University of Verona
Department of Computer Science
Graduate School of Natural Sciences and Engineering
PhD in Computer Science

EXplainable Artificial Intelligence: enabling AI in neurosciences and beyond

PhD Candidate: Federica Cruciani

Advisor: Prof. Gloria Menegaz

INF/01, XXXV cycle, 2023

Advisor:
prof. Gloria Menegaz

EXplainable Artificial Intelligence:
enabling AI in neuroscience and beyond - Cruciani Federica
PhD thesis
Università di Verona
Dipartimento di Informatica
Strada le Grazie 15, 37134 Verona, Italy
27 February 2023

To those who have always believed in me, even when I did not.

Abstract

The adoption of Artificial Intelligence (AI) models in medicine and neurosciences has the potential to play a significant role not only in bringing scientific advancements but also in clinical decision making. However, concerns mounts due to the eventual biases AI could have which could result in far-reaching consequences particularly in a critical field like biomedicine. It is challenging to achieve usable intelligence because not only it is fundamental to learn from prior data, extract knowledge and guarantee generalization capabilities, but also to disentangle the underlying explanatory factors in order to deeply understand the variables leading to the final decisions. There hence has been a call for approaches to open the AI 'black box' to increase trust and reliability on the decision-making capabilities of AI algorithms. Such approaches are commonly referred to as eXplainable Artificial Intelligence (XAI) and are starting to be applied in medical fields even if not yet fully exploited.

With this thesis we aim at contributing to enabling the use of AI in medicine and neurosciences by taking two fundamental steps: (i) practically pervade AI models with XAI (ii) Strongly validate XAI models.

The first step was achieved on one hand by focusing on XAI taxonomy and proposing some guidelines specific for the AI and XAI applications in the neuroscience domain. On the other hand, we faced concrete issues proposing XAI solutions to decode the brain modulations in neurodegeneration relying on the morphological, microstructural and functional changes occurring at different disease stages as well as their connections with the genotype substrate.

The second step was as well achieved by firstly defining four attributes related to XAI validation, namely *stability*, *consistency*, *understandability* and *plausibility*. Each attribute refers to a different aspect of XAI ranging from the assessment of explanations stability across different XAI methods, or highly collinear inputs, to the alignment of the

obtained explanations with the state-of-the-art literature. We then proposed different validation techniques aiming at practically fulfilling such requirements.

With this thesis, we contributed to the advancement of the research into XAI aiming at increasing awareness and critical use of AI methods opening the way to real-life applications enabling the development of personalized medicine and treatment by taking a data-driven and objective approach to healthcare.

Contents

Part V Conclusions and future directions

1

Introduction

Is eXplainable Artificial Intelligence (XAI) the enabling technology for the adoption of Artificial Intelligence (AI) in medicine?

AI systems are increasingly exploited in a myriad of sensitive domains, including medical and healthcare fields (e.g, disease diagnosis and progression modeling, image analysis, monitoring), with important ethical, legal, social and responsibility implications.

Starting from the definition of AI, it is conceivably the oldest field of computer science as it focuses on replicating cognitive capacities for resolving practical issues through the development of machines that can learn and reason like humans. One of the most significant branches of AI is Machine Learning (ML). Its goal is to create software that can automatically learn from historical data in order to gather expertise, and to gradually improve its learning behavior in order to produce predictions based on recent data [1]. The development of novel statistical learning techniques, the availability of large data sets, and the low cost of computation have all greatly advanced ML [2] resulting in one of today's most popular branches which is Deep Learning (DL), a family of deep neural network-based models [3] in which multiple layers of processing are used to extract progressively higher level features from data, including non linear intertwined interactions. In order to fully comprehend how the model came to its final decision, the inspection of all those layers for describing their relations would be unfeasible. For this reason, AI and in particular deep neural networks are often considered as 'black boxes', and concerns mount especially in the most critical fields such as medical applications where eventual biases can have far-reaching consequences. It was indeed recently emphasized that it is challenging to achieve usable intelligence because not only it is fundamental to learn from prior data, extract knowledge, generalize, and battle the curse of dimensionality, but also to disentangle the underlying explanatory

factors of the data in order to understand the context in an application domain, where to date a doctor-in-the-loop is necessary that is, of course, the case of AI application in medicine [4, 5]. Moreover, Medical experts have voiced their concern about the black box nature of deep learning which is however the current state of the art in medical image analysis. Furthermore, regulations such as the European Union's General Data Protection Regulation (GDPR, Article 15) require the right of patients to receive meaningful information about how a decision was rendered, hence introducing the right of explanation.

In this framework, making sense, understanding context, and making judgments in the face of uncertainty pose the biggest challenges [6]. There hence has been a call for approaches to open the black box and to increase the decision-making capabilities of ML and DL algorithms. Such approaches are commonly referred to as XAI [7, 8]. XAI has hence emerged with the goal of supplying new methodologies and algorithms to improve transparency and reliability to both the judgments made by predictive algorithms and the contributions and importance of individual features to the outcome [9, 10]. With this aim XAI could also be applied in a feedback loop allowing to increase models' performance given the explanations for the wrong classification or regressions. However, the field is still not mature, and there is a lack of consensus when referring to XAI in many respects, ranging from the taxonomy to the methods and to their validation and exploitability in real contexts. In this thesis, we aim to shed some light on all these aspects, contributing to the advancement of the field while proposing ad-hoc solutions for specific case studies. To this end, the first part of this thesis will present the state of the art on XAI as well as our view on the issue and fixing the terminology. In particular, we propose to differentiate explainability from interpretability with the former referring to the ability to decrypt the internal rules of the system, thus pointing to the so-called 'white box' or *ante-hoc* models, and the latter referring to XAI methods applied *post-hoc* on 'black box' models to derive explanations allowing to describe its function in human-comprehensible terms, supporting the identification of causal relationships.

Numerous studies have shown how essential XAI approaches are for personalized medicine, including tailored interventions and therapies [7, 11]. Moreover, as argued by Holziger and colleagues [4], through the adoption of XAI methods, human specialists have the ability on-demand to comprehend and retrace the machine decision process allowing the use of powerful AI systems to aid medical practitioners and, in some situations, even play a significant role in clinical decision making. We will indeed present some practical applications of XAI models facing real clinical problems within a guided tour illustrating the main concepts, bottlenecks and issues encountered in this context. Different scenarios will be considered including disease phenotypes classification

or feature-association studies relying both on directly explainable models and interpretability methods focusing on deep networks.

Moving further, trying to answer the leading question of this thesis, which is about enabling AI in medicine, a new core issue arises, which is models validation. If on one side it is necessary that XAI pervades AI models, on the other, it is also stringent a strong validation of XAI methods providing evidence that the outcomes of such models are non-biased by any factors such as the algorithms, data, and so on. The evaluation of explanation methods is still underinvestigated, however, since explainability is meant to increase confidence in artificial intelligence, it is vital to systematically analyze and compare explanation methods in order to ensure their accuracy [12]. Some approaches for XAI validation are already emerging aiming to propose a common and generalized approach or attributes XAI should respect in order to more strongly validate the obtained explanations. After a literature review, in thesis, we will propose four validation attributes that in our opinion a XAI method should fulfill and we will as well show how they were implemented in our works.

There are hence two steps to overcome for AI to be enabled in medicine: (1) The utilization of XAI methods which have a twofold role, allow clinicians to build trust on AI models enabling them for decision aiding and use them in a feedback loop where explanations aid in understanding how to improve model accuracy; (2) XAI model validation, which, due to the flourishing of diverse XAI methods, allows to properly choose the most reliable XAI methods, enhancing as well the trust in the obtained explanations.

