approach, all subjects are registered to a predefined template, allowing individual voxels to correspond with one another across samples. Several templates are available, with the Talairach template being one of the earliest [32]. The most commonly used templates today are those released by the Montreal Neurological Institute (MNI). The Colin 27 template is created by averaging 27 scans from the same individual [33]. The MNI 305 template aimed for increased generalizability and later became the foundation for what is now referred to as the MNI space [34]. Additionally, the ICBM 152 linear template represents the average brains of 152 young adults registered to the MNI 305 space [35].

3.1.2 Feature generation

Once the preprocessing of neuroimaging data is completed, it is possible to generate the structure-specific features. This typically involves either voxel-level analysis or extracting volumes of chosen brain areas.

Voxel-based morphometry

Voxel-Based Morphometry (VBM) reference a family of methods (which, however, may have a lot of different implementations), the purpose of which is to identify differences in the local amount of (mostly) the grey matter either on a level of the group differences or model its relationship with a certain covariate.

The pipeline consists of a few well-defined steps [36], [37], which are usually run on a group level, as they include registration of all subjects into a common space.

1. **Brain extraction** (see Preprocessing)

2. **Tissue segmentation** (see Preprocessing)

3. **Registration** – Is a crucial step in the VBM pipeline. All subjects are affine-registered into a standard space, and a study-specific template is often computed. The registration may also be performed only on a level of grey matter to eliminate artifacts. The registration process is crucial not only for its result but also for the transformation function itself. When registering a brain to a specific template, voxels must be expanded or contracted to fit the target volume. This transformation carries information about the proportion of grey matter volume change, which is relevant for the analysis.

4. **Modulation** – In this step, the information from transformation is combined with the result of registration. If the registration to the template was perfect, all the registered images would look the same, and the transformation would capture all of the information. However, registration is never perfect, which is reflected in the intensities of registered voxels. Thus, the registered images are multiplied with the transformation matrix in a modulation step.

5. **Smoothing** – Finally, the resulting modulated images are smoothened to decrease noise and increase alignment and sensitivity.

Volumetric features

The assessment of the volume of brain areas is usually a hierarchical process, which begins with the estimation of intracranial volume and the three main tissue types: grey matter, white matter, and CSF, which are subsequently split into smaller areas. These brain areas are defined and delineated using what is known as “brain atlases”.

Brain atlases serve as reference images of the “standard” brain, often referred to as the brain in the standard space, which we described in the context of inter-individual registration. In a brain atlas, each voxel is assigned a unique integer corresponding to

a specific part or region of the brain. These atlases can be constructed based on various criteria, such as anatomical similarity of brain structures [38], functional similarity (e.g., brain areas that exhibit synchronous or asynchronous activity) [39], or even genetic similarity [40]. The surface analysis allows for extracting specific features, including grey matter thickness, surface area, curvature, and normals (surface orientation). While there are atlases designed explicitly for surface analysis [41], [42], it is worth noting that most volumetric atlases can be projected onto the surface with some limitations and caveats.

3.2 Diffusion-weighted imaging

3.2.1 Preprocessing

Unlike structural images, the preprocessing of DWI is much more challenging and time-consuming. This modality is specifically prone to geometric distortions, and it is necessary to inspect the results of individual processing steps to ensure the quality of the final features.

Susceptibility induced distortions

In the DWI context, the acquisition speed plays a crucial role; therefore, an EPI or similar protocol is typically employed. However, the scanning sequence has to balance a trade-off between resolution and noise. Another challenge arises from the sensitivity of the EPI sequence to magnetic field inhomogeneities, which are most prominent near the sinuses. These inhomogeneities can cause significant distortion in the lower temporal areas of the image and are referred to as susceptibility-induced distortion. Fortunately, the nature of this artifact is well-known and understood. Specifically, the distortions occur along the phase-encoding direction, which remains constant throughout the acquisition. An additional image is acquired in the opposite direction to mitigate this artifact, effectively reversing the distortion. The pair of images can then be used to correct the distortion in all diffusion images.

Motion correction

Motion correction has a crucial impact on the acquisition of various neuroimaging features. However, it affects DWI and fMRI data slightly more than structural images due to the nature of their acquisition. DWI and fMRI generate a series of 4D images over time, whereas structural images typically consist of a single 3D image. In our discussion of fMRI data analysis, we will delve deeper into the process of motion correction, as in DWI,

motion correction is often combined with the correction of Eddy currents distortions (see below).

Eddy currents distortions

Eddy currents are probably the most notorious artifacts related to DWI. They result from the fast switching of the diffusion gradients and demonstrate as a shift between individual images [43]. The correction algorithms use various approaches rooted in field maps, model-driven approaches or image registration, the latter being the most popular and implemented in the FSL software tool [43].

Gibbs-ringing artifacts

Gibbs ringing artifacts, occur due to errors in the reconstruction process when converting the analogue signal to a digital image [44]. Specifically, they affect high-frequency components of the image, particularly at tissue boundaries or regions with sharp transitions. As a result, spurious oscillations appear around the boundaries, giving rise to the characteristic ringing appearance in the image.

3.2.2 Feature generation

The diffusion imaging data analysis was significantly transformed by introducing the diffusion tensor model [45], [46]. In essence, the model assumes one dominant direction of water molecules diffusion within each voxel, represented as a tensor. This tensor is effectively a symmetric, positive-definite 3x3 matrix. Its eigenvalues represent the magnitude or strength of diffusion, while the corresponding eigenvectors indicate the directionality.

Diffusivity measures

For practical purposes, it is often more convenient to describe the diffusion characteristics of a voxel using a single value rather than an entire matrix, which is why the diffusion metrics were developed. The simplest measure is Mean Diffusivity (MD) which represents an average magnitude of diffusion. It is computed as an average of all three eigenvalues (3.1) and, because higher values indicate less constrained diffusion, can indicate structural damage.

On the other hand, Fractional Anisotropy (FA) defines the degree of directionality (anisotropy) of diffusion (3.2) [47]. The higher the FA, the more restricted the diffusion in one direction. Unlike MD, the range of FA is restricted from zero to one, zero indicating isotropy[1].

[1] Isotropic movement of water is typical for CSF.

Additional measures as Axial Diffusivity (AD) and Radial Diffusivity (RD), can further complement the interpretation of MD and FA by providing specific directional information about diffusion along and perpendicular to the fibre tracts, respectively.

$$\mathrm{MD}(\hat{\lambda}) = \frac{\lambda_1 + \lambda_2 + \lambda_3}{3} \tag{3.1}$$

$$\mathrm{FA} = \sqrt{\frac{3}{2}} \frac{\sqrt{\sum_i^3 (\lambda_i - \hat{\lambda})^2}}{\sqrt{\sum_i^3 \lambda_i^2}} \tag{3.2}$$

Tract-based spatial statistics

The voxelwise fit of diffusion measures across white matter is a cornerstone of the most notorious method for analysing diffusion data: Tract-based Spatial Statistics (TBSS) [48]. The TBSS routine consists of three steps, the purpose of which is to register *all subjects in the analysis* into a common space and create *a skeleton* across which are the diffusivity measures compared.

1. **Registration** – FA (or another diffusivity measure) images are non-linearly registered to the chosen template, usually FMRIB58_FA standard-space image.

2. **Skeletonisation** – The nonlinear registration is applied, each subject is registered by affine registration to MNI152 space, and the mean skeletonised image is computed.

3. **Projection** – The mean FA skeleton is thresholded, and FA maps of individual subjects are projected onto the mean skeleton. The skeleton is then used for statistical evaluation.

Structural connectivity

The last method for generating diffusion features we will discuss is Structural Connectivity (SC). In this method, the structural connectivity of an individual's brain is represented as a symmetrical square matrix, where each row and column corresponds to a specific Region of Interest (ROI). The matrix elements provide information about the connectivity between different brain regions. To understand this method better, we need to explain the technique called tractography.

Tractography is a method that traces the orientation of diffusion in neighbouring voxels and connects them into tracts based on predefined criteria. These tracts can be seen as a model representing the white matter pathways in the brain, and their accuracy and complexity depend on various factors.

There are two main approaches to conducting tractography: deterministic and probabilistic [49]. In deterministic tractography, each voxel is assumed to have a single major direction of white matter tracts. However, this assumption may oversimplify the problem of fibre crossings, where multiple axon bundles pass through a voxel in different directions. Deterministic tractography selects a dominant direction for each voxel and produces the same result with each run.

On the other hand, probabilistic tractography assigns a probability to each direction of white matter within a voxel. The algorithm traces connections multiple times, sampling the probability at each voxel. This probabilistic approach is more computationally demanding and may yield slightly different results for the same subject each time it runs. Additionally, probabilistic tractography allows for more complex representations of diffusion directions, such as the use of fibre orientation distribution functions estimated using constrained spherical deconvolution [50], [51]. This technique involves deconvolving the measured signal with a response function. The response function is estimated using spherical basis functions, such as spherical harmonics, and it characterises the diffusion properties of a single fibre. For this purpose, voxels that are assumed to only contain axons aligned in a signal direction (usually, voxels in the corpus callosum or with high FA values) are used. Subsequently, the diffusion data are deconvolved with the response function, and the fibre orientation distribution function is obtained and refined using regularisation techniques to enhance estimation accuracy.

The choice between deterministic and probabilistic tractography has been a subject of debate, with arguments presented for each approach based on realism and suitability [49], [52], [53]. Deterministic tractography is faster and can be used with lower-quality data, but it simplifies the problem of crossing fibres. Despite being more computationally intensive, probabilistic tractography considers the uncertainty of fibre orientations and provides a more nuanced representation.

Regardless of the tractography approach, the process typically involves several steps:

1. **Preprocessing** – Denoising, unringing, motion and distortion correction, bias field correction (for more detail, see previous section on preprocessing)

2. **Brain-mask estimation**

3. **Estimation of diffusion orientation** – In the simplest case, it can be fitting a tensor to each voxel, or a) Response function estimation, b) Fibre orientation distribution estimation.

4. **Seeding and termination mask design** – Ensures anatomical validity of the reconstructed tracts. The tracts should start and end at a boundary between grey and white matter or, in a small number of cases, between grey matter and CSF.

5. **Creating streamlines** – Streamlines are created by connecting fibres in the neighbouring voxels based on specific criteria. There are several tractography algorithms available for each deterministic and probabilistic tractography. In case of a deeper interest in these techniques, we refer the reader to the documentation of the MRtrix3 software [54].

6. **Filtering streamlines (probabilistic tractography)** – The tractogram is recursively inspected, and possible faulty streamlines are discarded. Moreover, this step ensures no disproportion in the thickness of long and short tracts or in regions of crossing fibres [55].

7. **Computing the SC martix** – The structural connectivity can be defined in multiple ways depending on the preferences and research question. In the simplest case, it could be the number of tracts between two regions. However, this measure can sometimes be misleading and is often normalised by the size of the target regions, length, or inverse length of the streamlines. Another possible characteristic is the average of chosen diffusion metric along the fibres between the regions.

The elements of the structural connectivity matrix serve as features for subsequent analysis. It's important to note that, unlike TBSS, tractography takes place in the individual space, with only the registration of structural images and atlases to diffusion data for identifying individual grey-matter regions.

3.3 Functional magnetic resonance imaging

3.3.1 Preprocessing

The temporal nature of fMRI makes it an indispensable technique; however, it also introduces movement artifacts, which can significantly affect the analysis. The preprocessing depends partly on the experiment design, which generally falls into two categories: task-based fMRI and resting-state fMRI (rs-fMRI). Task-based fMRI investigates the nature and localisation of functional processes by having subjects perform specific tasks in the scanner. At the same time, rs-fMRI measures brain activity when subjects are "resting" in the scanner. In this thesis, we focus on resting-state fMRI and describe the processing and feature generation process for this design (although the preprocessing steps overlap between the two designs).

The decision to prioritise a resting-state design, both in this chapter and in the EEG section, is rooted in the fundamental research question addressed in this thesis: Can we relate cognitive or other clinical markers using neuroimaging data, even with limited or no

prior knowledge of the disease or its effects? This question is deliberately broad because there is currently a lack of well-established neuroimaging markers for most diseases and a lack of markers that effectively connect the measured brain activity with specific clinical markers of interest.

In such a context, relying solely on task-based designs would be constraining, as they require prior knowledge of the disease and present challenges in achieving consistent measurements across different scanning sites. Therefore, in this thesis, we opt for resting-state studies, which offer greater flexibility in the range of questions we can explore, the features we can generate, and the feasibility of conducting measurements across multiple institutions.

Susceptibility distortion correction

Susceptibility distortion correction in fMRI is analogous to the process described in the diffusion pipeline. However, since fMRI uses a different EPI sequence (gradient-echo) compared to diffusion-weighted imaging (spin-echo), there are not only distortions but also signal loss in specific regions near the sinuses. Similar to the diffusion step, it is desirable to acquire gradient-echo field map images to compensate for these distortions.

Motion correction

Motion artifacts are the most challenging issue in fMRI data analysis. While changes in BOLD signals are subtle, motion-induced changes are much more noticeable and can lead to false positive results. The most common approach to address motion in fMRI is to perform a series of registrations. Since the head only changes its position and not its shape, a rigid body transformation is sufficient. All images are registered to a specific reference image (e.g., the first, middle, or average image), generating a series of displacement parameters for transformations. These parameters can be utilised to regress out noise and identify outliers.

Slice timing correction

Neuroimaging data acquisition is conducted in the sequence of slices, capturing each slice at slightly different time points. While this temporal variation may not be significant for structural or diffusion data, it becomes critical for fMRI, which examines brain activity and its interregional correlations. Correcting for the minor time delays between slice acquisitions is essential to mitigate any potential biases arising during analysis.

Smoothing

Spatial smoothing is a common step in most pipelines, aimed at increasing the signal-to-noise ratio by averaging out noise. The size of the smoothing kernel is an arbitrary parameter that needs to be chosen as a trade-off between noise reduction and information loss. However, with advancements in data acquisition, this step is sometimes omitted.

Temporal filtering

Temporal filtering is a crucial step in both fMRI and EEG data analysis. Depending on the frequency nature of the signal of interest, it is possible to filter out signals that are not relevant, such as heartbeat or respiration. However, due to the low sampling frequency of the fMRI time series (2 seconds), the removal of these artifacts is more challenging due to aliasing. Nonetheless, methods that model physiological noise, mainly using general linear models, exist and are often used for this purpose, sometimes in combination with Independent Component Analysis (ICA) (see feature generation).

Intensity normalization

Intensity normalisation is a common step performed at the dataset level. It involves multiplying each subject's data by a scalar value to account for any between-subject differences that could potentially decrease the statistical power of the analysis. Normalising the intensities allows the data from different subjects to be more effectively compared and analysed.

Registration

Registration involves both within-subject and between-subject stages. Within-subject registration is particularly important as it aligns the structural and functional data, allowing for the use of the high spatial precision of structural imaging when performing tissue segmentation of fMRI. Since fMRI often focuses on studying the activity of grey matter, accurate tissue segmentation is crucial for precise analysis.

Depending on the specific processing toolbox, registration between subjects may be performed at the beginning or the end of the data processing pipeline. Comparably to VBM or TBSS, inter-subject registration is used to ensure voxel-level comparability across subjects for group-level testing.

3.3.2 Feature generation

Features derived from rs-fMRI usually quantify the activity of specific brain regions or the relationships between different regions. The metric used can vary, although research

shows that linear correlation is usually sufficient [56].

Amplitude of low-frequency fluctuations

The Amplitude of Low Frequency Fluctuations (ALFF) is a straightforward measure that can be computed at various levels, ranging from individual voxels to ROIs [57]. The time series of a voxel (or the average time series of an ROI) is first subjected to a band-pass filter to isolate frequencies between 0.01 Hz and 0.1 Hz. The power of the frequency spectrum within this range is then calculated and summed to obtain the ALFF value. To account for inter-subject differences, the global ALFF mean often normalises the values. ALFF changes have been observed in individuals with conditions such as schizophrenia [58], mild cognitive impairment [59], and have even been indicative of treatment response in depression [60].

Functional connectivity

Like structural connectivity, Functional Connectivity (FC) is represented by a symmetrical square matrix, where each row and column correspond to a brain region, and the elements represent the similarity measure of activity between regions.

The first step in the FC analysis involves selecting an atlas to parcellate the brain and define regions. A representative time series for each region is computed by averaging the time series of voxels within that region. These time series are preprocessed by detrending and band-pass filtering, typically within the 0.009-0.08 Hz range. Finally, the preprocessed time series are correlated using linear correlation, resulting in an FC matrix for each individual.

A related concept is *dynamical functional connectivity* [61], where instead of computing the FC matrix using the correlation of the entire acquired time series, the time series are split into shorter segments. For each segment, a separate FC matrix is computed, yielding a 4D dynamical connectivity matrix that enters the analysis.

Independent component analysis

ICA is a conceptual analogue of the Principal Component Analysis (PCA), but it is based on different assumptions and mathematical principles [62], [63]. Although not extensively used in this thesis, ICA is an important method in fMRI data analysis and deserves a brief mention. Unlike PCA, where the independent components are designed to be orthogonal to each other, the sources obtained through ICA are intended to be statistically independent. Additionally, ICA does not aim to maximise the variance in the first few sources but instead emphasises their separability. The result of ICA is

a set of *functional networks*, which are spatial maps (equal to the number of sources) representing regions whose activities are correlated over time. Currently, there are seven well-established cortical networks: the default mode network, the salience network, the limbic system, the dorsal attention network, the central executive network, the visual system, and the sensorimotor network [64]. Depending on the desired number of sources, these networks can be further divided into smaller networks. ICA is not only useful for analysing brain activity but also for data preprocessing. By including more sources, some can effectively model noise, which can then be regressed out of the data.

3.4 Electroencephalography

3.4.1 Preprocessing

EEG was developed as a non-invasive method for measuring electrical brain activity, offering an exceptional temporal resolution that facilitates the removal of known sources of bias (in contrast to fMRI), such as respiratory or cardiac interference. However, since the signal is measured only from the scalp, EEG is prone to various artifacts related to muscle movement and eye blinking that need to be corrected.

Similarly to fMRI, different study designs can be used in EEG experiments: task-based designs focus on time-locked EEG activity, known as event-related potentials, and resting designs. In this chapter, we will specifically discuss the preprocessing and analysis of the resting design.

Downsampling

The sampling frequency of the EEG device can range from 250 to 1000 Hz. However, for most research questions (unless studying high gamma activity), downsampling to 250 Hz is usually sufficient but unnecessary.

Filtering

To remove artifacts, a combination of filters is typically applied depending on the analysis. A high-pass filter around 0.5 or 0.1 Hz is commonly used to minimise slow drifts. In comparison, a low-pass filter is applied to remove frequencies above 40 Hz (unless the experiment focuses on a specific frequency band). Additionally, power line interference caused by the alternating current is mitigated by applying a notch filter at 50 or 60 Hz.

Channel inspection and interpolation

It is not uncommon for some electrodes to malfunction during EEG acquisition. This can be due to poor initial electrode-skin contact, gradual detachment from the scalp, or equipment malfunction. In such cases, the affected channel is excluded from the analysis. If desired, its activity can be estimated by interpolating the time-series data from neighbouring channels. When a significant number of channels are malfunctioning within a time window, that part of the recording may need to be discarded.

Artifacts identification and removal

Although we introduced the ICA as a method for identifying brain networks, it can also be effectively used for noise removal. By visually examining ICA components, channels exhibiting noise characteristics can be identified and regressed out. A rule of thumb is that components with time series spectra resembling a $1/f$ function are more likely to represent relevant signals, while muscle artifacts typically exhibit increased power in higher frequencies.

Rereferencing

The voltage measured in each EEG electrode is relative to a "reference electrode" chosen beforehand. Common reference locations include earlobes or mastoids (a bone behind the ear) due to their proximity to other electrodes and lower neural signal contamination. During analysis, it is common to re-reference the signals to other electrodes or to the average reference of all electrodes. The choice of the re-referencing method is extensively discussed in the EEG community and can influence the results and their interpretation.

3.4.2 Feature generation

Similar to other modalities, EEG allows us to generate a wide range of features, which are typically computed within specific frequency bands. Individual frequency bands play a distinct role in brain functioning, and the significance of each band concerning cognitive, motor, and resting processes is still an active area of research.

The main frequency bands and their associated characteristics are as follows:

- **δ band (2-4 Hz)**: First reported by Walter in 1936, δ band is linked to deep sleep and is prominent in infants up to one year of age [65].

- **ϑ band (4-8 Hz)**: Is observed during deep sleep and is also associated with memory processes [66].

- **α band (8-12 Hz)**: Was one of the first rhythms discovered by Berger when he introduced EEG [67]. It has been linked to attention processes and sensory stimulation.
- **β activity (15-30 Hz)**: Is more prominent during sensorimotor processes than during relaxed states [68]. It is high during postural maintenance and decreases during movement periods [69].
- **γ band (30-150 Hz)**: Requires high sampling frequencies for analysis. It involves perception, attention, memory, and motor control [70].

Spectral power

The most straightforward feature to extract from EEG is the spectral power within each frequency band. However, comparing power values between subjects can be challenging, so it is common to use relative band power instead.

Complex EEG features

In this work, we will not focus on more complex EEG features, but it's worth mentioning them briefly. Similarly to fMRI, EEG time series allow for functional connectivity estimation.

However, the limitations in spatial precision lead to the problem referred to as the *volume conduction problem*. The electrical signals from the brain travel through conductive tissues to reach the scalp, resulting in the same signal being detected by multiple electrodes and causing a high correlation between neighbouring electrodes [71]. Various approaches can be employed to address this issue, such as orthogonalising time series, ICA, applying spatial filters, or applying source-separation techniques like Low-Resolution Electromagnetic Tomography (LORETA), which generates a three-dimensional map representing estimated neural activity across different brain regions [72].

Once the challenge of volume conduction is mitigated, measures quantifying the degree of synchronisation between different brain areas can be computed, such as phase-locking value, phase difference, or coherence [73].

3.5 Multimodality in neuroimaging analysis

Throughout this chapter, we have already touched upon the concept of multimodality several times, but we haven't discussed it in detail. We saw that in the case of both DWI and fMRI, structural images play a role in the feature generation process. The high spatial precision of structural images aids in subject registration and tissue segmentation

of other lower-resolution modalities. This is already considered a form of modality fusion, specifically an asymmetric fusion, where one modality is used to enhance or refine features of another modality [74].

The advantages of multimodal analysis are increasingly apparent, although such studies remain less common compared to traditional unimodal analyses [75], [76]. Different neuroimaging modalities capture distinct information, such as white matter integrity in DWI and grey matter activity in fMRI. Combining these modalities gives us a more comprehensive understanding of neurobiological processes and their interactions. Moreover, the higher temporal and spatial resolutions of specific modalities can enhance the analysis of less precise modalities, improving localisation accuracy. Combining multiple modalities also increases the signal-to-noise ratio, thereby enhancing sensitivity and specificity and reducing noise.

However, there are significant challenges and reasons why multimodal analysis has not yet gained widespread popularity, extending beyond the complexities of data acquisition. In fact, coordinating data acquisition across different modalities can be time-consuming, costly, and logistically challenging [77]. Variability in acquisition protocols, scanner differences, and technical issues may introduce additional complexity and heterogeneity in the data, making integration and analysis more challenging. Furthermore, integrating data from multiple modalities is a complex task. As should be apparent from this and the previous chapter, each modality has its specific preprocessing steps, artifacts, and sources of noise. Aligning and combining data across modalities while maintaining the integrity of each modality is technically demanding, requiring specialised skills and imposing a substantial computational burden.

Additionally, a fundamental question arises regarding feature generation. Should features be designed by combining modalities from the outset, or should they be designed individually for each modality and combined in the analysis? These approaches are commonly referred to as data fusion and data integration, respectively.

Data integration

Data integration refers to the process of designing features individually for each modality and merging them at the level of analysis, which is more convenient. The differences in resolution, alignment, and information type among modalities can be quite significant, and attempting voxel-level fusion can be challenging and impractical while designing features separately offers flexibility. Additionally, the features are often designed at the ROI level, which reduces the dimensionality of the problem. For example, the widely used JHU atlas for white matter encompasses 48 regions, compared to the thousands of white matter voxels. However, this convenience can also be a double-edged sword. Depending on the

type of features, there may be an imbalance in the number of features per modality. For instance, the dimensionality of region-wise FC features is formed by the upper triangle of the FC matrix (in the case of standard Automated Anatomical Labelling (AAL) atlas with 90 regions, this results in 4,005 features), which has a significantly higher dimensionality than region-wise white matter features. This inequality can pose problems that need to be addressed during the analysis. Furthermore, when different brain atlases are used for different modalities (which is standard practice, with separate atlases for the grey and white matter), exploiting the underlying correlation structure of the data becomes more challenging.

Data fusion

There are two main branches of data fusion approaches: symmetric and asymmetric. In asymmetric fusion, one modality is used to enhance or refine features from another modality, such as using structural connectivity to constrain functional connectivity. This class also entails the most established multimodal pair of EEG and fMRI, where EEG is commonly used to model the neurophysiological correlates of fMRI [78]. The family of symmetric fusion approaches is larger and can be further divided into model-driven and data-driven methods.

Model-driven approaches rely on prior knowledge or hypotheses about the relationships between different brain regions. Examples of model-driven approaches include general linear models, dynamical causal models [79], or structural equation models [80]. These approaches reduce the dimensionality of the problem based on known relationships but will miss discovering new connections or dependencies.

Data-driven approaches, on the other hand, are hypothesis-free and aim to discover new relationships and dependencies between modalities. These approaches often rely on techniques such as ICA, canonical correlation analysis, or partial least squares. They mainly operate at the voxel level and seek to maximise the correlation between modalities, uncovering shared information and patterns across subjects. It is not our objective, nor is it necessary, to describe all the up-to-date algorithms designed to deal with the data-driven multimodal fusion, so here, we will only briefly describe one for the reader to have an idea about the general architecture of these methods.

Multimodal canonical correlation analysis involves first decomposing both modalities using methods like linear mixing models, PCA, or ICA [81]. The decomposition captures the main variance of each modality into a smaller number of features. Subsequently, the subject loadings are transformed using canonical correlation, which decomposes each matrix of loadings into components and mixing profiles (also referred to as canonical variates), describing the amount of each component in a subject. The correlation between

canonical variates identifies similar component profiles across subjects, creating a new set of features for analysis. Although this approach can enhance the signal-to-noise ratio, it may not fully utilise the multivariate data distribution. Handling more than two modalities and addressing mutual dependencies can be challenging.

Nevertheless, data fusion methods offer a valuable framework for integrating multimodal information and can be beneficial in multimodal neuroimaging analysis. It is necessary to mention that while this approach can effectively increase the signal-to-noise ratio, it may not fully exploit the multivariate distribution of the data. Additionally, extending these methods to more than two modalities and solving with mutual dependencies pose challenges. Nonetheless, data fusion methods provide a valuable framework for integrating multimodal information and can be useful in multimodal neuroimaging analysis.

Finally, a reader oriented in ML might be missing a significant branch of method in this chapter. Indeed, it is possible to attempt multimodal fusion also using deep learning methods [82]. However, because these approaches make it challenging and redundant to distinguish between the feature extraction stage and the classification or prediction stage, this topic will be addressed in the following chapter.

Chapter 4

Neuroimaging data analysis

In the preceding chapters, our focus has mostly been on elucidating the intricacies and techniques involved in acquiring and generating neuroimaging features. Here, we focus on analysing the data, providing an overview of the methods for assessing hypotheses and drawing meaningful conclusions. Initially, we will take a step back and explore the historical backdrop of neuroimaging data analysis, tracing its evolution over time, while in later sections, we will discuss the cutting-edge approaches used in contemporary neuroimaging research.

4.1 A brief history of the neuroimaging data analysis

As we find ourselves in the midst of the 21st century, scientific research has taken on a new dimension, particularly in the data-oriented branches. In these fields, the ability to master machine learning techniques has become almost synonymous with conducting scientific research. However, the allure of deep-learning methods has led us to sometimes overlook or disregard the (arguably) more rigorous statistical methods that precede these approaches. Therefore, it is important to review and reflect on the most pertinent ideas that have significantly impacted the field of neuroimaging data analysis.

In the early days of neuroscience, the predominant theory of functional processes in the brain was based on the concept of cerebral localisation, which had gained scientific support in the 19th century. This idea proposed that specific cognitive functions or behaviours were associated with particular regions of the brain. While ancient Greek philosophers first postulated that the brain was the seat of intelligence and sensation, it was not until the work of Paul Broca, a French physician, that the concept of cerebral localisation gained scientific grounding. Broca conducted post-mortem studies on the brains of patients with language impairments and identified an area in the left frontal lobe, now known as Broca's area, which, when damaged, caused speech difficulties [2].

The discovery was soon followed by the reports of Carl Wernicke, a German neurologist, who claimed to have identified an area of the brain responsible for understanding language – a posterior portion of the left temporal lobe [3]. These findings led to a surge of interest in the spatial localisation of brain functions during the 19th and 20th centuries. Many notable researchers made important contributions to this area, including Korbinian Brodmann, a German neurologist who believed that areas of unique cellular organisation would also be functionally specialised. As a result, he proposed a classification system for brain areas based on their unique cellular organization [83]. His maps, delineating 52 distinct brain regions, are still widely used today and considered the *lingua franca* for cortical localisation.

As non-invasive technology to study the brain became available, early computational methods focused on identifying anatomical areas involved in specific functions or undergoing particular changes. At the end of the 20th century, the natural choice of tools to analyse the data was rooted in classical statistics, which dominated the first half of the 20th century and became established in several fields, including psychology [84]–[86]. Probably the most notorious framework that leveraged classical statistics and is still used today was Statistical Parametric Mapping (SPM), introduced by Karl Friston in 1994 [87]. SPM implemented General Linear Models (GLM) for mass univariate testing, where statistical tests were conducted separately for each voxel in the brain to determine significant effects. This approach has multiple advantages as it is easy to understand and deploy, it is designed to localise an effect, and, compared to more sophisticated methods, it is interpretable.

Over time, the concept of functional localisation in the brain has faced growing scrutiny, challenging the notion that specific functions can be attributed to a single brain region [88]. What was once considered a fact started to appear more like wishful thinking among neuroscientists. Even Wernicke's groundbreaking report on speech comprehension has encountered significant challenges. It is now recognised that language comprehension involves a two-stage process [89], [90].

In the first stage, known as phoneme discrimination, we distinguish between the sounds of different words, such as "mad" and "sad". This initial stage intuitively involves the cortical auditory network, where the actual meaning of words or phonemes is irrelevant. The second stage, where meaning is assigned to words, is a separate process often called the "concept field". Therefore, it is overly simplistic to perceive speech recognition as a single process with distinct localization [88][1]. By the early 2000s, the neuroscience community recognised that "*Our brain is a network, and consists of spatially distributed, but functionally linked regions that continuously share information with each other.*" [91].

[1]Wernicke originally suggested this distinction; however, it was largely disregarded.

This perspective does not conflict with the theory of cerebral localisation, as highly specialised regions such as the visual or motor cortex exist. However, it expands upon this theory by acknowledging that these regions work collaboratively in a distributed manner to support various cognitive functions. The relationship between specialisation and distributed processing is not a simple dichotomy but a combination of both, exhibiting different degrees of specialisation and integration depending on the specific task.

Once acknowledging these new findings, the methodological limitations of GLM started to become apparent. It has been argued that GLM oversimplifies brain activity by testing each voxel separately, particularly in fMRI analysis [92]. Furthermore, the large number of tests inevitably leads to a high rate of false positive findings, although efforts such as the Threshold-free Cluster Enhancement (TFCE) correction have been developed to mitigate this issue [93].

Gradually, the field underwent a philosophical shift regarding the goals of neuroimaging data analysis. Instead of only assessing the existence of an effect, the focus shifted to predicting performance and generalising to out-of-sample data. Rather than conducting group-wise analyses, the applicability of models to single observations became desirable [94]. In other words, classical statistical tools were no longer sufficient, and statistical learning methods began to find their way into the field [95]. This marks the starting point for the subsequent sections of this chapter, where we will review influential approaches that have shaped the field since the 2000s.

4.2 Machine learning in neuroimaging – what to expect

This thesis aims to explore and develop analytical pipelines that can effectively classify and predict individual-level information from neuroimaging data, preferably integrating multiple modalities. In the existing literature, this task is often called the "*identification of neuroimaging biomarkers*" or "*neuroimaging fingerprinting*". This terminology correctly suggests that we aim to identify features derived from neuroimaging data that can provide insights into the clinical condition of the individuals being studied.

A biomarker is defined as a "*characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention*" [96]. Thus, the intention behind biomarkers is to obtain an unbiased and objective measure of the brain's condition. This is particularly relevant in psychiatric disorders, which are currently diagnosed only through subjective oral interviews that can be prone to biases. By using neuroimaging-based measures, we aim to complement or potentially improve the diagnostic process for these disorders.

However, biomarkers also hold value in neurological diseases, where having an objective assessment of the brain's state can inform treatment planning and provide insights into the expected longitudinal changes in patients.

Unfortunately, the conventional statistical approaches employed in biomarker design have certain limitations. Primarily focused on detecting significant differences between groups, classical statistics operates under the assumption that the group sample adequately represents the entire population. Consequently, the entire sample is often used to test hypotheses, assuming the results can be generalised. However, the absence of a well-defined hypothesis poses a challenge in biomarker discovery.

The process of biomarker identification is multi-staged, requiring the initial identification of patterns that meet the biomarker criteria, followed by an assessment of their generalisability. Neuroimaging data, coupled with the exploratory nature of the analysis, exacerbates the issue of dimensionality, while the sheer number of statistical tests conducted further compounds the problem. Traditional statistical tools struggle to address these challenges effectively.

However, in a different scientific niche, pattern recognition methods have emerged alongside statistics and are better equipped to handle exploratory analysis followed by validation. Leveraging the strengths of pattern recognition, we can overcome the limitations of classical statistics in biomarker discovery. Nevertheless, it is crucial to emphasise that these approaches are not intended to uncover the underlying mechanisms of the disease. They do not explain why and how the disease emerges or elucidate the mechanisms responsible for its progression. Instead, they serve as accurate "snapshots" of the brain's state, often without a specific hypothesis, which are then correlated with clinical variables.

4.3 The curse of dimensionality

One of the core problems related to neuroimaging data analysis, particularly in developing biomarkers, is the curse of dimensionality [97]. Due to the limited size of participants in many neuroimaging studies, there is a significant risk of overfitting when conducting exploratory analyses. Therefore, it is crucial to reduce the dimensionality of the problem as much as possible before the model-fitting stage of the analysis.

As discussed in the previous chapter, various strategies can be employed to tackle this issue. One approach is to generate features at different levels of resolution. For example, instead of considering every individual voxel in the brain, using a brain atlas can provide a lower-resolution solution that reduces the dimensionality. Additionally, leveraging domain knowledge can guide the selection of specific modalities or regions of the brain that are expected to be more impacted by the disease. Although this approach

is less common in the search for biomarkers of psychiatric diseases, it can be valuable in studies of neurodegenerative disorders.

Even after implementing these initial strategies, performing a separate dimensionality reduction step may still be necessary. There are different families of dimensionality reduction techniques, which can generally be divided into two categories: unsupervised and supervised methods [98]. Unsupervised techniques, such as PCA or ICA, aim to capture the inherent structure and patterns within the data without prior knowledge of class labels. On the other hand, supervised techniques, such as t-tests or regularisation methods, consider the class labels or outcome variables to guide the dimensionality reduction process.

Supervised dimensionality reduction

In supervised approaches, the objective is to determine the "usefulness" of features in terms of their ability to discriminate between groups or predict a specific outcome. That means that the algorithm optimises with respect to the data labels. Although these approaches can be powerful, proper nested cross-validation is crucial to prevent data leakage and double dipping. Mwangi categorises these approaches into three families: filter, wrapper, and embedded methods [98].

Filter methods

Filter methods are typically one-dimensional tests that assess whether a feature can effectively discriminate. Examples include t-tests, Analysis of Variance (ANOVA), Fisher score, or Pearson correlation coefficient. These methods select a limited number of the most significant features to include in the model. Their main advantage is their speed, simplicity, and interpretability, which are particularly valuable for biomarkers. However, conducting numerous tests increases the risk of false positives, which must be reflected in the threshold. Additionally, independent feature selection may lead to significant collinearity issues between selected features, posing challenges for certain classification and prediction algorithms.

Wrapper methods

Wrapper methods, such as recursive feature elimination or addition, consider the interaction between variables by selectively eliminating or adding features based on predefined criteria. In recursive feature elimination, the model is initially trained using all features, which are then ranked by their relevance in the model. At the end of an iteration, a percentage of the least influential features is discarded. The loop continues by fitting smaller

models and discarding more features until the stopping criterion is met. Unlike the previous family of methods, recursive methods incorporate feature interplay while maintaining interpretability, however, they can be computationally expensive, and the selected features may be unstable. This issue is usually addressed by constructing consensus maps describing the frequency of different features being chosen for the final model.

Embedded methods

Embedded methods refer to regularisation techniques, such as the Least Absolute Shrinkage and Selection Operator (LASSO) or elastic net. These methods optimise model coefficients to prevent overfitting by setting the coefficients of insignificant features to zero. However, LASSO may arbitrarily select a feature when dealing with correlated variables, leading to instability. Additionally, fit problems may arise when there is a large disproportion between the number of subjects and the feature dimension. The elastic net was designed to address these limitations, however, at the cost of adding a hyperparameter regulating the balance between LASSO and ridge penalty.

Partial least squares correlation also belongs to the family of embedded methods. Unlike the previous approaches, it does not use the subset of the original features. However, combining PCA and linear regression identifies latent variables, which capture the maximum covariance between a feature set and labels. Partial least squares correlation may be very efficient also in multimodal fusion. However, it is prone to outliers and overfitting and loses interpretability.

Unsupervised dimensionality reduction

Unsupervised dimensionality reduction algorithms offer robustness by not relying on data labels, although they may require larger datasets to identify underlying patterns accurately. PCA and ICA are common representatives of this class. PCA utilises singular value decomposition to transform the feature matrix into a new representation, where each column is a linear combination of the original columns, capturing most of the dataset's variation in the first few components. Additionally, the new feature vectors are orthogonal, which resolves issues of intercorrelation but sacrifices interpretability. ICA is akin to PCA and is frequently employed for extracting functional networks from fMRI or EEG data (see the previous chapter). Some researchers favour non-negative matrix factorisation as an alternative to PCA, which ensures non-negative components for improved interpretability [99]. However, this method requires non-negative values in the original feature matrix, posing challenges when analysing functional connectivity data, which include correlations with values ranging from -1 to 1 and are incompatible with non-negativity constraints.

In conclusion, due to the nature of neuroimaging data, the dimensionality reduction step can not be overlooked. Each method has its advantages and disadvantages, which need to be carefully considered based on the specific goals of the analysis. Decomposition-based methods are often effective in reducing dimensionality but can sacrifice interpretability. Conversely, feature selection methods may encounter instability issues due to the selection process or intercorrelation among features and can be computationally demanding.

4.4 A roadmap of models

The first step in algorithm selection is defining the specific task. It can generally be categorised as a classification or prediction task. Classification task involves the discrimination between different groups, such as patients and controls or distinct disease phenotypes. The prediction task aims to establish relationships between neuroimaging features and clinical scales. In some instances, there may be a third type of task, focusing on identifying new disease phenotypes using unsupervised clustering methods.

Classification tasks are typically considered relatively straightforward, aiming to detect differences between healthy individuals and patients. Often, this is only an initial step in the analytical pipeline, which serves to demonstrate the presence of discernible group differences, which is then followed by the more challenging task of predicting clinical scores. However, prediction tasks come with their own challenges, as clinical scores are often disease-specific and thus only available for patients, resulting in a smaller dataset for learning. Furthermore, the reliability of clinical measures can introduce uncertainties when used as the gold standard for modelling neuroimaging features [100].

The amount of available data is another critical factor to consider when selecting an algorithm. Data scarcity remains a bottleneck in using machine learning methods for neuroimaging data analysis. Studies have consistently shown a negative correlation between algorithm accuracy and sample size, suggesting that classifiers tend to perform worse with larger datasets [101]–[103]. This raises concerns regarding the generalisability of neuroimaging findings, underscores the risks of overfitting small datasets, and highlights the influence of scanner and site effects.

Lastly, the level of interpretability desired from the algorithm should also be taken into account. Simpler approaches tend to offer greater interpretability, although it is important to note that even these methods do not provide a comprehensive understanding of the underlying disease mechanisms. Nonetheless, the ability to formulate rules or identify features that contribute to successful classification can be advantageous when seeking insights into the analysis results.

Logistic regression

Logistic regression is a simple yet widely used classifier in neuroimaging data analysis. It models a linear combination of features and fits the binary data using maximum likelihood estimation with a log-link function. The resulting probabilities indicate the likelihood of each observation belonging to a specific class. During the validation process, a threshold is determined to classify unseen data based on these probabilities. Logistic regression is transparent, interpretable, and can be extended to classify into multiple classes. However, it may struggle with high-dimensional data. It has been successfully applied in various neuroimaging studies [102], [104]–[107], and its generalisation for prediction purposes is a well-known linear regression.

Linear discriminant analysis

Linear discriminant analysis is another popular method, which may be viewed as a supervised version of PCA. It can also serve as a dimensionality reduction technique. Under the assumptions of equal covariances and linear separability, it uses singular value decomposition to find a low-dimensional boundary that separates the classes. While linear discriminant analysis is effective due to its use of singular value decomposition, it sacrifices interpretability and has relatively strong assumptions. Its extension for prediction is quadratic discriminant analysis, but it did not gain widespread popularity in neuroimaging.

Support vector machines

Support Vector Machines (SVM) have been extensively used in neuroimaging studies, with over 70 out of 116 reviewed studies employing this method in 2015 [102]. SVM is advertised as particularly suitable for handling high-dimensional data with a large feature-to-sample ratio, which also offers a favourable balance between performance and the number of hyperparameters to tune. The algorithm constructs a boundary between classes based on extreme examples, known as support vectors, determined by regularisation criteria. Moreover, by employing kernel functions, the complexity of the boundary can be upgraded from linear to nonlinear. SVM provides some level of interpretability, but only to the extent where the samples used for constructing the boundary can be identified. Additionally, an extension of SVM called Support Vector Regression (SVR) exists for prediction purposes.

Decision trees

Decision trees are algorithms that recursively split data into smaller groups, optimising for the homogeneity of subgroups or, in the case of regression, the variance of the target variable. This approach is interpretable but prone to overfitting, especially with deep and complex trees. However, measures can be taken to mitigate this risk. Although individual decision trees are not extensively used in neuroimaging data analysis, ensemble methods based on decision trees have gained popularity in specific domains such as Alzheimer's disease [108]. A representative of an ensemble method are random forests, which construct multiple decision trees on bootstrapped subsamples of data and make predictions by majority voting. Alternatively, gradient boosting sequentially constructs decision trees to correct the mistakes of previous trees.

Multilayer perceptron

The multilayer perceptron is a fully connected feed-forward neural network, which opened the door for deep learning approaches in neuroimaging [109]. It consists of artificial neurons sorted into layers, aligned horizontally, where each neuron of the previous layer is connected to all neurons of the following layer. Thus the input to any neuron in a second layer is a weighted sum of the inputs from all neurons from the first layer. The weighting is unique for each connection and is tuned during learning. Similarly to a biological neuron, the neuron only passes information or "fires", if the input surpasses a threshold determined by an activation function. The first layer consists of input features, while the last one contains the number of neurons identical to the number of final classes. The number of hidden layers and units within them is arbitrary and can be tuned. Once the number of hidden layers is more than one, the neural network is called deep. Training a multilayer perceptron involves initialising the connection weights, feeding the network with random subsamples of data, evaluating the results, and updating the weights using the backpropagation algorithm. This process iterates until the desired performance or a specific number of learning rounds is achieved.

Neural networks, including multilayer perceptrons, are able to learn complex non-linear functions and can achieve remarkable performance given sufficient training data. They have shown success in applications such as glioma detection, although they have not significantly outperformed simpler methods [110]. However, their use in disease classification and outcome prediction is challenging. Especially psychiatric diseases often exhibit subtle differences between patients and controls, requiring even larger training datasets, which can be limiting for many researchers. Furthermore, achieving reliable performance requires tuning thousands of hyperparameters without overfitting.

Convolutional neural networks

Convolutional Neural Networks (CNNs) address the challenge of tuning numerous hyperparameters in fully connected networks. Initially developed for 2D image analysis, CNNs can be extended to analyse 3D structures suitable for neuroimaging data analysis. The input image is convolved with a set of kernels, sliding across the image to extract local dependencies and generate feature maps. Subsequently, subsampling is often performed to reduce the dimensionality of the feature maps, followed by additional rounds of convolutions and subsampling. Fully connected layers and a classification/prediction layer are typically included at the end of the network structure. Compared to multilayer perceptrons, CNNs have fewer parameters to tune due to the smaller size and stationary nature of the kernels.

While 2D CNNs have been previously used in neuroimaging research, a systematic review indicates that 3D networks outperform them [111]. However, despite the advantage of 3D networks, voxel-level SVMs still demonstrated comparable performance. Additionally, the review revealed instances of data leakage resulting from improper neural network implementations, emphasising the need for meticulous handling of such issues in any methodology, particularly in these advanced approaches.

The superiority of 3D networks over 2D is reasonably intuitive. In the 2D approach, a single slice of the brain is fed into the network's initial layer, requiring the network to either diagnose a disease based on a single slice or handle slices from different brain regions. In contrast, 3D net captures spatial dependencies but faces computational challenges when dealing with whole-brain inputs consisting of hundreds of slices. A possible solution to reduce computational requirements could be subsampling but at the risk of sacrificing information and precision.

Autoencoders

Autoencoders are a specific feed-forward network designed to learn a low-dimensional representation of the data rather than classification or prediction. They serve as dimensionality reduction techniques and are often integrated into larger AI frameworks. Autoencoders have input and output layers of the same dimension, with the hidden layers containing a decreasing number of neurons in the encoder section, minimising input dimensionality. The decoder section then reconstructs the original image, gradually increasing the number of neurons in the hidden layers. Various forms of autoencoders, including denoising, sparse, or convolutional autoencoders, emphasise different properties of the data [82].

Generative adversarial networks

Generative adversarial networks consist of two competing network blocks: the generator and the discriminator. The generator aims to produce realistic features, while the discriminator attempts to distinguish between true and generated feature vectors. Throughout the learning process, the generator improves its ability to produce realistic features, while the discriminator enhances its discriminatory capabilities [112]. Depending on their implementation and purpose, these networks can be supervised, semi-supervised, or unsupervised. When combined with a classifier, generative adversarial networks have been successful in classification tasks related to Alzheimer's disease, tumour detection, and simulating longitudinal brain aging [113]–[116]. These networks share disadvantages with other deep learning approaches: they require substantial amounts of data, which makes them primarily applicable to structural images, the most prevalent neuroimaging modality.

In summary, a wide range of algorithms from classical statistics, ML, and ML are suitable for neuroimaging data analysis. However, before choosing the "favourite" algorithm, it is necessary to recognise the implications of the "no free lunch" theorem, which reminds us that there is no one-size-fits-all algorithm that performs optimally for all problems [117]. In fact, the choice of the algorithm should consider various factors, including the specific characteristics of the data, the desired interpretability of the results, and the goals of the analysis. It is crucial to be aware of available options to avoid excessive dependence on a single method, enabling informed decision-making.

Chapter 5

The battle against confounders

In the previous chapters, we have dissected the steps involved in neuroimaging data analysis. However, it is important to recognise that additional factors come into play and complicate the analysis. One such underestimated factor is the presence of confounders. Confounders refer to features or effects that are correlated with the outcome but are unrelated to the effect being studied. Common examples include age, sex, and scanner effects. For instance, when examining the effect of stroke on cognition, we are interested in the impact **beyond** the natural cognitive decline associated with age. Failure to account for confounders can lead to dramatically different results [118], [119], as discussed in the second original article [120]. Therefore, it is crucial to address this issue appropriately.

5.1 Confounded

One straightforward approach to address confounders is to transform the original variable in a way that counters their influence. For example, regional volumes differ systematically between men and women due to differences in intracranial volume. Normalising the regional volumes by the intracranial volume represents the information as a fraction, effectively eliminating this bias. However, this approach is applicable only in specific and relatively trivial situations.

Another option is to include the confounder as a covariate in the analysis. Ideally, the variability in the target variable associated with the confounder will be accounted for by including it as a covariate, thereby eliminating its effect. However, complications can arise when there are strong intercorrelations among explanatory variables. If the confounder substantially impacts other explanatory variables, computational instability may occur, leading to erroneous associations.

A more drastic solution is to "regress out" the confounder. This involves two stages: First, the target variable is modelled as a function of the confounder, and the model is

fitted to obtain residuals. These residuals are then used as a new target variable for subsequent analysis. The same procedure is applied to the explanatory variables. While this approach effectively removes the confounder's influence, it can be overly aggressive and eliminate the very effect of interest if it is strongly associated with the confounder.

All of the aforementioned approaches for handling confounders are in use because they are conceptually understandable and can be integrated into the analytical pipeline. However, it is important to nest the control for variables within the validation cycle and perform it individually for each fold before training the classifier, which is sometimes disregarded [121]. However, a new approach to confounder control has recently emerged, which we will explore in the following section.

5.2 Normative modelling

With the explosion of available neuroimaging datasets, a new area of research known as "brain-age" studies has emerged. These studies aim to map the healthy development of the brain across the lifespan [118]. The idea behind the brain-age estimation is to compare an individual's brain to the expected brain development at their age, providing information on whether it appears older or younger than expected. However, the clinical benefits of this information have remained uncertain for some time.

One particular branch of brain-age research, known as normative modelling, has taken an interesting perspective on using large cohorts of healthy individuals [122], [123]. Normative modelling involves constructing a model of the brain (or, more specifically, brain-derived features), typically derived from neuroimaging data, as a function of clinical features such as age, sex, and scanning site. Various statistical methods, including hierarchical Bayesian regression and generalised additive models, are then used to fit the model to a reference cohort. Subsequently, the model is applied to new data of interest, often from patients, projecting their data onto the model [124]–[126]. The crucial output of this modelling process is not a direct prediction for each subject but rather the subject's position relative to the reference population, typically expressed as a z-score. By definition, the z-score combines the information about prediction with the prediction of the average subject, normalised by the population spread, thus allowing for the inter-subject comparison.

As outlined above, the essential part of normative modelling is a pool of healthy subjects, ideally uniformly distributed across age and sex, to reflect population distribution accurately. However, the access either to data or computational power to estimate these models may be limited for some researchers. Thus, having access to accurate pre-trained models, which may be adjusted for local samples, is necessary. Recently, some types

of pre-trained normative models were published and are now available for widespread use [127].

The deployment of pre-trained models has three stages. First, local data is preprocessed, preferably in line with the preprocessing applied to the original data the model was trained for. In the second phase, the model is adjusted for the local scanning site using a set of healthy controls' data. Finally, after the adjusting, new subjects' data are projected into the model space, and z-scores are obtained. The z-scores may be further used as new features for machine learning purposes or evaluated directly by, i.e. counting the number of subjects with extreme deviations.

Normative models have already found applications in various settings, providing valuable insights into the individual neuroanatomy of diseases such as Alzheimer's, schizophrenia, autism, and other neurological and mental disorders [123], [128]–[130].

Despite the demonstrated usefulness of normative models, there is still much room for improvement and further exploration. For instance, an extension of pre-trained models for longitudinal neuroimaging datasets has been lacking, which will be the focus of the final original chapter of this thesis.

Chapter 6

Review and replicate – a tale of sex and (EEG) power

6.1 Disclaimer

An original version of this work was accepted and published as an article in Frontiers in Neuroscience on the 27th of October 2020, under the title: Predicting Sex From EEG: Validity and Generalizability of Deep-Learning-Based Interpretable Classifier [131]. I would like to express my gratitude to my coauthors: Martin Brunovský, Martin Bareš, and Jaroslav Hlinka.

6.2 Introduction

In the introductory chapters, we dissected the plethora of classifiers used to analyse neuroimaging data and answer the question about the (non)existence of differences between clinical groups. The first original work of this thesis very much follows this narrative, albeit with a small spin on how we pose our question. First, we shall look for differences, not between diseased and healthy (as usual) but between males and females suffering from the same condition – clinical depression. Furthermore, unlike most to-date exploratory ML approaches, this work is hypothesis-driven because it is, to a degree, a replication study. Indeed, the poor reproducibility and generalisability of ML models have been denoted as the most significant pitfalls of ML [5], [132], which is one of the reasons we decided to conduct this study.

In 2018, van Putten *et al.* constructed a deep CNN to predict biological sex, analysing 1,308 clinical EEG recordings of healthy patients, with a reported accuracy of 81% [133]. To provide more insight, the authors performed a visualisation and analysis of the filters of all six convolutional layers of the network, discovering that the algorithm classified

preferably using the beta-band-derived features. In a subsequent step, they performed multivariate logistic regression using only the beta power from all channels and reached an accuracy of 70%. To the best of our knowledge, at the point of conducting this research, the report of Putten *et al.* was the only successful attempt to automatically discriminate biological sex from clinical quality EEG data.

Moreover, the beta activity being a biomarker of biological sex may appear as an inconsequential finding *per se*, if reproducible, it would point to potentially relevant biological sex-related differences in the processes generating the EEG signal. Understanding sex-related differences in EEG is important for quantitative EEG assessment in both research and clinical practice. To check this theory, one may also expect this difference to hold in patients with neurological or psychiatric disorders.

We consequently decided to examine this finding on patients suffering from a major depressive disorder. Major depressive disorder is a psychiatric condition that has been known for the alteration of the wake and sleep EEG patterns [134], [135], comprehensively summarised by [135]. They include relatively inconsistent reports of alpha asymmetry, elevated absolute and relative alpha activity, and further changes in the slow-wave activity. While the EEG changes in depression could, in principle, affect the accuracy of sex classification, the reports of alteration of the beta activity are relatively sparse. However, some authors indicated increased beta activity [136], [137].

Broader research has been done on identifying EEG activity alterations following the major depressive disorder treatment. There is evidence that antidepressant treatment changes the EEG patterns to an extent making the outcome of the treatment partially predictable [138]. As a change of pattern could negatively affect the performance of a biomarker, we decided to assess the performance of the beta-power independently for the EEG data acquired before and after antidepressant treatment.

Here, we present an independent validation of the interpretable hypothesis formed by [133] based on their deep network analysis of EEG data. Moreover, we construct univariate and multivariate families of classifiers based on the EEG beta-band power to assess the discriminative power of beta-power in EEG as a sex biomarker in a sample of patients suffering from a major depressive disorder. In order to control for the effect of treatment, we investigate the classification accuracy before and after the intervention.

6.3 Materials and methods

Participants

A total of 144 participants with major depressive disorder were recruited. For details of the sample and recruitment criteria, see previous full reports of the clinical analysis [139]–

[141]. The patients received four weeks of antidepressant treatment based on the decision of the psychiatrist. The distribution of treatments in the study was as follows: serotonin-norepinephrine reuptake inhibitors (53 patients); transcranial direct current stimulation (21 patients); repetitive transcranial magnetic stimulation (16 patients); selective serotonin reuptake inhibitors (16 patients); norepinephrine-dopamine reuptake inhibitors (11 patients); and other treatment (17 patients). Upon the initial preprocessing, we excluded ten patients due to technical difficulties with the EEG recordings; in six subjects, the recordings were distorted and not readable, and in four subjects, the recordings of three or more channels were silent. This resulted in the dataset of 134 patients (93 women) with a mean age of 46 years (std = 11.7, min = 18, max = 65). Every participant was recorded twice, before and after the treatment. Prior to the study, the patients were informed about the design of the study, and each participant provided his/her informed consent. The study was approved by the ethical committee of the Prague Psychiatric Centre/National Institute of Mental Health. The design and all procedures adhered to the latest version of the Declaration of Helsinki and ICH/Good Clinical Practice guidelines.

Data acquisition

We worked with 19 standard electrode positions that were common in all patients (while discarding from analysis any additional contacts available only in a subset of patients): Fp1, Fp2, F3, F4, C3, C4, P3, P4, O1, O2, F7, F8, T3, T4, T5, T6, Fz, Cz, and Pz. The EEG was recorded for 10 minutes in a sound-attenuated room with subdued lighting, with patients in a semirecumbent position and eyes closed in a maximally alert state. During the recording, the alertness was controlled. If the patterns of drowsiness appeared in the EEG, the subjects were aroused by acoustic stimuli.

Data processing

We adopted the EEGLab MATLAB toolbox for data processing. The cleaning process was inspired by the PREP pipeline [142], [143]. At first, the EEG was downsampled to 250 Hz. The initial and last 30 seconds of the recording were removed. Subsequently, the `clean rawdata` function was used. The function performs multiple operations: 1) Removes channels that have been flat for over five seconds. 2) Applies a high-pass filter with 0.5 Hz cutoff frequency (transition width of the IIR filter: 0.25, 0.75). 3) Rejects the channels correlated with the neighbouring channels less than a threshold (correlation threshold = 0.75). 4) Removes the bursts via Artifact subspace reconstruction – applies PCA decomposition to the channels in sliding window and rejects and reconstructs the components for which the standard deviation differs from the most representative part of the signal. The standard deviation threshold was set at 5 [144], [145]. 5) Removes the

unrepaired windows – a sliding window of one second and 66% overlap deletes the windows containing more than four "bad channels". The removed channels were interpolated using spherical interpolation. Finally, the data were re-referenced to average reference, and a low-pass FIR filter was applied with a 40 Hz threshold. For each channel, the relative β band power was computed by dividing the power in the β frequency range [12-25 Hz] by the sum of the power in the four key frequency bands used in the original study (δ [0.5-4 Hz], θ [4-8 Hz], α [8-12 Hz], and β [12-25 Hz]).

Analysis

In order to test the presence of global beta power differences, we conducted a nonparametric Wilcoxon two-sample test on the mean relative beta-band power (i.e., averaged across all electrodes). We subsequently repeated the test to assess the differences in each individual electrode and corrected for multiple testing using Bonferroni correction. To assess the classification power of the mean relative beta power, we performed logistic regression on this feature and constructed the Receiver Operator Characteristic (ROC) curve. To rule out any potential bias due to the inequality of the sex ratio in our data (although plain logistic regression is generally robust again this), for all main analyses conducted in this study, we constructed additional models adhering to the following approach: 1) Randomly sample 40 males and 40 females. 2) Perform the logistic regression. 3) Construct the ROC curve. 4) Repeat the subsampling 100 times. As a result of this approach, we present the mean ROC curve over all the iterations.

As a further step towards a potentially optimised classifier, multivariate logistic regression was applied in order to take advantage of the additional information that may have been present across the channels but could have been suppressed by using the average in the initial task. As in the univariate analysis, we constructed a full model including the relative beta power of all 19 channels and a separate averaged model for the sex-ratio balanced data. In order to evaluate and minimise the possibility of overfitting, the same procedure was repeated while applying a leave-one-out validation scheme.

We report the Area Under the Curve (AUC) and the highest overall accuracy across all thresholds that provided a true positive rate above and a false positive rate beneath 50%. The true and false positive rates are reported with respect to the prediction of the minor class in the data — men.

Moreover, the computations were performed twice for additional validation, once for the data acquired before and once after the antidepressant treatment. All statistical analyses were run using Matlab [146].

6.4 Results

The initial test of the global differences between men and women in relative beta power showed a significantly higher relative beta power in women both before and after the antidepressant treatment ($p < 0.001$). The difference was apparent across all 19 electrodes when investigating the individual channels (Figure 6.1).

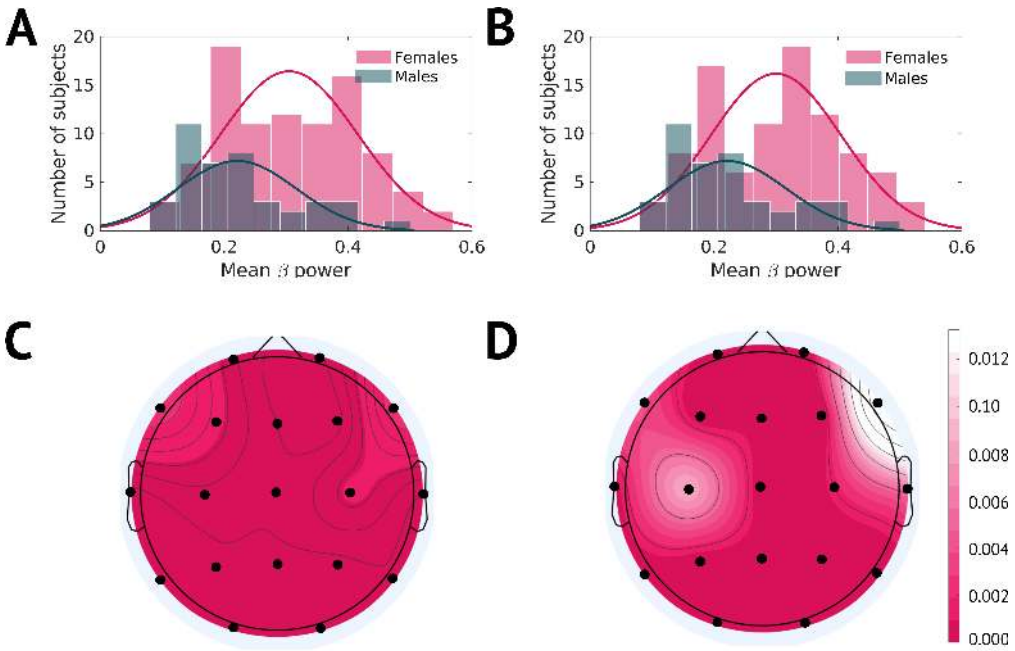

Figure 6.1: **The initial assessment of the beta-power difference between men and women.** **A** Depicts the histogram of the mean beta power in men and women before the antidepressant treatment and **B** after the treatment. **C** and **D** show the Bonferroni corrected p-values of the relative beta power differences across all electrodes before and after the treatment.

The use of one-dimensional logistic regression allowed powerful statistical evaluation of the full dataset without undergoing the risk of overfitting. The mean beta power feature generates the ROC curve with the AUC of 0.72 and 0.74 for the model before and after the treatment, respectively (Figure 6.2, Table 6.1). The highest accuracy (across thresholds for which the true positive rate was above and the false positive rate beneath 50%) was 77% and 70% for the treatment before and after, respectively.

The adoption of multivariate logistic models did not provide higher accuracy. Furthermore, the resulting AUC and accuracy substantially decreased after applying the leave-one-out validation, showing that the concern about overfitting was justified. Figure 6.3A shows the overfitted models where all data were used to build the model. In both cases (before and after the treatment), the AUC is above 0.8. However, applying out-of-sample prediction (Figure 6.3C), the AUC decreased to 0.64 and 0.72 for the results

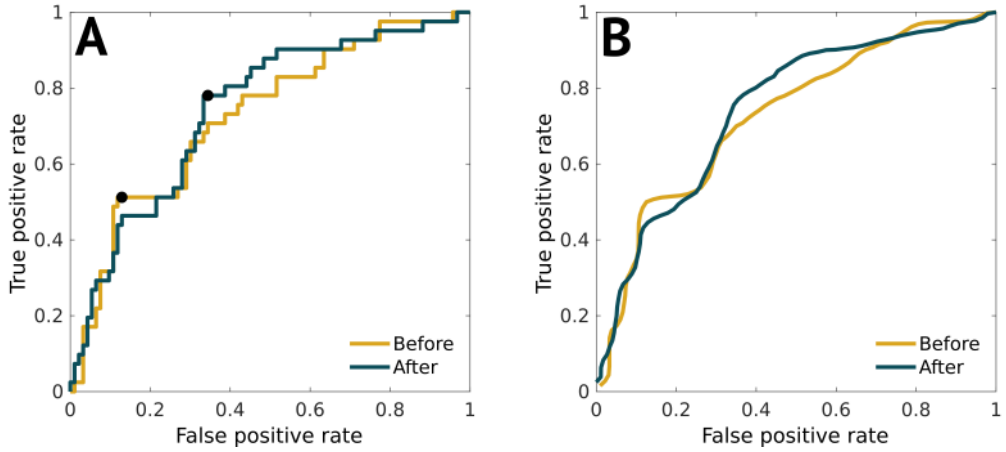

Figure 6.2: **The ROC curves of one-dimensional logistic regression.** Differences between men and women in mean relative beta power; before and after the treatment. **A:** The model was fitted on the whole dataset using the mean relative beta power. The black pointers indicate the position on the ROC curve, for which the overall accuracy is reported in Table 6.1. **B:** The model was fitted 100 times on a random balanced subsample of 80 patients, and the resulting ROC curves were averaged.

before and after the treatment, respectively, which is inferior to the initial grand mean approach. The subsampling procedure showed that the results on the whole dataset are not systematically biased by the majority class (Figure 6.3B and Figure 6.3D).

6.5 Discussion

Most authors agree that the ML approach to neuroscience has the potential to bring substantial advances to the field [5], [147]–[149]. Nevertheless, it has been rightly pointed out that the problematic reproducibility and interpretability of results limits their practical use [132]. Indeed, searching within black boxes allows us to identify features with high classification or prediction potential. However, our understanding of them is limited unless they are used in simpler, hypothesis-driven models [150]. Such simpler models are more comprehensible and often more neuroscientifically valid [12]. Although inferior in accuracy, they tend to be more robust, as they are less prone to overfitting due to the lower dimensionality [151]. To ensure the validity of simpler models, we need to conduct confirmatory studies that would investigate the findings reported by the ML on independent data [152].

In this work, we used the conclusions drawn by [133] from a deep learning study in a large sample of EEG data and decided to test for the relative beta-band power classification property with respect to biological sex. Working with the pre-defined hypothesis, we addressed two issues associated with the definition of biomarkers, namely testing the

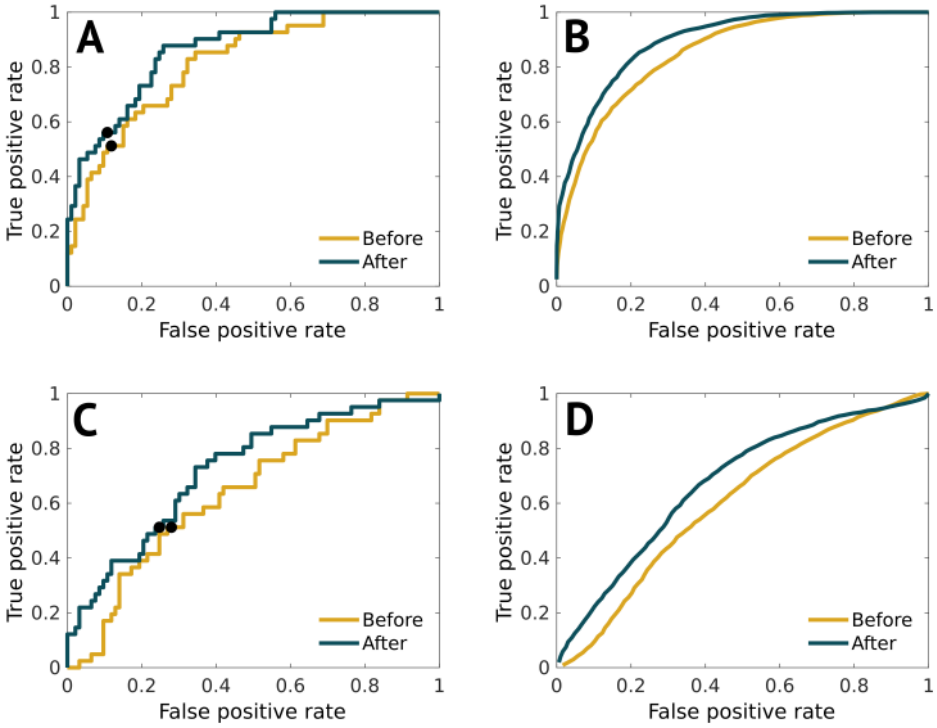

Figure 6.3: **The ROC curves of multivariate logistic regression.** Differences between men and women in relative beta power across electrodes; before and after the treatment. **A:** The model was fitted on the whole dataset using the relative beta power across all 19 electrodes. **B:** The model was fitted 100 times on a balanced random subset of 80 patients, and the resulting ROC curves were averaged. **C:** The model was fitted and evaluated using a leave-one-out validation scheme. **D:** The average of ROC was created using random sex-balanced subsets of the data and using a leave-one-out scheme to fit and evaluate the logistic regression. The black pointers in A and C indicate the position on the ROC curve, for which the overall accuracy is reported in Table 6.1.

results on an independent dataset and examining the robustness of the biomarker even in the presence of psychiatric disease.

In the statistical analysis, we focused solely on confirming the presence of the difference in the specific feature of *relative beta-band power* between men and women. This approach allowed us to minimise the amount of statistical testing, thus decreasing the probability of the occurrence of false positive findings. For this purpose, we have selected the logistic regression model as it is the model used in the original paper. The models containing only the mean relative beta-band power provided AUC above 0.72 and accuracy above 70% both before and after the treatment. Enhancing the models by using the relative beta-band powers from all individual channels did not significantly improve the diagnostic accuracy. In fact, the resulting multivariate models gained complexity due to the necessity to control for overfitting by out-of-sample testing without significantly improving the

Table 6.1: **The main results of the fitted models.** In non-balanced models, the overall accuracy is reported as the highest accuracy reached (assessed across all thresholds providing true positive rate above 50% and false positive rate below 50%); for the position of the points on the ROC curves, see Figures 6.2 and 6.3. For balanced models, the average of overall accuracies across 100 subsamples is reported.

	Area Under the Curve		**Overall accuracy for the chosen threshold**	
	Before	**After**	**Before**	**After**
Mean across the channels	0.7246	0.7425	76.87%	70.15%
Mean across the channels; balanced	0.7257	0.7257	69.14%	72.01%
All channels	0.8146	0.8652	77.61%	79.85%
All channels; balanced	0.8542	0.8941	79.09%	83.45%
All channels; leave-one-out	0.6420	0.7236	66.42%	68.66%
All channels; balanced; leave-one-out	0.5942	0.6481	61.47%	66.55%

predictions. Note that the AUC of the model using the mean beta power across channels is not prone to overfitting, as the only free parameter corresponds to the threshold that is varied across to provide the summary AUC measure.

Concerning classification accuracy, only the maximum across a range of thresholds is reported, while in practice, a specific working point is to be selected. However, the precise accuracy reached is meant to illustrate the strength of the differences rather than to aim towards devising a tool for diagnosing biological sex based on EEG. Rather, it suggests a substantial quantitative difference in the EEG signals between sexes that could point to some underlying differences in cognitive neurodynamics (see [133] for discussion of beta band differences to cognitive and emotional processing), or at the very least inform the EEG analysis practice of a potential confound of inter-subject analysis. Last but not least, it provides proof of principle and a springboard for the classification of clinically more relevant differences in EEG.

An interesting issue is that of using the multivariate or univariate model. In general, the accuracy reached by our one-dimensional model was consistent with the 70% accuracy reported by van Putten *et al.*. Note that we have used the same definition of the beta band as the authors of the original study (12-25 Hz). However, we decided to use relative spectral power, which should be robust with respect to interindividual and inter-session variability in the signal amplitude. To assess the robustness of the result, we also computed the logistic regression model on the averaged **absolute** beta power, which resulted in just a slight decrease in the overall accuracy of the classifier: 68% and 66% for the conditions before and after the treatment, respectively.

In principle, the multivariate model can potentially more sensitively fit more complex,

spatially dependent patterns of sexual dimorphism. On the other side, it is more prone to overfitting. The observed accuracy of the multivariate model is thus higher than for the univariate model, reaching up to 84 per cent accuracy. Conversely, the accuracy falls to 61–69 per cent with proper leave-one-out cross-validation. While, in general, the performance is comparable to that reported in the original work by [133], our results suggest that the use of a simpler and more robust univariate model based on the single feature of mean relative beta power is more accurate.

Of course, steps could be taken to improve multivariate models' accuracy, such as using dimensionality reduction methods or changing the modelling strategy to algorithms more suitable for high-dimensional data. In the case of dimensionality reduction, a prior decision on the method and the number of variables that ought to be present in the model is necessary. Furthermore, the method must be implemented correctly inside the cross-validation cycle to avoid double-dipping and prevent overoptimistic results [153]. To assess the role of the potential advantage of dimensionality reduction methods, we implemented a nonparametric Mann-Whitney test into leave-one-out cross-validation and compared the maximum accuracy reached across the number of channels used in the model. Overall, the condition of Bonferroni significance was not restrictive enough due to the widespread differences between males and females (see Figure 6.1), which resulted in the accuracy inferior to the one-dimensional mean of all features. However, reducing the number of variables to four or less improved the performance, and in the unbalanced dataset, even marginally outperformed the one-dimensional mean model.

Additionally, the SVM algorithm was used in order to compare the performance of full multivariate logistic regression models. Again, implemented in the cross-validation cycle, the SVM outperformed the mean logistic regression model on the data acquired after the treatment, but the classification accuracy on the data before the treatment was suboptimal, leading us to the conclusion that the logistic regression, used in the original study, was a suitable method for our experimental setting. Of course, while our results provided additional support concerning the validity of the original hypothesis, further re-validation and generalisation using independent datasets from both clinical groups and healthy subjects is warranted before widely utilised in practice.

We did not identify differences in the classification accuracy of the relative beta-band on data acquired before and after the subjects were given antidepressant treatment. In fact, the relative beta powers before and after the therapeutic intervention did not systematically change (paired t-test: $p = 0.1997$), and moreover, they were significantly correlated across subjects both in the mean (correlation of 0.8824, $p < 0.001$) as well as for all channels (mean correlation of 0.7798, std = 0.1018), supporting the existence of individually specific EEG signatures. Additionally, over half of the patients that were

incorrectly classified before the treatment were also misclassified based on the data after the treatment (19 out of 31). Our observation of a negligible effect of the antidepressant treatment on beta power is in line with the current literature. [154] described over 45 articles that derived quantitative EEG features in order to predict the depression treatment outcome. The most prevalent band-specific features were alpha-band activity, frontal theta activity, and theta cordance, whereas only one study reported decreased prefrontal delta and beta cordance in non-responders [155], which indicates that this band is not affected by treatment and thus does not play a role in treatment outcome prediction. Furthermore, to rule out any possible confounding effect of the different treatments on the relative beta power, we have tested for differences between groups using two-way ANOVA (accounting for sex and group and controlled for age) and observed no effect of group, both prior ($p = 0.53$) and after ($p = 0.62$) the treatment.

6.6 Conclusion

To summarise, in agreement with the explainable neuroscience framework, we followed up a previous deep-learning EEG study by testing for the presence of the inferred significant differences in the relative EEG beta-band power between men and women in an independent dataset. In order to test for the validity of this potential biomarker, we cautiously employed robust statistical approaches, which supported our hypothesis and provided classification accuracy of up to 77% in one-dimensional models. This illustrates the utility of explainable artificial intelligence approaches and independently supports a recent result concerning the sexual dimorphism of EEG signals.

Chapter 7

Cognition after stroke – (un)predictable?

7.1 Disclaimer

An original version of this work was accepted and published as an article in PLOS ONE on the 14th of April 2023, under the title: Structural connectivity-based predictors of cognitive impairment in stroke patients attributable to aging [120]. I would like to express my gratitude to my coauthors: David Kala, Jakub Kořenek, Veronika Matušková, Vojtěch Kumpošt, Lenka Svobodová, Jakub Otáhal, Antonín Škoch, Vlastimil Šulc, Anna Olšerová, Martin Vyhnálek, Petr Janský, Aleš Tomek, Petr Marusič, Přemysl Jiruška, and Jaroslav Hlinka.

7.2 Introduction

The previous chapter exemplified a typical hypothesis-driven classification task, where we formulated a hypothesis based on prior research and validated it using the available data. In this chapter, the hypothesis was only tentative, necessitating a more exploratory approach. Due to the limited sample size and the exploratory nature of the analysis, we decided to combine two modalities and focus on different levels of resolution, ranging from voxel-level analysis to examining graph connectivity metrics.

A common source of confusion in neuroimaging studies lies in the terminology surrounding prediction tasks. Immediate neuroimaging correlates are often referred to as "prediction," even when both the cognitive scores and neuroimaging data were measured on the same or closely related time points. This discrepancy in terminology does not present an inherent problem but requires readers to exercise caution when interpreting the findings. In this work, our primary objective was to explore the feasibility of predict-

ing cognitive scores measured one year after acquiring the neuroimaging data. Our focus was thus on the task of forecasting, specifically in the context of patients who experienced a stroke. The neuroimaging data were obtained within two weeks of the stroke, while the cognitive scores were measured one year later.

The global burden of stroke is increasing while the disease remains the second leading cause of death and disability, rising to 104.2 million prevalent cases worldwide in 2017 [156]. Despite the drop in incidence in developed countries, the number of cases is growing in low and middle-income states, reinforcing the need to understand the disease and the recovery process better [157].

Stroke is triggered by insufficient blood perfusion of the brain, which significantly affects patients and usually leads to considerable sensory-motor and cognitive disabilities [158]. The damage induced by stroke may be direct or indirect – through secondary degeneration.

Sensory-motor impairments following stroke are widely described and include hembody weakness, skin breaks, urinary tract or chest infections [159]–[161]. The treatment of muscle-restricted mobility usually consists of various forms of rehabilitation [162], [163]. The field has progressed so far as to construct predictive models to anticipate individual patient motor recovery potential [164], [165].

Contrarily, cognitive comorbidities of acute stroke, which include aphasia, loss of memory, orientation, and attention, although widely prevalent, are not as well understood and treated [166]–[168]. The current cognitive rehabilitation methods may be thus not optimally targeted [169]–[172]. As the treatment of post-stroke comorbidities presents a considerable social and economic burden [173], it is necessary to deepen our insight into the structural damage within the affected tissue, primarily the white matter [174]. The loss of white matter integrity is among the most direct consequences of stroke. Research concerning its impact on cognition has so far brought inconclusive results.

One of the well-established methods for studying white matter abnormalities is TBSS applied to FA maps or other white matter integrity metrics, linking the localised decrease in FA to the decline in various cognitive scales [175]–[179].

Another approach to investigate white matter integrity is to use a SC. Connectome describes the brain as a topologically complex interconnected network which balances regional and functional specialisation and integration [7]–[9]. This results in the coordination of processes across brain regions at low connection cost. However, it also implies that any dysfunction will spread through the network easily, possibly initiating pathological processes [180]–[182]. SC is determined by the model of white matter fibre pathways that physically connect predefined brain regions and is derived from the DWI using fibre tracking methods. Quantifying the relationships between the respective units of the brain

usually leads to constructing a connectivity matrix which describes the existence and, potentially, the magnitude of interconnection among all parts of the system and may be analysed (for more details, see Chapter 3).

Considering the uncertain effect of white matter integrity on cognition, the reports of the effect of structural disconnection are even more inconclusive. The ambivalence primarily originates in the differences among the study designs as well as in a wide variety of cognitive scales used [183], [184]. Moreover, the added potential of SC information provided on top of the usual lesion size for outcome prediction was investigated, so far, with contradictory results [185]–[187].

The analysis of connectivity matrices often employs so-called graph theoretical analysis. In this framework, the connectivity matrix is understood as an adjacency matrix of the graph [188]. The graph's vertices represent anatomically defined parts of the brain, and the edges are given by the matrix weights between individual regions. Graph-theoretical properties quantifying the topological features of the network are then determined. Usually, they include clustering coefficient, characteristic path length, small-world coefficient, centrality, efficiency, transitivity, assortativity, or rich club coefficient [189]–[191].

Using these measures, individual connectivity profiles may be derived to investigate healthy subjects and provide insights into networks damaged by either functional or structural disconnection. As stroke presents a violent disruption of the healthy network, clinical as well as empirical evidence suggests that investigation of the connectome or its parts could provide new insights into stroke-related comorbidities [192]–[194].

In summary, the relationship between white matter integrity and possible cognitive impairment following stroke is complex and has not yet been effectively explained. Numerous studies approached the topic using either white matter integrity measures such as FA or, more recently, analysed connectivity networks using structural neural paths derived from tractography [175], [184], [186]. However, the studies are not directly comparable, as they vary in design, methods for quantification and inference concerning white matter integrity disruption, cognitive scales used, and the interval between MRI and cognitive scales measurements with respect to the stroke date.

The time aspect is also particularly important, as each study may reflect a specific stage of white matter and cognitive recovery. In this work, unlike some previous studies that dealt with the immediate cognitive consequences of stroke, we focus on investigating the degree to which it is possible to predict *future* cognitive status (1 year after stroke) based on the white matter state measured within two weeks after the stroke. This task might be potentially more challenging but, conversely, more clinically relevant.

We decided to contribute to integrating the knowledge in this area by using three different methodologies for investigating the integrity of structural brain connectivity, par-

ticularly TBSS, the SC matrix estimated by tractography, and finally, graph-theoretical analysis thereof. Within each approach, we highlight specific methodological aspects and discuss their role in the analysis and interpretation of the results.

7.3 Materials and methods

Participants

Patients hospitalised with acute ischemic stroke between October 2015 and March 2017 were considered for the study. Within the acute phase of stroke (sudden onset language impairment, unilateral arm, leg, or face weakness), appropriate treatment was given (intravenous thrombolysis or/and mechanical thrombectomy) based on the decision of an on-call stroke specialist.

Subsequently, the patients were offered to participate in the study if they fulfilled the following criteria: age above 18 years, positive supratentorial acute ischemic lesion on admission (confirmed via MRI in the second week after stroke), and signed informed consent. The exclusion criteria included a history of epilepsy or acute symptomatic seizure preceding the current stroke, antiepileptic drug treatment planned for over two weeks after the stroke, history of clinical stroke, and contraindication to gadolinium administration. Furthermore, patients with other neurological (e.g., Alzheimer's disease, Parkinson's disease) or psychiatric comorbidities (e.g., bipolar disorder, major depression) possibly affecting cognition or brain tissue integrity were not included. Finally, approximately one year after the stroke, patients underwent a set of neuropsychological tests administered only to patients who did not clinically manifest aphasia.

The final dataset included 46 patients fulfilling all criteria (Fig. 7.1, Table 7.1). All volunteering patients gave written informed consent to participate in the study. Written informed consent was obtained directly from the included patients where possible. The level of information provided to patients matched their level of understanding as determined by the investigator. In large hemispheric infarction patients unable to understand or express themselves, the consent was given by a legally authorised representative (e.g., spouse or legal guardian) or physician not participating in the study team, in accordance with regional legal practice and regulations. The study was approved by the Ethics Committee of University Hospital Motol (Ref. number: EK-1091/14) and was conducted according to the Declaration of Helsinki's ethical principles.

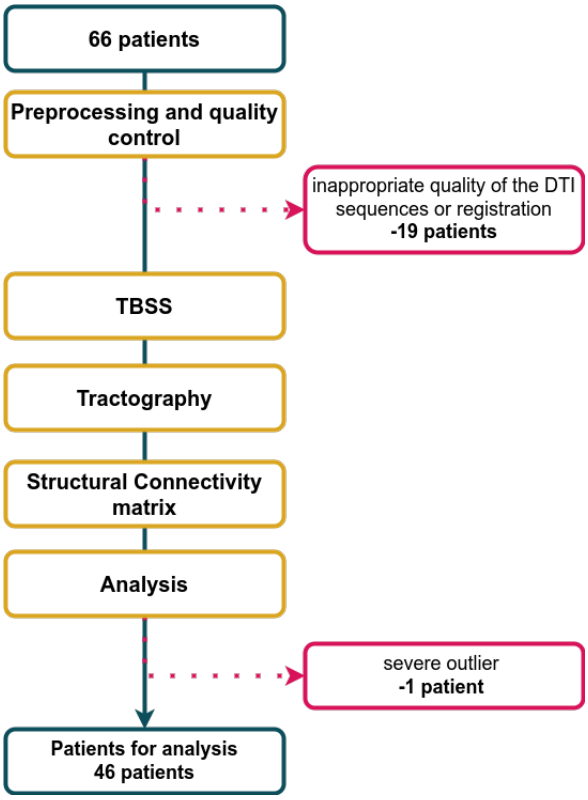

Figure 7.1: **An overview of the analytical process.** Initially the data of 66 patients were acquired. The data of 19 patients were discarded because of the unsuitable quality of the diffusion images or unsuccessful image registration. Moreover, we identified one severe outlier in neuropsychological performance, which was not included in the analysis. Overall, the data of 46 patients were analysed.

Cognitive function assessment

Patients' handedness was assessed during the acute phase using the Edinburgh handedness inventory. The rest of the neuropsychological scales were assessed on average 395 days after neuroimaging. The examination lasted approximately 45-60 minutes. Global cognitive performance was assessed by a Mini-mental state examination. The results of the neuropsychological assessment were converted to z-scores and combined into five cognitive domains by averaging the corresponding z-scores. Z-scores of the tests, in which a higher score indicated lower performance (Trail making test, Prague stroop test, Boston naming test), were inverted. The five cognitive domains were as follows: 1) Memory: Free and cued selective reminding test (Free recall, Total recall, Delayed free recall, Delayed total recall) [195]; 2) Executive functions: Trail making test part B, Phonemic verbal fluency (letters K, P, S), Similarities from the Wechsler adult intelligence scale-third edition, Prague stroop test [196], [197]; 3) Attention and working memory: Trail making test part A, Digit span forward and backwards from the Wechsler adult intelligence scale-third Edition [197]; 4) Language: Boston naming test (15-item version), Semantic verbal fluency

(animals) [197]; 5) Visuospatial functions: Visual object and space perception battery (number location), and Rey-Osterrieth complex figure test [198], [199]. Table A.1 shows a detailed description of the scores.

Table 7.1: **Description of the dataset.** Age and time between measurements are described as mean (std, min, max). Edinburgh handedness inventory is described as median (min, max). The time between measurements stands for the number of days between stroke and cognitive scales measurement.

	Men	**Women**	**All**
Number	24	22	46
Age (years)	66	66	66
	(9; 50; 88)	(11; 40; 86)	(10; 40; 88)
Time between	408	381	395
measurements (days)	(93; 340; 794)	(33; 342; 470)	(72; 340; 794)
Affected hemisphere	10/10/4	12/10/0	22/20/4
(R/L/Both)			
Edinburgh handedness	400	400	400
inventory	(-150, 400)	(-100, 400)	(-150, 400)

Data acquisition

MRI imaging was performed seven to twelve days after the onset of the symptoms using a 1.5 T magnetic resonance scanner (Philips Medical Systems). The acquisition protocol consisted of T1 and T2-weighted anatomical scans, FLAIR contrast, and DWI, with the following parameters: **3D T2 weighted:** TR 3200 ms, TE 263 ms, flip angle 90°, acquisition matrix 228×227, voxel size in mm 1.1x1.1x1.1, **3D T1 weighted:** TR 25 ms, TE 4.6 ms, FA 30°, acquisition matrix 220×198, voxel size in mm 1.1×1.1×1.1; **DWI:** TR 3157 ms, TE 94 ms, flip angle 90°, acquisition matrix 92×90, acquisition voxel size in mm 2.43×2.49×2.5, the reconstructed matrix dimension: 128×128 resulting in a reconstructed pixel of 1.75×1.75, no gaps, bipolar gradient sampling scheme, b 0 and 800 s/mm^2 (one b0 direction and 32 b800 directions).

Data processing

The DWI data were preprocessed using a combination of FSL software [200] and MRtrix3 [54]. First, all images were denoised, and Gibbs ringing artefacts were removed using the `dwidenoise` and `mrdegibbs` functions, respectively [44], [201]. Due to the lack of multiple b0 values, we employed the Synb0-DisCo algorithm [202] to synthesise an undistorted non-diffusion weighted image used as an anatomical target for distortion correction. Subsequently, the eddy current correction (`eddy`) was performed to address

geometrical distortions introduced by diffusion acquisition [43]. For the initial analysis, the FA, MD, AD, and RD maps were generated using the combination of `dwi2tensor` and `tensor2metric` commands, and the TBSS was employed to identify the regions of white matter related to the individual cognition scores [48]. The pipeline non-linearly aligns the FA maps onto a predefined template and subsequently affine aligns them to a standard MNI space where the image skeleton is created. We applied a threshold of 0.3 onto the mean skeleton, restricting the subsequent analysis only to the most dominant and well-aligned white matter tracts.

The construction of structural connectivity matrices based on deterministic tractography followed. We chose the Tax recursive calibration algorithm to determine the response function (`dwi2response`) and `dwi2fod csd` algorithm to estimate the fibre orientation distributions for spherical deconvolution [51], [203]. Deterministic tractography was performed using the `tckgen SD STREAM` algorithm with the following parameters: number of streamlines 10 million, step size 1, the maximum angle 60°, minimum length (in mm) 10, maximum length (in mm) 300; cutoff for terminating streamline FA < 0.1. [204], [205]. Seeding was performed homogeneously over white matter voxels. The tractograms were inspected, and a disproportional amount of 300 mm long tracts was discovered (as it was the maximum allowed length of the algorithm). To avoid a possible bias caused by these tracts, we decided to discard them. Finally, we registered the AAL atlas [38] in a two-step procedure: first, the standard MNI brain was registered to the T1 anatomical image using affine transformation with 12 degrees of freedom (`flirt`) available in the FSL library. Subsequently, the transformation that registered T1 to the b0 image of diffusion data was performed similarly. The individual structural connectivity matrices contained the absolute count of streamlines between each pair of regions.

Analysis

In the TBSS part of the analysis, we constructed a GLM to identify regions of white matter related to each cognitive scale. The statistics were corrected using TFCE [93], and a 5% level of significance was considered. We also considered variables suspected as potential confounds: age and hemisphere affected by stroke, which we included in the models as covariates.

In an exploratory analysis of the SC matrices, we initially constructed a median SC matrix, which served as a template defining typical brain structural connectivity. The median was chosen to avoid any bias towards the affected pairs of regions. Based on the template, we selected 5% of the most extensively connected pairs of regions and included them in further analyses. Thus, these strongest links define a backbone of the most substantial structural connections, including the most relevant 198 out of 3960 possible

pairs of regions. Apart from significantly decreasing the number of connections we tested, this approach also avoids superfluous analysis of potential false positive edges that might have arisen during tractography. We computed the Spearman correlation (R) of selected pairs of regions with age, and each of the six cognitive scales controlled for age. The p-values (p) obtained were corrected using False Discovery Rate (FDR) on the level of individual scales [206]. For clarity, we report the raw p-values of this analysis throughout, but only for those pairs of regions which survived the FDR correction.

Finally, we proceeded with the computation of graph-theoretical measures, which we correlated (using Spearman correlation) with clinical scores. We considered the following measures: assortativity, average strength, clustering coefficient, efficiency, graph energy, characteristic path length, rich-club coefficient, and transitivity. For comprehensiveness, we describe these measures in their binary form. Note that in this work, they have been appropriately adjusted and used in their weighted alternative [189].