The clinical outcomes of this thesis are neurodegenerative diseases, in particular Alzheimer's Disease (AD) and Multiple Sclerosis (MS). Neurodegeneration includes a heterogeneous group of disorders, characterized by the progressive degeneration of the structure and function of the central nervous system or peripheral nervous system. They occur when nerve cells in the brain or peripheral nervous system lose function over time and ultimately die. Although certain treatments may help relieve some of the physical or mental symptoms associated with neurodegenerative diseases no cures still exist. Of note, some studies were also performed based on brain aging and post-stroke rehabilitation. In this research area, there is growing evidence that multimodal brain imaging studies can aid in providing a more thorough understanding of the brain and its disorders. Brain imaging studies frequently gather data from a single subject using many Magnetic Resonance Imaging (MRI) modalities. For instance, they can tell us how brain structure influences brain function, how psychopathology affects them, and which functional or structural aspects of physiology may be responsible for human behavior and cognition. In this framework, the specific association between genetic measures and Imaging Derived Phenotype (IDP)s is, in particular, the target of Imaging Genetics (IG) which focuses on integrative studies to assess the influence of the genetic

architecture on brain structure and function, aiming at gaining new insights into the phenotypic characteristics and genetic mechanisms of the brain, and into their role in shaping normal and disordered brain conditions [13]. It has gained a central role in clinical research in recent years, as confirmed also by several articles highlighting the potential of meta-analyses of omics-wide association studies and imaging studies [13, 14]. This growing interest resulted in the increased availability of datasets containing both MRI acquisitions and DNA sequencing as well as gene expression values. Among these, concerning diseases, Alzheimer's Disease Neuroimaging Initiative (ADNI) is publicly available and comprehends a cohort of healthy controls, AD patients and subjects presenting Mild Cognitive Impairment (MCI) at different stages. The available data include MRI images, Positron Emission Tomography (PET), biospecimen, clinical as well as genetic data. The complementing information that can be extracted from this heterogeneous database is huge. For instance, structural Magnetic Resonance Imaging (sMRI) enables us to estimate the type of tissue for each voxel in the brain (Grey Matter (GM), White Matter (WM), and CerebroSpinal Fluid (CSF)); diffusion Magnetic Resonance Imaging (dMRI) can also provide information on the integrity of white matter tracts and structural connectivity; functional Magnetic Resonance Imaging (fMRI) measures the hemodynamic response related to neural activity in the brain dynamically; cognitive and clinical scores could shed light on behavioral information; and genetics itself would allow identifying genetic variants typical of each subject. With the focus still on XAI, in this thesis, we will present explainable models for IG considering multiple biomarkers as input and we will propose an interpretability framework for more complex data integration models.

1.1 Open questions and objectives

With this thesis, we aim at contributing to enabling the use of AI in medicine and neurosciences by following the two steps presented in the introduction. The first step is to practically pervade AI models with XAI, we hence firstly focused on XAI taxonomy and applications to describe the brain modulation in neurodegeneration relying on the morphological and microstructural changes occurring at different disease stages as well as their connection with genotype. The second step is XAI validation, indeed we proposed different validation techniques, as well as the practical application of the four proposed validation attributes that XAI models should satisfy.

More in detail, six main open questions emerge from the previous Sections regarding both XAI applications and heterogeneous data integration.

- XAI taxonomy is still highly confused in literature, hence limiting also its applicability to clinical problems;

- XAI is not yet fully exploited in brain imaging despite the emerging demand for techniques enabling the building of trust in clinical decision-aiding models;

- While new XAI methods are flourishing, especially in the field of computer vision, techniques to validate them and assess their stability are deemed as necessary to be applied in clinical contexts;

- Being brain diseases highly complex and not fully described by data, or features, coming from single acquisition modalities, heterogeneous data integration methods are lately being developed to account for the multi-faceted disease nature. In addition, dealing with a limited sample size while extracting still meaningful associations is still an open research question;

- Concerning input features, on the imaging side microstructure, as derived from dMRI data, is not yet fully exploited in the clinical contexts addressed in this thesis, both considering row data and classical or more advanced reconstruction models. Microstructural features are very powerful in describing brain tissues at a microscopic scale, thus injecting this information in both linear multivariate and deep models, has the potential to significantly enhance the performance. In particular, Region Of Interest (ROI), voxel or tract-based features, as well as 3D maps representing different diffusion-derived indices will be considered depending on the models;

- Moving to genetics, scores summarizing the disease status, still keeping enough information about subjects' mutations are needed in order to deal with small sample size data.

To summarize, the general aim of this thesis is indeed, firstly, the clarification of XAI taxonomy, proposing also a use case overview focusing on a specific brain imaging application. Then the exploitation of XAI to model multimodal and multidimensional data with neurodegeneration applications, concluding with the proposal of different techniques to validate XAI methods.

1.1.1 Thesis outline

The manuscript is structured in five parts.

Part I will present our view on XAI.

Chapter 2 clarifies various XAI aspects by merging the State-Of-the-Art (SOA) reviews on the topic, with a special focus on medical imaging and on XAI methods validation strategies.

Chapter 3 will then propose some guidelines on how to apply XAI methods for sample applications in brain imaging.

Part II will focus on the viability of XAI for multimodal visual inspection, focusing on feature visualization methods, with the clinical application of MS patients stratification.

Chapter 4 will present a multivariate approach trying to associate brain IDPs deriving from different modalities with cognitive assessment features.

Chapter 5 will propose a DL based approach allowing to skip the handcrafted feature selection, followed by the application of feature visualization techniques for interoperability. We indeed relied on 3D-Convolutional Neural Networks (CNN) which are powerful deep models allowing to account for the full spatial information of the input data, which is fundamental in particular when dealing with brain imaging. They are not directly explainable in terms of coefficients or parameters but allow the application of interpretability methods. Explanations for this model can be obtained in the form of saliency maps highlighting the relevance of each input voxel in the final decision. XAI was applied to the twofold aim of detecting which of the input maps better achieved the task of stratifying patients and which brain regions had the main role, also checking how the different inputs held different and complementary information better explaining the complexity of the disease in combination.

Part III will move a step forward towards XAI in data integration, adding genetics into the equation. Multivariate methods as well as generative models will be presented and applied to the decryption of the link between imaging and genetics in AD continuum. We started from the most simple yet explainable ones such as Partial Least Squares (PLS) which aims at multivariate modeling finding a latent representation where the covariance between input projections is maximized. This model is directly explainable since the fitted weights directly relate to the feature importance leading to the latent space generation. We indeed tested classical multivariate methods to derive a common latent space describing the patient state, firstly by considering only one imaging technique and summary scores on the genetic side, then relying on more complex models to account for multi-channel information derived both from different imaging techniques, while representing finer-grained information for genetics. The focus was still on either the explainability of the model, which was possible when considering simple multivariate methods, or their interpretation through the application of *post-hoc* interpretability analysis. This allowed the uncovering of the leading features for the generation of the common latent space, which was then exploited for more complex multimodal methods.

Chapter 6 will benchmark the use of PLS analysis relying on sMRI based IDPs and a set of Polygenic Risk Score (PRS) as genetic features.

Chapter 7 will introduce microstructure and functional MRI derived IDPs in two separate PLS models to study their association with a subset of the PRSs.

Chapter 8 will then explore a different set of gene-based genetic features allowing to deal with a small study cohort while expressing at gene level the genetic information compared with the PRS considered in the previous Chapters. Moreover, we proposed the validation of the obtained explanations through transcriptomic analysis.