Average strength is the average of all edge weights in the graph. *Graph energy* is the sum of the absolute values of the eigenvalues of its adjacency matrix. *Assortativity* describes a preference for graph nodes to attach to other similar nodes. Computationally, it is defined as the Pearson correlation coefficient of the node degree between pairs of linked nodes [190], i.e., assessing the connections to nodes similar in terms of connectivity degree or strength.

Clustering coefficient of vertex v is defined as the ratio of all triangles (cycles of length three) around vertex v to all possible triangles around v. The clustering coefficient of the entire graph is the average over all vertices. It quantifies the tendency of the graph to form clusters and is closely related to *transitivity*–the ratio of 3×number of all triangles in the graph to all possible triangles in the graph.

The *characteristic path length* of vertex v, is the average of all the shortest path lengths between the vertex v and the remaining vertices of the graph. Subsequently, the characteristic path length of the graph is the average of all characteristic path lengths of the vertices of the graph. Conversely, *efficiency* is an average of the inverse values of the shortest path lengths between all vertices in the graph. It measures the efficiency of information exchange between the vertices.

$\phi(k)$ is defined as the ratio of the present number of links to the maximum possible number of links between elements with node-degree at least k (in this study, we considered $k = 70$). In other words, $\phi(k)$ is the density of the subgraph induced by vertices of a degree greater than k. A generalisation for weighted graphs is described in detail in [207].

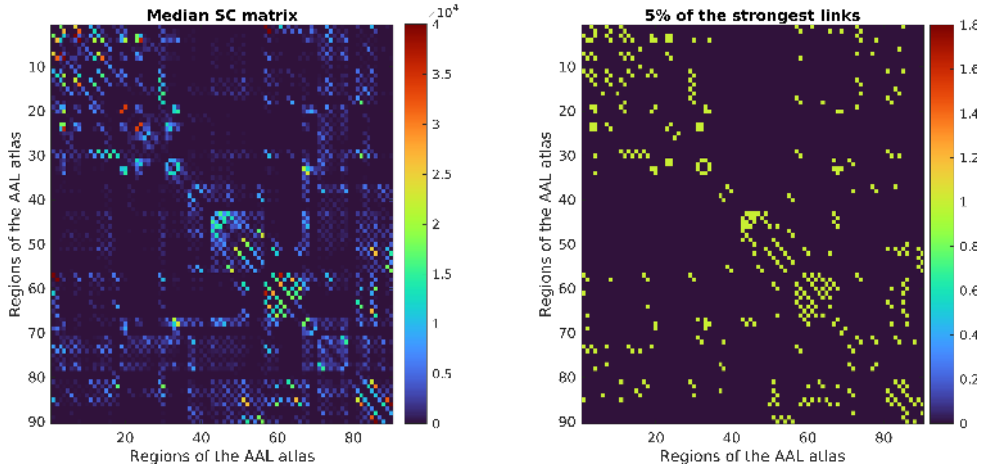

Figure 7.2: **Median Structural Connectivity (SC) matrix.** SC containing the median number of streamlines of 46 patients included in the analysis (left) and binary matrix with highlighted 5% of the strongest links (right). Each row/column represents an area of the AAL atlas (in the original order).

7.4 Results

The results of TBSS showed a positive correlation of FA with attention, executive functions, and memory (Fig 7.3). However, there was also a significant negative correlation with age. Notably, no statistically significant relationship was identified with cognitive scales after controlling for this variable. In contrast, the position of the stroke (in terms of the affected hemisphere) did not play a significant role in predicting cognitive status. We consequently disregarded this variable in further analyses. Mostly equivalent results were held for the rest of the diffusion metrics. In all cases, we observed a widespread negative correlation with executive functions and attention, none of which survived controlling for widespread positive correlation with age (Fig. A.1, Fig. A.2, Fig. A.3).

In the SC approach, a similar behaviour occurred. Specifically, we observed one pair of regions significantly correlated with memory, two pairs of regions significantly correlated with attention, and five pairs of regions correlated with visuospatial functions (all significant after the FDR correction). However, the age variable was again predictive of SC. Five pairs of regions showed a significant negative, and two pairs had a significant positive correlation. After controlling for age, we did not observe any link (belonging to the backbone) between the number of tracts and any cognitive scales that would survive the FDR correction.

In the case of graph-theoretical measures, we observed a slightly different behaviour. Relationship between the features and age was less predominant – only clustering coefficient, efficiency and the rich club were significantly correlated ($p = 0.0313$, $p = 0.0157$, $p = 0.0115$ respectively, Fig. 7.4). Moreover, none of the clinical scales was correlated

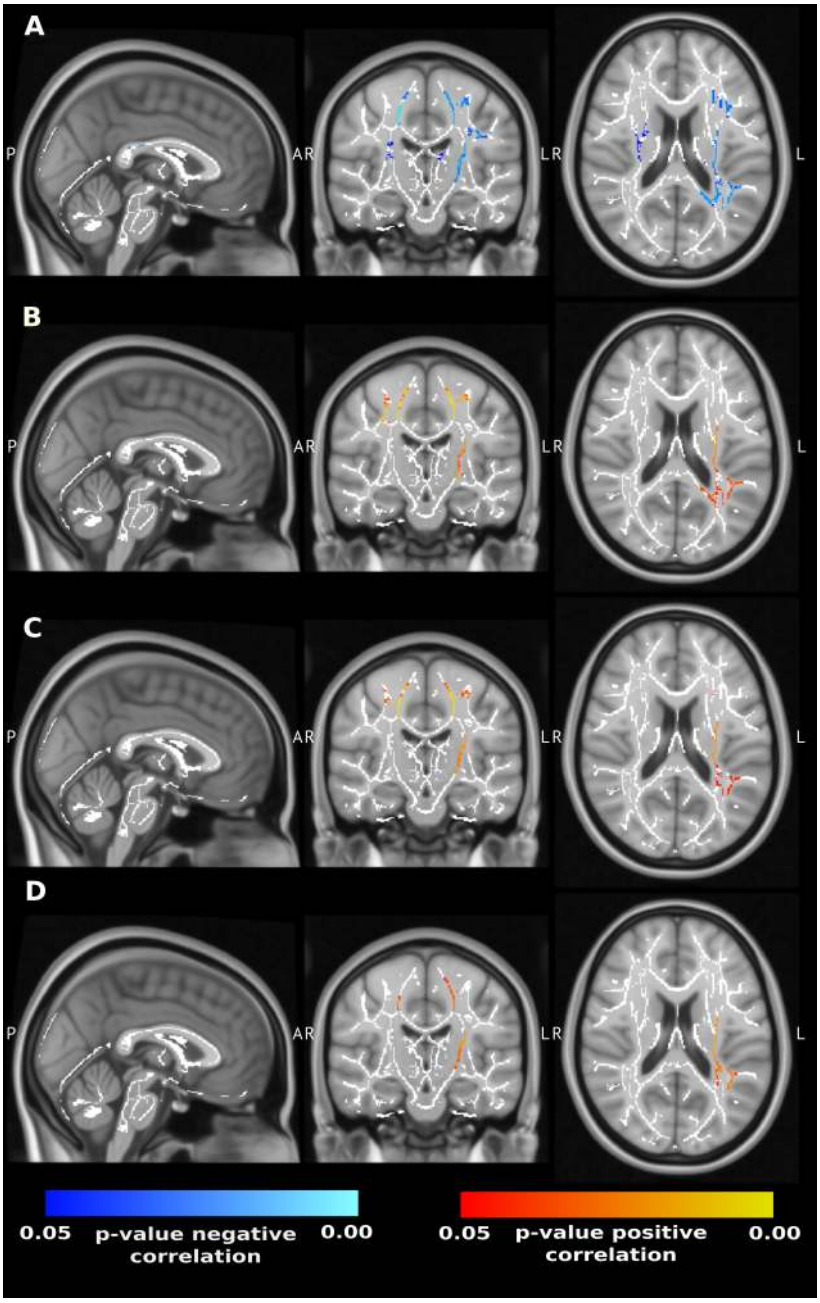

Figure 7.3: **The results of TBSS for fractional anisotropy (FA).** The blue colour scale signifies a negative correlation between FA and the clinical variable, red colour scale stands for positive correlation (either positive or negative correlation is depicted per clinical scale). **A:** Negative correlation of FA and age. **B:** Positive correlation between FA and executive functions. **C:** Positive correlation between FA and attention. **D:** Positive correlation between FA and memory. Note that the correlations shown in **B**, **C**, **C** did not persist (as statistically significant) after controlling for age, which was indeed detected as a significant analysis confound; see **A**.

with the features even before the age correction. This suggests that the graph-theoretical measures are less sensitive towards the effect of age than the previous methods; however, they are also less sensitive to the (future) cognition status.

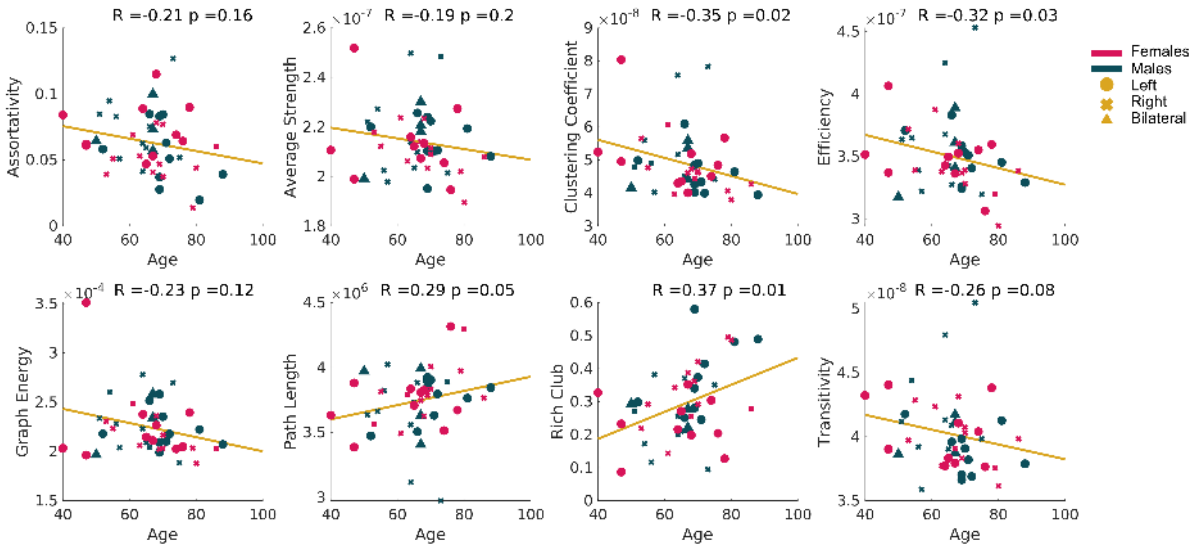

Figure 7.4: **The relationship of age and the graph-related measures.** Scatter plots of the graph-theoretical measures computed from the structural connectivity matrices of all subjects. Subjects are visually stratified according to sex (colour) and lesion location (marker).

7.5 Discussion

We examined associations between the condition of white matter acquired shortly after stroke and cognitive status measured with a one-year delay. The prolonged period between the two examinations is uncommon in this setup. Thus, the potential findings would reflect a predictive character of MRI features on later cognition rather than reflect an immediate relationship between brain damage and cognition. In contrast with previous works, we investigated different levels of resolution and methodology, namely: TBSS, statistical analysis of the SC matrices, and their graph-theoretical measures. Regardless of the analytical approach, we observed a strong effect of age, which was driving apparent correlations with the clinical scales. In all cases, correlations with clinical scales did not persevere after controlling for age.

TBSS analysis

The most significant correlate of all white matter-derived metrics in the TBSS analysis proved to be age. Additionally, executive functions, attention and memory score were significantly correlated with FA, MD, and RD. Only attention and memory were correlated with FA. However, after controlling for age, the effect in neither of the cognitive scales (and diffusion metric) was preserved. Indeed, cognitive decline is associated with ageing in healthy individuals [208], which suggests the effect of age on cognition would be relevant also in subjects after stroke. However, the effect of age is not always discussed or controlled

for in quantitative analyses, risking that its effect might be inappropriately assigned to other variables of interest [209], [210].

We also considered the effect of lesion laterality which was previously identified as an independent predictor of cognitive outcome after stroke [211]–[214]. However, the affected hemisphere did not play a significant role in the TBSS model, and we consequently disregarded it from further analyses. It is probable that the information about the position is already intrinsically present in the data in the form of reduced FA; alternatively, the (almost) binary nature of the position encoding might not have been precise enough to play a role in the analysis.

Dacosta *et al.* [176] studied 14 patients with right hemispheric stroke and found a significant decrease of FA in right brain anatomical areas compared to healthy controls. They demonstrated a relationship between cognitive functions and FA in several regions in both hemispheres. In another study [178], FA in the thalamus was associated with lower verbal fluency performance. However, in both cases, the sample size was relatively small– 17 patients in the latter study and subgroups as small as 8 or 6 subjects in the former study, and, notably, neither study discussed the effect of age nor reported controlling for it. We did not observe any comparable results in the language or executive functions domain, neither in the whole white matter skeleton analysis nor in a targeted analysis limiting the region only to the thalamic area as in the original study [178]. Of note, there is a range of relatively smaller differences between the studies, such as the time of acquisition, which was three months after stroke in the prior studies mentioned.

Zamboni *et al.* [215] studied the effect of acute stroke on early cognitive impairment (measured one month after the attack) represented by the Montreal cognitive assessment scale and Mini-mental state exam on over 400 patients. In this case, the Montreal cognitive assessment scale was correlated with reduced FA in the anterior tracts after controlling for the Mini-mental state exam.

In a longitudinal study following 117 stroke patients, 25 of whom were cognitively impaired, lower remote white matter integrity was associated with a worse long-term cognitive performance [175]. This result is specific as it reflects the relationship between cognition and white matter eleven years after the attack.

The reasons behind the absence of effect in the TBSS part of our study may be multifactorial. As argued above, the inconsistent practice concerning controlling for the effect of age may play a role in the heterogeneity of the previously reported results, as well as the time of DWI and cognitive scale assessment. In our case, the scans were taken in the subacute phase–within the second week after the stroke, whereas the cognitive scales were measured one year after, effectively attempting a more challenging medium-term cognitive outcome prediction rather than (almost) instantaneous correlation.

Structural connectivity analysis

On a SC level, we again detected pairs of regions significantly correlated with some of the clinical scales (memory, attention, and visuospatial functions), none of which persisted after controlling for age. We identified five pairs of regions significantly negatively correlated with age after the FDR correction, in line with the intuitive interpretation of the reduction of white matter tracts over time.

Additionally, two pairs of regions were identified to positively correlate with age. This observation is, to a degree, counterintuitive; however, such increases in the strength of shorter tracts might be a technical consequence of an overall white matter deterioration with age. In particular, such deterioration might complicate the correct tracing of longer tracts – thus disproportionately increasing the number of short tracts (when working with a fixed amount of tracts). Indeed the two pairs of tracts positively correlated with age in our dataset were on average shorter than those negatively correlated with age.

In related works which investigated SC in relationship to clinical outcomes such as aphasia, Yourganov *et al.* [184] constructed connectomes of 90 stroke patients scanned at least six, but on average, 42 months after the attack. The results highlighted the area of the temporoparietal junction and its connectivity as essential for language tasks, which was later supported by further analyses [216]. Apparently contradictory are the results from an extensive study that used the methodology of Yourganov *et al.* and acquired data from 818 patients suffering from aphasia approximately 58 months after a stroke [186]. The purpose of the study was to assess the added value of the structural disconnection information on top of the lesion load features to predict language score. No additional effect of the structural information was observed. Nevertheless, the study did not directly use the DWI data to evaluate the structural connectivity but rather imposed the disconnection defined by the lesion location on healthy subjects' tractography. This approach might thus disregard the remote structural changes caused by a stroke that would affect language performance.

None of the cognitive scales was significantly related to the SC in this framework. The reason behind the absence of a relationship may be tied to a common natural issue with such observational studies, that is, missing information on the exact cognitive scores before the stroke. Without the reference to the patient's cognitive performance before the stroke, the specific individual impact of stroke with respect to premorbid cognition can not be exactly inferred - this is a common problem for studies of stroke effects or other unexpected clinical events. Moreover, the potential presence of small vessel disease or other related conditions [217] could affect both the white matter and cognitive variables, leading to both spurious positive and false negative results depending on the specific effects.

Finally, in order to more directly connect and compare the results of the TBSS with our findings in structural connectivity, we extracted mean diffusion metrics along the backbone tracts in all patients and performed the same analysis (Spearman correlation with the FDR correction). Our findings further supported the results of TBSS in terms of finding significant correlations with age across backbone tracts, whereas finding no correlation with other cognitive scales after controlling for age (Fig. A.4). Overall, the effect of age on diffusion and the metrics derived thereof is a widely discussed topic; however, it is still not fully understood. There is sufficient evidence that the diffusion metrics are sensitive to age [218]; nevertheless, this effect is not necessarily homogeneous across the brain, but spatially varies [219]–[221]. Consequently, the sensitivity of the DWI-derived measures to age varies as well. In our study, FA, MD, and RD appeared to be more sensitive towards the effect of age than FA (in terms of the number of connections significantly correlated with age). This is consistent with the results of TBSS, where the effect of age was more widespread for other measures than for the FA.

Graph-theoretical measures analysis

In the final part of our analysis, the features represented by the graph-theoretical measures were less sensitive towards the effect of age than in the previous approaches. Despite the high degree of intercorrelation between the features, only the clustering coefficient, efficiency and rich club coefficient were significantly correlated with age. Additionally, no relationship between the features and clinical scales was found. There have been studies using other graph connectivity measures to study the effects of stroke [222]–[224]. Among the reported findings was the correlation of The national institute of health stroke scale with the betweenness centrality of the right pallidum and the clustering coefficient of the left superior occipital gyrus and a positive correlation between the nodal betweenness centrality of the posterior cingulate gyrus and Immediate recall [223], [224]. Upon replicating the measures, we did not observe any of the effects above, which, again, may be a consequence of the discrepancies in the designs and cognitive scales used. Notably, compared to the earlier discussed TBSS findings that we have not been able to confirm, these two studies included explicit control for age (as a key potential confound) and had a higher sample size ($N = 46$ and $N = 15$). Apart from some relatively minor technical differences and the ever-present chance of a false positive/negative result, the potential key factor behind the lack of replication of the observation of cognitive correlates of local graph theoretical measures reported by [223] is the temporal difference between the MRI and cognitive assessment–our study attempted one-year prediction, while the previous study apparently works with almost concurrent measurements.

Limitations

As was pointed out in the earlier parts of the article, the main impediment of our analysis is the lack of premorbid cognition scores for our participants, which affects our ability to unequivocally assign any observed relationship between white matter and cognition score solely to stroke. Unfortunately, this is the limitation of all studies discussing this topic and may only be solved by designing prospective trials focused on individuals at risk of stroke. A more tangible limitation of this work is the heterogeneity of the lesion location in participants. The inclusion criteria did not specify the position of the stroke. Consequently, it is possible that were the locations of the lesions more consistent across the dataset, more specific conclusions might have been drawn. However, to maximise the data size, we did not conduct a more position-specific analysis. Finally, as the diffusion acquisition protocol was optimised for widespread use in the hospital setup, we were limited by the methods which might be applicable to our data. In our case, we consciously used more conservative methods of fibre tracking, limited the analysis to mostly adjacent pairs of regions, and employed rigorous methods of statistical testing to minimise the possibility of obtaining false positive results.

7.6 Conclusion

In this work, we focused on investigating the relationship between white matter integrity in stroke patients and the prediction of cognitive status one year after the stroke event. Employing the TBSS analysis, we revealed that the cognitive correlates of white matter features in stroke patients could be attributed to the general effect of interindividual age differences, which was overlooked in some of the earlier studies.

From a methodological standpoint, this research required the use of asymmetrical multimodal fusion. Structural T1 images played a critical role in several stages, enhancing the registration, preprocessing, and analysis of the diffusion data: they were employed in the Synb0-DiscCo algorithm to synthesise undistorted images, serving as anatomical targets for distortion correction. Additionally, they contributed to refining white matter segmentation and atlas registration, facilitating the generation of high-quality structural connectivity matrices. Despite these methodological precautions, the dominance of age effects in the analysis hindered individual-level predictions for cognitive impairment.

Chapter 8

Together at last – uniting structure, function, and diffusion

8.1 Disclaimer

An original version of this work was accepted and published as an article in Brain Imaging and Behaviour on the 12th of November 2022, under the title: Multimodal-neuroimaging machine-learning analysis of motor disability in multiple sclerosis [225]. I would like to express my gratitude to my coauthors: Jan Mareš, Antonín Škoch, Jakub Kopal, Jaroslav Tintěra, Robert Dineen, Kamila Řasová, and Jaroslav Hlinka.

8.2 Introduction

In Chapter 3, we outlined the general structure of two philosophical techniques for handling multimodal data, specifically, multimodal integration and multimodal fusion. While integration approaches are easy for designing features, they increase dimensionality. On the other hand, the fusion strategies design features as a combination of modalities but are often tailored to combine only two of them. This study aims to integrate three separate modalities: fMRI, DWI, and structural T1. In light of the findings already presented in this thesis, we deliberately opted for simpler integration methodologies. This choice aimed to establish a robust baseline for effective comparison with fusion approaches. Interestingly, as this article unfolds, it becomes evident that these strategies have already demonstrated exceptional performance. Consequently, the need for fusion methods becomes redundant, at least for one of the two tasks.

In this case, our task is set in the context of multiple sclerosis, which is an autoimmune disease of the central nervous system. With an increase in the age-standardised prevalence, the latest epidemiological studies report 2.2 million cases worldwide, ranking it as the most

common demyelinating disease [226], [227].

Multiple sclerosis presents itself with a wide and heterogeneous spectrum of symptoms, of which motor impairment is the most dominant and restricting. Gait and postural control changes are present already in recently diagnosed minimally impaired patients and worsen over time [19], [228]. Reduction in mobility, together with fatigue, increase the risk of fall and injury, endangering the patient and reducing self-confidence [229]. More than 40% of people with multiple sclerosis report walking difficulties, which are identified as the most challenging disease aspect. However, the majority of people with multiple sclerosis do not voluntarily seek medical advice regarding walking impairments and detailed motor function evaluation is rarely performed in practice [230]–[232].

Measuring motor impairment

A range of specialised tests and questionnaires has been developed to quantify the degree of impairment in multiple sclerosis. The most extensively used is the Expanded Disability Status Scale (EDSS) [233], [234]. The assessment is administered by a clinician and ranges from zero to ten by 0.5 increments. Zero indicates no disability, ten signifies a death due to multiple sclerosis. However, EDSS is a general measure of disability, and even though it accounts for motor impairment to some degree (especially in the range EDSS 4-7), its sensitivity is limited [235], [236].

Reflecting the limitations of EDSS towards motoric impairment, two more questionnaires have been later created: Multiple Sclerosis Impact Scale (MSIS) and Multiple Sclerosis Walking Scale (MSWS) [237], [238]. MSIS consists of 29 questions concerning day-to-day life, whereas MSWS inquires specifically about walking disabilities. In both cases, the higher the score, the greater the impairment.

Another broadly used scale is Multiple sclerosis functional composite developed by the Multiple sclerosis society's clinical assessment task Force [239]. It consists of three functional tests, z-scores of which are combined into one coefficient. The tests are designed to rate leg, arm, and cognitive function using the Timed 25-foot walk test; 9-Hole peg test; and Paced auditory serial addition test, respectively. However, in most of the studies, the three tests are analysed separately for more intelligible interpretation. Other functional tests are Timed Up and Go (TUG) test, which requires the patient to get up of a chair, walk three meters, turn around, and sit back on the chair, and Berg Balance Scale (BBS) [240], [241]. The latter is the one scale present in this study, where the score declines with rising disability. Both tests have been declared valid for the assessments of disability in multiple sclerosis patients [242]–[244].

The role of imaging in multiple sclerosis

Linking the clinical status of the patient (as represented by a clinical scale) with the results of the MRI findings has been a long-standing issue in the research of multiple sclerosis due to the "clinical-MRI paradox" [245]. The predictive power of clinical impairment based on the findings from structural MRI or MR spectroscopy was not satisfactory in the early 2000s, even though some correlations between EDSS and the derived measures were reported at the time [246]–[248]. However, efforts are still invested into exhausting the potential of T1-derived measures in the prediction of impairment [249]. Other methods began to be used to understand the nature of the disease better [250].

Diffusion weighted imaging

Considering the white matter changes occurring in the brain in the course of the disease, the DWI proved to be an effective and widely used method for assessing the condition of the white matter [251]. Studies investigating the correlation between these measures and the clinical scales either produce voxelwise analysis [252]–[255], average the measures across ROIs [43], [256]–[261] or across the whole brain [262], [263].

Despite the benefits of the DWI, the relationship between its measures and EDSS has not yet been satisfactorily established. Some studies reported an existing correlation [252], [254], [261], [263], [264], whereas in others, no correlation was observed [251], [253], [255]–[260], [265]–[267].

Concerning more specific disability measures, Lowe *et al.* suggested that correlating pathway-specific disability measures with the whole brain lesion burden is a mistake and likely to decrease the correlation observed between lesion burden and functional loss [268]. Following this approach, he observed a significant correlation between the z-score of Multiple sclerosis functional composite and transverse diffusivity. Relationship between this scale and DWI measures was also examined by further studies, reporting mixed results [251], [257], [263], [269]. A study by Jakimovski *et al.* examined two disability measures: Expanded Timed Up and Go and Timed 25-Foot Walk Test, and reported that no association was observed [270].

Other modalities

Although multiple sclerosis was originally thought of as a white matter disease, there is growing evidence of grey matter being significantly affected [271]. Studies have reported differences in grey matter volume and other derived measures between patients and healthy controls as well as between the disease phenotypes [251], [272], [273]. Some of the derived measures were even reported to correlate with EDSS and other clinical im-

pairment scales. For this purpose, the structural images are analysed using either general measures of volume — grey matter volume, lesion volume; or more specific measures of the cortical thickness. In the most detailed approach, VBM may be applied [251].

fMRI is predominantly used to investigate brain plasticity and cognition impairment [274]–[276], somehow neglecting the potential relationship to disability measures. However, even though this relationship has been so far sparsely studied, several works reported promising correlations between disability scales and functional connectivity [264], [266], [277], [278].

In summary, the literature relating motor deficits to brain imaging in multiple sclerosis is limited, often focusing on single neuroimaging modality [266], [270], [272], [278]. Moreover, the studies generally feature a small sample size, low degree of impairment, and high number of statistical tests, weakening the findings reproducibility. These limitations call for methods that not only detect group differences but also may prospectively characterise and predict health development in a clinical setting.

Machine learning in multiple sclerosis

Progress has been made towards the application of ML in multiple sclerosis [279]. The most common task is the classification of subjects into patients and controls, sometimes also considering the degree of impairment and the phenotype of the disease [280]–[284]. Most of the classification approaches use the SVM [285], and although theoretically, no prior dimensionality reduction is needed, previous studies commonly employed Fisher score [286], which quantifies the discriminatory power of variables. However, the consequences of working only with a pre-selected set of features optimised for predictive power is often poorly discussed and, if not done carefully (i.e., blind to the test dataset), may lead to overfitting.

In this work, we initially test for the presence of differences in FA, FC, and regional grey matter volume between multiple sclerosis patients and controls. Following is the investigation of the correlation of FA with the clinical scales (within the patient group). Subsequently, we identify imaging biomarkers using ML to distinguish people with multiple sclerosis from controls and construct models for more subtle differences in motor performance among people with multiple sclerosis.

8.3 Materials and methods

Participants

People with multiple sclerosis were recruited from multiple sclerosis centres across the Czech Republic. The inclusion criteria were: positive diagnosis of multiple sclerosis [287]; spastic paraparesis as a prominent clinical feature; stable clinical status for at least three months preceding the study (determined by neurologist), EDSS $\leq$ 7.5. Participants with disturbed mobility for reasons unrelated to multiple sclerosis (e.g. fractures, pregnancy, stroke) were excluded. All multiple sclerosis phenotypes were accepted: relapsing-remitting; primary progressive; and secondary progressive. We analysed 64 participants with multiple sclerosis and 65 healthy controls statistically matched for age and sex (Table 8.1). Participants were informed about the experimental setup and provided written informed consent in accordance with the Declaration of Helsinki. The study design was approved by the Ethics Committee of the Faculty Hospital Královské Vinohrady.

Clinical assessment

Participants with multiple sclerosis underwent a set of standardised assessments with the BBS, the TUG test, MSIS, and MSWS.

Data acquisition

Imaging was performed with a 3 T magnetic resonance scanner (Siemens Trio Tim, Erlangen, Germany) using a 12-channel phased-array head coil. The protocol consisted of T1-weighted and T2-weighted anatomical scans, DWI, and rs-fMRI. The parameters were: **DWI:** TR 9,100 ms, TE 96 ms, FOV 260 $\times$ 211.25 mm, 64 contiguous axial slices, 2 mm thickness, b 0 and 1100 s/mm, 64 gradient directions, voxel size 2.03 $\times$ 2.03 $\times$ 2 mm; **Resting-State fMRI:** BOLD single-shot echo-planar images TR 2500 ms, TE 30 ms, flip angle 70°, 64 $\times$ 64 matrix, FOV 192 mm^2, 44 contiguous axial slices, 3 mm thick, 240 volumes, acquisition time 10 min; **T1 volumetric imaging:** TR 2300 ms, TE 4.63 ms, flip angle 10°, matrix 256 $\times$ 256, FOV 256 mm^2, 156 contiguous sagittal slices, 1 mm thick.

Data processing

DWI was preprocessed using the FSL tools (FMRIB Software Library v5.0, FMRIB, Oxford, UK) and MRtrix3 v3.0_rc3 [54]. The data were denoised (`dwidenoise`) [201], [288], Gibbs ringing artefacts were corrected (`mrdegibbs`) [44]. The volumes with low

Table 8.1: **Description of the dataset.** All values are listed as medians and ranges. RR – Relapsing-Remitting, PP – Primary Progressive, SP – Secondary Progressive, EDSS – Expanded Disability Status Scale, BBS – Berg Balance Scale, TUG – Timed Up and Go Test, MSIS – Multiple Sclerosis Impact Scale, MSWS – Twelve Item Multiple Sclerosis Walking Scale

Healthy controls			
	Women	**Men**	**All**
Number	32	33	65
Age	43 (31-68)	40 (31-68)	41 (31-68)
People with multiple sclerosis			
	Women	**Men**	**All**
Number	39	25	64
RR/SP/PP	23/14/1	15/6/4	38/20/5
Age	47 (22-70)	43 (29-68)	45 (22-70)
EDSS	4.5 (1.0-6.5)	4.5 (1.5-7.5)	4.5 (1.0-7.5)
BBS	44.0 (15-56)	35.0 (5-56)	42.0 (5-56)
TUG	10.0 (5-27)	13.0 (6-109)	11.0 (5-109)
MSIS	67.0 (63-100)	80.0 (34-114)	72.5 (33-114)
MSWS	34.0 (12-59)	39.0 (25-60)	36.5 (12-60)

quality (visually checked) were discarded. Subsequently, eddy-current-induced distortions and movement displacement were corrected by the `eddy` tool. Finally, we generated FA maps [45], [289] and applied TBSS [48]. We parcellated the resulting skeletonised images using the white matter ICBM-DTI-81 atlas [290] containing 48 regions and computed the mean FA for each region, which resulted in 48 FA features per subject.

The fMRI data were preprocessed using a combination of the SPM12 software package (Wellcome Department of Cognitive Neurology, London, UK); CONN toolbox (McGovern Institute for Brain Research, MIT, USA) running under MATLAB [146] and FSL routines. To form FC matrices, we cross-correlated ROI-based BOLD time series from 116 regions of the AAL atlas [38]. In line with the standard practice, we used the Pearson correlation coefficient [56], [291], providing 6,670 FC features per subject.

Segmentation of the T1 images into white matter, grey matter, and cerebrospinal fluid was a part of the CONN pipeline employed for fMRI processing. We computed the volumes of the GM areas defined by the AAL atlas, having 116 grey matter volume features per subject. Additionally, the T1 scans were processed using the VBM approach in CAT12 [292]. The preprocessing was performed using default settings and involved bias-field and noise correction, segmentation into grey and white matter, and registration using the DARTEL algorithm to a 1.5 mm isotropic adult template [293]. Total intracranial volume was estimated and used as a covariate in statistical analyses.

Analysis

Classical analytical approaches

We initially performed standard analyses on all modalities to identify structural and functional differences. TBSS skeletons of patients and controls were compared using permutation testing. The same method was used to identify the areas that would correlate with clinical scales. Significant results are reported at the 5% significance level, corrected for multiple testing with TFCE [93]. We compared the differences in FC between the groups by performing a two-sided non-parametric Mann-Whitney test on each pair of regions and performed the FDR correction of the obtained p-values. Finally, we assessed the differences in smoothed grey matter masks between the groups by one-sided two-sampled t-test (with the presumption of patients having a decrease in grey matter) with FDR correction.

Classification

For the classification task, we developed two strategies:

- Fisher score as a dimensionality reduction step, classification by support vector machines (FS-SVM)
- Principal component analysis as a dimensionality reduction step, classification by logistic regression (PCA-LR)

The first approach uses the SVM classifier that can, in principle, deal with high-dimensional data, although it may be prone to overfitting, especially when it is combined with informed feature selection procedure such as the commonly used Fisher score [286]. In this approach, the features are selected into the classifier based on their individual classification power on the data in the training set. The box constraint (parameter C) for SVM was fixed to 1.

The latter approach uses a classical linear regression in combination with a small number of PCA components of the original features. This dimension reduction approach avoids overfitting by being blind to the patient/control labels. Unlike in the Fisher score method, the transformation of the data is based on their explained variance in general and is not influenced by class labels.

For each strategy, we constructed four classifiers: based on FA, FC, grey matter volume, and their combination, using leave-one-out cross-validation (see Fig. 8.1). All neuroimaging features were transformed to z-scores prior to the dimensionality reduction step and model fitting.

To further assess how the classification accuracy depends on the number of features, we evaluated each classifier for a range of thresholds. For the FS-SVM combination, we included 1%, 10%, 25%, 50%, 75%, and 100% of the features with the top Fisher Score. For the low-dimensional PCA-LR, we included one to ten PCA components. We consider sensitivity, specificity, and overall accuracy as measures of classification quality and compare the classifiers using the McNemar test. All analyses were conducted using Matlab [146].

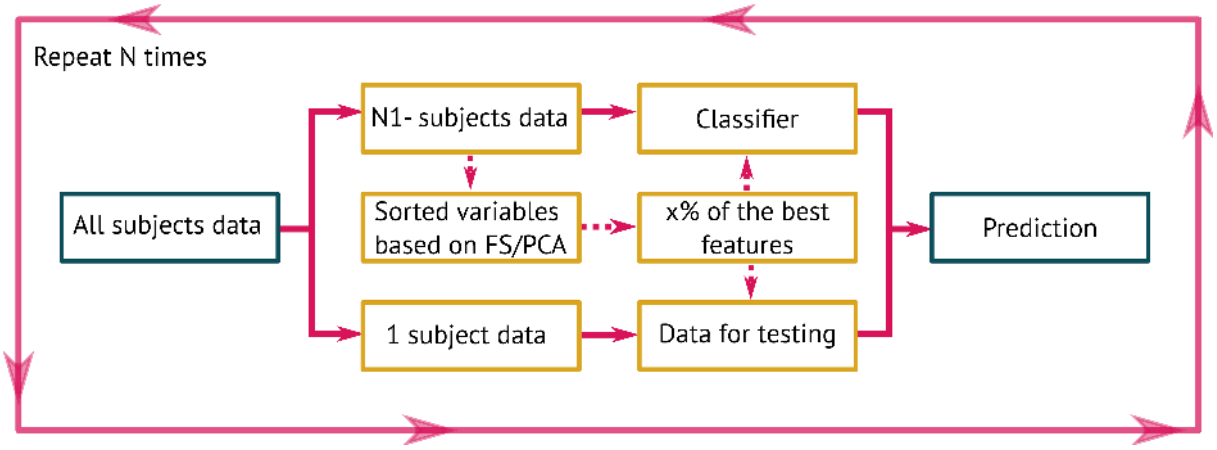

Figure 8.1: **A general scheme of the classification pipeline.** In every iteration, one subject is left out. Either Fisher score or PCA is computed on the training dataset, and the desired transformation (n terms of the number of components or features) is then applied. The classifier is trained using N-1 subjects and subsequently tested on the remaining subject. After N rounds, the classification accuracy is evaluated. Full lines on the image illustrate the data flow, whereas the dashed lines stand for data transformations.

Prediction of the motor impairment

We applied an analogous approach to the prediction of motor impairment. In particular, we replaced SVM with SVR and logistic regression with linear regression. For dimensionality reduction, we sorted the features according to the Spearman correlation with the predicted scale instead of using Fisher score. The use of PCA based on the variability explained remained unchanged. In each iteration, we performed a dimensionality reduction on N-1 patients, constructed a model and predicted the clinical scale score for the remaining patient. Model quality was assessed by the correlation between the prediction and the original scale. Prior to the regression analysis, Box-Cox normalisation was applied to the clinical scales. To avoid outlier and nongaussianity effects, Spearman correlation was employed for the final model fit assessment.

8.4 Results

TBSS analysis

There were widespread significant differences in the FA between patients and controls. The FA was higher in controls in 78% of the voxels in TBSS skeleton. For all clinical scales, the voxel-wise correlation was observed in the expected direction, that is, decreased FA with higher severity of symptoms (Figure 8.2). 18% of the voxels across the TBSS skeleton were significantly negatively correlated with the EDSS scale, predominantly in the corpus callosum, corona radiata, posterior thalamic radiation, and external capsule. In BBS, 40% of the voxels were significantly positively correlated across the whole skeleton. Finally, 8% of the voxels were significantly negatively correlated with TUG in the left posterior and right anterior part of the white matter. A wide but not significant negative correlation was observed across the whole skeleton in MSIS and MSWS.

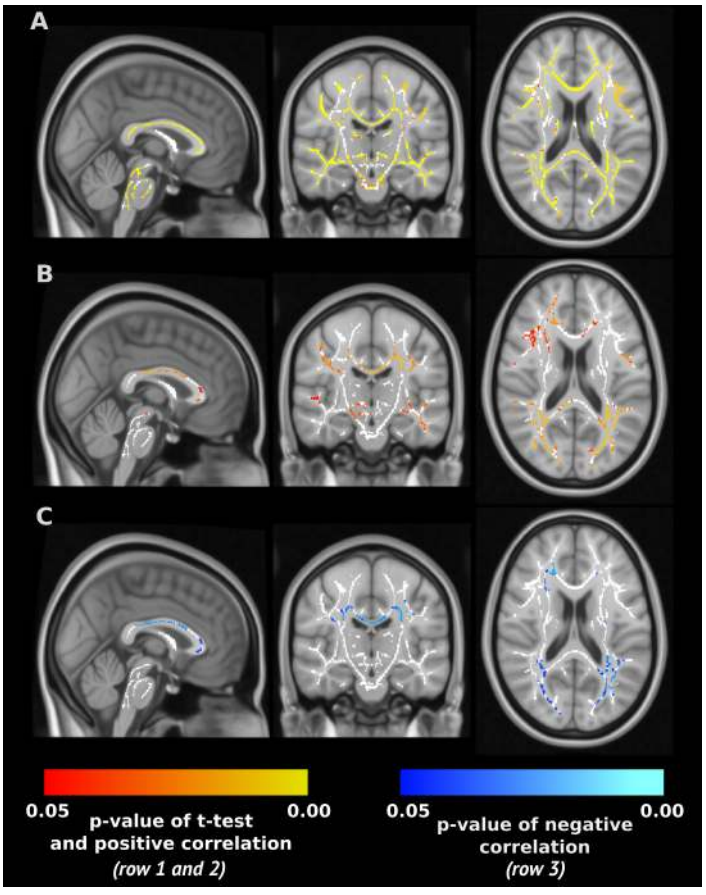

Figure 8.2: **The results of TBSS analysis. A:** TFCE corrected p-values of t-test between fractional anisotropy of patients and controls, yellow colour represents higher fractional anisotropy in controls. **B:** Positive correlation with the Berg balance scale. **C:** Negative voxel-wise correlation with the Expanded disability status scale.

FC analysis and VBM

The comparison of connectivity matrices between patients and control showed a significant difference in 87 pairs of regions (on 5% level of significance after FDR correction). In 84 cases, the FC was higher in controls. 50 most significantly different pairs are visualised in Figure 8.3. Finally, the VBM analysis confirmed a significant global decrease in grey matter volume in patients (Figure 8.4).

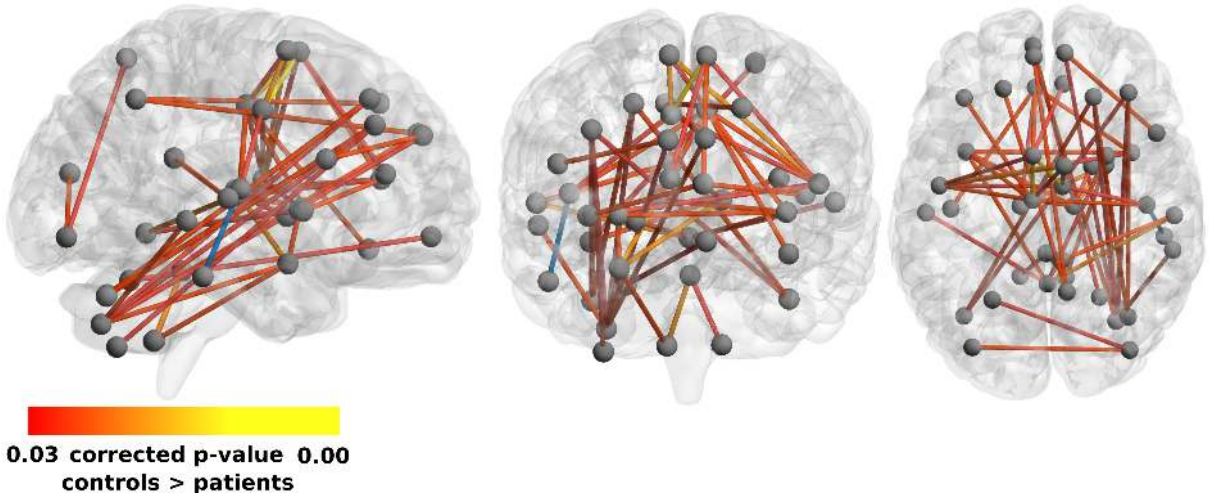

Figure 8.3: **Functional connectivity changes in multiple sclerosis.** 50 pairs of regions that most significantly differ between patients and controls are shown. Pairs of regions with higher functional connectivity in controls than patients are drawn in yellow to red; one pair of regions with higher functional connectivity in patients is drawn in blue.

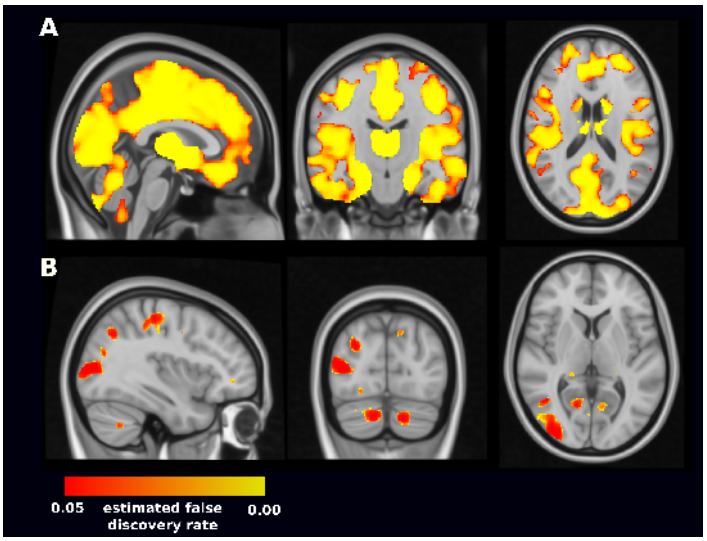

Figure 8.4: **Grey matter changes between patients and controls.** Decreases of grey matter volume in MS patients with respect to controls. To correct for multiple testing across voxels, thresholding controlling the false discovery rate at 0.05 is applied.

Classification

All neuroimaging modalities provided information sufficient to build a successful classifier, with accuracy ranging from 50 to 96 per cent (Fig. 8.5, Table B.1, Table B).

The SVM classifier based on FA performed the best; the baseline classifier consisting of a single FA feature reached an accuracy of 85%, whereas the accuracy for the full dataset was 96%. Albeit the accuracy was rising or stable over all the thresholds, the classifiers consisting of more than 25% of the features did not show significant performance differences (Fig. 8.5A, Fig. B.1).

The SVM classifiers based on FC or grey matter volume achieved about 70 per cent accuracy and were not generally improving with the number of features added.

We also assessed the possibility of further improving the classification by combining the features across modalities. The baseline classifier based on the combination of features from all modalities (containing 69 features) performed better than the corresponding classifier based on FA (90% vs 85% respectively); however, with a rising amount of data, the accuracy declined to 81% (full dataset).

In the low-dimensional PCA-LR approach, models using only the first PCA component of FA or grey matter volume reached an accuracy of 78% and 69%, respectively (Fig. 8.5B). The first component of FC provided insignificant classification; nevertheless, the third alone reached an accuracy of 70%. The lack of classification power of the first two components in fMRI suggests that, unlike the previous two modalities, the main source of variance does not relate strongly to the patients-controls differences. Notably, the three components with substantial classification power (the first FA, the first grey matter volume, and the third FC component) were significantly correlated (Spearman correlations: R(FA, grey matter volume) $= 0.6$, R(FA, FC) $= 0.47$, R(grey matter volume, FC) $= 0.46$). Therefore the combined model was not more effective.