Chapter 9 will finally present a framework for the application of XAI techniques on a multichannel Variational AutoEncoder (VAE) aiming at addressing both the study of the association between features and the generative performance of one channel from the others.

Part IV will be devoted to the challenges posed by the application and development of multiple interpretability methods. Since XAI is still in its early stages while multiple methods are being deployed, ways to assess attributes such as stability, robustness, reliability, etc. throughout the different possible choices are still being investigated.

Chapter 10 will present a new proxy to establish a stability criterion of XAI methods over feature collinearity in AD patients stratification.

Chapter 11 will present a study focusing on the comparison between different perturbation based methods applied to the problem of rehabilitation after stroke.

Chapter 12 will finally present a study on the comparison between different feature visualization methods applied to the problem of MS stage detection.

Part V will draw the conclusions of the work while opening to possible future research directions.

An extensive description of the background of the thesis will be given in Appendix A.

Part I

Explainable Artificial Intelligence (XAI): taxonomy and guidelines for its application in neuroimaging

2

XAI taxonomy in neuroimaging

eXplainable Artificial Intelligence (XAI) recently emerged as one of the hottest topics for understanding “the why and how" of the outcomes of Machine Learning (ML)/Deep Learning (DL) algorithms. However, this is still largely unexplored especially in the brain imaging field though it could help to disentangle the contributions of the different features shaping the final estimates, as well as to provide other hints about the subserving mechanisms that cannot be captured with traditional approaches. Before tackling the issue, we will try to elucidate one aspect that is still unclear in the literature, i.e. the difference between *explainability* and *interpretability*, which are still used quite interchangeably while subtending different concepts.

The concepts of *explainability* and *interpretability* are hard to encode and are usually considered interchangeable by ML researchers. Such ambiguity was also put forth by the query outcomes of the literature review. In fact, the keyword *explainability* did not return any result, while the keyword *interpretability* returned the papers discussed in this Section. A clear definition of such terms would be required in order to get to a common agreement on their meaning in this context so as to derive criteria for their assessment, either subjective or objective. Far from pretending to solve this issue, which requires philosophical thinking, we will shape the discussion on one possible signification of such terms as specified hereafter.

Following [9], *interpretability* is connected with the human intuition behind the outputs of a model, claiming that the more interpretable the model, the easier is to devise cause-and-effect relationships within the system input and output. This definition is strongly related to the concept of *causability*, quite relevant to the medical area, presented in [10]. Causability is defined as "the extent to which an explanation of a statement to human experts achieves a specified level of causal understanding with effectiveness, efficiency and satisfaction in a specified context of use". Instead, following [9], *explainability* would be associated with the decoding of the internal logic and mech-

Chapter2/Figures/Fig4.png

Fig. 2.1: Overview of the interpretability methods. a) Black-box models can obtain optimal predictions but they do not allow a complete understanding. The application of the interpretability methods allows to retrieve and interpret the most important features. b) Schematic representation of the difference between explainable and interpretable models. c) Taxonomy map of the interpretability methods classified in *ante-hoc* and *post-hoc*. Local, global, model agnostic and model specific attributes are exemplified as well as feature-probing properties.

anisms of a ML system. In particular, [10] defines *explainability* as highlighting the decision-relevant parts of the used representations of the algorithm and active parts in the algorithmic model, that either contribute to the model accuracy on the training set or to a specific prediction for one particular observation. It is hence not necessar-

ily related to human understanding. Therefore, regarding ML, interpretability does not axiomatically entail explainability and vice versa, following [9]. Figure 2.1 (b) tries to express the difference between these two concepts. Starting from the training data two directions can be followed to obtain model explanations: 1) using a directly explainable model, such as a decision tree or a linear regression model, for which the underlying logic is easy to follow and understand, and the explanation can be straightforwardly derived from the model coefficients; 2) applying a black box model (e.g. deep models such as Convolutional Neural Networks (CNN)s), followed by a *post-hoc* interpretability model to derive explanations, not necessarily requiring to understand the underlying model mechanism.

In order for a system to be interpreted, `explanations`, namely the outcomes of interpretability methods application, must be provided and the properties making an explanation effective to humans need to be defined. Holzinger *et al.*[10] states that directly understandable, hence interpretable for humans, are data, objects or any graphical representations $\leq \mathbb{R}^3$, such as images or text. Feature-probing methods provide explanations to enable model interpretation. Following [15], three feature properties are relevant: (i) feature stability, which is assessed through methods that measure how stable each feature contribution is over multiple models trained on held-out datasets using resampling methods or cross-validation; (ii) ranking of feature importance, obtained by assessing the impact of a feature on the prediction output; and (iii) feature visualization, that encompasses strategies providing a visual rendering of feature importance, such as saliency maps. While (ii) and (iii) aim at making the model outcomes humanly understandable, (i) can be considered as a way to assess the robustness (generalizability) of the solution. In this respect, bootstrap is usually employed in the training/validation phase.

In [16] the authors discussed the properties of models that might render them interpretable, highlighting that human decisions might admit *post-hoc* interpretability despite the black-box nature of human brains. One advantage of this reading of interpretability is that opaque models can be interpreted after the fact, and subtends a clear distinction with respect to explainability, which instead entails a clear understanding of the model's internal rules and functioning. In our work, we build on such a claim and assume that interpretability points to causability, while explainability means decoding the system's internal rules, they do not reciprocally entail and both fall under the XAI umbrella.

In what follows we also remind a few additional attributes of interpretability models that we consider to be relevant in this context. Interpretability models can be model-agnostic or model-specific. While the former tries to give some insights about the function underlying the model, regardless of the model structure, the latter can be applied

only to a specific prediction model or architecture. Moreover, interpretability models can be local or global, depending on the fact that the explanation concerns an individual prediction or small Sections of the whole model or the whole system, respectively.

Finally, another pertinent categorization proposed in [10] distinguishes between *post-hoc* and *ante-hoc* models. In our taxonomy, the first lie in the interpretable models while the latter are explainable models that also hold the interpretability property as they embed explainability directly into their structure.

2.1 XAI in neuroimaging: State-Of-the-Art overview

The interpretability step in ML or DL framework is still among the open issues and future direction. This was also pointed out in a very recent and complete review on ML and DL methods in brain disease diagnosis [5] where the authors claimed that recently XAI emerges as an oracle to make the Artificial Intelligence (AI)-based systems more transparent, even though not yet deeply exploited. However, it is a clear future research direction since explainable diagnosis will be the ultimate basis for reliable and trustworthy communications between medical experts and AI experts, which is highly important to transform the ML/DL-based brain disorder detection potentials into clinical practice. Jiang and colleagues arrived at the same conclusion in their review on predictive neuroimaging where they state XAI step has received much less attention in predictive neuroimaging, and they also provide some potential reasons such as the trend to reward higher prediction performance over neurobiologically meaningful interpretation [11].

In addition to pursuing higher predictive performance, determining which specific connections, regions, or functional networks contribute to prediction may significantly advance our knowledge of how the brain implements cognition and, more importantly, facilitate the translation of neuroimaging findings into clinical practice[4, 10]. Moreover, machine learning methods tend to be treated as a black box, which results in focusing on the highest possible predictive performance rather than mechanism understanding. This may lead to the current dilemma of researchers treating interpretation as a secondary goal, e.g., explaining feature importance in their own way and attempting to link with neurobiological significance in a relatively shallow manner without taking full advantage of interpretable models. In this regard, the arbitrary interpretation of models may make it hard to reveal the neural underpinnings of behavioral traits [11].