Moreover, the model consisting of the first, second, and third PCA component of FA reached an accuracy of 90%, not significantly worse than the full FA data SVM classifier ($p = 0.121$). Similarly, the classifiers based on the grey matter volume or FC did not perform significantly worse than their counterparts in FS-SVM models (except one and two-dimensional FC models). Nevertheless, their accuracy was inferior to that of FA.

Prediction of the motor impairment

In contrast to the classification for the prediction of motor impairment, the results were not dominated by the FA. The best results in the SVR prediction were reached when using the FC features (Fig. 8.6, Table B). Especially for the MSWS scale, the prediction was consistently significant, reaching up to $R = 0.79$. The FA and grey matter volume

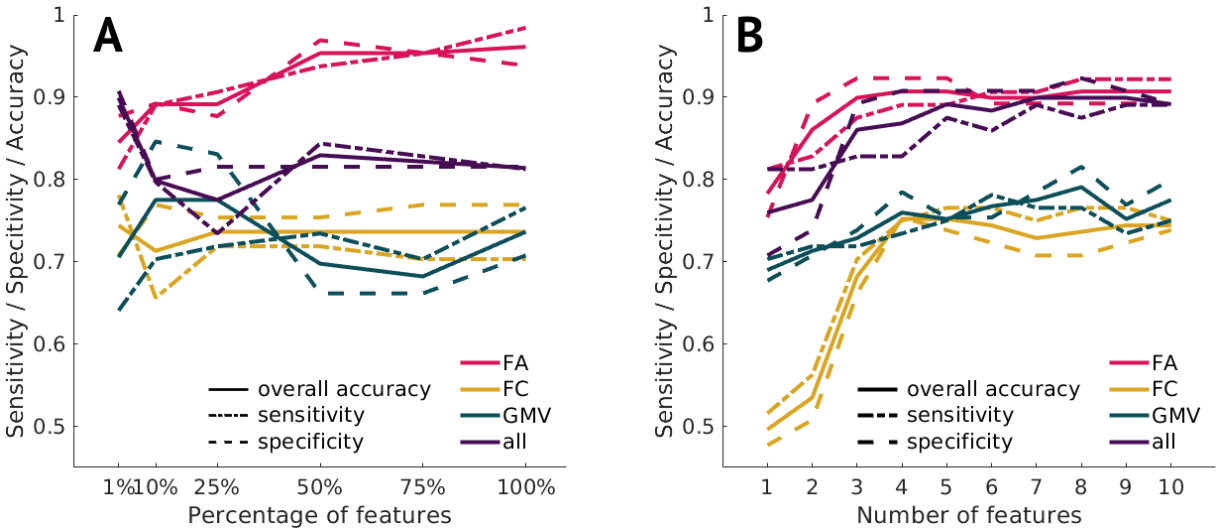

Figure 8.5: **Classification accuracy of Support Vector Machines classifiers (A) and Logistic Regression classifiers (B).** Overall accuracy (full line), sensitivity (dashed line with dots) and specificity (dashed line) for different thresholds. Classifiers based on fractional anisotropy (FA), functional connectivity (FC), and grey matter volume (GMV) correspond to pink, yellow and green lines, respectively; the classifier based on the combination of features corresponds to the violet colour.

prediction ability varied with the number of features included and did not exceed $R = 0.50$ for any combination of feature type, count and clinical scale. The results of PCA in combination with linear regression complement those reached by SVR (Fig. 8.7, Table B). While FC components were only successful in predicting the MSWS scale, when using the grey matter volume features, the low-dimensional model significantly predicted four out of five clinical scales (EDSS, BBS, TUG, and MSWS).

8.5 Discussion

General results

We have shown multimodal global distributed changes in brain white matter, grey matter and FC due to multiple sclerosis, generally in line with previous reports in literature mainly concerning single modalities. Moreover, we have observed significant relation of the white matter changes, as quantified by FA, to the motor or general impairment quantified by the BBS, TUG and EDSS scales. Importantly, we were able to assess and put the strength of these effects to test by constructing "diagnostic" classifiers with up to 96 per cent accuracy (cross-validated) and predictive models for the motor impairment from neuroimaging data reaching up to 0.79 correlation with the clinical scales.

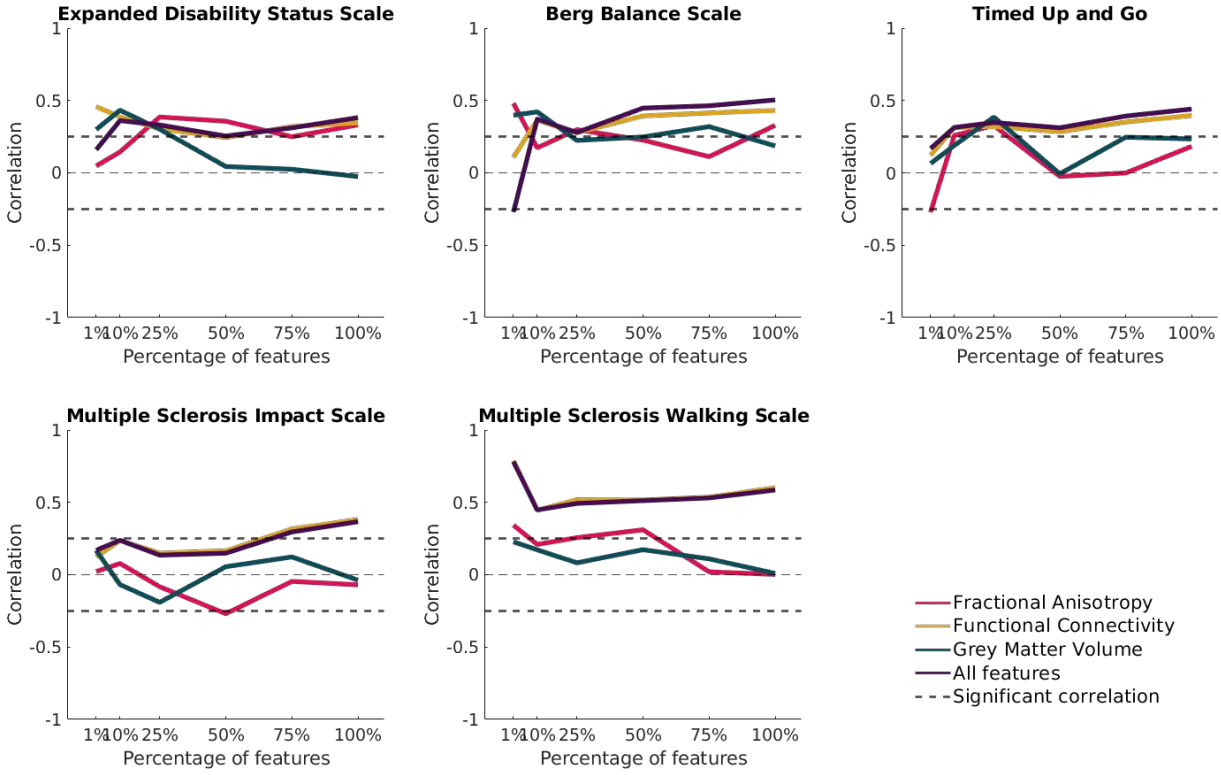

Figure 8.6: **The Spearman correlation of the support vector regression models prediction with the clinical scales.** Models based on fractional anisotropy, functional connectivity, and grey matter volume correspond to the pink, yellow and green lines, respectively; models based on the combination of all features correspond to the violet colour.

Classification

In our study, the model using only the regional average of FA in combination with SVM reached among the highest accuracy yet reported in the literature – 96% (sensitivity = 98%; specificity = 94%). The performance improved with the features added. These results justify SVM use for classification based on similar imaging datasets. However, the accuracy of low-dimensional PCA-LR is also relatively high - a simple three-dimensional model using the first three FA components reached 90%.

Interestingly, the combined models using FA, FC and grey matter volume were not superior to individual modality models. For SVM, this may be caused by the almost perfect accuracy for FA diminishing space for improvement. The lack of accuracy improvement in the low-dimensional scenario is due to the high correlation between the well-performing components of each modality - the three modalities thus did not provide a synergistic performance effect.

We observed similar performance of the approaches, contrasting with the commonly observed dominance of SVM in classification tasks. One might speculate concerning potential data overfitting in practice due to the class-informed choice of features in com-

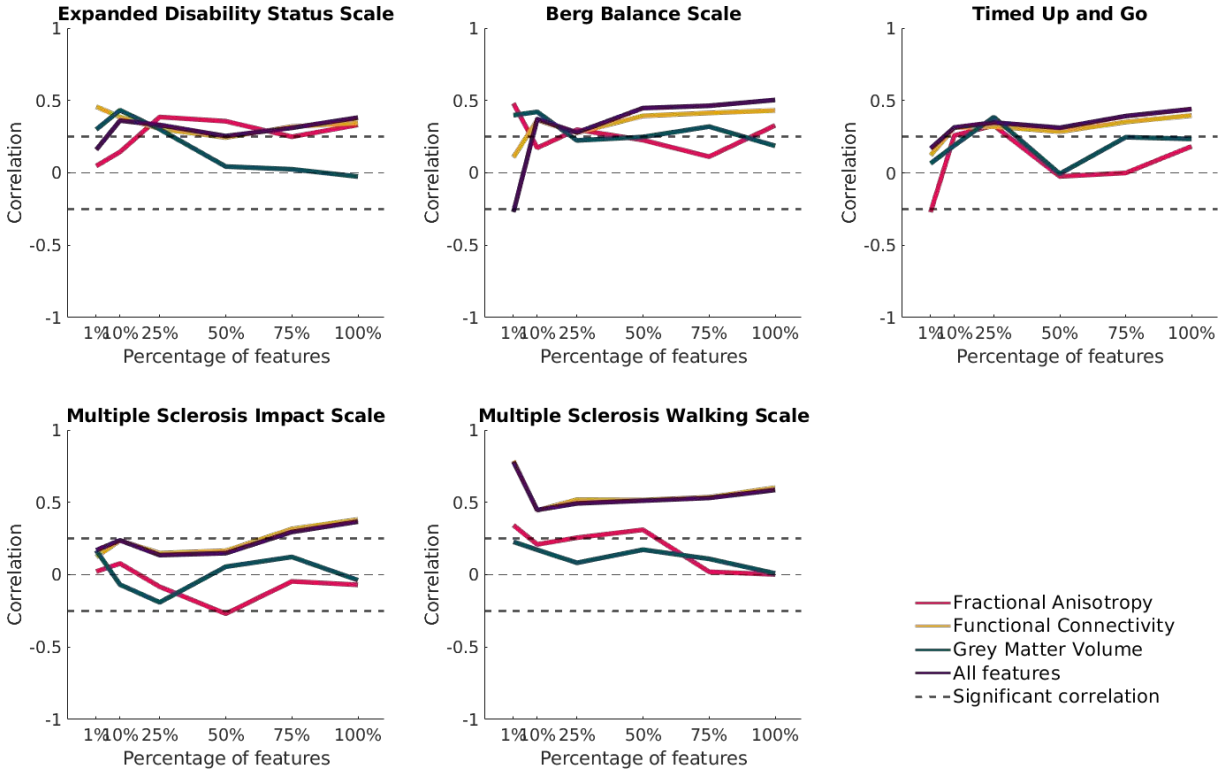

Figure 8.7: **The Spearman correlation of the linear regression models prediction with the clinical scales.** Models based on fractional anisotropy, functional connectivity, and grey matter volume correspond to pink, yellow and green lines, respectively; models based on the combination of all features correspond to the violet colour.

bination with a small data size - something we avoided by careful design of the pipeline. From this perspective, we would advocate the use of PCA as a class-independent data transformation method that is not prone to this type of bias. Moreover, it reduces the model dimensionality, providing more easily interpretable results.

Zurita *et al.* also used the SVM approach in combination with the Fisher score feature selection, reaching 88.9% classification accuracy between patients (EDSS higher than 1.5) and healthy controls, combining structural and functional connectivity. However, this classifier used almost 6,000 features, whereas our PCA-based classification reached 89.9% accuracy using only three. Contrary to our observation, their classification based on the FA was less accurate than using FC. The dominance of our FA-based model may stem from using robust regional averages and TBSS skeleton rather than noisy high-dimensional voxel-wise analysis. Our complementary analysis in which we applied SVM to the FC matrices generated using the Harvard-Oxford Atlas slightly outperformed the AAL atlas results, reaching 80% accuracy using 75% of features; still not matching the values reported [282]; additional investigation with liberal preprocessing only decreased the performance.

Another study [281] used SC matrices to derive graph-theoretical indices. SVM clas-

sification using combined features reached 92% precision comparing 12 subjects with clinically isolated syndrome to 24 healthy controls. While a small sample was analysed, the results line up with other literature exploring the classification potential of SC graph using other classification methods [283], [294].

Neuroimaging correlates of motor disability

We did not identify any study that applied ML analysis to explore the relationship of specialised motor impairment scales and brain imaging, although models have been proposed combining structural, functional and cognitive impairment [295], [296]. Tommasin *et al.* [264] reported that general disability (measured through EDSS) had a direct linear relationship with lesion load, the inverse of thalamic volume, and functional connectivity in bi-frontal region pairs. With a similar approach and atlas for FC quantification, we did not observe correlation with EDSS (or other clinical measures) after appropriate multiple comparison correction. However, the SVR consistently confirmed the significant predictive power of FC, indicating a presence of a more complex, multivariate relationship. Jakimovski *et al.* [270] studied the correlation between walking disability (represented by 25 Foot Walk and the expanded TUG) and selected MRI-derived measures and DWI-measures maps; the DWI measures were not associated with the scales. Using regression analysis, the best predictor for the expanded TUG proved to be the cortex volume, $R^2 = 0.176$, comparable with our results in TUG ($R^2 = 0.153$ with the first grey matter volume component). Steenwijk *et al.* [272] reported partial correlations between different measures of the corticospinal tract and disability. Although corticospinal tract FA was correlated neither with EDSS, nor with MSWS, there was a significant correlation for the cortical thickness of the cortical area connected to it.

Our results indicate that the first principal component of grey matter volume is significantly correlated with numerous clinical scales of motor impairment. The same holds - though to a smaller degree - for the first component of FA. For FC, the first and the second components played a role in regression, with the most noticeable results in the MSWS scale. Overall, the results of PCA-LinR are more reliable than those obtained by the SVR. The prediction quality did not vary significantly with dimensionality, whereas the SVR prediction using the FA and grey matter volume features varied depending on the number of features.

Classification of phenotypes

The study included numerous phenotypes of multiple sclerosis. Acknowledging that there are structural differences between the subtypes [296]–[299], we further investigated the

SVM's ability to separate the relapsing-remitting and the secondary progressive group. The highest accuracy of the FS-SVM classifier was using all the grey matter volume features – 75.9% (sensitivity 64.0%, specificity 81.6%). Considerable successes have been reported by groups which based the classifier on the graph representation of structural connectivity and combined the information with either SVM or graph CNN [281], [283], albeit the insufficient sample size calls for further research.

Limitations

The application of ML methods in neuroscience is generally limited by sample size. We analysed 64 multiple sclerosis patients and 65 healthy controls, which is a considerable size in this area of research. Nevertheless, the group of multiple sclerosis patients was not homogeneous in terms of the phenotype, potentially decreasing the classification power and accuracy of the regression models. Importantly, we took transparent precautions not to overfit the models; in particular, feature selection was carried out in a manner blind to the labels in the testing set. On the other hand, the degree of patients' impairment was, on average, higher than in other studies, which makes this project unique but also warrants caution when interpreting the comparison with other works in the classification task.

However, beyond the numerous advantages, there are limitations associated with ML applications. For SVM, the biggest concern may be associated with the dimensionality of the FC. The number of support vectors rose drastically with the number of features added to the model; nevertheless, the accuracy remained unchanged and inferior to the FA. The same applies to the PCA, where there were substantial differences between the variability explained by the first component among the three modalities. While the first component of grey matter volume explained 71.7%, for FC, it was only 8.6%. Thus, while we can generally recommend using logistic regression combined with prior dimension reduction by PCA, selecting the first component is not necessarily optimal.

Finally, in this work, we focused on the comparison of two specific analytical pipelines, for which we presented the results. However, as is the case in most data-oriented work, it is possible that better performance could be achieved using different algorithms. For example, as suggested by one of the reviewers, methods with regularisation might be applied, reducing the need for dimensionality reduction, although also making it more challenging to control for the number of features included. To explore this direction, we performed logistic regression with the LASSO regularisation known for the sparsity of its solutions; the observed performance is comparable to that logistic regression run on the PCA variables (Table B).

8.6 Conclusion

In this work, we conducted a multimodal analysis of neuroimaging data, investigating tasks of increasing complexity – from classification to prediction. Our approach stemmed from multimodal integration techniques (discussed in Chapter 3), as we aimed to combine three imaging modalities, making it a relatively novel and challenging task.

For each modality, we designed unique features, which were used for the application of multimodal ML models showing that the use of the synergy between the modalities remains a challenge in current settings, both due to potential redundancy between the features and problems with increasing the dimension of the feature set. Nevertheless, we showed that the white matter changes are specific and sensitive enough to provide 96% accuracy in recognising patients from healthy controls. We further proposed an approach to construct low-dimensional classifiers with competitive performance, paving the way for robust and interpretable clinical tools.

Finally, we fitted pr edictive models for the motor disability in multiple sclerosis, suggesting that while the disease itself is most apparent in decreased white matter integrity, the functional motor changes may indeed be more reflected by the accumulated cortical atrophy and changes in functional connectivity. These insights emphasise the significance of considering both structural and functional neuroimaging data in understanding motor disability.

Chapter 9

From neurological to psychiatric

9.1 Disclaimer

This chapter presents ongoing work in collaboration with the National Institute of Mental Health of the Czech Republic and the University of Pennsylvania. I would like to thank my collaborators, who made the data available to me and offered me the resources necessary for appropriate analysis: Filip Španiel, Christos Davatzikos, and Jaroslav Hlinka.

9.2 Introduction

In the preceding chapter, we successfully developed classifiers to differentiate between multiple sclerosis patients and healthy controls. Furthermore, we identified specific neuroimaging features that correlated with the level of motor disability. Nevertheless, we encountered a unique situation in the classification task. Specifically, the dominance of information carried by the DWI modality hindered the efficiency of multimodal integration. Despite this observation, other diseases may benefit more from multimodal analysis.

Neuropsychiatric diseases, such as schizophrenia, lack distinct neuroimaging profiles, thus presenting an opportunity for more promising multimodal fusion. Schizophrenia is a neurodevelopmental disorder with an onset typically occurring between 18 and 25 years of age [300]–[302]. The clinical presentation of the disease varies, with symptoms categorised into positive (e.g., paranoid delusions, auditory hallucinations) and negative (e.g., emotional withdrawal, lack of spontaneity) [303]. The precise aetiology is still unknown, but a combination of genetic and environmental factors is believed to be involved [304]–[308].

Despite significant clinical presentation, schizophrenia lacks definite and unique structural neuroimaging biomarkers, and the brain changes that occur are poorly understood. A meta-analysis of 246 neuroimaging studies focusing on 14 distinct brain structures unveiled significant differences in total brain volume, grey matter volume, particularly in

the frontal, prefrontal, and temporal lobes, and subcortical structures, such as smaller hippocampal volume. Notably, the most prominent effect sizes were associated with ventricular volume [309]. Longitudinal studies further supported the findings of grey matter loss and cortical thinning in children with schizophrenia as well as older patients [310], [311]. To this end, our analysis of longitudinal changes in first-episode schizophrenia aligns with these results, revealing thinner grey matter in patients compared to controls. However, we also observed a normalisation of grey matter thickness within the first year of treatment. For more detailed insights on this topic, refer to the following Chapter 11.

To gain a better understanding of the disease, some studies have explored the heterogeneity of symptoms to identify neuroimaging subtypes, resulting in the discovery of two distinct phenotypes [312]–[314]. However, the relationship between these phenotypes and the clinical profile remains unclear.

A significant breakthrough in schizophrenia research emerged with the adoption of connectomics approaches, which revealed widespread functional connectivity deficits associated with the disease [315]. Some functional connections have even been linked to anatomical changes, although the relationship between brain structure and function remains poorly understood. Furthermore, a systematic review identified abnormalities within and between regions of two functional networks [316], [317]. Nevertheless, similar to the case of structural changes, these functional abnormalities showed no direct links to cognition, leading to the hypothesis that they underlie mechanisms shared across various cognitive functions.

Heterogeneous symptoms of schizophrenia and the lack of neuroimaging signatures make it a perfect candidate for multimodal analysis. Indeed some research was already done to this end, especially in the domain of classification [318]–[323]. The combination of modalities varies widely among studies, with some incorporating task and rs-fMRI [322], while others integrate MEG with rs-fMRI [318], [319], or only combine structural and diffusion data [321]. Additionally, in terms of analytical approaches, the studies employed a broad spectrum, ranging from ridge regression to gradient-boosting random forests. It is essential to acknowledge that the studies significantly differ in terms of participant numbers and inclusion criteria, with only [321] and [323] focusing exclusively on first-episode schizophrenia cases. Nonetheless, a common thread among them is their achievement of final accuracies hovering around 80%.

This work aims to apply the classification methodology developed in the previous chapter to neuroimaging features derived from our dataset of 142 first-episode schizophrenia patients. Through this analysis, we hope to gain further insights into the potential of multimodal fusion in understanding and characterising schizophrenia.

9.3 Materials and methods

Participants

The clinical data used for the analysis were part of the Early Stages of Schizophrenia study [324]. We analysed data from 142 patients in the early stages of schizophrenia (51 females) and 84 controls (52 females). Patients and controls had on average age of 27 and 29 years, respectively. The average length from diagnosis to scanning was, on average, one month (std = 10).

Participants needed to be above 18 years of age and experiencing their initial psychiatric hospitalisation to be included in the study. They were diagnosed with either schizophrenia or acute and transient psychotic disorders and had untreated psychosis for less than 24 months. Upon admission, patients received medical treatment under their physician's guidance. Those with psychotic mood disorders were not part of the study.

Healthy controls over 18 years of age were recruited through advertisements unless: They had a personal history of any psychiatric disorder or had a positive family history of psychotic disorders in first- or second-degree relatives. Individuals from either group (patients or controls) with a history of neurological or cerebrovascular disorders or any MRI contraindications were excluded from the study.

The study was carried out in accordance with the latest version of the Declaration of Helsinki. Its design was reviewed and approved by the Research Ethics Board of the National Institute of Mental Health in Klecany, Czech Republic. Each participant received a complete description of the study and provided written informed consent.

Data acquisition

The data acquisition took place at the National Centre of Mental Health in Klecany, Czech Republic, using a Siemens MAGNETOM Prisma 3 T. Acquisition parameters for **T1-weighted** images using the MPRAGE sequence included 240 scans, with a slice thickness of 0.7 mm, RT of 2,400 ms, ET of 2.34 ms, inversion time of 1,000 ms, flip angle of 8°, and acquisition matrix of 320× 320 mm^2. For **Resting-State fMRI**, BOLD single-shot echo-planar images, the parameters were TR = 2,500 ms, TE = 30 ms, flip angle = 70°, 64 × 64 matrix, FOV = 192 mm^2, 44 contiguous axial slices, each 3 mm thick, 240 volumes, with an acquisition time of 10 minutes. For **DWI**, the parameters were TR = 8,200 ms, TE = 83 ms, flip angle = 90°, matrix = 98 × 98, FOV = 260 × 211.25 mm^2, 57 contiguous axial slices, each 2 mm thick, with b values of 0, 1,100, and 2,500 s/mm, 64 gradient directions, and a voxel size of 2 × 2 × 2 mm^3.

Data processing

The image processing followed procedures similar to those detailed in the preceding Chapter 9, employing the CONN and CAT12 toolboxes for functional and structural data preprocessing, respectively.

In terms of generated features, we used the average of DARTEL-registered structural images across regions of the AAL atlas. For fMRI, an FC matrix was constructed across 90 cortical regions of the AAL atlas. Additionally, the ALFF was computed, which has been noted to show differences in individuals with schizophrenia [58]. The advantage of ALFF lies in its lower dimensionality compared to functional connectivity and its potential to capture distinct information from that contained in FC.

For DWI data, a slightly different preprocessing pipeline was applied. DWI data underwent denoising [201] and reduction of Gibbs ringing artifacts [44]. Susceptibility-induced distortion was estimated by combining DWI data acquired with opposite phase encoding directions [325], [326]. The motion correction, including within-slice movement, eddy current-induced distortion, and outlier replacement, was performed using the `eddy_cuda` program [43]. The methods were a part of FSL [327] and MRtrix3 software [54].

Similarly to the previous chapter, the average FA across the JHU atlas was computed. Additionally, MD was calculated across the AAL atlas. Although the usage and interpretation of diffusivity features across grey matter is currently a topic of discussion, the inclusion of MD aimed to capture sensitivity to early microstructural cortical changes in grey matter, which was previously successfully used for Alzheimer's disease [328].

Analysis

The initial classifier followed the methodology of the preceding chapter, albeit with subtle modifications that took into account the specifics of the current dataset. Our validation process involved k-fold validation with $k = 40$ folds, maintaining a class ratio within each fold. Additionally, for evaluation we used balanced accuracy instead of the conventional accuracy or ROC metric to counteract any class imbalance.

We also introduced an additional integration strategy compared to the multiple sclerosis study. Unlike the previous approach centred on feature integration, this strategy operated at the classifier level. Specifically, we combined the classification scores from all three individual classifiers, summing these scores to derive the final score.

Regarding feature design, the first classifier resembled that of previous section, allowing for the comparative evaluation of its performance across diverse health conditions; however, we also designed a separate classifier containing features tailored to this condition (i.e., MD, FC, and ALFF).

9.4 Results

The initial SVM classifiers using features designed in the preceding chapter yielded the highest balanced accuracy of 78.3% for the voting strategy. This performance surpassed that of the best-performing unimodal classifier, which, in this instance, was based on FC (achieving 75.1% accuracy with the inclusion of the top 10% most discriminative features). Notably, the remaining two types of features – FA and grey matter volume, showed inferior performance and a lack of significant discriminatory capability.

When combining the schizophrenia-specific features, MD displayed a significant discriminatory potential, which was reflected in increased accuracy within the voting strategy, reaching 79.3%. Moreover, the performance of SVM remained relatively consistent regardless of the incremental addition of features.

The combined strategy involving PCA and logistic regression yielded results comparable to the SVM approach. FC maintained its dominance as the most effective feature, although the outcomes were more influenced by the number of incorporated components. It's worth noting that dimensionality reduction did not significantly improve the performance of FA, which remained underwhelming. In contrast, the classifier based 20 components of grey matter volumes achieved a balanced accuracy of 63%, which is reflected in the superiority of the voting approach over FC alone.

Finally, the best-performing classifier configuration was achieved through a voting ensemble composed of MD, FC, and ALFF classifiers, each of 20 PCA components, reaching a balanced accuracy of 80.8%. Interestingly, although the accuracy of FC-based classifiers was lower than in the SVM approach due to the enhanced performance of the remaining features, the voting strategy surpassed the efficacy of each modality.

9.5 Discussion

In this work, we decided to extend our exploration of the multimodal integration approach we previously employed for distinguishing between healthy controls and patients with multiple sclerosis. This time, we focused on a more intricate problem - the classification between healthy controls and individuals with a psychiatric disorder, specifically schizophrenia. This task posed greater complexity due to the fact that our cohort of patients was scanned shortly after their diagnosis, a time when distinct structural changes in the brain might not yet be prominently evident. Nevertheless, considering our aim to design neuroimaging biomarkers, it is essential to confront this task, as this is the very scenario to which such biomarkers would eventually be applied.

As anticipated, the outcomes of our classifiers significantly differed from those obtained

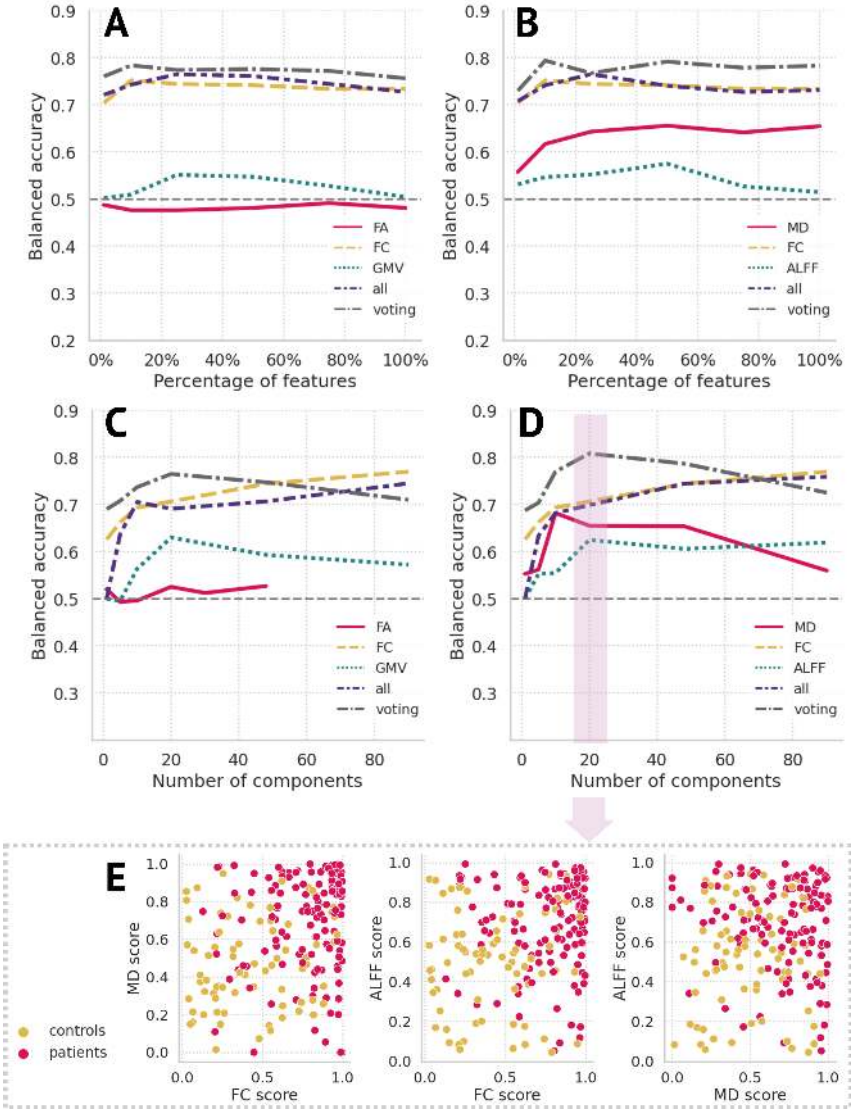

Figure 9.1: **The results of classifiers. A:** The SVM classifiers based on: FA — fractional anisotropy, FC — functional connectivity, GMV — grey matter volume, all — combination of all three, and the voting strategy. **B:** The SVM classifiers based on: MD — mean diffusivity, FC — functional connectivity, ALFF — the amplitude of low-frequency fluctuations, all — combination of all three, and the voting strategy. **C:** The logistic regression classifiers based on: FA — fractional anisotropy, FC — functional connectivity, GMV — grey matter volume, all — combination of all three, and the voting strategy. **D:** The logistic regression classifiers based on: MD — mean diffusivity, FC — functional connectivity, ALFF — the amplitude of low-frequency fluctuations, all — combination of all three, and the voting strategy. **E:** Details of classification scores which enter the voting strategy, depicted for logistic regression classifiers containing 20 PCA components.

in the context of multiple sclerosis. Primarily, fMRI emerged as the modality with the highest discriminatory capacity in schizophrenia classification. This observation aligns with the current narrative of schizophrenia as a disorder primarily rooted in disrupted (functional) connectivity [329]. Especially the rs-fMRI has been shown to be altered

in schizophrenia, with robust reductions also observed in other schizophrenia spectrum disorders [330].

In contrast to prior reports, the ALFF features exhibited a comparatively modest performance, with balanced accuracy hovering around 60%. This outcome may be partially attributed to the fact that the decrease in ALFF is reported mostly regionally and heterogeneously across the brain in the areas of bilateral postcentral gyri, occipital, sensorimotor cortices and others [331]–[333]. Consequently, this heterogeneity of ALFF changes across individuals might hinder its efficacy as a robust classifier. Nevertheless, it still increased the performance of the voting strategy.

Additionally, MD emerged as another efficient feature for classification, achieving an impressive balanced accuracy of almost 70% with just 10 PCA components. While MD has not been commonly used for schizophrenia classification, its extensive usage within Alzheimer's disease research further highlights its potential in differentiating clinical conditions [334]–[336].

Regarding the integration strategies, the inclusion of PCA components in the order of explained percentage of variance showed a trend similar to that observed in the context of multiple sclerosis. Its results largely aligned with the performance of the best-performing individual feature. This suggests that this criterion might not be optimally suited for feature integration. An additional integration technique was also applied, involving a single PCA performed on the entire feature set, followed by selecting the first n-components for classification. However, due to the highly comparable outcomes to the initial integration approach, these results have been omitted for the sake of conciseness.

Surprisingly, the voting strategy achieved remarkable results and holds potential for further enhancement by combining classifiers with varying numbers of features to participate in the voting process. For clarity, the voting here was exclusively based on classifiers with the same quantity of features. However, Figure 9.1D suggests that certain classifiers, such as the MD classifier with ten features, could potentially benefit the strategy.

To gain insight into the enhanced performance of the voting strategy, a scatter plot of scores from individual classifiers is presented in Figure 9.1E. Across various feature combinations, we observe mild to moderate separability coupled with some degree of correlation. The correlation coefficient between FC and MD scores is 0.29, while FC and ALFF exhibit a correlation of 0.27. Remarkably, MD and ALFF scores are not correlated ($R = 0.04$). This indicates that each classifier captures distinct information (to some extent), implying their potential to complement one another within a voting strategy.

Arguably, the methods we used for feature integration leveraged simplicity and more complex algorithms could be able to disentangle specifics of patients and controls more precisely. To check this, we used the `AutoGluon-Tabular` python package, which allows

for the construction of a diverse set of deep learning models [337]. We ran the default `Tabular` pipeline, automatically constructing 14 models. Balanced accuracy across 40 folds, where 20% of the data was used for testing, is shown in Figure 9.2. No algorithm outperformed our approach, although more tuning may be needed to fully exploit the potential of deep learning models.

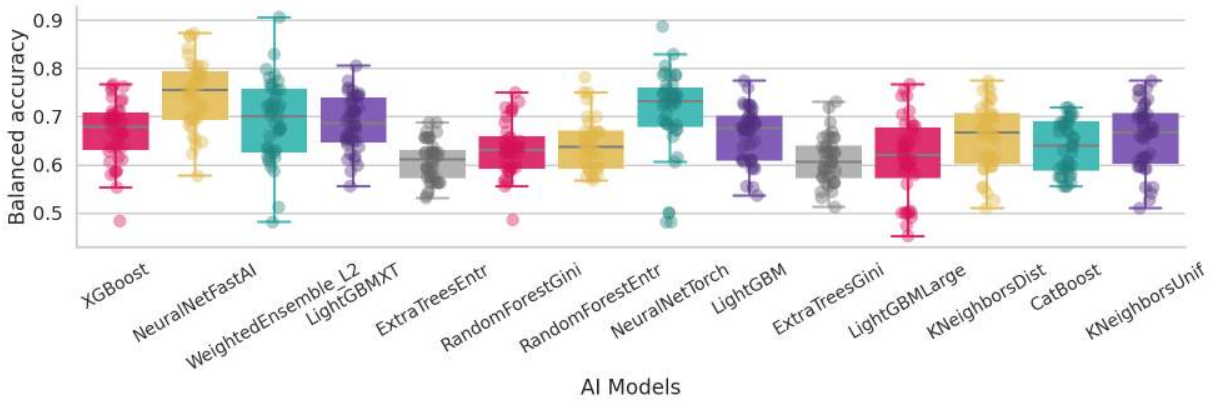

Figure 9.2: **AI models constructed using mean diffusivity, functional connectivity and amplitude of low-frequency fluctuations features.**

The outcomes of the most successful classifier, namely the voting strategy integrating three classifiers, each comprising 20 PCA components corresponding to their respective modalities, are comparable to the current state of the art. For instance, Zhuang *et al.* combined structural, functional, and diffusion data through SVM, yielding an accuracy of 82% [323]. This study aligns closely with ours regarding dataset age, which averaged 27 years. In another study, Liang *et al.* also investigated the data of first-episode schizophrenia patients, combining grey matter features, FA, and MD (in white matter) using gradient-boosting trees, reporting accuracy of 75% [321]. A more intricate approach by Hu *et al.* on an older cohort entailed a naive 3D convolutional network utilising structural and diffusion images, achieving an accuracy of 81% [338]. While making direct comparisons between studies can be complex due to cohort variations, particularly in terms of the duration since diagnosis, these findings underscore the ongoing challenge of identifying reliable biomarkers for schizophrenia.

As mentioned at the beginning, this project is an ongoing work, and further exploration is required before drawing definitive conclusions about the feasibility of classifying first-episode schizophrenic patients. We might identify other features with significant discriminatory potential; for instance, surface features could be more sensitive to early grey matter changes than the volumetric features we used. Alternatively, SC might prove more sensitive to white matter alterations compared to the FA. In addition, we should also explore pure fusion methods. However, their drawback is the focus on enhancing the signal-to-noise ratio by designing features that correlate across modalities. While effec-

tive, this approach overlooks the unique information embedded in specific features, which could be valuable for classification, as shown by the minimal correlation between ALFF and MD scores.

Finally, the natural progression of the classification task is the domain of prediction. In forthcoming investigations, we intend to develop a predictive algorithm aimed at forecasting the future cognitive status of patients based on their neuroimaging signatures.

9.6 Conclusion

In summary, this study extended the multimodal integration approach from the previous chapter to the more intricate task of early-stage schizophrenia classification. Our findings revealed that features based on fMRI discriminate the best, with FC being more effective than ALFF. Additionally, MD in grey matter showed considerable potential, though its application in this context has so far been limited. Furthermore, despite the simplicity of our integration strategies, the voting approach achieved promising results comparable with state-of-the-art classifiers.

Chapter 10

Intermezzo – more data!

One of the major challenges of incorporating AI into neuroimaging research is the scarcity of large-scale datasets required for training complex AI models. The high dimensionality of neuroimaging data necessitates a substantial amount of data to effectively train AI algorithms. Unfortunately, no single institution possesses a sufficiently diverse and high-quality dataset to construct such models independently. Therefore, data sharing has become indispensable. However, the field has faced reluctance to embrace data sharing for various reasons, including the financial burden associated with data acquisition and unresolved concerns regarding subject anonymisation.

Fortunately, recently, there has been a recognition of the importance of data sharing, leading to the emergence of consortia and databases aimed at fostering collaborations and facilitating data sharing across multiple scanning sites. Some databases are freely accessible, while others require written consent, active collaboration, or payment. Moreover, it is increasingly considered good practice to publish accompanying code and provide metadata (such as neuroimaging features) along with research publications.

Nevertheless, one might argue that due to the abovementioned challenges, researchers may be hesitant to share raw neuroimaging data that others could potentially use. To address this concern, academic journals have introduced a new type of publication known as "data papers" or "data reports" to incentivise researchers to publish articles alongside their raw data. *Nature* itself introduced a new brand journal into the family in 2016 titled Scientific Data, devoted solely to this purpose (although the topic is not specific to neuroimaging). The articles typically include discussions on the data acquisition process, recognition of previous work conducted with the data, and presentation of initial statistical analyses. Additionally, researchers may provide data at various preprocessing stages, ranging from raw to final extracted features. Thus, any further use of the data is properly cited, with the data owners having well-deserved recognition and researchers new data source.

We firmly believe in the importance of publishing data and accompanying code, which is why, in addition to our research endeavours, we have also published data papers and made second-level features for most of our works available. Here, we briefly describe two of our data papers as examples.

10.1 The impact of physical therapy in multiple sclerosis

Our first data paper titled: "Open Access: The Effect of Neurorehabilitation on Multiple Sclerosis—Unlocking the Resting-State fMRI Data" by Bučková *et al.* was published in 2021 in Frontiers in Neuroscience [339].

This article was complementary to our previous work on multiple sclerosis, which focused on cross-sectional classification and prediction results and was described in Chapter 8. However, that particular research branch originated from a larger project aimed at examining the effects of rehabilitation on multiple sclerosis, specifically on the neuroimaging features of individuals with the disease. While the clinical aspects and observed changes were addressed in a separate article [340], we noticed a significant scarcity of independent datasets for validating our findings. As a result, we decided to publish the raw fMRI time series of 60 individuals with multiple sclerosis, acquired before and after two months of neuroproprioceptive "facilitation and inhibition" treatment.

10.2 Ready-to-use structural connectivity data

Our second data paper was titled: "Human brain structural connectivity matrices–ready for modelling" by Škoch *et al.* and was published in 2022 in Scientific Data [341].

In contrast to the previous article, this paper emphasised the significance of second-level features rather than the raw data itself. Many researchers who focus on mathematical modelling but lack experience analysing neuroimaging data need prior information to model brain connectivity effectively [342]. It is crucial for these researchers to have access to such information to accurately represent the measured data in their models. Therefore, we decided to publish structural connectivity matrices of 88 healthy subjects alongside the raw data and a functioning pipeline for reproducibility purposes.

Chapter 11

Back to the future – longitudinal normative modelling

11.1 Disclaimer

An original version of this work is currently under review (August 2023) in IEEE Transactions on Medical Imaging, under the title: Using normative models pre-trained on cross-sectional data to evaluate longitudinal changes in neuroimaging data. It was also published on bioRxiv prior to submission [343]. I would like to express my gratitude to my coauthors: Charlotte Fraza, Rastislav Rehák, Marián Kolenič, Christian Beckmann, Filip Španiel, Andre Marquand, and Jaroslav Hlinka.

11.2 Introduction

The issue of inter-individual differences is one of the cornerstones of this thesis. It signifies the shift from pure statistics to machine learning, from global averages to subject-level prediction. And indeed, so far, the level of an individual was mostly addressed in the final analytical stage – the classification or prediction algorithm. However, as outlined in the final theoretical Chapter 5, which discussed the confounders, individual variability shouldn't only be considered on the level of analysis because their unique factors come to play sooner when controlling for variables like age, sex, scanning site and others. To this end, we introduced normative modelling as a promising technique and highlighted the advantages of pre-trained normative models.

This work dealt with the analysis of longitudinal datasets and aimed to accurately determine the significance of temporal changes observed in the brain. However, we noticed a lack of analytical methods tailored to this specific challenge. Consequently, we aimed to extend the applicability of pre-trained normative models to longitudinal data, offering a

novel approach to address inter-individual differences and reliably estimate the temporal dynamics of brain changes.

Longitudinal neuroimaging studies provide a unique opportunity to gain insight into the temporal dynamics of a disease, over and above the insights offered by cross-sectional studies. Efforts to acquire these datasets are non-trivial and require substantial time and funding while accounting for problems inherent to longitudinal studies, such as the dropout of subjects between time points, standardisation over multiple clinical centres, and changes in imaging technology over time. Considering this challenging task, it is of great consequence to have tools to analyse them effectively whilst also using more widely available cross-sectional data to refine inferences.

Despite the importance of having suitable tools for longitudinal data analysis, the development of appropriate methods is relatively scarce. Even though recent advances in the acquisition and publication of large neuroimaging datasets [344], [345] have significantly improved our understanding of population variation, the developed methodologies are largely focused on the cross-sectional nature of the data [346]. Indeed, these methods are essential for characterising individual differences; however, the vital factor of longitudinal change is largely neglected. Consequently, efforts should be made to tailor the standing methods modelling population variation for a longitudinal context to advance our understanding of disease progression, which is currently modelled mostly by mixed-effects models [347].

Longitudinal studies are conceptually well suited for normative modelling since they analyse individual trajectories over time. If adjusted appropriately, normative models could improve predictive accuracy and identify patterns of change, thereby enhancing our understanding of the disease. In contrast to traditional statistical methods that estimate the average change in the group, normative models could estimate individual deviations from healthy trajectories. This comprehensive approach would take into account both population heterogeneity and confounders, thus providing a more nuanced understanding of change over time.

Normative modelling is a relatively new area of research, and thus, despite its potential, longitudinal normative models have not been extensively explored [125], [348]. Indeed, virtually all large-scale normative models released to date are estimated on cross-sectional data [125], [127] and a recent report [348] has provided empirical data to suggest that such cross-sectional models may underestimate the variance in longitudinal data [348]. However, from a theoretical perspective, it is very important to recognise that cross-sectional models describe group-level population variation across the lifespan, where such group-level centiles are interpolated smoothly across time. It is well-known in the pediatric growth-charting literature (e.g., [349]) that centiles in such do not necessarily correspond

to individual level trajectories, rather it is possible that individuals cross multiple centiles as they proceed through development, even in the absence of pathology. Crucially, classical growth charts and current normative brain charts provide no information about how frequent such centile crossings are in general. In other words, they provide a *trajectory of distributions*, **not** a *distribution over trajectories*. There are different approaches to tackle this problem in the growth charting literature, including the estimation of "thrive lines" that map centiles of constant velocity across the lifespan and can be used to declare "failure to thrive" at the individual level see (e.g., [349]) for details. Unfortunately, this approach requires densely sampled longitudinal neuroimaging data to estimate growth velocity, which are currently unavailable across the human lifespan. Therefore, in this work, we adopt a different approach based on estimates of the uncertainty in the centile estimates themselves together with the uncertainty with which a point is measured (e.g., bounded by the test-retest reliability, noise etc.). By accounting for such variability, this provides a statistic to determine whether a centile crossing is large enough to be statistically different from the base level within the population.