In this State-Of-the-Art (SOA) description, we will review the main studies relying on XAI techniques in brain imaging for both disease detection and Imaging Genetics (IG) research with a particular focus on the methods implemented in this thesis.

2.1.1 Ante-hoc methods

Starting from the *ante-hoc* explainability models, this category encompasses methods that are *interpretable by design* meaning easily understandable by humans. The models in this category are usually known as intrinsic, transparent or white-box models. This class includes linear, decision trees, rule-based models and more complex models which are equally transparent and described in [9].

This class hence encompasses linear regression models, such as Ordinary Least Squares (OLS), Ridge and Least Absolute Shrinkage and Selection Operator (LASSO) which also hold the interpretability property, in that they can be directly interpreted in terms of their β coefficients both locally and globally. Assuming that the data has been standardized and the model contains no intercept term, large components of the β can be interpreted as features that are relevant to the regression task.

Linear regression models are sometimes preceded by linear latent variable models such as Principal Component Analysis (PCA), Independent Component Analysis (ICA) or their generalization in order to perform an initial feature selection or to find the 'modes' which embed the information deriving from either single modality or multimodal data in a smaller feature space. The so obtained features are then used as input to the prediction model. This does not compromise the interpretability property because the resulting models yield loading vectors for every component, which quantify the contribution of each feature to each component. These loading values are then used to assess the contribution to the model's outcome and thereby understand, for instance, which input features mostly contributed to it. Both the coefficients and the loadings can be directly visualized in a feature space that can also be a brain map.

2.1.2 Post-hoc methods

According to the very recent review performed by Van der Velden and colleagues [7] which considered more than 200 studies using *post-hoc* XAI methods in medical image analysis, neuroimaging resulted as one of the most frequent research areas to which XAI has been applied, with a total of more than 40 papers included until 2020. Despite the completeness and high clarity of their work, their focus is mainly on giving a comprehensive overview of the different XAI methods rather than focusing on the application to which they were fitted. Following the train of thought of the thesis in this Section, starting from the aforementioned review and adding more recent works, we will better detail the different problems and how XAI was applied to them, limiting the overview on Magnetic Resonance Imaging (MRI) based methods focusing on the brain.

Perturbation based methods

Perturbation based methods aim at finding a difference in the outcome based on a small permutation of the input. Some examples are Occlusion, Local Interpretable Model-Agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) which aim at building surrogate models to black-box ones to provide them interpretability.

Occlusion is a widely used approach due to its simplicity of application, and it was for example used in Feng *et al.* [17] for brain age prediction. They applied ablation analysis methods focusing on part of the input data, each time calculating a saliency map based on gradient.

Gaur and colleagues [18] proposed an explanation-driven DL model by utilizing a convolutional neural network, LIME, and SHAP for the prediction of discrete subtypes of brain tumors (meningioma, glioma, and pituitary) using an MRI image dataset.

SHAP alone was mainly applied to regression-based outcomes, such as brain age estimation. It was for example used to assess feature relevance on an adolescence cohort [19]. The authors used SHAP on multiple ML methods, revealing that the anatomical changes in a common set of regions drive model predictions of age, regardless of the model type. Dartora and colleagues [20] propose investigating ensemble models to classify groups in the aging cognitive decline spectrum by combining features extracted from single imaging modalities and combinations of imaging modalities (FluoroDeoxyGlucose (FDG)+AMY+MRI, and a Positron Emission Tomography (PET) ensemble). They applied SHAP with recursive feature elimination to evaluate the performance analysis of models using balanced accuracy before and after feature elimination.

Backpropagation or gradient-based methods

In general, visual explanations or saliency appeared to be the most popular interpretability methods. Approaches based on gradients still hold the leading positions, some examples are namely BackPropagation (BP) [21], Guided BackPropagation (GBP) [22], Class Activation Mapping (CAM) and all its variants [23, 24, 25].

BP, GBP and deconvolution methods were widely applied to brain imaging research. Some examples are the work of Gao and colleagues where deconvolution was applied to compare the features extracted through different models to decipher the behavior tasks from functional Magnetic Resonance Imaging (fMRI) recording during subjects performing different tasks.

Bohle *et al.* [26] moved a step forward, comparing various gradient based methods for the detection task of Alzheimer's Disease (AD) continuum. They appeared also to be widely applied in segmentation tasks, for example, the detection of enlarged perivascular spaces which are common in aging, and are considered a reflection of cerebral

small vessel disease, for which Dubost *et al.* [27] applied GBP to aid the computation of an automatic perivascular space score. Finally, backpropagation was also applied to a generative framework to predict the PET-derived myelin content map from multimodal MRI in Multiple Sclerosis (MS) by Wei *et al.* [28].

CAM and its variations appeared as the most popular approach in brain imaging. The application fields ranged from tumor detection or segmentation to disease detection, to the generation of synthetic data for unbalanced classes. For example, Chrakaborty *et al.* [29] proposed one of the first approaches in exploiting CAM to check the goodness of Parkinson's Disease (PD) classification by observing the resulting relevant Region Of Interest (ROI)s. Grad-CAM methods were then proposed to overcome CAM limitations regarding noise and architecture limitations. This was used for example by Zhang *et al.* [30] to classify MS types relying on MRI images. However, one main drawback of CAM techniques is that they extract feature maps from either the final layer or a single intermediate layer to create the discriminative maps and then interpolate to upsample to the original image resolution. The subject specific localization is coarse and unable to capture subtle abnormalities. To mitigate this, still aiming at PD detection, Shinde *et al.* [31] proposed a CNN based discriminative localization model which accounts for layers from each resolution facilitating a comprehensive map that can delineate the pathology for each subject by combining low-level, intermediate as well as high-level features from the CNN directly providing the discriminative map in the resolution of the original image.

Decomposition based methods

To overcome CAM limitations and obtain fine-grained subject-specific heatmaps Layerwise Relevance Propagation (LRP) [32] came out and were widely applied for disease detection. A few examples are the works of Bohole *et al.* [26] where LRP maps were compared with other feature visualization through saliency maps techniques to detect AD, while Eitel *et al.* [33] applied them to uncover brain regions leading the differentiation between MS and healthy subjects. LRP was also very recently adopted for AD classification in Deatsch *et al.* [34] to evaluate the best imaging modality between T1-weighted (T1-w) and FDG PET over longitudinal data revealing that PET-trained methods outperformed MRI-trained ones, particularly when adding the longitudinal information, which instead had no influence on MRI-trained models. Finally, LRP was also applied by Dang and Chaudhury [35] to the estimation of brain connectivity score starting from MRI in order to determine the contribution of the higher order connectivity between two brain regions.

Other XAI methods

Other interesting XAI techniques are trainable attention models proposed in [36], which used trained attention to further amplify relevant areas and suppress irrelevant areas. Dubost *et al.* [37] exploited this method to detect brain lesions due to cerebrovascular diseases. They proposed a new weakly supervised detection method using neural networks, that computes attention maps revealing the locations of brain lesions. These attention maps are computed using the last feature maps of a segmentation network optimized only with global image-level labels. The proposed method can generate attention maps at full input resolution without the need for interpolation during preprocessing, which allows small lesions to appear in attention maps. Lian *et al.*[38] exploited this method to predict dementia status from brain MRI. They proposed a multi-task weakly supervised attention network to jointly predict multiple clinical scores from the baseline MRI data, by explicitly considering the individual specificities of different subjects. Leveraging a fully trainable dementia attention block, they claim that their method can automatically identify subject-specific discriminative locations from the whole-brain MRI for end-to-end feature learning and multi-task regression. For the purpose of creating saliency maps, Zintgraf *et al.* [39] modified prediction difference analysis [40]. Prediction difference analysis, which measures how the prediction changes if the pixel is considered unknown, assigns a significance value to each pixel if each pixel in an image is considered a feature. This was further developed by Zintgraf *et al.* [39] by adding conditional sampling, which limited the analysis to pixels that are difficult to predict by simply examining nearby pixels, and by adding multivariable analysis, which involved analyzing patches of connected pixels rather than individual pixels. They provided a comparison of brain MRI results between Human Immunodeficiency Virus (HIV) patients and healthy controls to provide an explanation of the classifier's choice. On various scales, Seo *et al.* [41] combined superpixels (or supervoxels for 3D) with prediction difference analysis. Since they follow image edges, these multi scale supervoxel-based saliency maps offered explanations that the authors described as visually appealing. A classifier could distinguish between people with AD and healthy controls by using the regions that the saliency maps identified as informative.

Finally, an XAI method that offers textual descriptions is a textual explanation. Such descriptions range from really basic traits to whole medical reports. They are beginning to be used in medical imaging as well mainly dealing with X-ray or Computed Tomography (CT) imaging for chest-based research. To the best of our knowledge, they have not yet been applied to brain MRI studies, opening the way to new research paths.

2.2 XAI validation and assessment methods

As XAI in medicine is in an early stage of the investigation, some issues have still to be addressed. Firstly, a deep validation of the XAI methods and secondly the increase in the simplicity of explanations while maintaining an elevated level of performance.

For the first criticism, it is important to provide a solid validation of the outcomes at a clinical level. For example, Lorenzi and colleagues [42] applied functional prioritization meaning that the candidate genetic variants resulting from Partial Least Squares (PLS) were subsequently screened for functional relevance by querying high-dimensional gene expression databases, such as the Genotype Tissue Expression project (GTEX) (gtexportal.org) [43] to be strongly validated.

Despite that there is no consensus regarding a validation proxy or protocol to evaluate explainability methods[44], many proxies and attributes have been proposed for the assessment of the explainability of different models. For example, Sundararajan *et al.* [45] proposed two axioms to evaluate explainability methods for Deep Neural Network (DNN) models, sensitivity and implementation of invariance. The first one has a double claim, firstly if there are two different predictions and they have the same input but are different in only one feature, that feature should not have a zero attribution; secondly if a model does not depend on some features to predict a value, then the importance of these features should be zero. The second axiom, implementation of invariance, argues that if two models are identical, trained on the same task and have an identical prediction, then they should have the same attribution for their networks. Montavon, *et al.* [46] defined other two attributes for XAI explanations which are explanation selectivity and explanation continuity. Explanation selectivity argues that if the relevant features that are identified by the explainability method are then removed from the model, there should be a sharp reduction in the model performance. Explanation continuity states instead that if two data points are equivalent, their explanation of the prediction should also be equivalent.

Focusing on the XAI method itself and not on the explanations, Silva *et al.* [47] proposed three complementary Cs to evaluate an explainability method. The three Cs are Correctness, Completeness and Compactness. Correctness implies that the explainability method should be accurate and measure the accuracy. Completeness indicates that it should be possible to apply explainability when the audience can verify its validation and it is quantified by the fraction of the training data set covered by the explanation. Finally, Compactness argues that the time needed to understand the explainability should be proportional to its length. Other two approaches were suggested by Herman *et al.*[48]. The first one is to conduct an experiment using simulated data with known characteristics to validate the correctness of the explainability method, while the sec-

ond one is to test the consistency and stability of the explainability method using a well-defined metric.

Moving to human interpretation of XAI explanations, Lipton [16] suggested three attributes to evaluate any explainability method that are simulatability, algorithmic transparency and decomposability. Simulatability measures to what extent a human can repeat or simulate the experiment based on the provided explanations. Algorithmic transparency measures how humans can fully understand a predictive model. Finally, decomposability quantifies the ability of the explainability method to explain each part of the model and its function (informative features, monotonic relationship, parameters, the output, etc)

Finally, a common criticism of DL and ML in general is ignoring the wisdom and expertise of hypothesis-driven research, which is the benchmark in the medical field. One of the emerging approaches is in fact to use expert human knowledge in combination with XAI to develop an interpretable model. This is consistent with the theory for which a DL framework could be reflected by performing XAI on a learner with a complete feature set, have the DL model generate results consistent with domain expertise and have the DL re-perform its calculations excluding the rules judged by the experts as superfluous [49]. As Smucny and colleagues [49] state, in fMRI this could be achieved by focusing on rules that involve brain regions known to be associated with the cognitive process of interest, an approach which could be easily tailored also on MRI and diffusion Magnetic Resonance Imaging (dMRI) fields.

The need of making the explanations more humanely understandable is also present in the very recent manifesto on explainability for artificial intelligence in medicine proposed by Combi and colleagues [50] in which they consider usefulness and usability as two of the four principles for XAI, together with interpretability and understandability. Though their distinction between explainability and interpretability is inherently different from our view shown in this Chapter, confirming that a clear definition of concepts is still not achieved, of interest is the introduction of the specific requirement of usefulness and usability in XAI methods. Usability refers to the ease with which a user can learn to operate, prepare inputs for, and interpret outputs of a system or component. Usefulness, on the other hand, is seen as the practical worth or applicability of a XAI system. Both hence relate to human utilization and exploitation of the explanations obtained through XAI methods.

2.2.1 Attributes for XAI validation

The desirable attributes of the obtained explanation, either from explainable or interpretable methods, can be summarized into four main classes which, in our view, are sta-

bility, consistency, plausibility and understandability, which encompass the attributes proposed in the previous Section.

The *stability* answers to the question '*Given similar instances, are also the relative explanations similar?*'. The concept of similarity is founded on two elements: (i) Instances must be close in the feature space; (ii) model predictions must be close. Indeed, a similarity based just on the values of the characteristics would not suffice to produce identical explanations. For instance, examining two sides of a model's decision boundary might result in distinct predictions and, consequently, explanations. However, feature values may be very similar. Stability is often not addressed among the theoretical assumptions on which the existing explainability approaches rely. Therefore, it does not always follow. Consequently, it is essential to develop a measure for evaluating this element. Among the proxies proposed above stability encloses the sensitivity [45] as well as the explanation selectivity and explanation continuity, which all represent different shades of stability. Also, the proxy proposed by Herman *et al.* [48] referring to the testing on simulated data can be fully framed in the stability analysis.

The second attribute is *consistency*, answering the question '*Do different explainability methods give, on average, similar explanations?*' In fact, it is important that given an instance and a prediction, different XAI methods should return similar explanations, hence feature importance should not depend on the chosen XAI method or its assumptions. This attributes encompasses the implementation of invariance proposed in [45], which even if referring to different prediction models instead of different XAI methods, is still an approach to test for consistency from a different point of view.

We proposed to add also *plausibility* among the validation attributes which assume a fundamental role, especially concerning XAI in the medical field. Plausibility answers to the question '*Do the obtained explanations sufficiently associate with what is known from literature?*'. This assumes a fundamental role since, in order to enable AI in medicine, clinicians must obtain confirmation that the results are, at least partially, in line with what is already present in the literature and with the prior knowledge on the faced problem. This has a double outcome. First, the literature validation of the obtained results, and second the gain in the reliability of new results, pointing in different directions which could really give a strong and original contribution to the research.