We stress that our aim is not to build a longitudinal normative model *per se*. Considering the much greater availability of cross-sectional data relative to longitudinal data, we, instead, leverage existing models constructed from densely sampled cross-sectional populations and provide methods for applying these to longitudinal cohorts. We argue that although these models lack explicit intra-subject dynamics, they contain sufficient information to enable precise assessments of changes over time. Nevertheless, including longitudinal data into existing models largely estimated from cross-sectional data is also an important goal and can be approached with hierarchical models [350]; however, we do not tackle this problem here.

Our approach requires (i) a probabilistic framework to coherently manage uncertainty and (ii) cross-sectional models estimated on large reference cohort to accurately estimate population centiles. To this end, we utilise the Warped Bayesian linear regression normative model [124] as a basis for our work. Training these models requires significant amounts of data and computational resources, limiting their use for smaller research groups. However, the availability of pre-trained models has made them more accessible to researchers from a wider range of backgrounds, as reported by Rutherford et al. [127].

In summary, we propose a framework for using pre-trained normative models to evaluate longitudinal studies. We briefly present the existing model and derive a novel set of difference ("*z-diff*") scores for statistical evaluation of change between measurements. We then describe its implementation and showcase its practical application to an in-house longitudinal dataset of 98 patients in the early stages of schizophrenia who underwent MRI examinations shortly after being diagnosed and one year after.

11.3 Materials and methods

Model formulation

Original model for cross-sectional data

The original model [124] we use for pre-training is developed for a cross-sectional database $\mathbf{Y} = (y_{nd}) \in \mathbb{R}^{N \times D}$ and $\mathbf{X} = (x_{nm}) \in \mathbb{R}^{N \times M}$, where y_{nd} is the d-th Image-Derived Phenotype (IDP) of the n-th subject and x_{nm} is the m-th covariate (e.g., age or sex) of the n-th subject.

The IDPs are treated separately, so we focus on a single IDP d. To simplify notation, this IDP-specific index is dropped, and we denote $\mathbf{y} = (y^{(1)}, \ldots, y_N)^T$. The observations are assumed to be independent (across n). We model the distribution of $\varphi(y_n)$ conditional on exogenous variables $\mathbf{x}_n = (x_{n1}, \ldots, x_{nM})^T$, vector of parameters $\mathbf{w}$, and a hyper-parameter β as

$$\varphi(y_n)|\mathbf{x}_n; \mathbf{w}; \beta \sim \mathcal{N}(\mathbf{w}^T \phi(\mathbf{x}_n), \beta^{-1}), \tag{11.1}$$

where:

- $\varphi(y_n)$ is the original variable y_n transformed by a warping function φ, which is parametrised by hyper-parameters $\boldsymbol{\gamma}$ (e.g., composition of SinhArcsinh warping functions, which performed well in prior work [124]); this transformation is used to accommodate non-Gaussian errors in the original space of dependent variables;

- $\phi(\mathbf{x}_n) \in \mathbb{R}^K$ is a common B-spline basis expansion of the original independent variables $\mathbf{x}_n$ (specifically, cubic splines with three evenly spaced knot points [124]) to capture non-linear relationships, appended with site dummies to accommodate site level effects;

- β is a precision of measurements treated as a hyper-parameter (the framework in [124] also allows for site-specific precision terms, which we don't use here).

We write this as

$$\varphi(y_n) = \mathbf{w}^T \phi(\mathbf{x}_n) + \varepsilon_n, \ \varepsilon_n \sim \mathcal{N}(0, \beta^{-1}), \tag{11.2}$$

where ε_n are independent from $\mathbf{x}_n$ and across n. We further denote $\boldsymbol{\Lambda}_\beta = \beta \mathbf{I} \in \mathbb{R}^{N \times N}$ and the design matrix $\boldsymbol{\Phi} = (\phi(\mathbf{x}_n)_k) \in \mathbb{R}^{N \times K}$.

The estimation of parameters $\mathbf{w}$ is performed by empirical Bayesian methods. In particular, prior about $\mathbf{w}$

$$\mathcal{N}(0, \boldsymbol{\Lambda}_\alpha^{-1}), \ \boldsymbol{\Lambda}_\alpha = \alpha \mathbf{I} \tag{11.3}$$

is combined with the likelihood function to derive the posterior

$$\mathbf{w}|\mathbf{y}, \mathbf{\Phi}; \alpha, \beta, \boldsymbol{\gamma} \sim \mathcal{N}(\bar{\mathbf{w}}, \mathbf{A}^{-1}), \tag{11.4}$$

$$\mathbf{A} = \mathbf{\Phi}^T \mathbf{\Lambda}_\beta \mathbf{\Phi} + \mathbf{\Lambda}_\alpha, \tag{11.5}$$

$$\bar{\mathbf{w}} = \mathbf{A}^{-1} \mathbf{\Phi}^T \mathbf{\Lambda}_\beta \mathbf{y}. \tag{11.6}$$

The hyper-parameters $\alpha, \beta, \boldsymbol{\gamma}$ are estimated by maximising the warped marginal log-likelihood.

The predictive distribution of $\varphi(y)$ for a subject with $\mathbf{x}$ is

$$\mathcal{N}(\bar{\mathbf{w}}^T \phi(\mathbf{x}), \phi(\mathbf{x})^T \mathbf{A}^{-1} \phi(\mathbf{x}) + \beta^{-1}). \tag{11.7}$$

Finally, the z-score characterising the position of this subject within a population is

$$z = \frac{\varphi(y) - \bar{\mathbf{w}}^T \phi(\mathbf{x})}{\sqrt{\phi(\mathbf{x})^T \mathbf{A}^{-1} \phi(\mathbf{x}) + \beta^{-1}}}, \tag{11.8}$$

where $\varphi(y)$ is the realized warped observation of IDP d for this subject.

Note that formulae (11.7) and (11.8) implicitly evaluate only (potentially new) subjects measured at sites already present in the original database $\mathbf{y}$, $\mathbf{\Phi}$. If we want to evaluate subjects measured at a new site, we will have to run an adaptation procedure to account for the effect of the new site. This adaptation procedure is described elsewhere and is readily accessible for use online (see [127]). In short, a sample of a reference (healthy) cohort measured on the same scanner as the population of interest is needed to accommodate a site-specific effect. Using the reference cohort, the parameters of the original pre-trained model are adapted and only then the sample of interest is analysed (Figure 11.1).

In the following section, we develop a procedure that allows us to extend the original cross-sectional framework pre-trained on database $\mathbf{y}$, $\mathbf{\Phi}$ to a new longitudinal dataset for evaluation of changes in regional brain thickness.

Adaptation to longitudinal data

We aim to adapt the original model [124], leveraging its pre-trained parameters, to design a z-score (for clarity further referred to as *z-diff* score) for a change between visits, based on which we could detect large changes in regional brain thickness.

To utilise the normative model of the healthy population pre-trained on cross-sectional data in the longitudinal setup, we have to make an assumption about the trajectory of healthy controls. The natural first approximate assumption is that a healthy subject does not deviate substantially from their position within the population as time progresses, so

the observed position changes between the visits of a healthy control subject are assumed to stem from observation noise (due to technical or physiological factors) and are therefore constrained by the test-retest reliability of the measurement. Note that this does not imply that a healthy subject does not change over time, but rather change that occurs follows approximately the centile of distribution at which the individual is placed. We acknowledge that this is a relatively strong assumption, but it is reasonable to assume that a subject's brain activity or structure will remain relatively stable over short time periods (years, but not necessarily decades), and any significant changes in the pattern compared to the normative reference database may indicate disease progression or response to intervention [123]. Also, it is very important to recognise that this model does not constrain a given subject to follow a population-level centile trajectory exactly because the model includes an error component ($\xi^{(i)}$) that allows for individual level deviations from the population centile.

We connect to the cross-sectional model by invoking the above assumption through reinterpretation of the error term $\varepsilon \sim \mathcal{N}(0, \beta^{-1})$. In particular, we decompose it to a subject-specific factor η and a measurement error $\xi^{(i)}$ in the i-th visit. While η is considered constant across the visits (of a healthy cohort subject) and captures the subject's position within the population, $\xi^{(i)}$ is specific for the i-th visit and captures a combination of movement, acquisition, processing noise, etc. Hence, the model for the i-th visit of a subject with given covariates $\mathbf{x}^{(i)}$ is

$$\begin{aligned} \varphi(y^{(i)}) &= \mathbf{w}^T \phi(\mathbf{x}^{(i)}) + \eta + \xi^{(i)} \\ \eta &\sim \mathcal{N}(0, \sigma_\eta^2) \\ \xi^{(i)} &\sim \mathcal{N}(0, \sigma_\xi^2) \\ \beta^{-1} &= \sigma_\eta^2 + \sigma_\xi^2 \end{aligned} \tag{11.9}$$

where η, $\xi^{(i)}$, and $\mathbf{x}^{(i)}$ are mutually independent for a given i, and $\xi^{(i)}$ and $\xi^{(j)}$ are independent for $i \neq j$. Note that we dropped the subject-specific index n introduced in (11.1). This change in notation should force the reader to distinguish between data used for training the original pre-trained model and a new set of longitudinal data that is going to be used for estimating the longitudinal change.

In our longitudinal data, we are interested in the change for a given individual across two visits. According to model (11.9), the difference in $\varphi(y)$ between visits 1 and 2, $\varphi(y^{(2)}) - \varphi(y^{(1)})$, for a subject with covariates $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$ is given by

$$\varphi(y^{(2)}) - \varphi(y^{(1)}) = \mathbf{w}^T[\phi(\mathbf{x}^{(2)}) - \phi(\mathbf{x}^{(1)})] + \xi^{(2)} - \xi^{(1)} \tag{11.10}$$

with $\xi^{(2)} - \xi^{(1)} \sim \mathcal{N}(0, 2\sigma_\xi^2)$. We use the posterior distribution of $\mathbf{w}$ (and hyperparameters $\alpha, \beta, \boldsymbol{\gamma}$) estimated on the original cross-sectional database $\mathbf{y}, \boldsymbol{\Phi}$ [124]. Therefore, the posterior predictive distribution for the difference $\varphi(y^{(2)}) - \varphi(y^{(1)})$ for our subject is (for more detailed derivation, please refer to the supplement C)

$$\mathcal{N}\left(\bar{\mathbf{w}}^T[\phi(\mathbf{x}^{(2)}) - \phi(\mathbf{x}^{(1)})], [\phi(\mathbf{x}^{(2)}) - \phi(\mathbf{x}^{(1)})]^T \mathbf{A}^{-1}[\phi(\mathbf{x}^{(2)}) - \phi(\mathbf{x}^{(1)})] + 2\sigma_\xi^2\right). \quad (11.11)$$

Finally, the z-score for the difference in $\varphi(y)$ between visits 1 and 2 is

$$z_{diff} = \frac{[\varphi(y^{(2)}) - \varphi(y^{(1)})] - \bar{\mathbf{w}}^T[\phi(\mathbf{x}^{(2)}) - \phi(\mathbf{x}^{(1)})]}{\sqrt{[\phi(\mathbf{x}^{(2)}) - \phi(\mathbf{x}^{(1)})]^T \mathbf{A}^{-1}[\phi(\mathbf{x}^{(2)}) - \phi(\mathbf{x}^{(1)})] + 2\sigma_\xi^2}}, \quad (11.12)$$

where $\varphi(y^{(2)}) - \varphi(y^{(1)})$ is the realized change in the warped observations of IDP d for this subject. Since this z_{diff} score is standard normal for the population of healthy controls, any large deviations may be used to detect suspicious changes.

The primary role of adaptation of the pre-trained cross-sectional model to longitudinal data is to account for the measurement noise variance σ_ξ^2. From the posterior predictive distribution (11.11), we have

$$\begin{aligned}&\mathrm{E}\Big[\left(\varphi(y^{(2)}) - \varphi(y^{(1)}) - \mathrm{E}[\varphi(y^{(2)}) - \varphi(y^{(1)})|\mathbf{x}^{(1)}, \mathbf{x}^{(2)}; \mathbf{y}, \boldsymbol{\Phi}; \alpha, \beta, \boldsymbol{\gamma}]\right)^2 \Big| \mathbf{x}^{(1)}, \mathbf{x}^{(2)}; \mathbf{y}, \boldsymbol{\Phi}; \alpha, \beta, \boldsymbol{\gamma}\Big] \\ &\quad = [\phi(\mathbf{x}^{(2)}) - \phi(\mathbf{x}^{(1)})]^T \mathbf{A}^{-1}[\phi(\mathbf{x}^{(2)}) - \phi(\mathbf{x}^{(1)})] + 2\sigma_\xi^2. \quad (11.13)\end{aligned}$$

Hence, by the Law of Iterated Expectations (to integrate out $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$), we obtain

$$\begin{aligned}&\mathrm{E}\Big[\left(\varphi(y^{(2)}) - \varphi(y^{(1)}) - \bar{\mathbf{w}}^T[\phi(\mathbf{x}^{(2)}) - \phi(\mathbf{x}^{(1)})]\right)^2 \\ &\quad - [\phi(\mathbf{x}^{(2)}) - \phi(\mathbf{x}^{(1)})]^T \mathbf{A}^{-1}[\phi(\mathbf{x}^{(2)}) - \phi(\mathbf{x}^{(1)})] \Big| \mathbf{y}, \boldsymbol{\Phi}; \alpha, \beta, \boldsymbol{\gamma}\Big] = 2\sigma_\xi^2. \quad (11.14)\end{aligned}$$

Therefore, we estimate $2\sigma_\xi^2$ by the sample analogue of the left-hand side in (11.14). Specifically, we devote a subsample C of our controls to adaptation, and we compute

$$\begin{aligned}\widehat{2\sigma_\xi^2} = \frac{1}{|C|} \sum_{k \in C} \Big[&\Big(\varphi(y_k^{(2)}) - \varphi(y_k^{(1)}) - \bar{\mathbf{w}}^T[\phi(\mathbf{x}_k^{(2)}) - \phi(\mathbf{x}_k^{(1)})]\Big)^2 \\ &- [\phi(\mathbf{x}_k^{(2)}) - \phi(\mathbf{x}_k^{(1)})]^T \mathbf{A}^{-1}[\phi(\mathbf{x}_k^{(2)}) - \phi(\mathbf{x}_k^{(1)})]\Big]. \quad (11.15)\end{aligned}$$

Moreover, another useful feature of longitudinal data is that $[\phi(\mathbf{x}_k^{(2)}) - \phi(\mathbf{x}_k^{(1)})]$ is negligible (especially with stable covariates, like sex and age). Sex (typically) does not change across the two visits, and age relatively little (in our target application) with respect to the full span of ageing. Consequently, $[\phi(\mathbf{x}_k^{(2)}) - \phi(\mathbf{x}_k^{(1)})]^T \mathbf{A}^{-1}[\phi(\mathbf{x}_k^{(2)}) - \phi(\mathbf{x}_k^{(1)})]$

in (11.15) is negligible in adult cohorts but must be treated with caution in developmental or ageing groups. Finally, it is apparent from (11.5) that $\mathbf{A}$ scales with a number of subjects and its inverse will be negligible for substantial training datasets, such as the one that was used for pre-training.

Implementation

To implement the method (Fig. 11.1), we used the **PCN toolkit**. Running the adaptation pipeline [127] returns all intermediate results, which is convenient for analysis.

Participants

Early stages of schizophrenia data

The clinical data used for the analysis were part of the Early Stages of Schizophrenia study [324]. We analysed data from 98 patients in the early stages of schizophrenia (38 females) and 67 controls (42 females) (Table 11.1). For the details on the inclusion criteria, ethical approval and acquisition parameters, please, refer to Chapter 9.

Preprocessing and analysis

All T1 images were preprocessed using the Freesurfer v.(7.2) `recon-all` pipeline. Preprocessing was performed twice, in cross-sectional and longitudinal [351] settings, to evaluate the effect of preprocessing. In line with [127], we performed a simple quality control procedure whereby we excluded all subjects having a rescaled Euler number greater than ten were labelled outliers and were not included in the analysis (Table 11.1) (see [127] and [350] for further details). The pre-trained model used for adaptation was the `lifespan_58K_82_sites` [127].

We evaluated the effect of preprocessing by running normative models for both visits and both types of preprocessing. We then tested for the difference in the variance of the difference of the cross-sectional z-scores $z^{(2)} - z^{(1)}$ in held-out controls.

After preprocessing, patient data were projected into the adapted normative model (median Rho across all IDP was 0.3 and 0.26 for the first and the second visit, respectively – see Fig. C.1). For each subject and visit, we obtained cross-sectional z-score, as well as the underlying values needed for its computation, particularly $\phi(y)$, and $\bar{\mathbf{w}}^T\varphi(\mathbf{x})$. We conducted a cross-sectional analysis of the original z-scores to evaluate each measurement independently. We used the Wilcoxon test to determine whether the z-scores of the patients were significantly different from zero and corrected for multiple tests using the Benjamini-Hochberg FDR correction at the 5% level of significance.

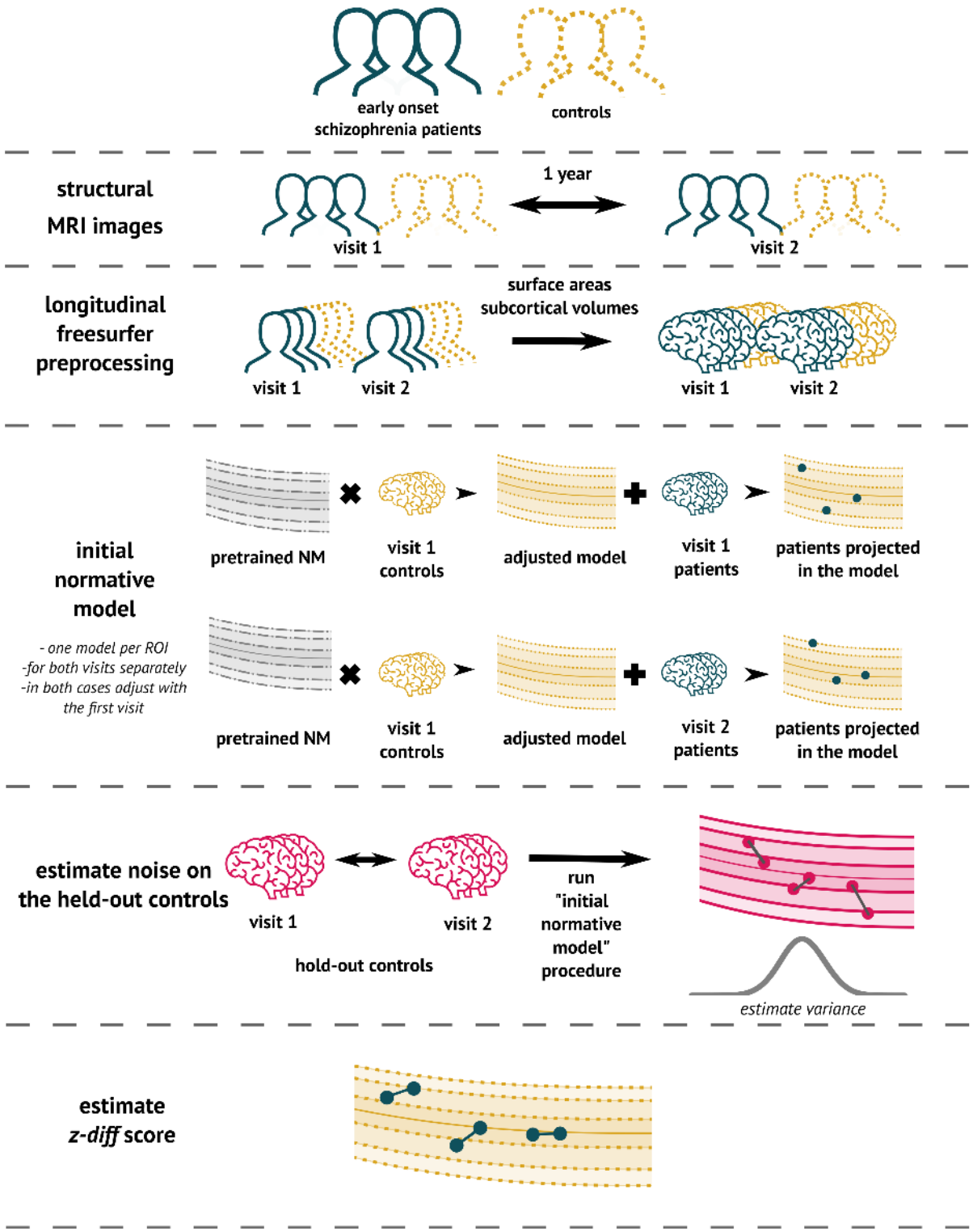

Figure 11.1: **The overview of the analytical pipeline for our schizophrenia patients.** First, data are preprocessed using Freesurfer's longitudinal pipeline. Subsequently, the pre-trained models are adjusted to a local sample of healthy controls. The site-specific measurement noise variance $\sigma^2_{\xi_s}$ in healthy subjects is estimated using held-out controls, and finally, the *z-diff* score is computed.

Table 11.1: **Description of the dataset.**

	Patients	**Controls**
N (% females)	98 (39%)	67 (63%)
Age, median (min, max), years	27 (18, 46)	29 (18, 54)
Interval between visits, median (min, max), years	1.1 (0.9,2.7)	1.2 (0.9, 3)
Diagnosis (only for patients)		
Schizophrenia	53	
Brief psychotic disorder	45	
Length of disease, median (min, max), months	4 (1,21)	
Clinical scales (only for patients)	**Visit 1**	**Visit 2**
PANSS sum, median (min, max)	53 (30, 94)	44 (30, 84)
PANSS Positive Symptoms, median (min, max)	11 (7, 21)	8 (7, 26)
PANSS Negative Symptoms, median (min, max)	14.5 (7, 30)	11.5 (7, 24)
GAF, median (min, max)	70 (25, 100)	80.5 (40, 98)

Subsequently, following (11.12), we derived the *z-diff* scores of change between visits. We conducted two analyses, one to investigate the group-level effect and another to link the *z-diff* to the changes in clinical scales.

At a group level, we identified regions with *z-diff* scores significantly different from zero using the Wilcoxon test, accounting for multiple comparisons using the Benjamini-Hochberg FDR correction. Additionally, we performed a more traditional longitudinal analysis. As all visits were approximately one-year apart, we conducted an analysis of covariance (ANCOVA). The ANCOVA model combines GLM and ANOVA. Its purpose is to examine whether the means of a dependent variable (thickness in V2) are consistent across levels of a categorical independent variable (patients or controls) while accounting for the influences of other variables (age, gender, and thickness in V1). We conducted a separate test for each IDP and controlled the relevant p-values across tests using the FDR correction.

For linking the *z-diff* score to clinical change, we transformed the *z-diff* score across all IDPs using PCA to decrease the dimensionality of the data as well as to avoid fishing. We ran PCA with 10 components and, using Spearman correlation, related the scores with changes in the Positive and Negative Syndrome Scale (PANSS) and Global Assessment of Functioning (GAF) scale.

11.4 Results

Effect of preprocessing

After running the normative models in both visits, we visually observed a decrease in variance between the two visits in longitudinal preprocessing compared to cross-sectional (Figure 11.2). More specifically, we calculated the mean of the difference between z-scores of V^2 and V^1 for each individual IDP, stratified by preprocessing and group, across all subjects. We then visualised the distribution of these means using a histogram (Figure 11.2C). Alternatively, we also computed the mean difference between z-scores of V^2 and V^1 across all IDPs for each subject and plotted a histogram of these values. Based on these results, we continued the analysis using longitudinal preprocessing for all data, although we also compared our results to the results derived from cross-sectional preprocessing.

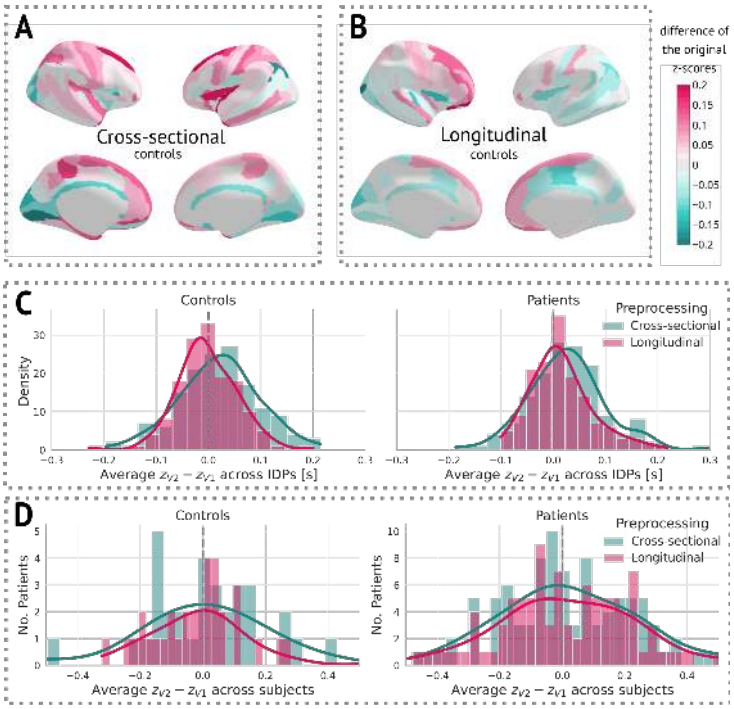

Figure 11.2: **The effect of preprocessing across all subjects and IDPs. A:** Cross-sectional preprocessing: Heatmap of the difference of the original z-scores $(z^{(2)} - z^{(1)})$ on held-out controls. **B:** Longitudinal preprocessing: Heatmap of the difference of the original z-scores $(z^{(2)} - z^{(1)})$ on held-out controls. **C:** Histogram of the average $(z^{(2)} - z^{(1)})$ across all IDPs stratified by health status and preprocessing. **D:** Histogram of the average $(z^{(2)} - z^{(1)})$ of each subject stratified by health status and preprocessing.

Cross-sectional results

At a group level, patients had significantly lower thicknesses in most areas compared to healthy populations. In particular, this difference was distinct even in the first visit, indicating structural changes prior to diagnosis (Figure 11.3).

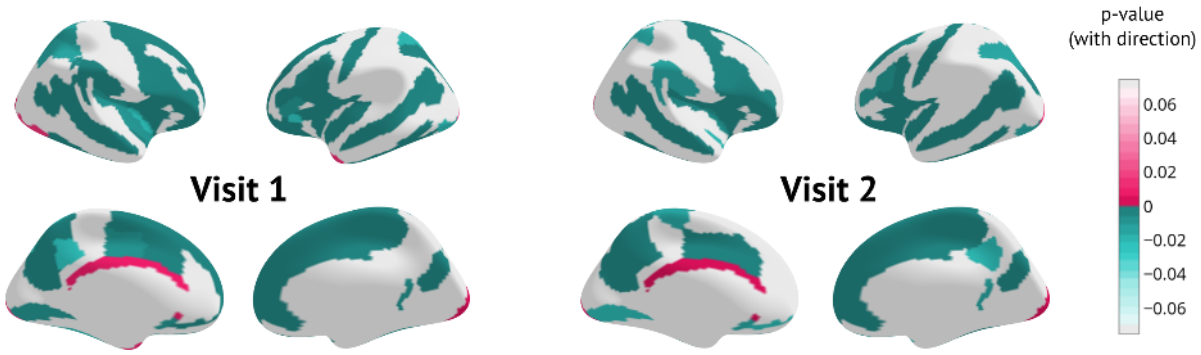

Figure 11.3: **Cross-sectional results for each visit separately.** p-values surviving Benjamini-Hochberg correction. The sign indicates the direction of change (negative means lower thickness in patients).

Longitudinal results and patterns of change

A longitudinal analysis that evaluated the amount of structural change between the two visits showed a significant cortex normalisation of several frontal areas, namely the right and left superior frontal sulcus, the right and left middle frontal sulcus, the right and left middle frontal gyrus, and the right superior frontal gyrus (Figure 11.4).

In terms of linking change in clinical scores with changes in *z-diff* scores, each of the two scales was well correlated with a different component. The first PCA component, which itself reflected the average change in global thickness across patients, was correlated with the change in GAF score, whereas the second component significantly correlated with the change in PANSS score (see Fig. 11.5).

11.5 Discussion

Longitudinal neuroimaging studies allow us to assess the effectiveness of interventions and gain deeper insights into the fundamental mechanisms of underlying diseases. Despite the significant expansion of our knowledge regarding population variation through the availability of publicly accessible neuroimaging data, this knowledge, predominantly derived from cross-sectional observations, has not been adequately integrated into methods for evaluating longitudinal changes.

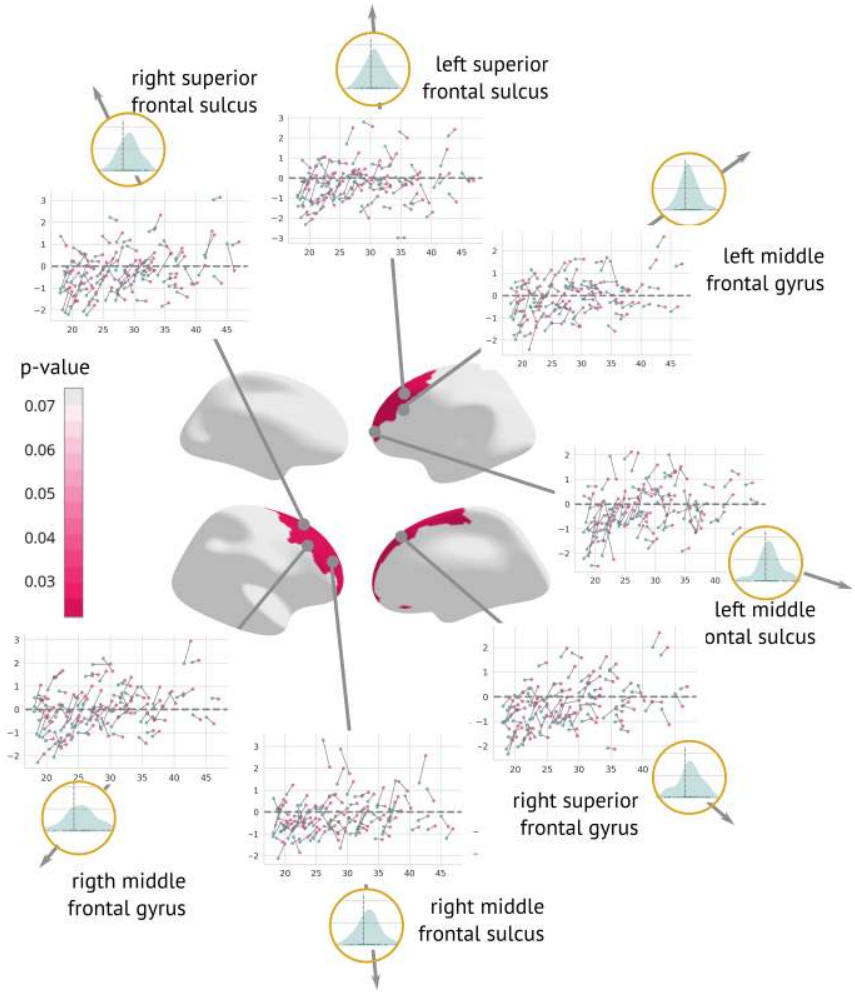

Figure 11.4: **Regions significantly changed between the visits.** Map of regions significantly changed between the two visits (centre). Each region is described using a scatterplot of z-scores across all patients for both visits (the x-axis describes age, and the y-axis depicts the z-score. Blue dots represent the first and pink dots represent the second visit). The Grey dashed line highlights z=0. Histograms in the golden circles depict the distribution of the *z-diff* score.

We propose an analytical framework that builds on normative modelling and generates unbiased features that quantify the degree of change between visits whilst capitalising on information extracted from large cross-sectional cohorts.

Theoretical derivation

Our approach is rooted in the normative modelling method based on Bayesian regression [124], the pre-trained version of which recently became available [127]. We theoretically showed that the estimation of longitudinal changes using the results of the normative model is readily available based on a preexisting cross-sectional normative model and only requires a set of healthy controls on which the variance of healthy change might be estimated. We denoted the score obtained after running the procedure as a *z-diff* score,

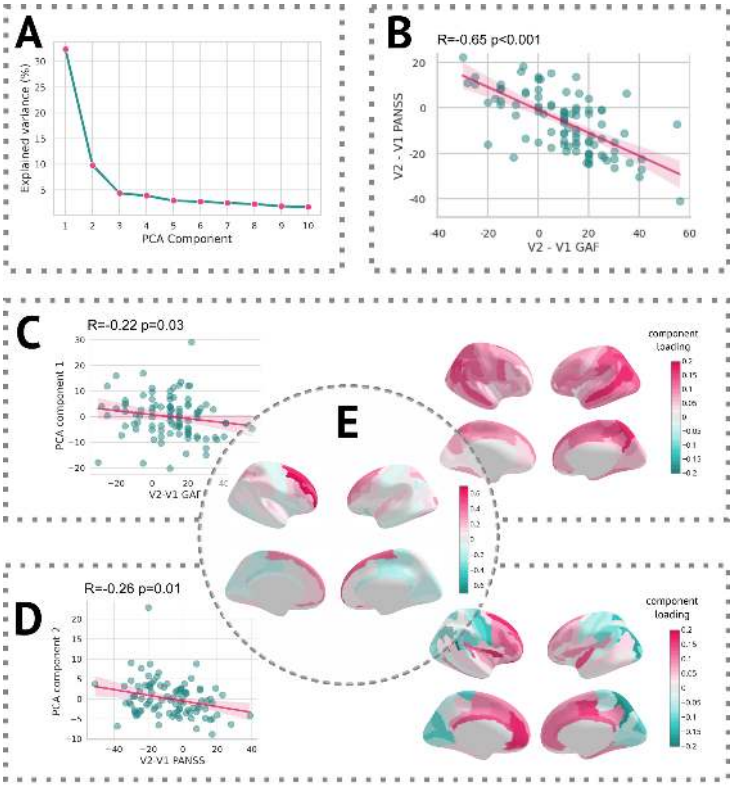

Figure 11.5: **Results of the PCA analysis. A:** Scree plot of the explained variance of PCA components. **B:** Scatterplot of change in the GAF scale vs. the change in the PANSS scale **C Left:** Scatter plot of the first PCA component and difference in the GAF scale. **C Right:** Heatmap of PCA loadings for the first component. **D Left:** Scatter plot of the second PCA component and difference in the PANSS scale. **D Right:** A Heatmap of PCA loadings for the second component. **E:** Average *z-diff* score.

which quantifies the probability that such a difference could have occurred between visits even in the absence of true brain change.

The core assumption of the method is that, on average, healthy controls do not deviate from their trajectory between visits. This assumption stems from the essential idea of normative modelling that, with respect to the reference population, the position of a healthy subject does not significantly deviate over time. To verify this assumption, we used the data of 33 healthy controls which were originally used for the site-specific adaptation (for more details, see the discussion part on implementation) and computed their *z-diff* scores. After averaging these scores across all subjects, the *z-diff* score of no region was statistically significant from zero (after FDR correction). However, as pointed out by a recent work studying the effect of cross-sectional normative models on longitudinal predictions, the cross-sectionally derived population centiles *by design* lack information about longitudinal dynamics. Consequently, what may appear as a population-level trajectory does not necessarily align with individual subjects' actual trajectories.

To address this limitation, our model considers the variation in individual centiles.

This is achieved by estimating the impact of noise and reliability, which manifests as apparent crossings of centiles observed in healthy controls. Naturally, by incorporating this element of uncertainty, the model's ability to detect subjects who experienced substantial changes in their trajectory over time decreases. As evident from the clinical findings, only a fraction of subjects were identified as having undergone significant changes (Fig. C.2). However, at the group level, the significance of the observed changes persisted. Hence, while we adopt a cautious approach when assessing individual changes, the method still effectively identifies group-level changes.

Furthermore, unlike in [348], our approach does not aim to predict individual trajectories but rather to quantify whether the observed changes over time exceed what would be expected by chance.

Implementation

At the implementation level, our model requires two stages of adaptation: site-specific adaptation, as presented in [127], and a second level, where we compute the variance of healthy change (noise) in healthy controls. However, if the number of longitudinal controls is limited, the site-specific adaptation may be omitted. The purpose of site-specific adaptation is to generate unbiased cross-sectional z-scores that are zero-centred, with a variance of one for healthy controls. However, in the case of longitudinal analysis, the offset and normalisation constant are irrelevant since they will be identical for both visits. Therefore, the estimation of healthy change is the only essential factor in producing the *z-diff* score. Note that in this scenario, the cross-sectional result should not be interpreted.

Clinical results

Examination of the preprocessed data showed the superiority of the longitudinal pipeline in terms of consistency over time. As our objective was to minimise the intra-individual variance, we preferred and primarily reported the longitudinal approach in further analyses. However, to assess the added benefit of the preprocessing, we also computed the core results (regions that significantly changed in time) for the cross-sectional data. The significant results were mostly consistent with a longitudinal pipeline: Six out of seven originally significant regions were still statistically significant (with the exception of the right middle frontal sulcus), and three other regions were labelled significant: the left superior frontal gyrus, the right inferior frontal sulcus, and the right medial or olfactory orbital sulcus (Fig. C.3). Therefore, it is also possible to use cross-sectional preprocessing for longitudinal analysis; however, at the cost of increased between-visit variance and

consequently decreased power (in comparison to the longitudinal preprocessing).

The observation of cortical normalisation between the visits of early schizophrenia patients is, to a degree, counterintuitive and inconsistent with other works, which mostly report grey matter thinning. However, a meta-analysis of 50 longitudinal studies examining individuals with a heightened risk of psychosis revealed that 15 of the 19 studies indicated deviations in grey matter developmental trajectories between those with persistent symptoms and those whose symptoms resolved [352]. The authors propose that grey matter developmental trajectories may return to normal levels in individuals in the high-risk remitting group by early adulthood, whereas neurological irregularities may continue to advance in those whose symptoms do not resolve. Although our cohort had already received a diagnosis of schizophrenia, it is possible that early identification and treatment supported these compensatory mechanisms, as demonstrated by the normalisation of grey matter thickness in frontal regions. Notably, the affected regions also increased in raw grey matter thickness (as measured in mm, see Fig. C.4).

Furthermore, we observed significant correlations between the PCA components of the *z-diff* score and changes in clinical scales, as illustrated in Fig. 11.5. Notably, each clinical scale exhibited distinct associations with separate PCA components, despite substantial intercorrelations (Fig. 11.5 (B)).

The first PCA component, which predominantly captured global changes in grey matter thickness, displayed a negative correlation with improvements in the GAF score (Fig. 11.5 (C)). This unexpected inverse relationship would suggest that patients who demonstrated clinical improvement over time exhibited a more pronounced decrease in grey matter thickness, as quantified by the *z-diff* score. However, further investigation revealed that this correlation was primarily driven by the patients' GAF scores in the initial visit. Specifically, the correlation between GAF scores at the first visit, and the first PCA component yielded a coefficient of $R = 0.19$ ($p = 0.06$), whereas the correlation with scores at the second visit was $R =$ -0.10 ($p = 0.31$). These findings suggest that lower GAF scores during the initial visit are predictive of subsequent grey matter thinning.

Conversely, the interpretation of the second PCA component, significantly correlated with changes in the PANSS score, was more straightforward (Fig. 11.5 (D)). The observed normalisation of grey matter thickness in frontal areas was positively correlated with improvements in the PANSS scale, indicating that symptom amelioration was accompanied by the normalisation of grey matter thickness in these regions.

Finally, we conducted an analysis of change using conventional statistical approaches to compare the results with normative modelling. Out of 148 areas tested by ANCOVA, six were statistically significant. However, after controlling for multiple comparisons, no IDP persisted. This result highlights the advantages of normative models and shows

improved sensitivity of our method in comparison with more conventional approaches.

Limitations

Estimating the intra-subject variability is a complex task that might be affected by acquisition and physiological noise. Assumptions must be made about the longitudinal behaviour of healthy subjects. The former problem is unavoidable, whereas the latter might be addressed by constructing longitudinal normative models. However, the project necessary for such a task would have to map individuals across their lifespan consistently. The efforts to create such a dataset are already in progress through projects like the ABCD study [353], but much more data are still needed to construct a full-lifespan longitudinal model.

Additionally, our clinical results may be affected by selection bias, where subjects experiencing a worsening of their condition dropped out of the study, whereas patients with lower genetic risk or more effective treatment continued to participate.

11.6 Conclusion

In conclusion, we have successfully developed a modality-blind method that uses pretrained normative models to detect longitudinal changes in neuroimaging data. In line with the scope of this thesis, our modelling approach allows for statistical inference on the level of individual, paving the way towards individualised medicine. It offers a user-friendly implementation and has demonstrated its effectiveness through a comprehensive analysis. We showcased the method on our internal longitudinal dataset of first-episode schizophrenia patients, where we observed significant grey matter changes in the frontal lobe over time. Furthermore, the approach surpassed the sensitivity of conventional statistical approaches, highlighting the importance of considering individual variability in the early stages of the analysis.

Chapter 12

Conclusion

In this thesis, I have described the state of the research in (multimodal) neuroimaging data analysis, emphasising the methods used and applied in its various stages. In the theoretical part, I dissected the most pertinent processing and analytical stages, starting with data acquisition, followed by the most common preprocessing and feature design steps, until the analysis itself. I have also raised topics of confounders and possible strategies to overcome them. Where appropriate, I have critically evaluated these methods, objectively stating their strengths and limitations with respect to the purpose of the analysis. In the body of the thesis, I have presented my original works, in which I developed analytical methods and approaches most suitable for a diverse set of research questions.

The notion of classification and prediction is a relatively recent concept in neuroimaging data analysis. This field was dominated by classical statistics for a long time, satisfied with making statements about the group-level difference instead of attempting a subject-level classification and prediction. However, with the rising amount of available data and scientists able to appropriately analyse them, we pose more challenging questions, answering of which require more advanced methodologies.

Indeed, the question at the centre of this thesis focuses on the possibilities of individual-level classification and prediction using neuroimaging data. However, as I tried to demonstrate in the theoretical part of the thesis, there is much more than meets the eye in this process. Proper preprocessing is particularly crucial, as incorrect handling can lead to flawed analyses with inconclusive or mystifying results. It is thus essential to understand and approach feature design and data analysis as interconnected processes.

Later chapters in the theoretical part of this thesis explored the importance and relevance of dimensionality reduction methods and, finally, classification and prediction algorithms. Despite the rise of AI in neuroimaging analysis, dimensionality reduction remains a pivotal and frequently undervalued stage, which can significantly affect its results, as was shown in some of my original works. Additionally, selecting appropriate classification and

prediction algorithms is crucial and must be carefully considered in light of our research objectives. However, among all analytical stages, this aspect is the most susceptible to passing trends and fashion, often leading to suboptimal outcomes.

Challenges in the use of AI in neuroimaging data analysis

The adoption of AI in neuroimaging data analysis faces significant obstacles, such as the lack of interpretability of models, challenges in transferring models to unseen data from different scanning sites, and the scarcity of independent replication studies [103], [354]. In the first original article, all of these issues were followed up upon [131]. Using the results of a previous report which employed a deep neural network to discriminate between males and females [133], I constructed a much simpler classifier, verifying and strengthening the original claim on an independent data set. This result has not only clinical implications in terms of accounting for sex covariates in EEG data analyses but also highlights and reinforces the importance of returning to simple approaches due to their interpretability, robustness, and transferability. Nevertheless, the pursuit of interpretability in AI models should not be abandoned. The field of explainable AI holds immense potential in neuroscience, offering valuable insights that can deepen our understanding of the brain. As such, it remains a pivotal and dynamic area of research for the future [148].

The double meaning of prediction

Neuroscience attracts researchers from diverse backgrounds, leading to both innovative ideas and occasional confusion. While machine learning prediction refers to an algorithm predicting unseen data in a clinical context, it also involves forecasting future outcomes. These definitions are not contradictory, with the latter being a specific case of the former; however, accurate longitudinal prediction is extremely challenging and remains a significant milestone yet to be reached in neuroscience research. The second original article provided valuable insights into the complexities of longitudinal prediction in the elderly, where age plays a crucial role. Studying the contribution of age revealed the need for comprehensive control of covariates, a topic often overlooked but with a substantial impact on neuroimaging analysis results [121]. The effect of age significantly propagated into all stages of analysis, and as we showed, when not addressed properly, it can lead to misleading results [120].

This project raised several issues that deserve further investigation. The lack of a pre-morbid baseline is a serious limitation when attempting longitudinal prediction in sudden diseases like stroke. This could theoretically be partially addressed by normative models for cognition, which would model clinical scores before the attack and provide a probabilistic baseline for the analysis. Additionally, extending the analysis to include

functional connectivity could offer valuable insights into specific pathways that play a role in influencing long-term cognition.

All for one and one for all

Combining two modalities for classification or prediction is a common practice in neuroimaging data analysis [74]. However, the fusion of three modalities is relatively rare due to challenges related to the subjects-to-features ratio and the high level of expertise required for preprocessing and feature design. In the third article, I explored the potential of modality integration to classify multiple sclerosis patients and controls while also exploring the relationship between neuroimaging features and clinical scales of motor impairment in patients [225].

The findings revealed an intriguing aspect of multimodal analysis: when one modality significantly dominates in informativeness, incorporating additional modalities may not significantly improve results. In fact, if the extra features are too noisy or overwhelming in number, they could even lower the classification accuracy. However, applying a reasonable dimensionality reduction method can mitigate these adverse effects, resulting in the multimodal analysis performing on par with the best modality. Nevertheless, asserting that one modality suffices for all tasks would be misleading. For instance, diffusion-weighted imaging features proved highly efficient in the classification task but were not useful for prediction, where structural and functional features dominated.

The significance of multimodal fusion was further highlighted in the follow-up phase, where the integration of modalities was performed at the classifier level rather than the feature level. Through the integration of classifiers from diverse features using a voting mechanism, the final classifier achieved superior performance. This finding is even more encouraging because of the complexity of the task, which was set in the context of an early psychiatric disease with heterogeneous symptomatology and subtle changes.

One of the most significant challenges of multimodal analysis is addressing the vast differences between the natural dimensionalities among modalities. fMRI, in particular, often overwhelms other modalities with the features due to its temporal nature. Researchers have employed various techniques to handle this issue, such as averaging fMRI time series or using features like the amplitude of low-frequency fluctuations, yielding varying degrees of success [74]–[76]. Another promising solution involves intelligent data subsampling using algorithms based on determinantal point processes or similar models [355].

New generation models for neuroimaging data analysis

The publication of pre-trained normative models strives to overcome the major challenges in neuroimaging data analysis – the lack of data and appropriate methods for transfer learning [123]. Researchers train normative models on large samples comprising tens of thousands of subjects (which are out of reach for most research groups) and share pre-trained models alongside code for transferring the models to local datasets, empowering all researchers with transferable models [127], [356].

Normative models model the distribution of healthy populations across covariates to generate an unbiased set of features, which can be further used for various analytical setups. My work in this area focused on developing a method that leverages these pre-trained models for evaluating changes in longitudinal neuroimaging data [343]. The method is modality-blind, requiring only a pre-trained normative model, which can be fitted to any modality. It holds promise for enhancing longitudinal analysis, the methods for which are not developed as actively as in the case of cross-sectional design, despite the abundance of resources invested in acquiring them.

The natural progression of my research on longitudinal normative models leads to the aspiration of creating a fully longitudinal normative model, which would encompass subject-level dynamics and integrate temporal information. While this endeavour appears ambitious, given the complexities of both methodology and data availability, the field is rapidly evolving, which increases the chances of it being feasible in the foreseeable future.

In summary, the goals of this thesis were successfully achieved through an exploration of advanced methodologies in neuroimaging data analysis and their critical application in addressing challenges in successful classification and prediction. The theoretical part of the thesis systematically dissected the core stages of neuroimaging data analysis, from data acquisition to feature design, while also addressing the impact of confounders and the significance of dimensionality reduction and analytical methods. In the applied part of the thesis, I thoroughly investigated the integration of multiple neuroimaging modalities to capture complex brain patterns and relationships. The use of modality fusion in classification tasks revealed that dominant modalities can significantly influence results, while dimensionality reduction techniques can mitigate the impact of noisy features. Furthermore, I addressed the challenges tied to longitudinal prediction, revealing the importance of proper control for covariates to avoid misleading results.

Throughout these past years, it has been a pleasure and a privilege to be a part of the dynamic and ever-evolving neuroimaging community. It is an exciting field, constantly expanding and transforming as researchers unveil novel projects and ideas. Beyond the integration of neuroimaging modalities, we are witnessing an exciting convergence with

other domains, such as genetic data, receptor distribution, and metabolic profiles, elevating the concept of multimodal fusion to new heights. Consequently, the need for advanced analytical methods to handle and combine this vast amount of information will be more critical than ever before. As I conclude this thesis, I am hopeful that my efforts have contributed to the advancement of neuroscience, providing valuable insights into the complexity and possibilities of this captivating field.

Appendix A

Appendix – Chapter 7

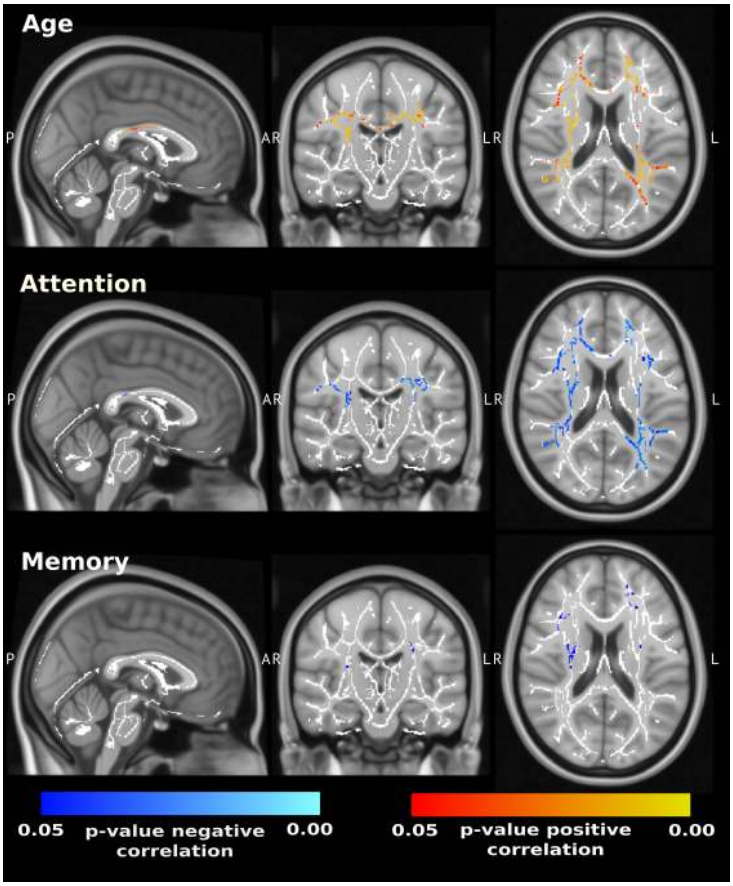

Figure A.1: **The results of TBSS for axial diffusivity (AD).** Blue colour scale signifies a negative correlation between AD and the clinical variable, red colour scale stands for a positive correlation. Either a positive or negative correlation is depicted per clinical scale). We observed: **Age:** Widespread positive correlation of AD and age. **Attention:** Global negative correlation between AD and attention. Negative correlation between AD and executive functions. **Memory:** Scattered negative correlation between AD and memory. Localized negative correlation between AD and language. Note that the correlations with clinical scales did not persist (as statistically significant) after controlling for age.

Table A.1: **The results of clinical tests across 46 patients.** FCSRT — Free and Cued Selective Reminding Test, P-VF — Phonemic verbal fluency, PST — Prague Stroop Test, TMT — Trail Making Test, DS — Digit Span, BNT — Boston Naming Test, S-VF — Semantic verbal fluency, ROCF — Rey-Osterrieth Complex Figure, Visuospatial functions — VOSP

	Men mean (sd, min, max)	**Women** mean (sd, min, max)	**All** mean (sd, min, max)
Mini-mental state examination	27.21 (2.5, 18, 30)	27.68 (1.6, 24, 30)	27.43 (2.1, 18, 30)
FCSRT free recall	28 (8, 11, 39)	32 (4, 25, 40)	30 (6, 11, 40)
FCSRT total recall	47 (1, 43, 48)	48 (1,46,48)	48 (1,43,48)
FCSRT delayed free recall	9 (3, 3, 15)	11 (2, 8, 15)	10 (3, 3, 15)
FCSRT delayed total recall	16 (0, 15, 16)	16 (0, 15, 16)	16 (0, 15, 16)
TMT B (time to completion)	182 (142, 50, 500)	168 (128, 54, 500)	176 (134, 50, 500)
P-VF	32 (13, 11, 64)	39 (14, 19, 72)	35 (14, 11, 72)
Similarities	22 (5, 8, 29)	20 (5, 10, 28)	21 (5, 8, 29)
PST - colour (time to completion)	38 (17, 22, 88)	41 (18, 19, 93)	40 (17, 19, 93)
TMT A (time to completion)	51 (26, 23, 123)	43 (15, 20, 68)	47 (22, 20, 123)
DS forward	8 (2, 4, 12)	8 (2, 6, 13)	8 (2, 4, 13)
DS backward	5 (2, 2, 8)	6 (2, 3, 11)	5 (2, 2, 11)
BNT-15 (number of errors)	2 (1, 0, 5)	2 (2, 0, 6)	2 (2, 0, 6)
S-VF animals	20 (5, 11, 29)	22 (6, 11, 32)	21 (6, 11, 32)
ROCF copy	27 (3, 21, 32)	27 (5, 14, 35)	27 (4, 14, 35)
VOSP number location	9 (1, 7, 10)	9 (2, 1, 10)	9 (2, 1, 10)

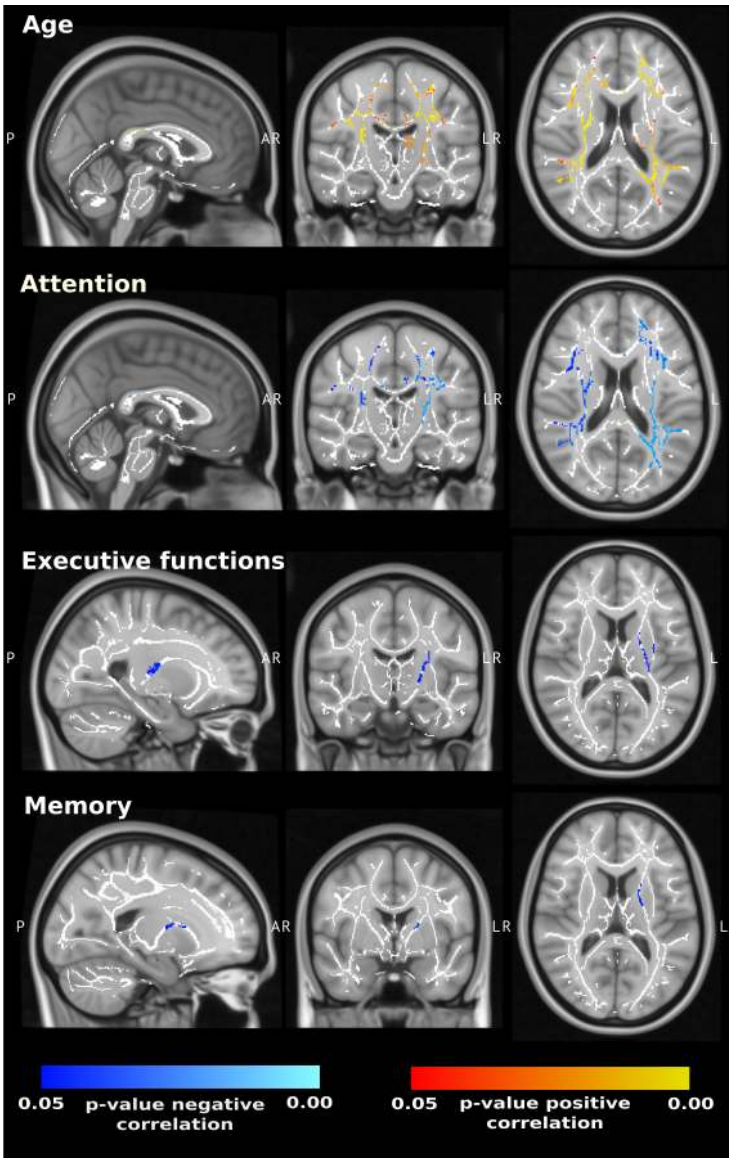

Figure A.2: **The results of TBSS for mean diffusivity (MD).** Blue colour scale signifies negative correlation between MD and the clinical variable, red colour colour scale stands for a positive correlation. Either positive or negative correlation is depicted per clinical scale). We observed: **Age:** Widespread positive correlation of MD and age. **Attention:** Global negative correlation between MD and attention. **Executive functions:** Negative correlation between MD and Executive functions. Scattered negative correlation between MD and language. **Memory:** Localised negative correlation between MD and memory. Scattered negative correlation between MD and language. Note that the correlations with clinical scales did not persist (as statistically significant) after controlling for age.

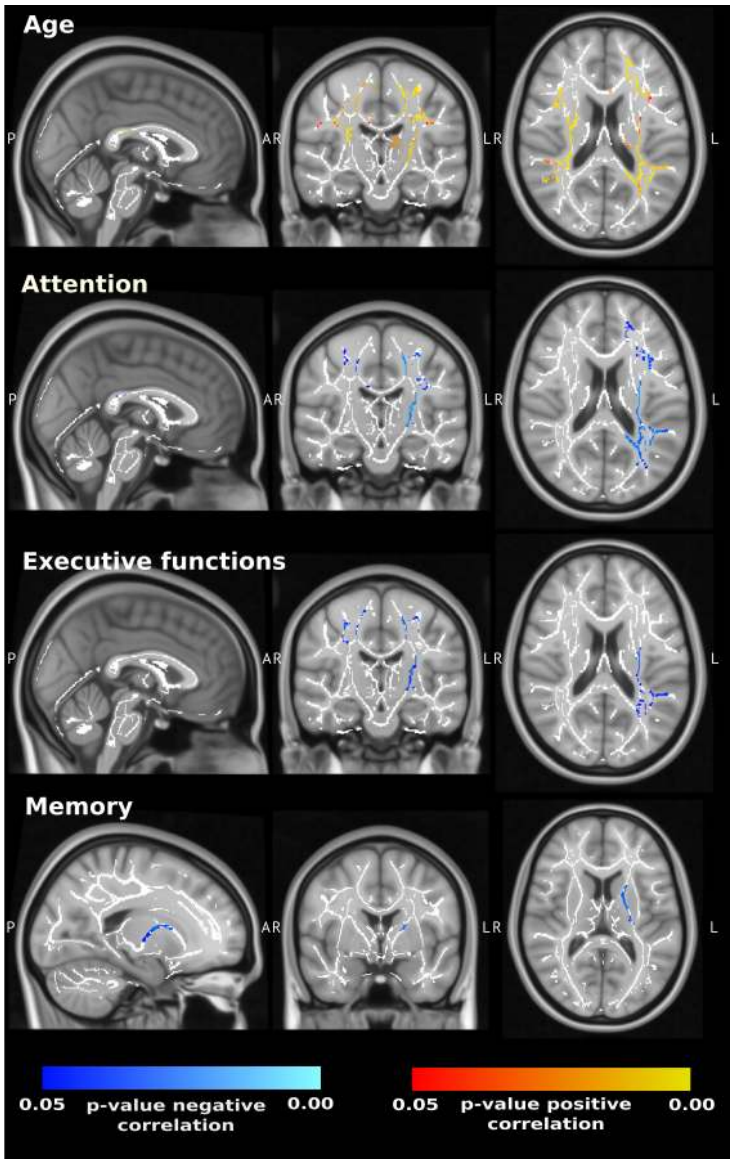

Figure A.3: **The results of TBSS for radial diffusivity (RD).**Blue colour scale signifies negative correlation between RD and the clinical variable, red colour colour scale stands for positive correlation. Either positive or negative correlation is depicted per clinical scale). We observed: **Age:** Widespread positive correlation of RD and age. **Attention:** Negative correlation between RD and attention predominantly in the left hemisphere. **Executive functions:** Scattered negative correlation between RD and executive functions. Scattered negative correlation between RD and Language. **Memory:** Localised negative correlation between RD and memory. Note that the correlations with clinical scales did not persist (as statistically significant) after controlling for age.

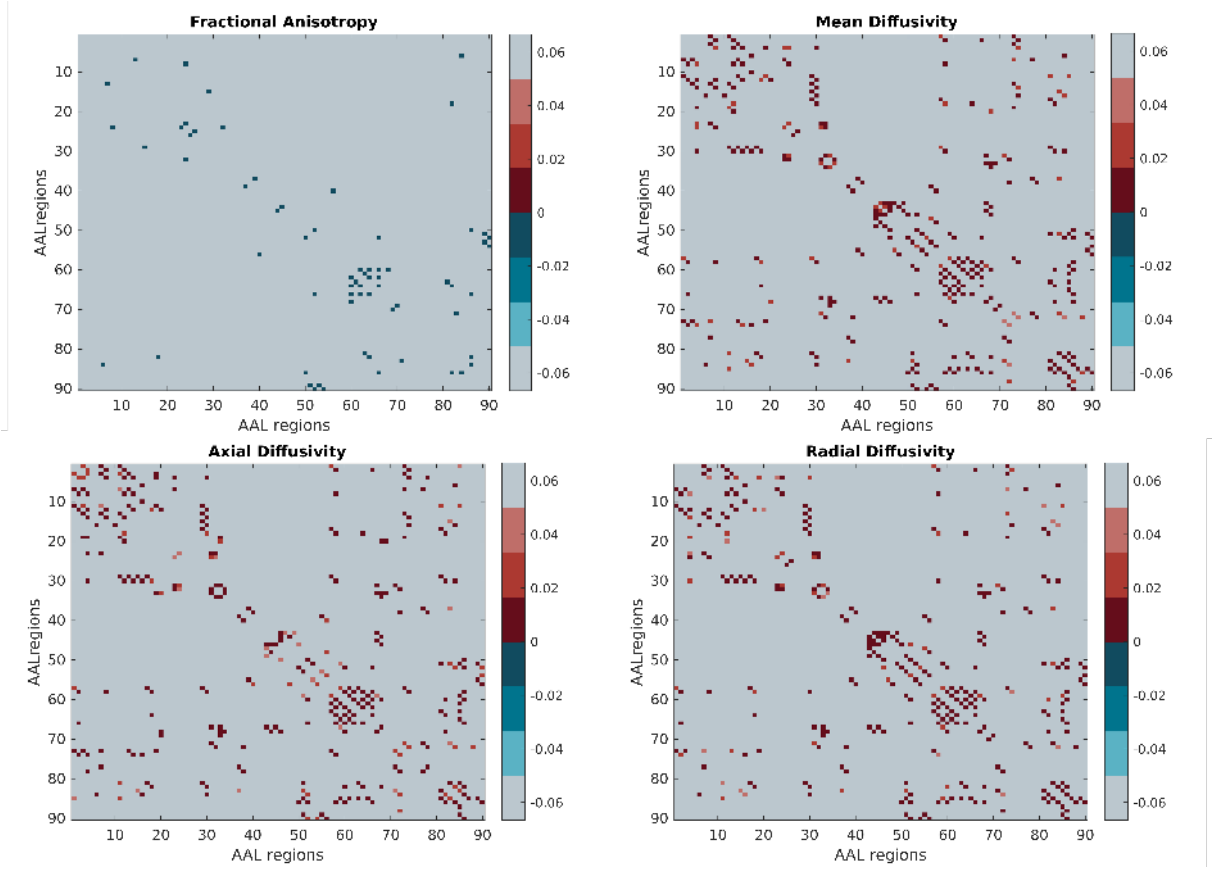

Figure A.4: **Correlation of fractional anisotropy mean diffusivity, axial, and radial diffusivity with age along the backbone tracts.** We extracted tracts between each pair of regions along the backbone and computed average FA and MD. The figures depict the FDR-corrected p-values of Spearman correlations of these values with age. The colour indicates whether the correlation is positive or negative. Note that no pair of regions was significantly correlated with any diffusivity metrics after controlling for age.

Appendix B

Appendix – Chapter 8

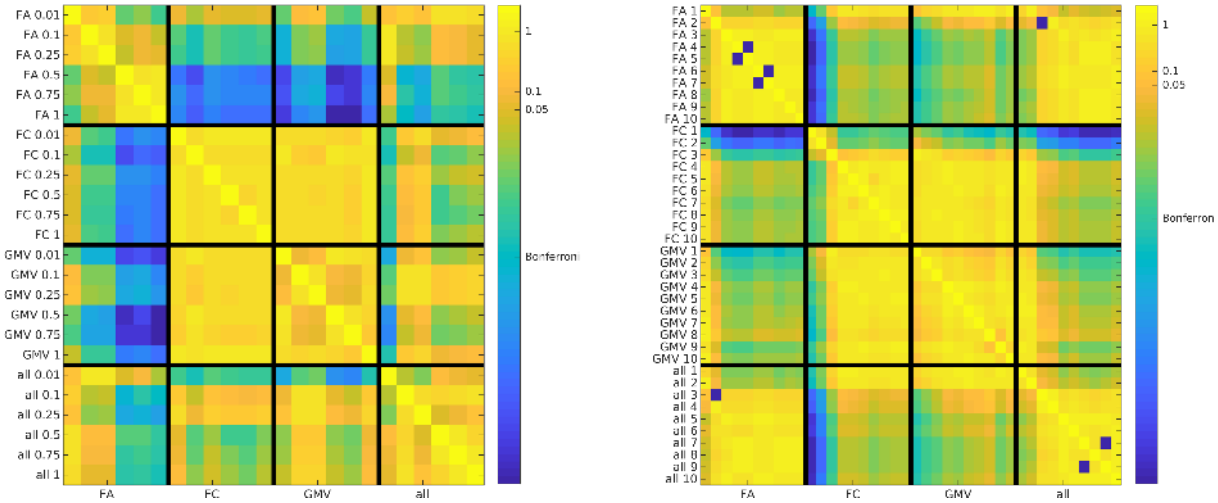

Figure B.1: **P-values of McNemar test between all thresholds and classifiers.** Support Vector Machines (left), Logistic Regression (right). Threshold values in Support Vector Machines stand for the percentage of all features; in Logistic regression, they represent the number of PCA components added to the model. FA – Fractional Anisotropy; FC – Functional Connectivity; GMV – Grey Matter Volume; all – combination of all three modalities.

Table B.1: **The results of the Support Vector Machine models across modalities for all thresholds.** FA — Fractional Anisotropy; FC — Functional Connectivity; GMV — Grey Matter Volume; all — combination of all three modalities.

		Fisher Score and Support Vector Machines percentage of features					
		1%	**10%**	**25%**	**50%**	**75%**	**100%**
FA	**sensitivity**	81.2%	89.1%	90.6%	93.8%	95.3%	98.4%
	specificity	87.7%	89.2%	87.7%	96.9%	95.4%	93.8%
	overall accuracy	84.5%	89.1%	89.1%	95.3%	95.3%	96.1%
FC	**sensitivity**	78.1%	65.6%	71.9%	71.9%	70.3%	70.3%
	specificity	70.8%	76.9%	75.4%	75.4%	76.9%	76.9%
	overall accuracy	74.4%	71.3%	73.6%	73.6%	73.6%	73.6%
GMV	**sensitivity**	64.1%	70.3%	71.9%	73.4%	70.3%	76.6%
	specificity	76.9%	84.6%	83.1%	66.2%	66.2%	70.8%
	overall accuracy	70.5%	77.5%	77.5%	69.8%	68.2%	73.6%
all	**sensitivity**	89.1%	79.7%	73.4%	84.4%	82.8%	81.2%
	specificity	90.8%	80.0%	81.5%	81.5%	81.5%	81.5%
	overall accuracy	89.9%	79.8%	77.5%	82.9%	82.2%	81.4%

Table B.2: **The results of the Logistic Regression models combined with Principal Component Analysis features across modalities for all thresholds.** FA — Fractional Anisotropy; FC — Functional Connectivity; GMV — Grey Matter Volume; all — combination of all three modalities; sens. — sensitivity; spec. — specificity; oa — overall accuracy

		Principal Component Analysis and Logistic Regression number of features									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
FA	**sens.**	81.2%	82.8%	87.5%	89.1%	89.1%	90.6%	90.6%	92.2%	92.2%	92.2%
	spec.	75.4%	89.2%	92.3%	92.3%	92.3%	89.2%	89.2%	89.2%	89.2%	89.2%
	oa	78.3%	86.0%	89.9%	90.7%	90.7%	89.9%	89.9%	90.7%	90.7%	90.7%
FC	**sens.**	51.6%	56.2%	70.3%	75.0%	76.6%	76.6%	75.0%	76.6%	76.6%	75.0%
	spec.	47.7%	50.8%	66.2%	75.4%	73.8%	72.3%	70.8%	70.8%	72.3%	73.8%
	oa	49.6%	53.5%	68.2%	75.2%	75.2%	74.4%	72.9%	73.6%	74.4%	74.4%
GMV	**sens.**	70.3%	71.9%	71.9%	73.4%	75.0%	78.1%	76.6%	76.6%	73.4%	75.0%
	spec.	67.7%	70.8%	73.8%	78.5%	75.4%	75.4%	78.5%	81.5%	76.9%	80.0%
	oa	69.0%	71.3%	72.9%	76.0%	75.2%	76.7%	77.5%	79.1%	75.2%	77.5%
all	**sens.**	81.2%	81.2%	82.8%	82.8%	87.5%	85.9%	89.1%	87.5%	89.1%	89.1%
	spec.	70.8%	73.8%	89.2%	90.8%	90.8%	90.8%	90.8%	92.3%	90.8%	89.2%
	oa	76.0%	77.5%	86.0%	86.8%	89.1%	88.4%	89.9%	89.9%	89.9%	89.1%

Table B.3: **The results of the Support Vector Regression models across modalities for all thresholds.** FA — Fractional Anisotropy; FC — Functional Connectivity; GMV — Grey Matter Volume; all — combination of all three modalities; EDSS — Expanded Disability Status Scale; BBS — Berg Balance Scale; TUG — Timed Up and Go Test; MSIS — Multiple Sclerosis Impact Scale; MSWS — Twelve Item Multiple Sclerosis Walking Scale.

		Fisher Score and Support Vector Regression percentage of features					
		1%	**10%**	**25%**	**50%**	**75%**	**100%**
EDSS	**FA**	0.05	0.14	0.39	0.36	0.25	0.33
	FC	0.46	0.39	0.30	0.24	0.32	0.35
	GMV	0.30	0.43	0.30	0.04	0.02	-0.03
	all	0.16	0.36	0.33	0.25	0.31	0.38
BBS	**FA**	0.48	0.17	0.30	0.23	0.11	0.33
	FC	0.11	0.36	0.28	0.39	0.41	0.43
	GMV	0.40	0.42	0.22	0.25	0.32	0.19
	all	-0.27	0.37	0.28	0.45	0.46	0.50
TUG	**FA**	-0.27	0.26	0.33	-0.02	0.00	0.18
	FC	0.12	0.31	0.32	0.28	0.35	0.40
	GMV	0.06	0.19	0.39	-0.01	0.25	0.23
	all	0.17	0.31	0.35	0.31	0.39	0.44
MSIS	**FA**	0.02	0.08	-0.08	-0.27	-0.05	-0.07
	FC	0.12	0.24	0.15	0.17	0.32	0.38
	GMV	0.17	-0.07	-0.19	0.06	0.12	-0.04
	all	0.17	0.24	0.14	0.15	0.30	0.37
MSWS	**FA**	0.34	0.21	0.26	0.31	0.02	0.00
	FC	0.79	0.44	0.52	0.52	0.54	0.60
	GMV	0.23	0.17	0.08	0.17	0.11	0.01
	all	0.78	0.45	0.49	0.51	0.53	0.59

Table B.4: **The results of the Linear Regression models combined with Principal Component Analysis components across modalities for all thresholds.** FA — Fractional Anisotropy; FC — Functional Connectivity; GMV — Grey Matter Volume; all — combination of all three modalities; EDSS — Expanded Disability Status Scale; BBS — Berg Balance Scale; TUG — Timed Up and Go Test; MSIS — Multiple Sclerosis Impact Scale; MSWS — Twelve Item Multiple Sclerosis Walking Scale.

		Principal Component analysis and Linear Regression number of features									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
EDSS	**FA**	0.31	0.31	0.33	0.31	0.27	0.23	0.21	0.21	0.14	0.13
	FC	0.19	0.18	0.17	0.16	0.11	0.06	0.07	0.07	0.09	0.11
	GMV	0.42	0.40	0.39	0.36	0.32	0.32	0.29	0.26	0.27	0.26
	all	0.31	0.32	0.29	0.33	0.36	0.37	0.35	0.34	0.36	0.35
BBS	**FA**	0.24	0.22	0.28	0.27	0.28	0.28	0.41	0.42	0.39	0.39
	FC	-0.03	0.15	0.15	0.26	0.31	0.28	0.27	0.32	0.31	0.32
	GMV	0.46	0.43	0.43	0.40	0.38	0.36	0.32	0.28	0.40	0.40
	all	0.24	0.25	0.19	0.18	0.24	0.24	0.27	0.27	0.26	0.25
TUG	**FA**	0.18	0.24	0.29	0.24	0.19	0.17	0.28	0.26	0.22	0.23
	FC	0.20	0.21	0.20	0.25	0.25	0.20	0.19	0.23	0.22	0.20
	GMV	0.43	0.41	0.43	0.38	0.33	0.37	0.35	0.34	0.35	0.35
	all	0.18	0.19	0.17	0.23	0.25	0.21	0.20	0.19	0.14	0.10
MSIS	**FA**	0.16	0.07	0.00	-0.03	-0.02	-0.06	-0.09	-0.09	-0.04	-0.02
	FC	-0.03	0.07	0.09	0.19	0.11	0.08	0.09	0.05	0.02	0.01
	GMV	0.07	-0.01	-0.06	-0.07	-0.13	-0.14	-0.09	-0.09	-0.13	-0.14
	all	0.15	0.17	0.08	0.13	0.06	0.06	0.07	0.08	0.05	0.03
MSWS	**FA**	0.14	0.01	0.13	0.11	0.09	0.03	0.12	0.10	0.11	0.10
	FC	0.28	0.29	0.28	0.34	0.37	0.35	0.34	0.32	0.28	0.30
	GMV	0.28	0.29	0.27	0.24	0.21	0.15	0.07	0.11	0.11	0.11
	all	0.14	0.15	0.03	0.11	0.14	0.13	0.14	0.12	0.12	0.03

Table B.5: **The results of the LASSO Logistic Regression.** LASSO had been fitted separately for every fold in leave-one-out cross-validation, and the best model was chosen in each fold using Bayesian Information Criterion. FA – Fractional Anisotropy; FC – Functional Connectivity; GMV – Grey Matter Volume; all – combination of all three modalities.

	sensitivity	**specificity**	**accuracy**	**AUC**	**median no. variables across folds**
FA	93.8%	93.9%	93.8%	0.97	4
FC	85.9%	66.2%	76.0%	0.78	6
GMV	85.9%	73.9%	79.8%	0.85	4
all	84.4%	92.3%	88.4%	0.94	7

Appendix C

Appendix – Chapter 11

Posterior predictive distribution for difference between visits

Here we derive the posterior predictive distribution for the difference $\varphi(y^{(2)}) - \varphi(y^{(1)})$. The argument is standard. Denote $\Delta_{\mathbf{x}} = \phi(\mathbf{x}^{(2)}) - \phi(\mathbf{x}^{(1)})$ and $\Delta_y = \varphi(y^{(2)}) - \varphi(y^{(1)})$. Since $\Delta_{\mathbf{x}}^T \mathbf{w} | \mathbf{x}^{(1)}, \mathbf{x}^{(2)}; \mathbf{y}, \mathbf{\Phi}; \alpha, \beta, \boldsymbol{\gamma} \sim \mathcal{N}(\Delta_{\mathbf{x}}^T \bar{\mathbf{w}}, \Delta_{\mathbf{x}}^T \mathbf{A}^{-1} \Delta_{\mathbf{x}})$ and $\Delta_y | \mathbf{x}^{(1)}, \mathbf{x}^{(2)}; \mathbf{w} \sim \mathcal{N}(\Delta_{\mathbf{x}}^T \mathbf{w}, 2\sigma_\xi^2)$, the posterior predictive density is

$$
\begin{aligned}
&f(\Delta_y | \mathbf{x}^{(1)}, \mathbf{x}^{(2)}; \mathbf{y}, \mathbf{\Phi}; \alpha, \beta, \boldsymbol{\gamma}) = \\
&\quad = \int f_{\mathcal{N}(\Delta_{\mathbf{x}}^T \mathbf{w}, 2\sigma_\xi^2)}(\Delta_y | \mathbf{x}^{(1)}, \mathbf{x}^{(2)}; \mathbf{w}) \cdot f_{\mathcal{N}(\Delta_{\mathbf{x}}^T \bar{\mathbf{w}}, \Delta_{\mathbf{x}}^T \mathbf{A}^{-1} \Delta_{\mathbf{x}})}(\Delta_{\mathbf{x}}^T \mathbf{w} | \mathbf{x}^{(1)}, \mathbf{x}^{(2)}; \mathbf{y}, \mathbf{\Phi}; \alpha, \beta, \boldsymbol{\gamma}) \, \mathrm{d}(\Delta_{\mathbf{x}}^T \mathbf{w}) \\
&\quad = \int f_{\mathcal{N}(0, 2\sigma_\xi^2)}(\Delta_y - \Delta_{\mathbf{x}}^T \mathbf{w} | \mathbf{x}^{(1)}, \mathbf{x}^{(2)}; \mathbf{w}) \cdot f_{\mathcal{N}(\Delta_{\mathbf{x}}^T \bar{\mathbf{w}}, \Delta_{\mathbf{x}}^T \mathbf{A}^{-1} \Delta_{\mathbf{x}})}(\Delta_{\mathbf{x}}^T \mathbf{w} | \mathbf{x}^{(1)}, \mathbf{x}^{(2)}; \mathbf{y}, \mathbf{\Phi}; \alpha, \beta, \boldsymbol{\gamma}) \, \mathrm{d}(\Delta_{\mathbf{x}}^T \mathbf{w}).
\end{aligned}
$$

This has the convolution form of the densities of $\mathcal{N}(0, 2\sigma_\xi^2)$ and $\mathcal{N}(\Delta_{\mathbf{x}}^T \bar{\mathbf{w}}, \Delta_{\mathbf{x}}^T \mathbf{A}^{-1} \Delta_{\mathbf{x}})$. It is known to produce the density of $\mathcal{N}(\Delta_{\mathbf{x}}^T \bar{\mathbf{w}}, \Delta_{\mathbf{x}}^T \mathbf{A}^{-1} \Delta_{\mathbf{x}} + 2\sigma_\xi^2)$ (by completion to squares in the exponent).

Quality of fit across regions of interest

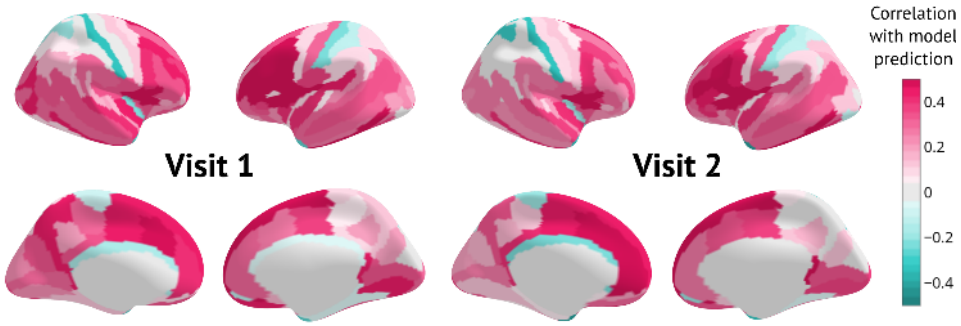

Figure C.1: **Quality of fit as measured by correlation for the first and the second visit.**(controls which have not been used for site-specific adjustment).

Comparison of preprocessing

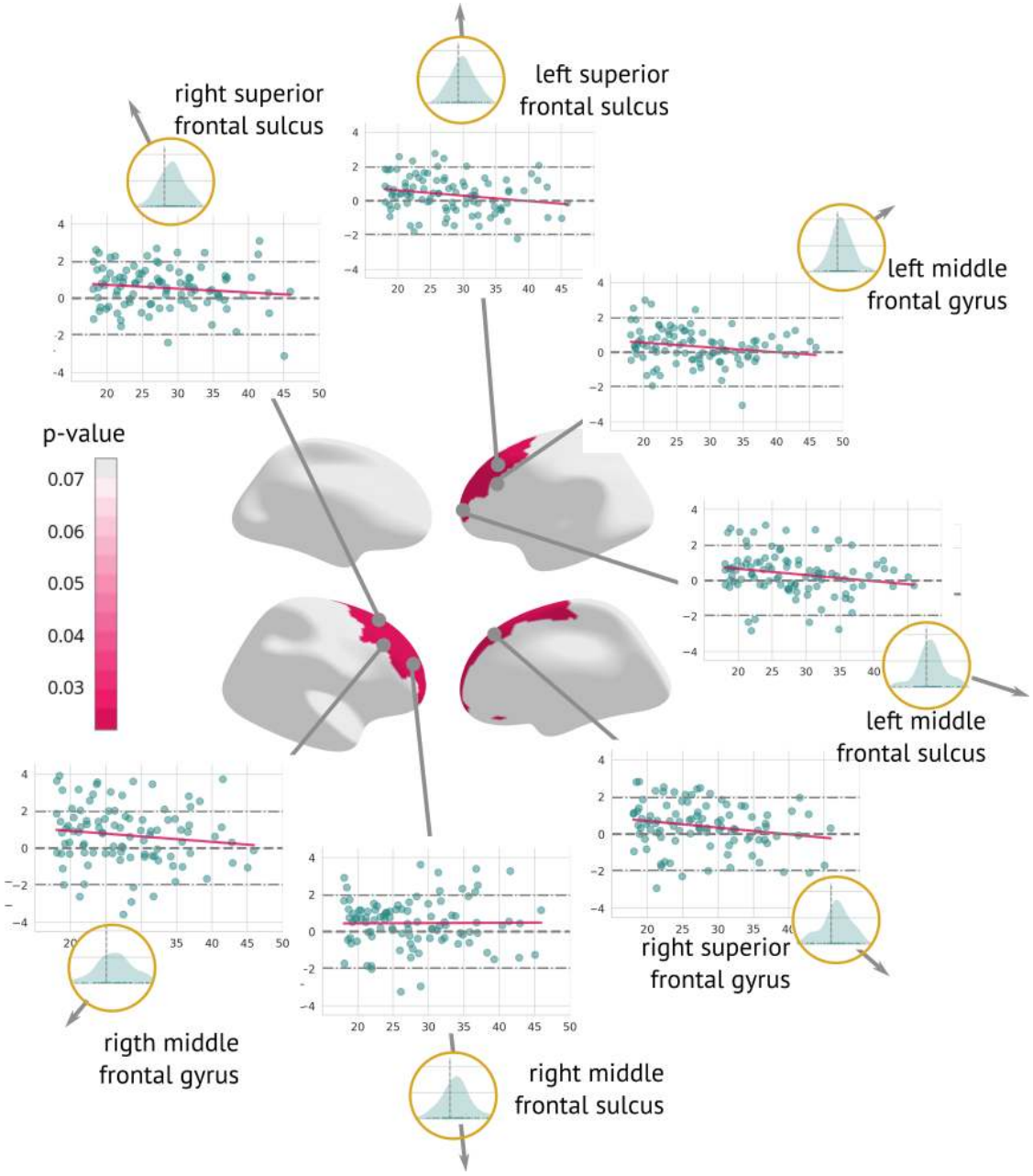

Figure C.2: **Regions significantly changed between the visits (longitudinal preprocessing).** Map of regions significantly changed between the two visits (centre). Each region is described using a scatterplot of *z-diff* across all patients for both visits (the x-axis describes age, and the y-axis depicts the *z-diff*. Blue dots represent individual patients and the pink line shows a trend of *z-diff* change). The Grey dashed line highlights z=0; and $z = \pm 1.96$. Histograms in the golden circles depict the distribution of the *z-diff* score.

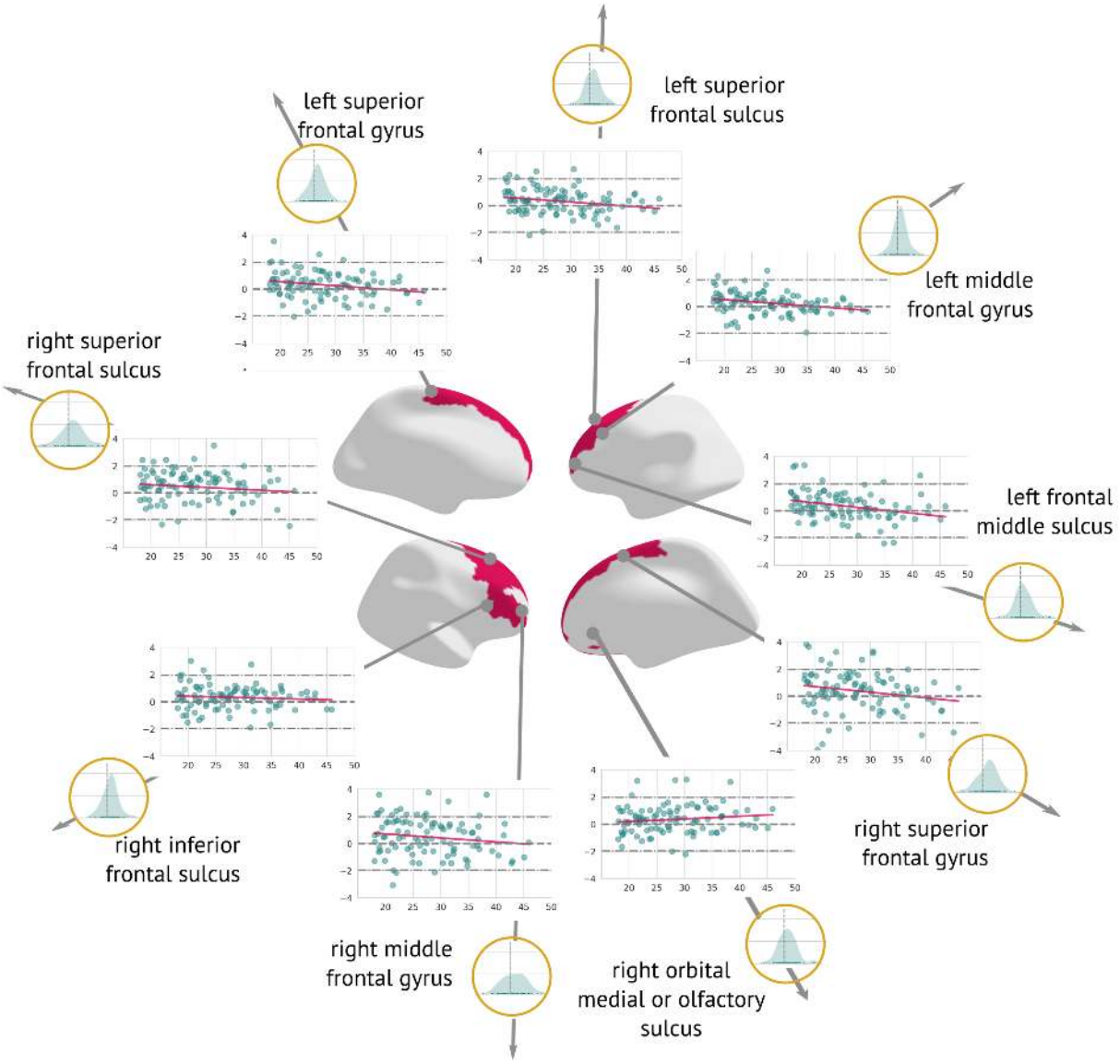

Figure C.3: **Regions significantly changed between the visits (cross-sectional preprocessing).** Map of regions significantly changed between the two visits (centre). Each region is described using a scatterplot of *z-diff* scores across all patients for both visits (the x-axis describes age, and the y-axis depicts the *z-diff* score. The Grey dashed line highlights z=0; and $z = \pm 1.96$. Histograms in the golden circles depict the distribution of the *z-diff* score.

Raw changes observed in significant regions

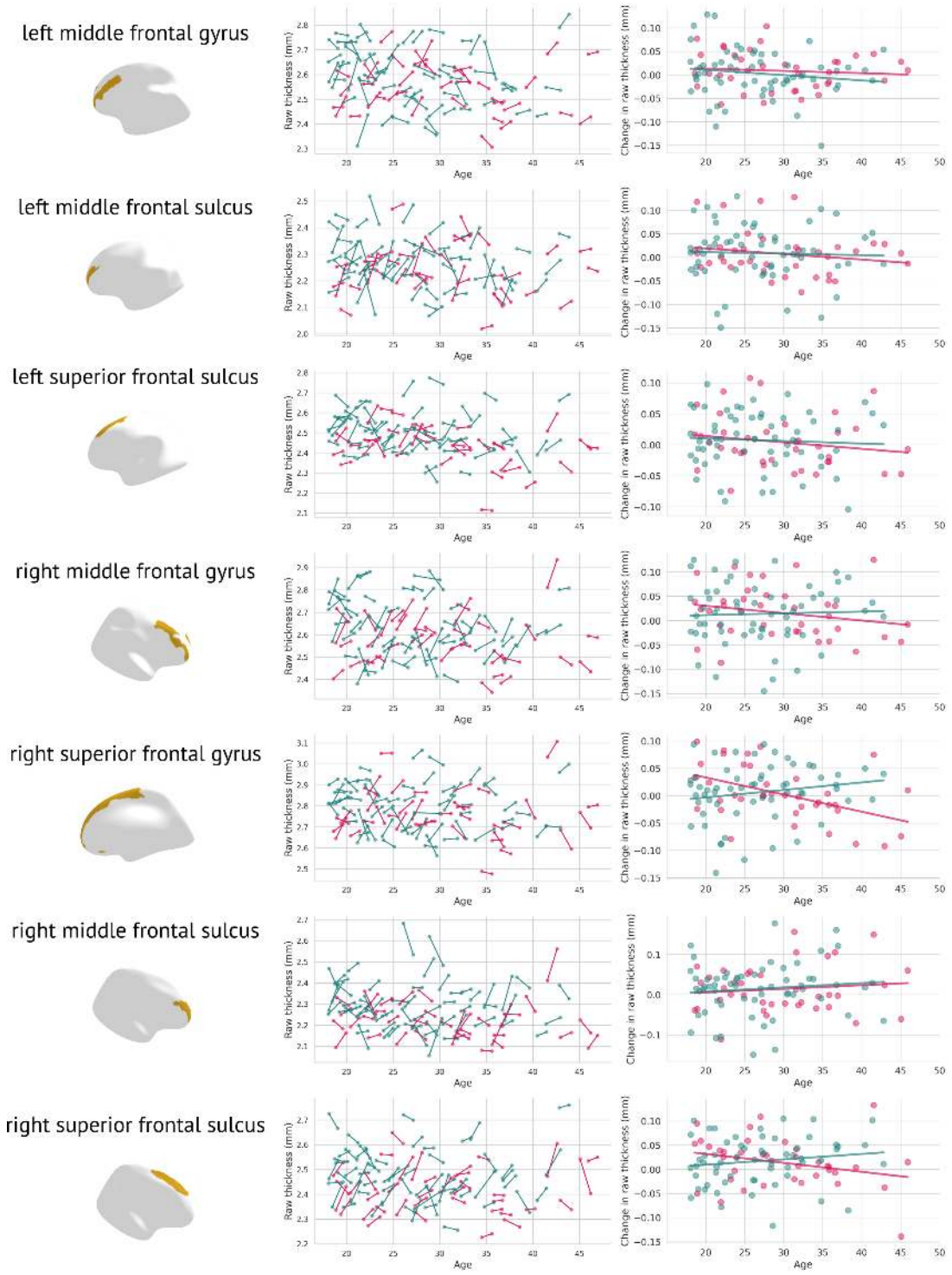

Figure C.4: **Raw changes in grey matter thickness.** Each significantly changed region is presented twice, once as a scatter plot containing the original grey matter thickness for both visits (left); females are plotted in pink, males in blue. The figure on the right depicts V2-V1 in raw thicknesses (separately for females – pink, and males – blue).