Finally, the *understandability* answers to the question '*Are the obtained explanations easily understandable by humans?*' For instance, even if the employed models are basic, there is no assurance that the explanations will be easily understandable as well. A basic linear regression with 50 features reveals an issue in fact the number of features has a significant impact on explanations. Importantly, in medicine, a good explanation might vary depending on the area of expertise of the person receiving it. For instance, a radiologist or a researcher specializing in medical picture analysis may find a visual

description specifying the location of the illness to be adequate. Oncologists, neurologists, and hematologists would most likely welcome the addition of XAI to their clinical decision-making framework. This framework would also include the patient's medical history, prior and current treatments, treatment choices, and anticipated consequences or outcomes. It is hence fundamental to have the doctor in the loop in order to choose the best XAI approach allowing to obtain the most useful explanations. Understandability is inherently different from interpretability which indeed indicates a class of XAI methods while understandability refers to human comprehension of the obtained explanations. The completeness and the compactness defined in [47], as well as the three metrics proposed in Lipton *et al.* [16] which are all human-related, namely the simulatability, the algorithmic transparency and the decomposability all fall in this class.

2.3 Conclusions

It is clear that there is still a lack of consensus in the literature on the taxonomy of XAI methods. With this Chapter, we clarified the main concepts related to XAI, as well as differentiating interpretability from explainability, each relating to clear and different XAI approaches and not being used as synonyms. It is indeed fundamental to differentiate between methods that are explainable by design (explainable) and methods that can be interpreted *post-hoc* (interpretable), also to better understand when to use one or the other approach. We also moved a step further defining four validation attributes that encompass the different proxies present in literature namely stability, consistency, plausibility, and understandability which could allow the evaluation of the quality of the explanations provided by existing XAI methods. Importantly, our contribution is the addition of plausibility to the list which is particularly relevant in the medical field. Stated that explainability methods are vital to gain a deep understanding of ML and DL model predictions their application must be faced with caution, verifying the presence of the four characteristics presented above in order to be practically used and earn reliable and meaningful insights on the approached problem.

Part of the work presented in this Chapter was recently published in [51].

3

General pipeline for XAI application to brain imaging

In this Chapter we aim at proposing a general pipeline for eXplainable Artificial Intelligence (XAI) application to medical imaging problems. Despite the multiple reviews on the argument [7, 9], an overview of the steps needed starting from the definition of the problem and concluding with XAI methods is still not present in literature. Here we will try to elucidate some key steps which we consider as fundamental for a correct application of Artificial Intelligence (AI) models in neuroimaging, with a focus also on the input Imaging Derived Phenotype (IDP)s and the specific models considered. We also add the XAI step proposing some example situations for when to use which method. We will consider the problem of predicting Brain Aging (BA) as our working example. Part of this Chapter was published in our recent review on XAI in BA.[51].

3.1 Use case: Brain aging

The study of BA has recently gained attention in the scientific community since developing accurate biomarkers for BA relying on neuroimaging data in combination with *ad-hoc* statistical analyses would open new perspectives in different domains, allowing to disentangle age-related from disease-specific changes and to track the disease progression at the single-subject level [52]. The prediction model, generally trained on large samples of controls, is fed with candidates endophenotypes and outputs the estimated, or predicted, age [53]. The so-called delta or gap is then defined, given by the difference between the predicted and the chronological age [54]. These resulting delta (hereafter referred to as brain-PAD [55]) reflects individual's deviation from the population norm, highlighting accelerated aging (positive delta) or resilience to aging (negative delta) [54], thus informing on the brain health status. brain-PAD measures are of value for assessing normal aging and disease, with recent studies revealing patterns of faster aging in several neurological and psychiatric pathologies, even prior to overt

disease manifestations [53, 52]. An important example can be found in the context of neurodegenerative conditions, where an initial study by [56] demonstrated significant differences between brain-PAD scores of controls/stable Mild Cognitive Impairment (MCI) and Alzheimer's Disease (AD) patients, and a more accurate prediction of conversion to AD when using the brain-PAD scores rather than neuropsychological tests. Different solutions have been proposed for tackling this problem, from the choice of the endophenotypes to the methodologies proposed for predicting BA. In the current State-Of-the-Art (SOA), such methodologies range from classical linear regression to Machine Learning (ML) models working with single/multi-modal data. The advent of publicly available large repositories of heterogeneous data called for new methods allowing to cope with high data dimensionality, Deep Learning (DL) being first in line [57]. This made stringent the issue of explainability/interpretability of the models outcomes especially considering the lack of ground-truth that is inherent to BA estimation.

3.2 XAI pipeline and guidelines

The initial steps required for the XAI process are presented in what follows, aiming at clarifying through SOA examples, the fundamental steps for XAI application in brain imaging studies. A similar analysis was recently proposed in [58], however multiple steps that we included were not present in their proposed pipeline, as well as the brain age literature examples. Interestingly, they propose XAI as a step of a feedback loop where firstly the chosen models is debugged and then tested extracting also explanations, at this point the end user can give feedback about the decisions obtained and the parameters which can be re-tested by retraining the model.

In our view, XAI has a twofold aim. The first is to obtain explanations for the models' decision once the AI model is well tested and validated and the achieved accuracy is satisfactory. The second aim is to use XAI in a feedback loop where, when dealing with a poorly performing model having no hints about the reason, XAI can help for checking the presence of issues related to the input data and also to select which features excluding/including in the following tests. In what follows we will present a possible XAI pipeline describing all the necessary steps enabling the building of a correct AI model, including XAI steps.

A. Data definition and pre-processing

This initial phase describes the choice of the appropriate data for the selected task, data cleansing, recovery/imputation, and top feature analysis. Some steps could be the processing of inaccurate, duplicate, corrupted, or incomplete datasets, whereas based on

the adopted model, data imputation refers to the replacement of missing data with substitute values, if the model does not account for missing data.

For BA, in the current literature, most predictive models rely on T1-weighted (T1-w) images as inputs [55], given their larger availability, reliability and easiness of interpretation. Depending on the granularity and on the given framework, several features are generally extracted from the T1-w images to be used as predictors in BA models. The easiest solution consists in using the raw whole-brain T1-w data, avoiding the step of feature engineering. Conversely, more recently there was a shift from voxel-based towards region-based approaches and extracted summary statistics for different Regions Of Interest (Region Of Interest (ROI)s), in particular cortical thickness, surface area and volume.

Nevertheless, neuroimaging modalities other than conventional T1-w Magnetic Resonance Imaging (MRI) can complement the picture provided by these data and inform on other relevant aspects, such as tissue microstructure or brain functioning. From diffusion Magnetic Resonance Imaging (dMRI) for example it would be possible to extract IDPs from diffusion maps obtained after model fitting, for which details will be given in Appendix A. Starting from these maps, the IDPs generally extracted are represented by the mean values calculated over WM maps or along different tracts, the latter identified either with tract-based spatial statistics (TBSS) [59, 60, 54, 55, 61] or tractography [62, 60, 55].

Besides analysing the brain architecture, important information on its functionality can be extracted by relying on functional MRI based on the Blood Oxygenation Level Dependent contrast (BOLD) functional Magnetic Resonance Imaging (fMRI) and Arterial Spin Labelling (ASL). fMRI scans can either be acquired during the execution of a given task (task-fMRI) or while resting (Resting State functional Magnetic Resonance Imaging (rs-fMRI)). Among the works here selected, task-related IDPs were used only within the United Kingdom Biobank (UKB) framework [60, 54, 55], and were represented by activation measures in regions derived by the group-level maps (i.e., median and 90th percentile for both the percent signal change and z-statistics). Conversely, several IDPs based on rs-fMRI data have been explored in this context. While measures related to the amplitude of low-frequency fluctuations and regional homogeneity were reported in a single study [61], features describing the Functional Connectivity (FC) patterns were usually employed in such studies. Moreover, in this step standardization and deconfounding could be applied. Those actions can be very crucial as well as highly influence the obtained results. Deconfounding is really common particularly in medical image analysis, yet no agreement is still present in the community.