Bibliography

[1] S. R. y. Cajal, "La fine structure des centres nerveux. the croonian lecture", *Proc. R. Soc. Lond*, vol. 55, pp. 444–468, 1894.

[2] P. Broca, "Remarques sur le siège de la faculté du langage articulé, suivies d'une observation d'aphémie (perte de la parole)", *Bulletin et Memoires de la Societe anatomique de Paris*, vol. 6, pp. 330–357, 1861.

[3] C. Wernicke, "Der aphasische Symptomenkomplex", in *Der aphasische Symptomencomplex: Eine psychologische Studie auf anatomischer Basis*, C. Wernicke, Ed., Berlin, Heidelberg: Springer, 1974, pp. 1–70.

[4] T. J. Sejnowski, C. Koch, and P. S. Churchland, "Computational neuroscience", *Science*, vol. 241, no. 4871, pp. 1299–1306, 1988.

[5] T. J. Sejnowski, P. S. Churchland, and J. A. Movshon, "Putting big data to good use in neuroscience", *Nature Neuroscience*, vol. 17, no. 11, pp. 1440–1441, 2014.

[6] F. Beltrame and S. Koslow, "Neuroinformatics as a megascience issue", *IEEE Transactions on Information Technology in Biomedicine*, vol. 3, no. 3, pp. 239–240, 1999.

[7] E. Bullmore and O. Sporns, "Complex brain networks: Graph theoretical analysis of structural and functional systems", *Nature Reviews Neuroscience*, vol. 10, no. 3, pp. 186–198, 2009.

[8] S. Achard and E. Bullmore, "Efficiency and cost of economical brain functional networks", *PLOS Computational Biology*, vol. 3, no. 2, e17, 2007.

[9] D. Meunier, R. Lambiotte, A. Fornito, K. Ersche, and E. T. Bullmore, "Hierarchical modularity in human brain functional networks", *Frontiers in Neuroinformatics*, vol. 3, 2009.

[10] A. W. K. Yeung, T. K. Goto, and W. K. Leung, "The changing landscape of neuroscience research, 2006–2015: A bibliometric study", *Frontiers in Neuroscience*, vol. 11, 2017.

[11] X. Chen, J. Chen, G. Cheng, and T. Gong, "Topics and trends in artificial intelligence assisted human brain research", *PLOS ONE*, vol. 15, no. 4, e0231192, 2020.

[12] C.-W. Woo, L. J. Chang, M. A. Lindquist, and T. D. Wager, "Building better biomarkers: Brain models in translational neuroimaging", *Nature Neuroscience*, vol. 20, no. 3, pp. 365–378, 2017.

[13] M. F. Bear, B. W. Connors, and M. A. Paradiso, *Neuroscience: Exploring the Brain, Fourth Edition by Mark F. Bear, Barry W. Connors, Michael A. Paradiso (2015) Hardcover*, 4th edition. Philadelphia: Jones & Bartlett Learning, 2015.

[14] P. Lambin, E. Rios-Velazquez, R. Leijenaar, *et al.*, "Radiomics: Extracting more information from medical images using advanced feature analysis", *European Journal of Cancer*, vol. 48, no. 4, pp. 441–446, 2012.

[15] V. Kumar, Y. Gu, S. Basu, *et al.*, "Radiomics: The process and the challenges", *Magnetic Resonance Imaging*, vol. 30, no. 9, pp. 1234–1248, 2012.

[16] M. Elmahdy and R. Sebro, "Radiomics analysis in medical imaging research", *Journal of Medical Radiation Sciences*, vol. 70, no. 1, pp. 3–7, 2023.

[17] J. M. Teno, "Garbage in, Garbage out—Words of Caution on Big Data and Machine Learning in Medical Practice", *JAMA Health Forum*, vol. 4, no. 2, e230397, 2023.

[18] R. Ansorge and M. Graves, *The physics and mathematics of MRI*. Morgan & Claypool Publishers, 2016.

[19] A. Martín-Araguz, C. Bustamante-Martínez, M. T. Emam-Mansour, and J. M. Moreno-Martínez, "[Neuroscience in ancient Egypt and in the school of Alexandria]", *Revista De Neurologia*, vol. 34, no. 12, pp. 1183–1194, 2002.

[20] R. Behling, "X-ray sources: 125 years of developments of this intriguing technology", *Physica Medica*, vol. 79, pp. 162–187, 2020.

[21] P. C. Lauterbur, "Magnetic resonance zeugmatography", *Pure and Applied Chemistry*, vol. 40, no. 1-2, pp. 149–157, 1974.

[22] L. F. Haas, "Hans Berger (1873–1941), Richard Caton (1842–1926), and electroencephalography", *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 74, no. 1, pp. 9–9, 2003.

[23] A. J. Farrall, "Magnetic resonance imaging", *Practical Neurology*, vol. 6, no. 5, pp. 318–325, 2006.

[24] D. G. Mitchell and M. Cohen (Ph.D.), *MRI Principles*. Saunders, 2004.

[25] J. Hornak, *The Basics of MRI*, https://www.cis.rit.edu/htbooks/mri/index.html, 2020.

[26] P. Mansfield, "Multi-planar image formation using nmr spin echoes", *Journal of Physics C: Solid State Physics*, vol. 10, no. 3, p. L55, 1977.

[27] M. Jenkinson, M. Chappell, M. Jenkinson, and M. Chappell, *Introduction to Neuroimaging Analysis*. Oxford, New York: Oxford University Press, 2017.

[28] J. Bijsterbosch, S. M. Smith, C. F. Beckmann, J. Bijsterbosch, S. M. Smith, and C. F. Beckmann, *Introduction to Resting State fMRI Functional Connectivity*. Oxford, New York: Oxford University Press, 2017.

[29] M. X. Cohen, *Analyzing Neural Time Series Data: Theory and Practice*. 2014.

[30] A. Hoopes, J. S. Mora, A. V. Dalca, B. Fischl, and M. Hoffmann, "SynthStrip: Skull-stripping for any brain image", *NeuroImage*, vol. 260, p. 119 474, 2022.

[31] J. Juntu, J. Sijbers, D. Van Dyck, and J. Gielen, "Bias Field Correction for MRI Images", in *Computer Recognition Systems*, M. Kurzyński, E. Puchała, M. Woźniak, and A. żołnierek, Eds., Berlin, Heidelberg: Springer, 2005, pp. 543–551.

[32] J Talairach, "3-dimensional proportional system; an approach to cerebral imaging. co-planar stereotaxic atlas of the human brain", *Thieme*, pp. 1–122, 1988.

[33] C. J. Holmes, R. Hoge, L. Collins, R. Woods, A. W. Toga, and A. C. Evans, "Enhancement of MR Images Using Registration for Signal Averaging", *Journal of Computer Assisted Tomography*, vol. 22, no. 2, p. 324, 1998.

[34] A. C. Evans, D. Collins, and B. Milner, "An MRI-based stereotactic atlas from 250 young normal subjects", *The Society for Neuroscience Abstracts*, 1992.

[35] J Mazziotta, A Toga, A Evans, *et al.*, "A probabilistic atlas and reference system for the human brain: International Consortium for Brain Mapping (ICBM).", *Philosophical Transactions of the Royal Society of London. Series B*, vol. 356, no. 1412, pp. 1293–1322, 2001.

[36] J. Ashburner and K. J. Friston, "Unified segmentation", *NeuroImage*, vol. 26, no. 3, pp. 839–851, 2005.

[37] C. Gaser, R. Dahnke, P. M. Thompson, F. Kurth, E. Luders, and A. D. N. Initiative, *CAT – A Computational Anatomy Toolbox for the Analysis of Structural MRI Data*, 2022.

[38] N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, *et al.*, "Automated Anatomical Labeling of Activations in SPM Using a Macroscopic Anatomical Parcellation of the MNI MRI Single-Subject Brain", *NeuroImage*, vol. 15, no. 1, pp. 273–289, 2002.

[39] A. Schaefer, R. Kong, E. M. Gordon, *et al.*, "Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI", *Cerebral Cortex (New York, N.Y.: 1991)*, vol. 28, no. 9, pp. 3095–3114, 2018.

[40] E. H. Shen, C. C. Overly, and A. R. Jones, "The Allen Human Brain Atlas: Comprehensive gene expression mapping of the human brain", *Trends in Neurosciences*, vol. 35, no. 12, pp. 711–714, 2012.

[41] C. Destrieux, B. Fischl, A. Dale, and E. Halgren, "Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature", *NeuroImage*, vol. 53, no. 1, pp. 1–15, 2010.

[42] R. S. Desikan, F. Ségonne, B. Fischl, *et al.*, "An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest", *NeuroImage*, vol. 31, no. 3, pp. 968–980, 2006.

[43] J. L. R. Andersson and S. N. Sotiropoulos, "An integrated approach to correction for off-resonance effects and subject movement in diffusion MR imaging", *NeuroImage*, vol. 125, pp. 1063–1078, 2016.

[44] E. Kellner, B. Dhital, V. G. Kiselev, and M. Reisert, "Gibbs-ringing artifact removal based on local subvoxel-shifts", *Magnetic Resonance in Medicine*, vol. 76, no. 5, pp. 1574–1581, 2016.

[45] P. J. Basser, J. Mattiello, and D. LeBihan, "MR diffusion tensor spectroscopy and imaging", *Biophysical Journal*, vol. 66, no. 1, pp. 259–267, 1994.

[46] C. Pierpaoli and P. J. Basser, "Toward a quantitative assessment of diffusion anisotropy", *Magnetic Resonance in Medicine*, vol. 36, no. 6, pp. 893–906, 1996.

[47] P. J. Basser and C. Pierpaoli, "Microstructural and Physiological Features of Tissues Elucidated by Quantitative-Diffusion-Tensor MRI", *Journal of Magnetic Resonance, Series B*, vol. 111, no. 3, pp. 209–219, 1996.

[48] S. M. Smith, M. Jenkinson, H. Johansen-Berg, *et al.*, "Tract-based spatial statistics: Voxelwise analysis of multi-subject diffusion data", *NeuroImage*, vol. 31, no. 4, pp. 1487–1505, 2006.

[49] T. Sarwar, K. Ramamohanarao, and A. Zalesky, "Mapping connectomes with diffusion MRI: Deterministic or probabilistic tractography?", *Magnetic Resonance in Medicine*, vol. 81, no. 2, pp. 1368–1384, 2019.

[50] J.-D. Tournier, F. Calamante, D. G. Gadian, and A. Connelly, "Direct estimation of the fiber orientation density function from diffusion-weighted MRI data using spherical deconvolution", *NeuroImage*, vol. 23, no. 3, pp. 1176–1185, 2004.

[51] J.-D. Tournier, F. Calamante, and A. Connelly, "Robust determination of the fibre orientation distribution in diffusion MRI: Non-negativity constrained super-resolved spherical deconvolution", *NeuroImage*, vol. 35, no. 4, pp. 1459–1472, 2007.

[52] Y. Lilja, M. Ljungberg, G. Starck, K. Malmgren, B. Rydenhag, and D. T. Nilsson, "Visualizing Meyer's loop: A comparison of deterministic and probabilistic tractography", *Epilepsy Research*, vol. 108, no. 3, pp. 481–490, 2014.

[53] R. E. Smith, F. Calamante, and A. Connelly, "Mapping connectomes with diffusion MRI: Deterministic or probabilistic tractography?", *Magnetic Resonance in Medicine*, vol. 83, no. 3, pp. 787–790, 2020.

[54] J.-D. Tournier, R. Smith, D. Raffelt, *et al.*, "MRtrix3: A fast, flexible and open software framework for medical image processing and visualisation", *NeuroImage*, vol. 202, p. 116 137, 2019.

[55] R. E. Smith, J.-D. Tournier, F. Calamante, and A. Connelly, "SIFT: Spherical-deconvolution informed filtering of tractograms", *NeuroImage*, vol. 67, pp. 298–312, 2013.

[56] J. Hlinka, M. Paluš, M. Vejmelka, D. Mantini, and M. Corbetta, "Functional connectivity in resting-state fMRI: Is linear correlation sufficient?", *NeuroImage*, vol. 54, no. 3, pp. 2218–2225, 2011.

[57] B. Biswal, F. Zerrin Yetkin, V. M. Haughton, and J. S. Hyde, "Functional connectivity in the motor cortex of resting human brain using echo-planar mri", *Magnetic Resonance in Medicine*, vol. 34, no. 4, pp. 537–541, 1995.

[58] J. Turner, E. Damaraju, T. Van Erp, *et al.*, "A multi-site resting state fMRI study on the amplitude of low frequency fluctuations in schizophrenia", *Frontiers in Neuroscience*, vol. 7, 2013.

[59] Y. Han, J. Wang, Z. Zhao, *et al.*, "Frequency-dependent changes in the amplitude of low-frequency fluctuations in amnestic mild cognitive impairment: A resting-state fMRI study", *NeuroImage*, vol. 55, no. 1, pp. 287–295, 2011.

[60] W.-b. Guo, F. Liu, Z.-m. Xue, *et al.*, "Alterations of the amplitude of low-frequency fluctuations in treatment-resistant and treatment-response depression: A resting-state fMRI study", *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, vol. 37, no. 1, pp. 153–160, 2012.

[61] A. A. Ioannides, "Dynamic functional connectivity", *Current Opinion in Neurobiology*, vol. 17, no. 2, pp. 161–170, 2007.

[62] P. Comon, "Independent Component Analysis", in Elsevier, 1992, p. 29.

[63] C. F. Beckmann, M. DeLuca, J. T. Devlin, and S. M. Smith, "Investigations into resting-state connectivity using independent component analysis", *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 360, no. 1457, pp. 1001–1013, 2005.

[64] B. T. Thomas Yeo, F. M. Krienen, J. Sepulcre, *et al.*, "The organization of the human cerebral cortex estimated by intrinsic functional connectivity", *Journal of Neurophysiology*, vol. 106, no. 3, pp. 1125–1165, 2011.

[65] W. G. Walter, "The Location of cerebral Tumors by Electro-encephalography", *The Lancet*, vol. 228, no. 5893, pp. 305–308, 1936.

[66] W. Klimesch, "EEG alpha and theta oscillations reflect cognitive and memory performance: A review and analysis", *Brain Research Reviews*, vol. 29, no. 2, pp. 169–195, 1999.

[67] H. Berger, "Über das Elektrenkephalogramm des Menschen", *Archiv für Psychiatrie und Nervenkrankheiten*, vol. 87, no. 1, pp. 527–570, 1929.

[68] K. Takahashi, M. Saleh, R. Penn, and N. Hatsopoulos, "Propagating Waves in Human Motor Cortex", *Frontiers in Human Neuroscience*, vol. 5, 2011.

[69] B. E. Kilavik, M. Zaepffel, A. Brovelli, W. A. MacKay, and A. Riehle, "The ups and downs of beta oscillations in sensorimotor cortex", *Experimental Neurology*, vol. 245, pp. 15–26, 2013.

[70] C. Amo, L. de Santiago, R. Barea, A. López-Dorado, and L. Boquete, "Analysis of Gamma-Band Activity from Human EEG Using Empirical Mode Decomposition", *Sensors (Basel, Switzerland)*, vol. 17, no. 5, p. 989, 2017.

[71] C. Brunner, M. Billinger, M. Seeber, T. R. Mullen, and S. Makeig, "Volume Conduction Influences Scalp-Based Connectivity Estimates", *Frontiers in Computational Neuroscience*, vol. 10, 2016.

[72] R. D. Pascual-Marqui, C. M. Michel, and D. Lehmann, "Low resolution electromagnetic tomography: A new method for localizing electrical activity in the brain", *International Journal of Psychophysiology*, vol. 18, no. 1, pp. 49–65, 1994.

[73] S. Basharpoor, F. Heidari, and P. Molavi, "EEG coherence in theta, alpha, and beta bands in frontal regions and executive functions", *Applied Neuropsychology: Adult*, vol. 28, no. 3, pp. 310–317, 2021.

[74] J. Sui, T. Adali, Q. Yu, J. Chen, and V. D. Calhoun, "A review of multivariate methods for multimodal fusion of brain imaging data", *Journal of Neuroscience Methods*, vol. 204, no. 1, pp. 68–81, 2012.

[75] V. D. Calhoun and J. Sui, "Multimodal Fusion of Brain Imaging Data: A Key to Finding the Missing Link(s) in Complex Mental Illness", *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, vol. 1, no. 3, pp. 230–244, 2016.

[76] S. Liu, W. Cai, S. Liu, *et al.*, "Multimodal neuroimaging computing: A review of the applications in neuropsychiatric disorders", *Brain Informatics*, vol. 2, no. 3, pp. 167–180, 2015.

[77] J. D. Van Horn and A. W. Toga, "Human Neuroimaging as a "Big Data" Science", *Brain imaging and behavior*, vol. 8, no. 2, pp. 323–331, 2014.

[78] S. Ulmer and O. Jansen, *fMRI: Basics and Clinical Applications*. Springer Nature, 2020.

[79] K. J. Friston, L. Harrison, and W. Penny, "Dynamic causal modelling", *NeuroImage*, vol. 19, no. 4, pp. 1273–1302, 2003.

[80] L. Astolfi, F. Cincotti, D. Mattia, *et al.*, "Estimation of the effective and functional human cortical connectivity with structural equation modeling and directed transfer function applied to high-resolution EEG", *Magnetic Resonance Imaging*, vol. 22, no. 10, pp. 1457–1470, 2004.

[81] N. Correa, T. Adalı, and V. D. Calhoun, "Performance of blind source separation algorithms for fMRI analysis using a group ICA method", *Magnetic Resonance Imaging*, vol. 25, no. 5, pp. 684–694, 2007.

[82] A. Shoeibi, M. Khodatars, M. Jafari, *et al.*, "Diagnosis of brain diseases in fusion of neuroimaging modalities using deep learning: A review", *Information Fusion*, vol. 93, pp. 85–117, 2023.

[83] M. Loukas, C. Pennell, C. Groat, R. S. Tubbs, and A. A. Cohen-Gadol, "Korbinian Brodmann (1868-1918) and his contributions to mapping the cerebral cortex", *Neurosurgery*, vol. 68, no. 1, 6–11; discussion 11, 2011.

[84] R. A. Fisher and W. A. Mackenzie, "Studies in crop variation. II. The manurial response of different potato varieties", *The Journal of Agricultural Science*, vol. 13, no. 3, pp. 311–320, 1923.

[85] R. A. Fisher, "Design of Experiments", *British Medical Journal*, vol. 1, no. 3923, p. 554, 1936.

[86] J. Neyman, E. S. Pearson, and K. Pearson, "IX. On the problem of the most efficient tests of statistical hypotheses", *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, vol. 231, no. 694-706, pp. 289–337, 1932.

[87] K. J. Friston, A. P. Holmes, K. J. Worsley, J.-P. Poline, C. D. Frith, and R. S. J. Frackowiak, "Statistical parametric maps in functional imaging: A general linear approach", *Human Brain Mapping*, vol. 2, no. 4, pp. 189–210, 1994.

[88] J. R. Binder, "Current Controversies on Wernicke's Area and its Role in Language", *Current Neurology and Neuroscience Reports*, vol. 17, no. 8, p. 58, 2017.

[89] A. D. Friederici and I. Wartenburger, "Language and brain", *WIREs Cognitive Science*, vol. 1, no. 2, pp. 150–159, 2010.

[90] A. D. Friederici, "The Brain Basis of Language Processing: From Structure to Function", *Physiological Reviews*, vol. 91, no. 4, pp. 1357–1392, 2011.

[91] M. P. van den Heuvel and H. E. Hulshoff Pol, "Exploring the brain network: A review on resting-state fMRI functional connectivity", *European Neuropsychopharmacology*, vol. 20, no. 8, pp. 519–534, 2010.

[92] C. Davatzikos, "Why voxel-based morphometric analysis should be used with great caution when characterizing group differences", *NeuroImage*, vol. 23, no. 1, pp. 17–20, 2004.

[93] S. M. Smith and T. E. Nichols, "Threshold-free cluster enhancement: Addressing problems of smoothing, threshold dependence and localisation in cluster inference", *NeuroImage*, vol. 44, no. 1, pp. 83–98, 2009.

[94] D. Bzdok, "Classical Statistics and Statistical Learning in Imaging Neuroscience", *Frontiers in Neuroscience*, vol. 11, 2017.

[95] T. Hastie, J. Friedman, and R. Tibshirani, *The Elements of Statistical Learning*. New York, NY: Springer, 2001.

[96] "Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework", *Clinical Pharmacology & Therapeutics*, vol. 69, no. 3, pp. 89–95, 2001.

[97] R. Bellman, *Adaptive Control Processes: A Guided Tour*. Princeton University Press, 1961. JSTOR: `j.ctt183ph6v`.

[98] B. Mwangi, T. S. Tian, and J. C. Soares, "A Review of Feature Reduction Techniques in Neuroimaging", *Neuroinformatics*, vol. 12, no. 2, pp. 229–244, 2014.

[99] A. Sotiras, S. M. Resnick, and C. Davatzikos, "Finding imaging patterns of structural covariance via Non-Negative Matrix Factorization", *NeuroImage*, vol. 108, pp. 1–16, 2015.

[100] M. Gell, S. B. Eickhoff, A. Omidvarnia, *et al.*, *The Burden of Reliability: How Measurement Noise Limits Brain-Behaviour Predictions*, 2023.

[101] M. R. Arbabshirani, S. Plis, J. Sui, and V. D. Calhoun, "Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls", *NeuroImage*, vol. 145, no. Pt B, pp. 137–165, 2017.

[102] T. Wolfers, J. K. Buitelaar, C. F. Beckmann, B. Franke, and A. F. Marquand, "From estimating activation locality to predicting disorder: A review of pattern recognition for neuroimaging-based psychiatric diagnostics", *Neuroscience & Biobehavioral Reviews*, vol. 57, pp. 328–349, 2015.

[103] R. J. Janssen, J. Mourão-Miranda, and H. G. Schnack, "Making Individual Prognoses in Psychiatry Using Neuroimaging and Machine Learning", *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, vol. 3, no. 9, pp. 798–808, 2018.

[104] D. Sun, T. G. M. van Erp, P. M. Thompson, *et al.*, "Elucidating a Magnetic Resonance Imaging-Based Neuroanatomic Biomarker for Psychosis: Classification Analysis Using Probabilistic Brain Atlas and Machine Learning Algorithms", *Biological Psychiatry*, vol. 66, no. 11, pp. 1055–1060, 2009.

[105] D. L. Murdaugh, S. V. Shinkareva, H. R. Deshpande, J. Wang, M. R. Pennick, and R. K. Kana, "Differential Deactivation during Mentalizing and Classification of Autism Based on Default Mode Network Connectivity", *PLOS ONE*, vol. 7, no. 11, e50064, 19. 11. 2012.

[106] L. Q. Uddin, K. Supekar, C. J. Lynch, *et al.*, "Salience network-based classification and prediction of symptom severity in children with autism", *JAMA psychiatry*, vol. 70, no. 8, pp. 869–879, 2013.

[107] A. N. Goldstein-Piekarski, M. S. Korgaonkar, E. Green, *et al.*, "Human amygdala engagement moderated by early life stress exposure is a biobehavioral target for predicting recovery on antidepressants", *Proceedings of the National Academy of Sciences of the United States of America*, vol. 113, no. 42, pp. 11 955–11 960, 2016.

[108] A. Sarica, A. Cerasa, and A. Quattrone, "Random Forest Algorithm for the Classification of Neuroimaging Data in Alzheimer's Disease: A Systematic Review", *Frontiers in Aging Neuroscience*, vol. 9, 2017.

[109] F. Eitel, M.-A. Schulz, M. Seiler, H. Walter, and K. Ritter, "Promises and pitfalls of deep neural networks in neuroimaging-based psychiatric research", *Experimental Neurology*, vol. 339, p. 113 608, 2021.

[110] Q. D. Buchlak, N. Esmaili, J.-C. Leveque, C. Bennett, F. Farrokhi, and M. Piccardi, "Machine learning applications to neuroimaging for glioma detection and classification: An artificial intelligence augmented systematic review", *Journal of Clinical Neuroscience*, vol. 89, pp. 177–198, 2021.

[111] J. Wen, E. Thibeau-Sutre, M. Diaz-Melo, *et al.*, "Convolutional neural networks for classification of Alzheimer's disease: Overview and reproducible evaluation", *Medical Image Analysis*, vol. 63, p. 101 694, 2020.

[112] I. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, "Generative Adversarial Nets", in *Advances in Neural Information Processing Systems*, vol. 27, Curran Associates, Inc., 2014.

[113] W. Yu, B. Lei, M. K. Ng, A. C. Cheung, Y. Shen, and S. Wang, "Tensorizing GAN With High-Order Pooling for Alzheimer's Disease Assessment", *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 9, pp. 4945–4959, 2022.

[114] Z. Yang, I. M. Nasrallah, H. Shou, *et al.*, "A deep learning framework identifies dimensional representations of Alzheimer's Disease from brain structure", *Nature Communications*, vol. 12, no. 1, p. 7065, 2021.

[115] C. Han, L. Rundo, R. Araki, *et al.*, "Combining Noise-to-Image and Image-to-Image GANs: Brain MR Image Augmentation for Tumor Detection", *IEEE Access*, vol. 7, pp. 156 966–156 977, 2019.

[116] T. Xia, A. Chartsias, C. Wang, and S. A. Tsaftaris, "Learning to synthesise the ageing brain without longitudinal data", *Medical Image Analysis*, vol. 73, p. 102 169, 2021.

[117] C. Davatzikos, "Machine learning in neuroimaging: Progress and challenges", *NeuroImage*, vol. 197, pp. 652–656, 2019.

[118] J. H. Cole, R. E. Marioni, S. E. Harris, and I. J. Deary, "Brain age and other bodily 'ages': Implications for neuropsychiatry", *Molecular Psychiatry*, vol. 24, no. 2, pp. 266–281, 2019.

[119] A. A. Chen, J. C. Beer, N. J. Tustison, *et al.*, "Mitigating site effects in covariance for machine learning in neuroimaging data", *Human Brain Mapping*, vol. 43, no. 4, pp. 1179–1195, 2022.

[120] B. R. Bučková, D. Kala, J. Kořenek, *et al.*, "Structural connectivity-based predictors of cognitive impairment in stroke patients attributable to aging", *Plos one*, vol. 18, no. 4, e0280892, 2023.

[121] L. Snoek, S. Miletić, and H. S. Scholte, "How to control for confounds in decoding analyses of neuroimaging data", *NeuroImage*, vol. 184, pp. 741–760, 2019.

[122] A. F. Marquand, I. Rezek, J. Buitelaar, and C. F. Beckmann, "Understanding Heterogeneity in Clinical Cohorts Using Normative Models: Beyond Case-Control Studies", *Biological Psychiatry*, vol. 80, no. 7, pp. 552–561, 2016.

[123] A. F. Marquand, S. M. Kia, M. Zabihi, T. Wolfers, J. K. Buitelaar, and C. F. Beckmann, "Conceptualizing mental disorders as deviations from normative functioning", *Molecular Psychiatry*, vol. 24, no. 10, pp. 1415–1424, 2019.

[124] C. J. Fraza, R. Dinga, C. F. Beckmann, and A. F. Marquand, "Warped Bayesian linear regression for normative modelling of big data", *NeuroImage*, vol. 245, p. 118 715, 2021.

[125] R. a. I. Bethlehem, J. Seidlitz, S. R. White, *et al.*, "Brain charts for the human lifespan", *Nature*, vol. 604, no. 7906, pp. 525–533, 2022.

[126] S. M. Kia, H. Huijsdens, S. Rutherford, *et al.*, *Federated Multi-Site Normative Modeling using Hierarchical Bayesian Regression*, 2021.

[127] S. Rutherford, C. Fraza, R. Dinga, *et al.*, "Charting Brain Growth and Aging at High Spatial Precision", Cold Spring Harbor Laboratory, Tech. Rep., 2021, ch. New Results, p. 2021.08.08.455487.

[128] W. H. L. Pinaya, C. Scarpazza, R. Garcia-Dias, *et al.*, "Using normative modelling to detect disease progression in mild cognitive impairment and Alzheimer's disease in a cross-sectional multi-cohort study", *Scientific Reports*, vol. 11, no. 1, p. 15 746, 2021.

[129] T. Wolfers, J. Rokicki, D. Alnæs, *et al.*, "Replicating extensive brain structural heterogeneity in individuals with schizophrenia and bipolar disorder", *Human Brain Mapping*, vol. 42, no. 8, pp. 2546–2555, 2021.

[130] M. Zabihi, M. Oldehinkel, T. Wolfers, *et al.*, "Dissecting the Heterogeneous Cortical Anatomy of Autism Spectrum Disorder Using Normative Models", *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, vol. 4, no. 6, pp. 567–578, 2019.

[131] B. Bučková, M. Brunovský, M. Bareš, and J. Hlinka, "Predicting sex from EEG: Validity and generalizability of deep-learning-based interpretable classifier", *Frontiers in Neuroscience*, vol. 14, 2020.

[132] T. Carlson, E. Goddard, D. M. Kaplan, C. Klein, and J. B. Ritchie, "Ghosts in machine learning for cognitive neuroscience: Moving from data to theory", *NeuroImage*, vol. 180, pp. 88–100, 2018.

[133] M. J. A. M. van Putten, S. Olbrich, and M. Arns, "Predicting sex from brain rhythms with deep learning", *Scientific Reports*, vol. 8, no. 1, p. 3069, 2018.

[134] R. Thibodeau, R. Jorgensen, and S. Kim, "Depression, Anxiety, and Resting Frontal EEG Asymmetry: A Meta-Analytic Review", *Journal of Abnormal Psychology*, vol. 115, no. 4, pp. 715–729, 2006.

[135] S. Olbrich, R. van Dinteren, and M. Arns, "Personalized Medicine: Review and Perspectives of Promising Baseline EEG Biomarkers in Major Depressive Disorder and Attention Deficit Hyperactivity Disorder", *Neuropsychobiology*, vol. 72, no. 3-4, pp. 229–240, 2015.

[136] A. L. Lieber and L. S. Prichep, "Diagnosis and subtyping of depressive disorders by quantitative electroencephalography: I. Discriminant analysis of selected variables in untreated depressives", *Hillside Journal of Clinical Psychiatry*, vol. 10, no. 1, pp. 71–83, 1988.

[137] V. Knott, C. Mahoney, S. Kennedy, and K. Evans, "EEG power, frequency, asymmetry and coherence in male depression", *Psychiatry Research: Neuroimaging*, vol. 106, no. 2, pp. 123–140, 2001.

[138] A. S. Widge, M. T. Bilge, R. Montana, *et al.*, "Electroencephalographic Biomarkers for Treatment Response Prediction in Major Depressive Illness: A Meta-Analysis", *American Journal of Psychiatry*, vol. 176, no. 1, pp. 44–56, 2018.

[139] M. Bareš, M. Brunovský, T. Novák, *et al.*, "The change of prefrontal QEEG theta cordance as a predictor of response to bupropion treatment in patients who had failed to respond to previous antidepressant treatments", *European Neuropsychopharmacology*, vol. 20, no. 7, pp. 459–466, 2010.

[140] M. Bareš, M. Brunovský, T. Novák, *et al.*, "QEEG Theta Cordance in the Prediction of Treatment Outcome to Prefrontal Repetitive Transcranial Magnetic Stimulation or Venlafaxine ER in Patients With Major Depressive Disorder", *Clinical EEG and Neuroscience*, vol. 46, no. 2, pp. 73–80, 2015.

[141] M. Bareš, T. Novák, M. Kopeček, M. Brunovský, P. Stopková, and C. Höschl, "The effectiveness of prefrontal theta cordance and early reduction of depressive symptoms in the prediction of antidepressant treatment outcome in patients with resistant depression: Analysis of naturalistic data", *European Archives of Psychiatry and Clinical Neuroscience*, vol. 265, no. 1, pp. 73–82, 2015.

[142] N. Bigdely-Shamlo, T. Mullen, C. Kothe, K.-M. Su, and K. A. Robbins, "The PREP pipeline: Standardized preprocessing for large-scale EEG analysis", *Frontiers in Neuroinformatics*, vol. 9, 2015.

[143] A. Delorme and S. Makeig, "EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis", *Journal of Neuroscience Methods*, vol. 1, no. 134, pp. 9–21, 2004.

[144] T. Mullen, C. Kothe, Y. M. Chi, *et al.*, "Real-time modeling and 3D visualization of source dynamics and connectivity using wearable EEG", in *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2013, pp. 2184–2187.

[145] M. Plechawska-Wojcik, M. Kaczorowska, and D. Zapala, "The Artifact Subspace Reconstruction (ASR) for EEG Signal Correction. A Comparative Study", in *Information Systems Architecture and Technology: Proceedings of 39th International Conference on Information Systems Architecture and Technology – ISAT 2018*, Cham: Springer International Publishing, 2019, pp. 125–135.

[146] MATLAB, *version 9.5.0 (R2018b)*. Natick, Massachusetts: The MathWorks Inc., 2018.

[147] N. Vogt, "Machine learning in neuroscience", *Nature Methods*, vol. 15, no. 1, pp. 33–33, 2018.

[148] W. Samek, T. Wiegand, and K.-R. Müller, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models", *arXiv:1708.08296*, 2017. arXiv: 1708.08296.

[149] C. P. Langlotz, B. Allen, B. J. Erickson, *et al.*, "A Roadmap for Foundational Research on Artificial Intelligence in Medical Imaging", *Radiology*, vol. 291, no. 3, pp. 781–791, 2019.

[150] J. I. Glaser, A. S. Benjamin, R. Farhoodi, and K. P. Kording, "The roles of supervised machine learning in systems neuroscience", *Progress in Neurobiology*, vol. 175, pp. 126–137, 2019.

[151] R. Whelan and H. Garavan, "When Optimism Hurts: Inflated Predictions in Psychiatric Neuroimaging", *Biological Psychiatry*, vol. 75, no. 9, pp. 746–748, 2014.

[152] N. Yahata, K. Kasai, and M. Kawato, "Computational neuroscience approach to biomarkers and treatments for mental disorders", *Psychiatry and Clinical Neurosciences*, vol. 71, no. 4, pp. 215–237, 2017.

[153] T. Maggipinto, R. Bellotti, N. Amoroso, *et al.*, "Dti measurements for alzheimer's classification", *Physics in Medicine & Biology*, vol. 62, no. 6, p. 2361, 2017.

[154] E. C. Wade and D. V. Iosifescu, "Using Electroencephalography for Treatment Guidance in Major Depressive Disorder", *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, vol. 1, no. 5, pp. 411–422, 2016.

[155] M. Arns, W. H. Drinkenburg, P. B. Fitzgerald, and J. L. Kenemans, "Neurophysiological predictors of non-response to rTMS in depression", *Brain Stimulation*, vol. 5, no. 4, pp. 569–576, 2012.

[156] R. V. Krishnamurthi, T. Ikeda, and V. L. Feigin, "Global, Regional and Country-Specific Burden of Ischaemic Stroke, Intracerebral Haemorrhage and Subarachnoid Haemorrhage: A Systematic Analysis of the Global Burden of Disease Study 2017", *Neuroepidemiology*, vol. 54, no. 2, pp. 171–179, 2020.

[157] J. Kim, T. Thayabaranathan, G. A. Donnan, *et al.*, "Global Stroke Statistics 2019", *International Journal of Stroke*, p. 1 747 493 020 909 545, 2020.

[158] WHO Monica Project Principal Investigators, "The world health organization monica project (monitoring trends and determinants in cardiovascular disease): A major international collaboration", *Journal of Clinical Epidemiology*, vol. 41, no. 2, pp. 105–114, 1988.

[159] Davenport R.J., Dennis M.S., Wellwood I., and Warlow C.P., "Complications After Acute Stroke", *Stroke*, vol. 27, no. 3, pp. 415–420, 1996.

[160] H. S. Jørgensen, H. Nakayama, H. O. Raaschou, and T. S. Olsen, "Recovery of walking function in stroke patients: The copenhagen stroke study", *Archives of Physical Medicine and Rehabilitation*, vol. 76, no. 1, pp. 27–32, 1995.

[161] Barker William H. and Mullooly John P., "Stroke in a Defined Elderly Population, 1967-1985", *Stroke*, vol. 28, no. 2, pp. 284–290, 1997.

[162] P. Langhorne, F. Coupar, and A. Pollock, "Motor recovery after stroke: A systematic review", *The Lancet Neurology*, vol. 8, no. 8, pp. 741–754, 2009.

[163] P. Langhorne, J. Bernhardt, and G. Kwakkel, "Stroke rehabilitation", *The Lancet*, vol. 377, no. 9778, pp. 1693–1702, 2011.

[164] C. Rosso and J.-C. Lamy, "Prediction of motor recovery after stroke: Being pragmatic or innovative?", *Current Opinion in Neurology*, vol. 33, no. 4, pp. 482–487, 2020.

[165] Stinear Cathy M., Byblow Winston D., Ackerley Suzanne J., Barber P. Alan, and Smith Marie-Claire, "Predicting Recovery Potential for Individual Stroke Patients Increases Rehabilitation Efficiency", *Stroke*, vol. 48, no. 4, pp. 1011–1019, 2017.

[166] T. K. Tatemichi, D. W. Desmond, Y. Stern, M. Paik, M. Sano, and E. Bagiella, "Cognitive impairment after stroke: Frequency, patterns, and relationship to functional abilities.", *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 57, no. 2, pp. 202–207, 1994.

[167] M. L. Seghier, E. Patel, S. Prejawa, *et al.*, "The PLORAS Database: A data repository for Predicting Language Outcome and Recovery After Stroke", *NeuroImage*, vol. 124, pp. 1208–1212, 2016.

[168] G. M. S. Nys, M. J. E. van Zandvoort, P. L. M. de Kort, B. P. W. Jansen, L. J. Kappelle, and E. H. F. de Haan, "Restrictions of the Mini-Mental State Examination in acute stroke", *Archives of Clinical Neuropsychology*, vol. 20, no. 5, pp. 623–629, 2005.

[169] R. das Nair, H. Cogger, E. Worthington, and N. B. Lincoln, "Cognitive rehabilitation for memory deficits after stroke", *Cochrane Database of Systematic Reviews*, no. 9, 2016.

[170] T. Loetscher, K.-J. Potter, D. Wong, and R. das Nair, "Cognitive rehabilitation for attention deficits following stroke", *Cochrane Database of Systematic Reviews*, no. 11, 2019.

[171] A. Bowen, C. Hazelton, A. Pollock, and N. B. Lincoln, "Cognitive rehabilitation for spatial neglect following stroke", *Cochrane Database of Systematic Reviews*, no. 7, 2013.

[172] K. D. Cicerone, Y. Goldin, K. Ganci, *et al.*, "Evidence-Based Cognitive Rehabilitation: Systematic Review of the Literature From 2009 Through 2014", *Archives of Physical Medicine and Rehabilitation*, vol. 100, no. 8, pp. 1515–1533, 2019.

[173] S. Rajsic, H. Gothe, H. H. Borba, *et al.*, "Economic burden of stroke: A systematic review on post-stroke care", *The European Journal of Health Economics*, vol. 20, no. 1, pp. 107–134, 2019.

[174] Y. Wang, G. Liu, D. Hong, F. Chen, X. Ji, and G. Cao, "White Matter Injury in Ischemic Stroke", *Progress in neurobiology*, vol. 141, pp. 45–60, 2016.

[175] Schaapsmeerders Pauline, Tuladhar Anil M., Arntz Renate M., *et al.*, "Remote Lower White Matter Integrity Increases the Risk of Long-Term Cognitive Impairment After Ischemic Stroke in Young Adults", *Stroke*, vol. 47, no. 10, pp. 2517–2525, 2016.

[176] R. Dacosta-Aguayo, M. Graña, M. Fernández-Andújar, *et al.*, "Structural Integrity of the Contralesional Hemisphere Predicts Cognitive Impairment in Ischemic Stroke at Three Months", *PLOS ONE*, vol. 9, no. 1, e86119, 2014.

[177] M. R. Etherton, O. Wu, P. Cougo, *et al.*, "Integrity of normal-appearing white matter and functional outcomes after acute ischemic stroke", *Neurology*, vol. 88, no. 18, pp. 1701–1708, 2017.

[178] M. Fernández-Andújar, F. Doornink, R. Dacosta-Aguayo, *et al.*, "Remote thalamic microstructural abnormalities related to cognitive function in ischemic stroke patients", *Neuropsychology*, vol. 28, no. 6, pp. 984–996, 2014.

[179] Z. Keser, E. L. Meier, M. D. Stockbridge, B. L. Breining, R. Sebastian, and A. E. Hillis, "Thalamic Nuclei and Thalamocortical Pathways After Left Hemispheric Stroke and Their Association with Picture Naming", *Brain Connectivity*, vol. 11, no. 7, pp. 553–565, 2021.

[180] E. Bullmore and O. Sporns, "The economy of brain network organization", *Nature Reviews Neuroscience*, vol. 13, no. 5, pp. 336–349, 2012.

[181] S. V. Buldyrev, R. Parshani, G. Paul, H. E. Stanley, and S. Havlin, "Catastrophic cascade of failures in interdependent networks", *Nature*, vol. 464, no. 7291, pp. 1025–1028, 2010.

[182] A. Fornito, A. Zalesky, and M. Breakspear, "The connectomics of brain disorders", *Nature Reviews Neuroscience*, vol. 16, no. 3, pp. 159–172, 2015.

[183] L. Bonilha, T. Nesland, C. Rorden, P. Fillmore, R. P. Ratnayake, and J. Fridriksson, *Mapping Remote Subcortical Ramifications of Injury after Ischemic Strokes*, Research Article, 2014.

[184] G. Yourganov, J. Fridriksson, C. Rorden, E. Gleichgerrcht, and L. Bonilha, "Multivariate Connectome-Based Symptom Mapping in Post-Stroke Patients: Networks Supporting Language and Speech", *Journal of Neuroscience*, vol. 36, no. 25, pp. 6668–6679, 2016.

[185] A. Kuceyeski, B. B. Navi, H. Kamel, *et al.*, "Structural connectome disruption at baseline predicts 6-months post-stroke outcome", *Human Brain Mapping*, vol. 37, no. 7, pp. 2587–2601, 2016.

[186] T. M. H. Hope, A. P. Leff, and C. J. Price, "Predicting language outcomes after stroke: Is structural disconnection a useful predictor?", *NeuroImage: Clinical*, vol. 19, pp. 22–29, 2018.

[187] S. Saxena, Z. Keser, C. Rorden, *et al.*, "Disruptions of the Human Connectome Associated With Hemispatial Neglect", *Neurology*, vol. 98, no. 2, e107–e114, 2022.

[188] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D. U. Hwang, "Complex networks: Structure and dynamics", *Physics Reports*, vol. 424, no. 4, pp. 175–308, 2006.

[189] M. Rubinov and O. Sporns, "Complex network measures of brain connectivity: Uses and interpretations.", eng, *NeuroImage*, vol. 52, pp. 1059–69, 3 2010.

[190] M. E. J. Newman, "Assortative mixing in networks", *Physical Review Letters*, vol. 89, no. 20, 2002, ISSN: 1079-7114.

[191] V. Colizza, A. Flammini, M. A. Serrano, and A. Vespignani, "Detecting rich-club ordering in complex networks", *Nature Physics*, vol. 2, no. 2, pp. 110–115, Feb. 2006, ISSN: 1745-2481.

[192] J.-S. Lim and D.-W. Kang, "Stroke Connectome and Its Implications for Cognitive and Behavioral Sequela of Stroke", *Journal of Stroke*, vol. 17, no. 3, pp. 256–267, 2015.

[193] C. Grefkes and G. R. Fink, "Connectivity-based approaches in stroke and recovery of function", *The Lancet Neurology*, vol. 13, no. 2, pp. 206–216, 2014.

[194] A. R. Carter, G. L. Shulman, and M. Corbetta, "Why use a connectivity-based approach to study stroke and recovery of function?", *NeuroImage*, vol. 62, no. 4, pp. 2271–2280, 2012.