Confounds

Confounds are variables which might introduce spurious associations between independent variables and strongly bias the resulting estimates. The criteria defining the role of confounds are far from trivial and are heavily context-dependent. In the neuroimaging and in general in neurodegeneration framework, for instance, age plays the role of variable of interest but also of a possible confound. These variables can be assigned to some common categories depending on their nature, such as subject-specific features (e.g. age, gender, education, intracranial volume, APOE), scanner / acquisition / processing parameters (e.g. centre, coil, head motion) and non-linear interaction terms [63]. Once the confounds are defined, common practice is to regress them out from the data as a pre-processing step (deconfounding) or add them as regressors in all the analyses. In the specific case of DL and 3D Convolutional Neural Networks (CNN)s, besides the deconfounding strategy, two main approaches were observed: 1) testing with a linear regression model the main effects of the covariates on the cross-validated brain-PAD estimates [19]; 2) adding them as inputs to the final CNN layer [64]. These play as additional constraints during training to limit the solution space of the network enforcing the net to more accurately capture the relevant factors and their interactions. All these aspects deserve further investigations, as they would have a great impact on the statistical power of the analysis as well as on the outcomes of the association analyses.

B. Feature selection

This step can be considered as optional since it highly depends on the chosen data and model. Dimensionality reduction is usually applied before feeding the IDPs to the prediction model. This is generally based on Principal Component Analysis (PCA) [56, 65, 66], Canonical Correlation Analysis (CCA) [66] and Independent Component Analysis (ICA) [67, 60] to find the 'modes' which embed the information deriving from either single or multi-modal data in a smaller feature space. These were recently complemented by feature handcrafting as in [59]. It might be useful to remind that while PCA projects the data along the dimensions of maximum variance, CCA maximises a given similarity measure, most commonly the correlation, among the data [66]. When ICA is applied for dimensionality reduction, data are projected in a space where the components are assumed to be non-Gaussian and as much independent as possible. Such data projection can provide additional information on the population regarding its intrinsic variance or similarity. The latent variable models yield loading vectors for every component, which quantify the contribution of each feature to each component. This could be also considered a preliminary step towards explainability since the obtained feature grouping

more than increasing the performance of the model, the associated loading still allow to understand which feature contributed mostly to each latent component.

C. Data modeling

In this step the aim is the choice of the data modeling technique to be applied. In particular, if the addressed problem is quite simple linear approaches which are intrinsically explainable could be the right choice. The choice of the model depends on multiple factors such as data availability, feature number, multimodality, data input type etc.

For example, for BA the classical linear regression approach is still widely used in literature. From a design perspective, most BA studies use large training sets of subjects within a supervised learning framework to build the age prediction model having brain IDPs as independent (predictor) variables along with chronological age as the dependent (outcome) variable. Simple Linear Regression (SLR) and its extension to accommodate multiple predictors (Multiple Linear Regression (MLR)) have been initially proposed as simple though effective methods to model their statistical relationship.

However, more recently, CNN architectures were proposed to estimate BA using 3D T1w images (for more details on the CNN please refer to the Supplementary Materials-S3). While numerous variants of CNNs are present in the current literature, the solutions explored in the BA context so far are mostly based on Visual Geometry Group (VGG) Network and Residual Neural Networks (ResNet) architectures.

In order to reduce the model reliance on pre-processing steps such as image realignment or registration, all the studies tend to apply only minimal pre-processing to the input data. In addition, different regularization and data augmentation strategies, including dropout, data rotation, translation, mirroring, scaling or addition of random noise, are usually applied during the training phase to avoid overfitting and improve generalization.

D. Training, validation and testing

In order to validate an AI model, most studies employ a cross-validation approach in which a proportion of the samples from the entire training group is left-out (typically ranging between 10 and 20%) whilst the remaining largest portion is used to train the model. This is then applied to the left-out group to predict the individual ages. This operation is performed until the whole set of disjoint partitions has been explored. Spacifically for BA prediction, the model performance is evaluated relying on predefined measures, typically: Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and Pearson correlation coefficient (r-value) between the estimated BA and the chronological age. Whilst these measures are largely used to assess the accuracy of the models,

they should be interpreted with caution especially when comparing the results across studies as they are affected by several factors, including the age range of the sample which could lead to changes in the performance.

Usually, the best model is retained for testing on an unseen set of samples (testing set), comprising healthy and/or diseased subjects, and generating individual predictions. This operation allows to further validate the model and to prove its generalizability across several samples, possibly coming from different sources and databases.

E. Explanations

This phase gives an explanation for each choice so that the algorithm's significance and behavior may be understood. The explanation includes extensive justifications for every model decisions, including preprocessing, prediction method, classification, evaluation, and conclusion. As the explanations comprise the core content of XAI, it increases the end user, domain experts, and client acceptance of the deployed system.

Ante-hoc models

- `When to use them?`
  As presented in the previous Chapter, *ante-hoc* XAI methods are deeply linked to the data analysis method, since they do not require an additional step after model fitting. For their nature they are highly model specific, in fact the most common *ante-hoc* methods are linear regression, logistic classification or decision trees for which the explanations comes directly from the analysis of the coefficient or the obtained rules. Other methods such as PCA, ICA for feature reduction are still *ante-hoc* explainable methods since they allow to retrieve feature importance by observing weights and loadings. Such methods can be applied when, given the addressed problem, they can reach satisfying performance not calling for the adoption of more complex non linear and deep models. In this case explainability is hence achieved through constraints imposed on the complexity of the ML model.
- `How are they adopted in BA?`
  In [68], PCA was employed as a preliminary step for their linear model relying on fMRI features. The latent variables corresponded to FC networks describing connectivity patterns. A similar approach was presented by Smith et al. [60] which used PCA and ICA to extract 62 modes of subjective variability, acting as aging brainprints. Each mode represented different aspects of BA, showing distinct patterns of functional/structural brain changes as well as selective associations with genetics, lifestyle, cognition, physical measures and disease. In the scalar-on-image procedure proposed by Palma and colleagues [69] the obtained eigenfunctions encoded

the main differences among healthy, MCI and AD subjects. They observed, for instance, that the first eigenfunction allowed distinguishing the lateral ventricles from the rest of the brain, concluding that the scores for this eigenfunction could be correlated with diagnosis and chronological age. Finally [66] directly employed CCA to visualize the features significantly contributing to BA prediction.