[195] K. Horáková, H. Štěpánková, O. Bezdíček, and M. Kopeček, "Kontrolované učení ve starším věku", *Československá psychologie*, vol. 61, no. 3, pp. 213–229, 2017.

[196] O. Bezdicek, J. Lukavsky, H. Stepankova, *et al.*, "The Prague Stroop Test: Normative standards in older Czech adults and discriminative validity for mild cognitive impairment in Parkinson's disease", *Journal of Clinical and Experimental Neuropsychology*, vol. 37, no. 8, pp. 794–807, 2015.

[197] T. Nikolai, H. Stepankova, M. Kopecek, Z. Sulc, M. Vyhnalek, and O. Bezdicek, "The Uniform Data Set, Czech Version: Normative Data in Older Adults from an International Perspective", *Journal of Alzheimer's Disease*, vol. 61, no. 3, pp. 1233–1240, 2018.

[198] K. Drozdová, H. Stepankova Georgi, J. Lukavsky, O. Bezdicek, and M. Kopecek, "Normative Data for the Rey- Osterrieth Complex Figure Test in Older Czech Adults", *Česká a Slovenská neurologie a neurochirurgie*, vol. 78/111, pp. 542–549, 2015.

[199] E. Warrington and M. James, *The Visual Object and Space Perception Battery*. Bury St Edmunds : Thames Valley Test Company, 1991.

[200] M. W. Woolrich, S. Jbabdi, B. Patenaude, *et al.*, "Bayesian analysis of neuroimaging data in FSL", *NeuroImage*, vol. 45, no. 1, Supplement 1, S173–S186, 2009.

[201] J. Veraart, D. S. Novikov, D. Christiaens, B. Ades-aron, J. Sijbers, and E. Fieremans, "Denoising of diffusion MRI using random matrix theory", *NeuroImage*, vol. 142, pp. 394–406, 2016.

[202] K. G. Schilling, J. Blaber, Y. Huo, *et al.*, "Synthesized b0 for diffusion distortion correction (Synb0-DisCo)", *Magnetic Resonance Imaging*, vol. 64, pp. 62–70, 2019.

[203] C. M. W. Tax, B. Jeurissen, S. B. Vos, M. A. Viergever, and A. Leemans, "Recursive calibration of the fiber response function for spherical deconvolution of diffusion MRI data", *NeuroImage*, vol. 86, pp. 67–80, 2014.

[204] J.-D. Tournier, F. Calamante, and A. Connelly, "MRtrix: Diffusion tractography in crossing fiber regions", *International Journal of Imaging Systems and Technology*, vol. 22, no. 1, pp. 53–66, 2012.

[205] N. Sinha, Y. Wang, J. Dauwels, *et al.*, "Computer modelling of connectivity change suggests epileptogenesis mechanisms in idiopathic generalised epilepsy", *NeuroImage: Clinical*, vol. 21, p. 101 655, 2019.

[206] Y. Benjamini and Y. Hochberg, "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing", *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.

[207] J. Alstott, P. Panzarasa, M. Rubinov, E. T. Bullmore, and P. E. Vértes, "A unifying framework for measuring weighted rich clubs", *Scientific Reports*, vol. 4, no. 1, p. 7258, Dec. 2014, ISSN: 2045-2322.

[208] K. A. Kuznetsova, S. M. Maniega, S. J. Ritchie, *et al.*, "Brain white matter structure and information processing speed in healthy older age", *Brain Structure and Function*, vol. 221, no. 6, pp. 3223–3235, 2016.

[209] J. Barnes, G. R. Ridgway, J. Bartlett, *et al.*, "Head size, age and gender adjustment in MRI studies: A necessary nuisance?", *NeuroImage*, vol. 53, no. 4, pp. 1244–1255, 2010.

[210] C. S. Hyatt, M. M. Owens, M. L. Crowe, N. T. Carter, D. R. Lynam, and J. D. Miller, "The quandary of covarying: A brief review and empirical examination of covariate use in structural neuroimaging studies on psychological variables", *NeuroImage*, vol. 205, p. 116 225, 2020.

[211] Munsch Fanny, Sagnier Sharmila, Asselineau Julien, *et al.*, "Stroke Location Is an Independent Predictor of Cognitive Outcome", *Stroke*, vol. 47, no. 1, pp. 66–73, 2016.

[212] L. Zhao, A. Wong, Y. Luo, *et al.*, "The Additional Contribution of White Matter Hyperintensity Location to Post-stroke Cognitive Impairment: Insights From a Multiple-Lesion Symptom Mapping Study", *Frontiers in Neuroscience*, vol. 12, 2018.

[213] G. Nys, M. Van Zandvoort, P. De Kort, B. Jansen, E. De Haan, and L. Kappelle, "Cognitive disorders in acute stroke: Prevalence and clinical determinants", *Cerebrovascular Diseases*, vol. 23, no. 5-6, pp. 408–416, 2007.

[214] A. Jaillard, S. Grand, J. F. Le Bas, and M. Hommel, "Predicting cognitive dysfunctioning in nondemented patients early after stroke", *Cerebrovascular Diseases*, vol. 29, no. 5, pp. 415–423, 2010.

[215] Zamboni Giovanna, Griffanti Ludovica, Jenkinson Mark, *et al.*, "White Matter Imaging Correlates of Early Cognitive Impairment Detected by the Montreal Cognitive Assessment After Transient Ischemic Attack and Minor Stroke", *Stroke*, vol. 48, no. 6, pp. 1539–1547, 2017.

[216] J. Del Gaizo, J. Fridriksson, G. Yourganov, *et al.*, "Mapping Language Networks Using the Structural and Dynamic Brain Connectomes", *eNeuro*, vol. 4, no. 5, 2017.

[217] van der Flier Wiesje M., van Straaten Elizabeth C.W., Barkhof Frederik, *et al.*, "Small Vessel Disease and General Cognitive Function in Nondisabled Elderly", *Stroke*, vol. 36, no. 10, pp. 2116–2120, 2005.

[218] G. Beaudet, A. Tsuchida, L. Petit, *et al.*, "Age-related changes of peak width skeletonized mean diffusivity (psmd) across the adult lifespan: A multi-cohort study", *Frontiers in psychiatry*, vol. 11, p. 342, 2020.

[219] A. Behler, J. Kassubek, and H.-P. Müller, "Age-Related Alterations in DTI Metrics in the Human Brain—Consequences for Age Correction", *Frontiers in Aging Neuroscience*, vol. 13, 2021.

[220] T. D. Faizy, C. Thaler, G. Broocks, *et al.*, "The Myelin Water Fraction Serves as a Marker for Age-Related Myelin Alterations in the Cerebral White Matter – A Multiparametric MRI Aging Study", *Frontiers in Neuroscience*, vol. 14, 2020.

[221] C. J. Molloy, S. Nugent, and A. L. Bokde, "Alterations in diffusion measures of white matter integrity associated with healthy aging", *The Journals of Gerontology: Series A*, vol. 76, no. 6, pp. 945–954, 2021.

[222] D. Pustina, H. B. Coslett, L. Ungar, *et al.*, "Enhanced estimations of post-stroke aphasia severity using stacked multimodal predictions", *Human Brain Mapping*, vol. 38, no. 11, pp. 5603–5615, 2017.

[223] L. Shi, D. Wang, W. C. W. Chu, *et al.*, "Abnormal Organization of White Matter Network in Patients with No Dementia after Ischemic Stroke", *PLOS ONE*, vol. 8, no. 12, e81388, 2013.

[224] J. Zhang, Y. Zhang, L. Wang, *et al.*, "Disrupted structural and functional connectivity networks in ischemic stroke patients", *Neuroscience*, vol. 364, pp. 212–225, 2017.

[225] B. Rehák Bučková, J. Mareš, A. Škoch, *et al.*, "Multimodal-neuroimaging machine-learning analysis of motor disability in multiple sclerosis", *Brain Imaging and Behavior*, vol. 17, no. 1, pp. 18–34, 2023.

[226] M. T. Wallin, W. J. Culpepper, E. Nichols, *et al.*, "Global, regional, and national burden of multiple sclerosis 1990–2016: A systematic analysis for the Global Burden of Disease Study 2016", *The Lancet Neurology*, vol. 18, no. 3, pp. 269–285, 2019.

[227] E. Leray, T. Moreau, A. Fromont, and G. Edan, "Epidemiology of multiple sclerosis", *Revue Neurologique*, vol. 172, no. 1, pp. 3–13, 2016.

[228] J. M. Burschka, P. M. Keune, U. Menge, U. H.-v. Oy, P. Oschmann, and O. Hoos, "An exploration of impaired walking dynamics and fatigue in Multiple Sclerosis", *BMC Neurology*, vol. 12, no. 1, p. 161, 2012.

[229] C. Giannì, L. Prosperini, J. Jonsdottir, and D. Cattaneo, "A systematic review of factors associated with accidental falls in people with multiple sclerosis: A meta-analytic approach", *Clinical Rehabilitation*, vol. 28, no. 7, pp. 704–716, 2014.

[230] C Heesen, J Böhm, C Reich, J Kasper, M Goebel, and SM Gold, "Patient perception of bodily functions in multiple sclerosis: Gait and visual function are the most valuable", *Multiple Sclerosis Journal*, vol. 14, no. 7, pp. 988–991, 2008.

[231] P. van Asch, "Impact of Mobility Impairment in Multiple Sclerosis 2 - Patients' Perspectives", *European Neurological Review*, vol. 6, no. 2, pp. 115–120, 2011.

[232] N. G. LaRocca, "Impact of Walking Impairment in Multiple Sclerosis", *The Patient: Patient-Centered Outcomes Research*, vol. 4, no. 3, pp. 189–201, 2011.

[233] J. F. Kurtzke, "Rating neurologic impairment in multiple sclerosis: An expanded disability status scale (EDSS).", *Neurology*, vol. 33, no. 11, pp. 1444–1452, 1983.

[234] S. Meyer-Moock, Y.-S. Feng, M. Maeurer, F.-W. Dippel, and T. Kohlmann, "Systematic literature review and validity evaluation of the Expanded Disability Status Scale (EDSS) and the Multiple Sclerosis Functional Composite (MSFC) in patients with multiple sclerosis", *BMC Neurology*, vol. 14, no. 1, p. 58, 2014.

[235] B. C. Kieseier and C. Pozzilli, "Assessing walking disability in multiple sclerosis", *Multiple Sclerosis Journal*, vol. 18, no. 7, pp. 914–924, 2012.

[236] J. Hobart, J. Freeman, and A. Thompson, "Kurtzke scales revisited: The application of psychometric methods to clinical intuition", *Brain*, vol. 123, no. 5, pp. 1027–1040, 2000.

[237] J. Hobart, D. Lamping, R. Fitzpatrick, A. Riazi, and A. Thompson, "The Multiple Sclerosis Impact Scale (MSIS-29)A new patient-based outcome measure", *Brain*, vol. 124, no. 5, pp. 962–973, 2001.

[238] J. Hobart, A. Riazi, D. Lamping, R. Fitzpatrick, and A. Thompson, "Measuring the impact of MS on walking ability: The 12-Item MS Walking Scale (MSWS-12)", *Neurology*, vol. 60, no. 1, pp. 31–36, 2003.

[239] J. S. Fischer, R. A. Rudick, G. R. Cutter, and S. C. Reingold, "The Multiple Sclerosis Functional Composite measure (MSFC): An integrated approach to MS clinical outcome assessment", *Multiple Sclerosis Journal*, vol. 5, no. 4, pp. 244–250, 1999.

[240] D. Podsiadlo and S. Richardson, "The Timed "Up & Go": A Test of Basic Functional Mobility for Frail Elderly Persons", *Journal of the American Geriatrics Society*, vol. 39, no. 2, pp. 142–148, 1991.

[241] K. Berg, S. Wood-Dauphine, J. I. Williams, and D. Gayton, "Measuring balance in the elderly: Preliminary development of an instrument", *Physiotherapy Canada*, vol. 41, no. 6, pp. 304–311, 2009.

[242] D. Cattaneo, A. Regola, and M. Meotti, "Validity of six balance disorders scales in persons with multiple sclerosis", *Disability and Rehabilitation*, vol. 28, no. 12, pp. 789–795, 2006.

[243] D. Cattaneo, J. Jonsdottir, and S. Repetti, "Reliability of four scales on balance disorders in persons with multiple sclerosis", *Disability and Rehabilitation*, vol. 29, no. 24, pp. 1920–1925, 2007.

[244] Y. C. Learmonth, L. Paul, A. K. McFadyen, P. Mattison, and L. Miller, "Reliability and clinical significance of mobility and balance assessments in multiple sclerosis", *International Journal of Rehabilitation Research*, vol. 35, no. 1, p. 69, 2012.

[245] M. Filippi, M. A. Rocca, and M. Rovaris, "Clinical trials and clinical practice in multiple sclerosis: Conventional and emerging magnetic resonance imaging technologies", *Current Neurology and Neuroscience Reports*, vol. 2, no. 3, pp. 267–276, 2002.

[246] P. D. Molyneux, G. J. Barker, F. Barkhof, *et al.*, "Clinical-MRI correlations in a European trial of interferon beta-1b in secondary progressive MS", *Neurology*, vol. 57, no. 12, pp. 2191–2197, 2001.

[247] M. Filippi, M. Cercignani, M. Inglese, M. A. Horsfield, and G. Comi, "Diffusion tensor magnetic resonance imaging in multiple sclerosis", *Neurology*, vol. 56, no. 3, pp. 304–311, 2001.

[248] N. D. Stefano, S. Narayanan, G. S. Francis, *et al.*, "Evidence of Axonal Damage in the Early Stages of Multiple Sclerosis and Its Relevance to Disability", *Archives of Neurology*, vol. 58, no. 1, pp. 65–70, 2001.

[249] M. A. Rocca, G. Comi, and M. Filippi, "The Role of T1-Weighted Derived Measures of Neurodegeneration for Assessing Disability Progression in Multiple Sclerosis", *Frontiers in Neurology*, vol. 8, p. 433, 2017.

[250] C. Enzinger and F. Fazekas, "Measuring Gray Matter and White Matter Damage in MS: Why This is Not Enough", *Frontiers in Neurology*, vol. 6, 2015.

[251] E. Sbardella, N. Petsas, F. Tona, *et al.*, "Assessing the Correlation between Grey and White Matter Damage with Motor and Cognitive Impairment in Multiple Sclerosis Patients", *PLOS ONE*, vol. 8, no. 5, e63250, 2013.

[252] S. Cader, H. Johansen-Berg, M. Wylezinska, *et al.*, "Discordant white matter N-acetylasparate and diffusion MRI measures suggest that chronic metabolic dysfunction contributes to axonal pathology in multiple sclerosis", *NeuroImage*, vol. 36, no. 1, pp. 19–27, 2007.

[253] A. Giorgio, J. Palace, H. Johansen-Berg, *et al.*, "Relationships of brain white matter microstructure with clinical and MR measures in relapsing-remitting multiple sclerosis", *Journal of Magnetic Resonance Imaging*, vol. 31, no. 2, pp. 309–316, 2010.

[254] Y. Liu, Y. Duan, Y. He, *et al.*, "Whole brain white matter changes revealed by multiple diffusion metrics in multiple sclerosis: A TBSS study", *European Journal of Radiology*, vol. 81, no. 10, pp. 2826–2832, 2012.

[255] S. D. Roosendaal, J. J. G. Geurts, H. Vrenken, *et al.*, "Regional DTI differences in multiple sclerosis patients", *NeuroImage*, vol. 44, no. 4, pp. 1397–1403, 2009.

[256] O. Ciccarelli, D. J. Werring, C. A. Wheeler-Kingshott, *et al.*, "Investigation of MS normal-appearing brain using diffusion tensor MRI with clinical correlations", *Neurology*, vol. 56, no. 7, pp. 926–933, 2001.

[257] C. M. Griffin, D. T. Chard, O. Ciccarelli, *et al.*, "Diffusion tensor imaging in early relapsing-remitting multiple sclerosis", *Multiple Sclerosis Journal*, vol. 7, no. 5, pp. 290–297, 2001.

[258] K. C. Kern, J. Sarcona, M. Montag, B. S. Giesser, and N. L. Sicotte, "Corpus callosal diffusivity predicts motor impairment in relapsing–remitting multiple sclerosis: A TBSS and tractography study", *NeuroImage*, vol. 55, no. 3, pp. 1169–1177, 2011.

[259] W. Tian, T. Zhu, J. Zhong, *et al.*, "Progressive decline in fractional anisotropy on serial DTI examinations of the corpus callosum: A putative marker of disease activity and progression in SPMS", *Neuroradiology*, vol. 54, no. 4, pp. 287–297, 2012.

[260] S. Llufriu, Y. Blanco, E. Martinez-Heras, *et al.*, "Influence of Corpus Callosum Damage on Cognition and Physical Disability in Multiple Sclerosis: A Multimodal Study", *PLOS ONE*, vol. 7, no. 5, e37167, 2012.

[261] M. Deppe, K. Tabelow, J. Krämer, *et al.*, "Evidence for early, non-lesional cerebellar damage in patients with multiple sclerosis: DTI measures correlate with disability, atrophy, and disease duration", *Multiple Sclerosis Journal*, vol. 22, no. 1, pp. 73–84, 2016.

[262] E. Sbardella, F. Tona, N. Petsas, and P. Pantano, "DTI Measurements in Multiple Sclerosis: Evaluation of Brain Damage and Clinical Implications", *Multiple Sclerosis International*, p. 671 730, 2013.

[263] M. Onu, A. Roceanu, U. Sboto-Frankenstein, *et al.*, "Diffusion abnormality maps in demyelinating disease: Correlations with clinical scores", *European Journal of Radiology*, vol. 81, no. 3, e386–e391, 2012.

[264] S. Tommasin, L. De Giglio, S. Ruggieri, *et al.*, "Relation between functional connectivity and disability in multiple sclerosis: A non-linear model", *Journal of Neurology*, vol. 265, no. 12, pp. 2881–2892, 2018.

[265] O. Andersen, A. Hildeman, M. Longfils, *et al.*, "Diffusion tensor imaging in multiple sclerosis at different final outcomes", *Acta Neurologica Scandinavica*, vol. 137, no. 2, pp. 165–173, 2018.

[266] E. Sbardella, F Tona, N Petsas, *et al.*, "Functional connectivity changes and their relationship with clinical disability and white matter integrity in patients with relapsing–remitting multiple sclerosis", *Multiple Sclerosis Journal*, vol. 21, no. 13, pp. 1681–1692, 2015.

[267] F. Fink, J. Klein, M. Lanz, *et al.*, "Comparison of Diffusion Tensor-Based Tractography and Quantified Brain Atrophy for Analyzing Demyelination and Axonal Loss in MS", *Journal of Neuroimaging*, vol. 20, no. 4, pp. 334–344, 2010.

[268] M. J. Lowe, C. Horenstein, J. G. Hirsch, *et al.*, "Functional pathway-defined MRI diffusion measures reveal increased transverse diffusivity of water in multiple sclerosis", *NeuroImage*, vol. 32, no. 3, pp. 1127–1133, 2006.

[269] X. Lin, C. R. Tench, P. S. Morgan, G. Niepel, and C. S. Constantinescu, "'Importance sampling' in MS: Use of diffusion tensor tractography to quantify pathology related to specific impairment", *Journal of the Neurological Sciences*, vol. 237, no. 1, pp. 13–19, 2005.

[270] D. Jakimovski, B. Weinstock-Guttman, J. Hagemeier, *et al.*, "Walking disability measures in multiple sclerosis patients: Correlations with MRI-derived global and microstructural damage", *Journal of the Neurological Sciences*, vol. 393, pp. 128–134, 2018.

[271] M. Calabrese, R. Magliozzi, O. Ciccarelli, J. J. G. Geurts, R. Reynolds, and R. Martin, "Exploring the origins of grey matter damage in multiple sclerosis", *Nature Reviews Neuroscience*, vol. 16, no. 3, pp. 147–158, 2015.

[272] M. D. Steenwijk, M. Daams, P. J. W. Pouwels, *et al.*, "What Explains Gray Matter Atrophy in Long-standing Multiple Sclerosis?", *Radiology*, vol. 272, no. 3, pp. 832–842, 2014.

[273] A. Ceccarelli, M. A. Rocca, E. Pagani, *et al.*, "A voxel-based morphometry study of grey matter loss in MS patients with different clinical phenotypes", *NeuroImage*, vol. 42, no. 1, pp. 315–322, 2008.

[274] H. M. Genova, J. F. Sumowski, N. Chiaravalloti, G. Voelbel, and J. Deluca, "Cognition in multiple sclerosis: A review of neuropsychological and fMRI research", *Frontiers in Bioscience*, vol. 14, no. 5, pp. 1730–1744, 2009.

[275] R. A. Dineen, J. Vilisaar, J. Hlinka, *et al.*, "Disconnection as a mechanism for cognitive dysfunction in multiple sclerosis", *Brain*, vol. 132, no. 1, pp. 239–249, 2009.

[276] Y. Zhou, M. Milham, X.-N. Zuo, *et al.*, "Functional Homotopic Changes in Multiple Sclerosis with Resting-State Functional MR Imaging", *American Journal of Neuroradiology*, vol. 34, no. 6, pp. 1180–1187, 2013.

[277] A. Faivre, A. Rico, W. Zaaraoui, *et al.*, "Assessing brain connectivity at rest is clinically relevant in early multiple sclerosis", *Multiple Sclerosis Journal*, vol. 18, no. 9, pp. 1251–1258, 2012.

[278] F. Tona, N. Petsas, E. Sbardella, *et al.*, "Multiple Sclerosis: Altered Thalamic Resting-State Functional Connectivity and Its Effect on Cognitive Function", *Radiology*, vol. 271, no. 3, pp. 814–821, 2014.

[279] I. S. Stafford, M. Kellermann, E. Mossotto, R. M. Beattie, B. D. MacArthur, and S. Ennis, "A systematic review of the applications of artificial intelligence and machine learning in autoimmune diseases", *npj Digital Medicine*, vol. 3, no. 1, pp. 1–11, 2020.

[280] K. Bendfeldt, S. Klöppel, T. E. Nichols, *et al.*, "Multivariate pattern classification of gray matter pathology in multiple sclerosis", *NeuroImage*, vol. 60, no. 1, pp. 400–408, 2012.

[281] G. Kocevar, C. Stamile, S. Hannoun, *et al.*, "Graph Theory-Based Brain Connectivity for Automatic Classification of Multiple Sclerosis Clinical Courses", *Frontiers in Neuroscience*, vol. 10, 2016.

[282] M. Zurita, C. Montalba, T. Labbé, *et al.*, "Characterization of relapsing-remitting multiple sclerosis patients using support vector machine classifications of functional and diffusion MRI data", *NeuroImage: Clinical*, vol. 20, pp. 724–730, 2018.

[283] A. Marzullo, G. Kocevar, C. Stamile, *et al.*, "Classification of Multiple Sclerosis Clinical Profiles via Graph Convolutional Neural Networks", *Frontiers in Neuroscience*, vol. 13, 2019.

[284] F. Eitel, E. Soehler, J. Bellmann-Strobl, *et al.*, "Uncovering convolutional neural network decisions for diagnosing multiple sclerosis on conventional mri using layer-wise relevance propagation", *NeuroImage: Clinical*, vol. 24, p. 102 003, 2019.

[285] V. Vapnik and O. Chapelle, "Bounds on Error Expectation for Support Vector Machines", *Neural Computation*, vol. 12, no. 9, pp. 2013–2036, 2000.

[286] X. He, D. Cai, and P. Niyogi, "Laplacian Score for Feature Selection", in *Advances in Neural Information Processing Systems 18*, Y. Weiss, B. Schölkopf, and J. C. Platt, Eds., MIT Press, 2006, pp. 507–514.

[287] C. H. Polman, S. C. Reingold, B. Banwell, *et al.*, "Diagnostic criteria for multiple sclerosis: 2010 Revisions to the McDonald criteria", *Annals of Neurology*, vol. 69, no. 2, pp. 292–302, 2011.

[288] J. Veraart, E. Fieremans, and D. S. Novikov, "Diffusion MRI noise mapping using random matrix theory", *Magnetic Resonance in Medicine*, vol. 76, no. 5, pp. 1582–1593, 2016.

[289] C.-F. Westin, S. Peled, H. Gubjartsson, R. Kikinis, and F. A. Jolesz, "Geometrical diffusion measures for mri from tensor basis analysis", 1997.

[290] S. Mori, S. Wakana, P. C. M. van Zijl, and L. M. Nagae-Poetscher, *MRI Atlas of Human White Matter*. Elsevier, 2005.

[291] D. Hartman, J. Hlinka, M. Paluš, D. Mantini, and M. Corbetta, "The role of nonlinearity in computing graph-theoretical properties of resting-state functional magnetic resonance imaging brain networks", *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 21, no. 1, p. 013 119, 2011.

[292] S. S. Keller and N. Roberts, "Voxel-based morphometry of temporal lobe epilepsy: An introduction and review of the literature", *Epilepsia*, vol. 49, no. 5, pp. 741–757, 2008.

[293] M. A. Yassa and C. E. L. Stark, "A quantitative evaluation of cross-participant registration techniques for MRI studies of the medial temporal lobe", *NeuroImage*, vol. 44, no. 2, pp. 319–327, 2009.

[294] T. Charalambous, C. Tur, F. Prados, *et al.*, "Structural network disruption markers explain disability in multiple sclerosis", *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 90, no. 2, pp. 219–226, 2019.

[295] Schoonheim, M.M., Geurts, J.J.G., Barkhof, F., *et al.*, "The limits of functional reorganization in multiple sclerosis", *Neurology*, vol. 74, no. 16, pp. 1246–1247, 2010.

[296] M. M. Schoonheim, K. A. Meijer, and J. J. G. Geurts, "Network Collapse and Cognitive Impairment in Multiple Sclerosis", *Frontiers in Neurology*, vol. 6, 2015.

[297] M. Bozzali, M. Cercignani, M. P. Sormani, G. Comi, and M. Filippi, "Quantification of Brain Gray Matter Damage in Different MS Phenotypes by Use of Diffusion Tensor MR Imaging", *American Journal of Neuroradiology*, vol. 23, no. 6, pp. 985–988, 2002.

[298] M. Rovaris, M. Bozzali, G. Iannucci, *et al.*, "Assessment of Normal-Appearing White and Gray Matter in Patients With Primary Progressive Multiple Sclerosis: A Diffusion-Tensor Magnetic Resonance Imaging Study", *Archives of Neurology*, vol. 59, no. 9, pp. 1406–1412, 2002.

[299] M. Filippi and M. A. Rocca, "MR Imaging of Multiple Sclerosis", *Radiology*, vol. 259, no. 3, pp. 659–681, 2011.

[300] T. R. Insel, "Rethinking schizophrenia", *Nature*, vol. 468, no. 7321, pp. 187–193, Nov. 2010.

[301] H. Jaaro-Peled, A. Hayashi-Takagi, S. Seshadri, A. Kamiya, N. J. Brandon, and A. Sawa, "Neurodevelopmental mechanisms of schizophrenia: Understanding disturbed postnatal brain maturation through neuregulin-1–ErbB4 and DISC1", *Trends in Neurosciences*, vol. 32, no. 9, pp. 485–495, Sep. 2009.

[302] D. A. Lewis and P. Levitt, "Schizophrenia as a Disorder of Neurodevelopment", *Annual Review of Neuroscience*, vol. 25, no. 1, pp. 409–432, 2002.

[303] S. R. Kay, A. Fiszbein, and L. A. Opler, "The positive and negative syndrome scale (panss) for schizophrenia", *Schizophrenia bulletin*, vol. 13, no. 2, pp. 261–276, 1987.

[304] J. Tang, Y. Fan, H. Li, *et al.*, "Whole-genome sequencing of monozygotic twins discordant for schizophrenia indicates multiple genetic risk factors for schizophrenia", *Journal of Genetics and Genomics*, vol. 44, no. 6, pp. 295–306, Jun. 2017.

[305] H. J. Sørensen, E. L. Mortensen, J. Schiffman, J. M. Reinisch, J. Maeda, and S. A. Mednick, "Early developmental milestones and risk of schizophrenia: A 45-year follow-up of the Copenhagen Perinatal Cohort", *Schizophrenia Research*, vol. 118, no. 1, pp. 41–47, May 2010.

[306] K. Dean and R. M. Murray, "Environmental risk factors for psychosis", *Dialogues in Clinical Neuroscience*, vol. 7, no. 1, pp. 69–80, Mar. 2005.

[307] C. Morgan and H. Fisher, "Environment and schizophrenia: Environmental factors in schizophrenia: Childhood trauma—a critical review", *Schizophrenia bulletin*, vol. 33, no. 1, pp. 3–10, 2007.

[308] J Vilain, A.-M. Galliot, J Durand-Roger, *et al.*, "[Environmental risk factors for schizophrenia: a review]", *L'Encephale*, vol. 39, no. 1, pp. 19–28, Feb. 2013.

[309] S. S. Kuo and M. F. Pogue-Geile, "Variation in fourteen brain structure volumes in schizophrenia: A comprehensive meta-analysis of 246 studies", *Neuroscience & Biobehavioral Reviews*, vol. 98, pp. 85–94, Mar. 2019.

[310] T. Paus, M. Keshavan, and J. N. Giedd, "Why do many psychiatric disorders emerge during adolescence?", *Nature Reviews Neuroscience*, vol. 9, no. 12, pp. 947–957, Dec. 2008.

[311] N. Gogtay, A. Sporn, L. S. Clasen, *et al.*, "Comparison of Progressive Cortical Gray Matter Loss in Childhood-OnsetSchizophrenia With That in Childhood-Onset Atypical Psychoses", *Archives of General Psychiatry*, vol. 61, no. 1, pp. 17–22, Jan. 2004.

[312] N. Honnorat, A. Dong, E. Meisenzahl-Lechner, N. Koutsouleris, and C. Davatzikos, "Neuroanatomical heterogeneity of schizophrenia revealed by semi-supervised machine learning methods", *Schizophrenia Research*, Machine Learning in Schizophrenia, vol. 214, pp. 43–50, Dec. 2019.

[313] D. B. Dwyer, G. B. Chand, A. Pigoni, *et al.*, "Psychosis brain subtypes validated in first-episode cohorts and related to illness remission: Results from the PHENOM consortium", *Molecular Psychiatry*, pp. 1–10, May 2023.

[314] G. B. Chand, D. B. Dwyer, G. Erus, *et al.*, "Two distinct neuroanatomical subtypes of schizophrenia revealed using machine learning", *Brain*, vol. 143, no. 3, pp. 1027–1038, 2020.

[315] A. Fornito, A. Zalesky, C. Pantelis, and E. T. Bullmore, "Schizophrenia, neuroimaging and connectomics", *NeuroImage*, Connectivity, vol. 62, no. 4, pp. 2296–2314, Oct. 2012.

[316] J. M. Sheffield and D. M. Barch, "Cognition and resting-state functional connectivity in schizophrenia", *Neuroscience & Biobehavioral Reviews*, vol. 61, pp. 108–120, Feb. 2016.

[317] D. M. Barch, D. Pagliaccio, and K. Luking, "Mechanisms underlying motivational deficits in psychopathology: Similarities and differences in depression and schizophrenia", *Behavioral neuroscience of motivation*, pp. 411–449, 2016.

[318] M. S. Cetin, J. M. Houck, V. M. Vergara, R. L. Miller, and V. Calhoun, "Multimodal based classification of schizophrenia patients", in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Aug. 2015, pp. 2629–2632.

[319] M. S. Cetin, J. M. Houck, B. Rashid, *et al.*, "Multimodal Classification of Schizophrenia Patients with MEG and fMRI Data Using Static and Dynamic Connectivity Measures", *Frontiers in Neuroscience*, vol. 10, 2016.

[320] C. Cabral, L. Kambeitz-Ilankovic, J. Kambeitz, *et al.*, "Classifying schizophrenia using multimodal multivariate pattern recognition analysis: Evaluating the impact of individual clinical profiles on the neurodiagnostic performance", *Schizophrenia bulletin*, vol. 42, no. suppl_1, S110–S117, 2016.

[321] S. Liang, Y. Li, Z. Zhang, *et al.*, "Classification of First-Episode Schizophrenia Using Multimodal Brain Features: A Combined Structural and Diffusion Imaging Study", *Schizophrenia Bulletin*, vol. 45, no. 3, pp. 591–599, Apr. 2019.

[322] R. Salvador, E. Canales-Rodríguez, A. Guerrero-Pedraza, *et al.*, "Multimodal Integration of Brain Images for MRI-Based Diagnosis in Schizophrenia", *Frontiers in Neuroscience*, vol. 13, 2019.

[323] H. Zhuang, R. Liu, C. Wu, *et al.*, "Multimodal classification of drug-naïve first-episode schizophrenia combining anatomical, diffusion and resting state functional resonance imaging", *Neuroscience Letters*, vol. 705, pp. 87–93, Jul. 2019.

[324] F. Spaniel, J. Tintera, J. Rydlo, *et al.*, "Altered Neural Correlate of the Self-Agency Experience in First-Episode Schizophrenia-Spectrum Patients: An fMRI Study", *Schizophrenia Bulletin*, vol. 42, no. 4, pp. 916–925, 2016.

[325] J. L. Andersson, S. Skare, and J. Ashburner, "How to correct susceptibility distortions in spin-echo echo-planar images: Application to diffusion tensor imaging", *Neuroimage*, vol. 20, no. 2, pp. 870–888, 2003.

[326] S. Skare and R. Bammer, "Jacobian weighting of distortion corrected epi data", *Stockholm, Sweden*, 2010.

[327] S. M. Smith, M. Jenkinson, M. W. Woolrich, *et al.*, "Advances in functional and structural mr image analysis and implementation as fsl", *Neuroimage*, vol. 23, S208–S219, 2004.

[328] P. S. J. Weston, T. Poole, J. M. Nicholas, *et al.*, "Measuring cortical mean diffusivity to assess early microstructural cortical change in presymptomatic familial Alzheimer's disease", *Alzheimer's Research & Therapy*, vol. 12, no. 1, p. 112, Sep. 2020.

[329] C. Zhuo, J. Zhu, W. Qin, *et al.*, "Functional connectivity density alterations in schizophrenia", *Frontiers in Behavioral Neuroscience*, vol. 8, p. 404, Nov. 2014.

[330] K. C. Skåtun, T. Kaufmann, N. T. Doan, *et al.*, "Consistent Functional Connectivity Alterations in Schizophrenia Spectrum Disorder: A Multisite Study", *Schizophrenia Bulletin*, vol. 43, no. 4, pp. 914–924, Jul. 2017.

[331] P. Wang, J. Yang, Z. Yin, *et al.*, "Amplitude of low-frequency fluctuation (ALFF) may be associated with cognitive impairment in schizophrenia: A correlation study", *BMC Psychiatry*, vol. 19, no. 1, p. 30, Jan. 2019.

[332] Y. Xu, C. Zhuo, W. Qin, J. Zhu, and C. Yu, "Altered Spontaneous Brain Activity in Schizophrenia: A Meta-Analysis and a Large-Sample Study", *BioMed Research International*, vol. 2015, p. 204 628, 2015.

[333] M. J. Hoptman, X.-N. Zuo, P. D. Butler, *et al.*, "Amplitude of low-frequency oscillations in schizophrenia: A resting state fMRI study", *Schizophrenia research*, vol. 117, no. 1, p. 13, Mar. 2010.

[334] J. Henf, M. J. Grothe, K. Brueggen, S. Teipel, and M. Dyrba, "Mean diffusivity in cortical gray matter in Alzheimer's disease: The importance of partial volume correction", *NeuroImage : Clinical*, vol. 17, pp. 579–586, Oct. 2017.

[335] P. S. Weston, I. J. Simpson, N. S. Ryan, S. Ourselin, and N. C. Fox, "Diffusion imaging changes in grey matter in alzheimer's disease: A potential marker of early neurodegeneration", *Alzheimer's research & therapy*, vol. 7, pp. 1–8, 2015.

[336] P. Lee, H.-R. Kim, Y. Jeong, and for the Alzheimer's Disease Neuroimaging Initiative, "Detection of gray matter microstructural changes in Alzheimer's disease continuum using fiber orientation", *BMC Neurology*, vol. 20, no. 1, p. 362, Oct. 2020.

[337] N. Erickson, J. Mueller, A. Shirkov, *et al.*, *AutoGluon-Tabular: Robust and Accurate AutoML for Structured Data*, 2020. arXiv: 2003.06505.

[338] M. Hu, X. Qian, S. Liu, *et al.*, "Structural and diffusion MRI based schizophrenia classification using 2D pretrained and 3D naive Convolutional Neural Networks", *Schizophrenia Research*, vol. 243, pp. 330–341, May 2022.

[339] B. Bučková, J. Kopal, K. Řasová, J. Tintěra, and J. Hlinka, "Open Access: The Effect of Neurorehabilitation on Multiple Sclerosis—Unlocking the Resting-State fMRI Data", *Frontiers in Neuroscience*, vol. 15, p. 615, 2021.

[340] K. Řasová, B. Bučková, T. prokupisová, *et al.*, "A three-arm parallel-group exploratory trial documents balance improvement without much evidence of white matter integrity changes in people with multiple sclerosis following two months ambulatory neuroproprioceptive "facilitation and inhibition" physical therapy", *European Journal of Physical and Rehabilitation Medicine*, vol. 57, no. 6, pp. 889–99, 2021.

[341] A. Škoch, B. Rehák Bučková, J. Mareš, *et al.*, "Human brain structural connectivity matrices–ready for modelling", *Scientific Data*, vol. 9, no. 1, p. 486, 2022.

[342] G. Deco, A. R. McIntosh, K. Shen, *et al.*, "Identification of optimal structural connectivity using functional connectivity and neural modeling", *Journal of Neuroscience*, vol. 34, no. 23, pp. 7910–7916, 2014.

[343] B. Rehak Buckova, C. Fraza, R. Rehak, *et al.*, "Using normative models pre-trained on cross-sectional data to evaluate longitudinal changes in neuroimaging data", *bioRxiv*, pp. 2023–06, 2023.

[344] C. M. Williams, H. Peyre, R. Toro, and F. Ramus, "Neuroanatomical norms in the UK Biobank: The impact of allometric scaling, sex, and age", *Human Brain Mapping*, vol. 42, no. 14, pp. 4623–4642, 2021.

[345] S. G. Heeringa and P. A. Berglund, *A Guide for Population-based Analysis of the Adolescent Brain Cognitive Development (ABCD) Study Baseline Data*, 2020.

[346] K. Franke and C. Gaser, "Ten Years of BrainAGE as a Neuroimaging Biomarker of Brain Aging: What Insights Have We Gained?", *Frontiers in Neurology*, vol. 10, 2019.

[347] C. Schuster, M. Elamin, O. Hardiman, and P. Bede, "Presymptomatic and longitudinal neuroimaging in neurodegeneration—from snapshots to motion picture: A systematic review", *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 86, no. 10, pp. 1089–1096, 2015.

[348] M. A. Di Biase, Y. E. Tian, R. A. I. Bethlehem, *et al.*, "Mapping human brain charts cross-sectionally and longitudinally", *Proceedings of the National Academy of Sciences*, vol. 120, no. 20, e2216798120, 2023.

[349] T. Cole, "The development of growth references and growth charts", *Annals of Human Biology*, vol. 39, no. 5, pp. 382–394, 2012.

[350] S. M. Kia, H. Huijsdens, S. Rutherford, *et al.*, "Closing the life-cycle of normative modeling using federated hierarchical Bayesian regression", *PLOS ONE*, vol. 17, no. 12, e0278776, 8. 12. 2022.

[351] M. Reuter, N. J. Schmansky, H. D. Rosas, and B. Fischl, "Within-subject template estimation for unbiased longitudinal image analysis", *Neuroimage*, vol. 61, no. 4, pp. 1402–1418, 2012.

[352] K. Merritt, P. Luque Laguna, A. Irfan, and A. S. David, "Longitudinal Structural MRI Findings in Individuals at Genetic and Clinical High Risk for Psychosis: A Systematic Review", *Frontiers in Psychiatry*, vol. 12, 2021.

[353] B. J. Casey, T. Cannonier, M. I. Conley, *et al.*, "The Adolescent Brain Cognitive Development (ABCD) study: Imaging acquisition across 21 sites", *Developmental Cognitive Neuroscience*, vol. 32, pp. 43–54, 2018.

[354] M. Ienca and K. Ignatiadis, "Artificial Intelligence in Clinical Neuroscience: Methodological and Ethical Challenges", *AJOB Neuroscience*, vol. 11, no. 2, pp. 77–87, 2020.

[355] A. Kulesza, B. Taskar, *et al.*, "Determinantal point processes for machine learning", *Foundations and Trends® in Machine Learning*, vol. 5, no. 2–3, pp. 123–286, 2012.

[356] S. Rutherford, S. M. Kia, T. Wolfers, *et al.*, "The Normative Modeling Framework for Computational Psychiatry", Cold Spring Harbor Laboratory, Tech. Rep., 2021, ch. New Results, p. 2021.08.08.455583.

Author's publications

List of author's publications related to the doctoral thesis

[J1] **B. Bučková**, M. Brunovský, M. Bareš, and J. Hlinka, "Predicting sex from EEG: Validity and generalizability of deep-learning-based interpretable classifier", *Frontiers in Neuroscience*, vol. 14, 2020.

[J2] **B. Bučková**, J. Kopal, K. Řasová, J. Tintěra, and J. Hlinka, "Open Access: The Effect of Neurorehabilitation on Multiple Sclerosis—Unlocking the Resting-State fMRI Data", Frontiers in *Neuroscience*, vol. 15, p. 615, 2021.

[J3] A. Škoch, **B. Rehák Bučková**, J. Mareš, *et al.*, "Human brain structural connectivity matrices–ready for modelling", *Scientific Data*, vol. 9, no. 1, p. 486, 2022.

[J4] K. Řasová, **B. Bučková**, T. Prokopiusova, *et al.*, "A three-arm parallel group exploratory trial documents balance improvement without much evidence of white matter integrity changes in people with multiple sclerosis following two months ambulatory neuroproprioceptive "facilitation and inhibition" physical therapy", *European Journal of Physical and Rehabilitation Medicine*, vol. 57, no. 6, pp. 889–99, 2021.

[J5] **B. Rehák Bučková**, D. Kala, J. Kořenek, *et al.*, "Structural connectivity-based predictors of cognitive impairment in stroke patients attributable to aging", *Plos one*, vol. 18, no. 4, e0280892, 2023.

[J6] **B. Rehák Bučková**, J. Mareš, A. Škoch, *et al.*, "Multimodal-neuroimaging machine learning analysis of motor disability in multiple sclerosis", *Brain Imaging and Behavior*, vol. 17, no. 1, pp. 18–34, 2023.

Citations in Web of Knowledge (except self-citations)

Journal paper [J1]

[J1 Cit1] C. A. Ellis, M. S. Sendi, R. Zhang, *et al.*, "Novel methods for elucidating modality importance in multimodal electrophysiology classifiers", *Frontiers in Neuroinformatics*, vol. 17, p. 1 123 376, 2023.

[J1 Cit2] D. W. Joyce, A. Kormilitzin, K. A. Smith, and A. Cipriani, "Explainable artificial intelligence for mental health through transparency and interpretability for understandability", *npj Digital Medicine*, vol. 6, no. 1, p. 6, 2023.

[J1 Cit3] M. Khayretdinova, A. Shovkun, V. Degtyarev, A. Kiryasov, P. Pshonkovskaya, and I. Zakharov, "Predicting age from resting-state scalp eeg signals with deep convolutional neural networks on td-brain dataset", *Frontiers in Aging Neuroscience*, vol. 14, p. 1367, 2022.

[J1 Cit4] Y. Holler, S. T. Jónsdóttir, A. H. Hannesdóttir, and R. P. Ólafsson, "Eeg-responses to mood induction interact with seasonality and age", *Frontiers in Psychiatry*, vol. 13, p. 1774, 2022.

[J1 Cit5] C. Babiloni, G. Noce, R. Ferri, *et al.*, "Resting state alpha electroencephalography rhythms are affected by sex in cognitively unimpaired seniors and patients with alzheimer's disease and amnesic mild cognitive impairment: A retrospective and exploratory study", *Cerebral Cortex*, vol. 32, no. 10, pp. 2197–2215, 2022.

Journal paper [J2]

[J2 Cit1] N. A. Roman, V. I. Tuchel, C. Nicolau, O.-D. Grigorescu, and R. Necula, "Functional electrostimulation in patients affected by the most frequent central motoreuron disorders—a scoping review", *Applied Sciences*, vol. 13, no. 6, p. 3732, 2023.

[J2 Cit2] S. MD and A. CR, "Noise invariant convolution neural network for segmentation of multiple sclerosis lesions from brain magnetic resonance imaging.", *International Journal of Online & Biomedical Engineering*, vol. 18, no. 13, 2022.

Journal paper [J3]

[J3 Cit1] E. Capobianco and M. Dominietto, "Assessment of brain cancer atlas maps with multimodal imaging features", *Journal of Translational Medicine*, vol. 21, no. 1, pp. 1–11, 2023.

[J3 Cit2] W. Zúñiga-Galindo and B. Zambrano-Luna, "Hierarchical wilson–cowan models and connection matrices", *Entropy*, vol. 25, no. 6, p. 949, 2023.

Journal paper [J4]

[J4 Cit1] E. Warmerdam, M. Schumacher, T. Beyer, *et al.*, "Postural sway in parkinson's disease and multiple sclerosis patients during tasks with different complexity", *Frontiers in Neurology*, vol. 13, p. 857 406, 2022.