Post-hoc models

- `When to use them?`
  In this class fall the majority of XAI methods since *post-hoc* models are applied after training of a complex model like a deep neural network or random forest for which there is not necessarily access to the internal structure, commonly referred as 'black box' models. With *post-hoc* methods it is hence possible to obtain explanations of existing models and are referred to as *post-hoc* explanations. Based on the input data different explanations can be obtained. In neuroimaging the most common data types are either images/volumes or tabular data. The most common approaches for tabular data are feature importance based either in permutation on perturbation of the input, while in the case of image/volumes saliency based explanations assume the central role, either obtained through gradient backpropagation, decomposition, input perturbation among the others.
- `How are they adopted in BA?`
  - *Permutation-based feature importance*. Among the works exploiting the full volumetric information, either considering T1-w volumes or other acquisitions, [70] applied permutation feature importance to analyze the importance of different White Matter (WM)/Grey Matter (GM) brain regions by quantifying their contribution to their 3D-CNN predictions.
    [71] used permutation feature importance, measured as an increase of Mean Square Error (MSE), to interpret the outcomes of a Random Forest (RF) model including T1-w, dMRI and fMRI tabular IDPs revealing that the model integrating all modalities was mostly driven by the cortical thickness, T1-w/T2-weighted (T2-w) ratio and subcortical volumes. A similar approach was followed in [61] where the feature importance was calculated over the reduction of the R^2 of the considered regression models which were Ridge regression, Support Vector Regression, Gaussian Process Regression and Deep Neural Network (DNN). Finally, Engemann et al. [72] proposed a stacked model composed by Ridge regression and RF and interpreted the model through permutation feature importance over the RF model given tabular metrics of fMRI, Magnetoencephalography (MEG) and structural Magnetic Resonance Imaging (sMRI). They were able to unravel the presence of an additive component between MEG and fMRI phenotypes.

- *Perturbation-based feature importance*. SHapley Additive exPlanations (SHAP) model is among the most famous perturbation-based feature importance methods, being model agnostic and hence applicable to any regression method. It was employed for example in [19] on multiple ML methods, importantly revealing that the anatomical changes in a common set of regions drive model predictions of age, regardless of the model type. The regions found following the most important features reflected developmental growth patterns of the cortex in childhood and adolescence. On the contrary, input perturbation can be also applied by training the model multiple times, each time with a different subset of features. This approach was followed in [73], the authors assessed the specificity of spatial brain-PAD patterns by training prediction models using each time only a subset of tabular features derived from occipital, frontal, temporal, parietal, cingulate, insula and cerebellum regions, respectively. They highlighted some differential spatial patterns across the eleven different clinical groups that were analyzed. This analysis allowed to uncover that relative aging across regions showed opposite patterns in neurodevelopmental (schizophrenia) versus neurodegenerative (Multiple Sclerosis (MS) or dementia) disorders.
- *Saliency maps*. Starting from gradient based methods, [74] exploited SmoothGrad to produce explanation maps for their CNN model. Relying on a large data sample, they were able to create aggregated population-based explanation maps. The similarity between each pair of group explanation maps was assessed and clustering was applied to highlight the brain regions that contributed the most to age prediction. Such regions showed the highest correlation to the brain-PAD, indicating the specificity of the derived maps to their model. Grad-Class Activation Mapping (CAM) approach was instead followed by both Wang et al. [75] and Feng et al. [17] retrieved from their T1-w-based 3D-CNN to show the relative importance of different regions for BA prediction. In detail, [75] found that while the network looks at the entire GM, the attention pattern is quite complex, suggesting that brain-PAD is more related to specific features than to global measures of GM volume when predicting BA. Feng et al. [17] moved a step forward. Besides the *post-hoc* Grad-CAM derived saliency maps, they applied ablation analysis methods, which indeed fall into input perturbation, focusing on part of the input data. They highlighted patterns of neuroanatomical contributions of normal aging providing evidence for the prominence of frontal regions in all age epochs in the adult lifespan.

F. Validation of XAI method

Evaluation of XAI as presented in Chapter 2, Section 2.2 is not yet a standard technique in medical image analysis articles. Stability, consistency, understandability, and plausibility should be tested for the specific XAI method chosen. For stability, tests such as evaluating the explanations for similar instances could be performed, as well as removing some features from the input data and recalculating the feature importance to assess the stability of the obtained ranking. Concerning consistency, multiple XAI methods can be applied to the same model in order to evaluate their consistency on the same problem. Finally, plausibility should be tested by searching for validation in the literature of the obtained results. This last attribute is really important in medicine since in order to build trust in AI literature support is always highly preferred by clinicians. Concerning understandability, the chosen XAI methods should give explanations with which it is possible to easily deal with, such as saliency maps or feature ranking. For example, explanations referring to the model structure or connections, despite being interesting are not directly understandable and hence useful for interpreting the model. Useful strategies to validate XAI models will be presented in Part IV.

G. Re-evaluation feedback loop

When explanations are obtained, if the model performance is not satisfactory, the end user can interact with the algorithm by giving the system the necessary feedback for each choice and parameter used, which can then be examined and changed in the following tests. Sample situations can be realizing that the model is making decisions based on the background or some biases present in the input images, or on trivial input features. As a result, it improves future versions of training data and weight augmentation while also facilitating usability and incorporating the end user into the system.

3.3 Conclusions

In this Chapter we proposed a guideline that could help to unravel the huge world of XAI, with a particular focus on medical framework. Even if our use case was BA estimation, we described the general issues linked to data processing, feature extraction, model selection and finally XAI application giving useful and valid hints expendable in any other XAI application.

Part of this Chapter was recently published in [51].

Part II

XAI for subject stratification in Multiple Sclerosis

4

Explainable multivariate modeling suggests the link between brain microstructure and cognitive impairment in Multiple Sclerosis

In this Chapter we will present our preliminary work on Multiple Sclerosis (MS). MS is a disease affecting Grey Matter (GM), usually devoted to executive functions possibly susceptible of cognitive impairment. Potentially, diffusion Magnetic Resonance Imaging (dMRI) can highlight microstructural changes associated with cognitive impairment. Aiming at shading lights on the joint variation between the cognitive assessment in MS and the dMRI derived GM microstructural alterations, we fitted a Partial Least Squares (PLS) regression to data collected on a cohort of 36 patients. Results showed that 45% variation of the data can be explained by an anti-correlation between anisotropy and restriction dMRI features, and diffusivity ones, together with relevant neuropsychological tests scores. Moreover, the data projected to the PLS derived latent space were distinguishable between cognitively impaired and preserved individuals, with a model significance $p < 0.05$.

4.1 Introduction

MS is a chronic, inflammatory, neurodegenerative disease of the Central Nervous System (CNS), characterized by the accumulation of White Matter (WM) and GM damage. It affects the brain and the spinal cord, resulting in physical and cognitive disability due to the damage of the myelin sheath wrapping WM axons as well as neurodegeneration and axonal loss [76].

Different studies considered MS as WM disorder, considering WM derived features for MS analysis and staging [77, 78] but more recently it has been hypothesized that the distinction might be related to the appearance of lesions in the GM whose impairment has been found to be associated with the early onset of the pathology [79, 80].

Along with the disease's impact on the physical disability, another debilitating consequence is the cognitive impairment that affects from 43% to 70% of the patients [81].

It has been shown that an early diagnosis of cognitive impairment is particularly important because it has several consequences, e.g., it is able to predict the conversion to definite MS [82], etc.

The cognitive impairment in MS, has been usually associated to a macro-scale brain damage such as atrophy, whose information is extracted from classical T1-weighted (T1-w) Magnetic Resonance Imaging (MRI) [83]. Compared to classical T1-w MRI, MRI has the advantage to depict tissue alterations at a micro-scale level of detail. In particular, it captures the signal emitted by the water particles that diffuse within the structures formed by the different coexisting neuronal cells. Signal reconstruction models, such as Diffusion Tensor Imaging (DTI) MRI [84] or more advanced 3D Simple Harmonics Oscillator based Reconstruction and Estimation (3D-SHORE) [85], allow to extract several maps describing the underlying microstructural properties of the brain.

Exploring the link between possible GM alterations in terms of dMRI derived features and cognitive assessment of MS patients may reveal insightful aspects of the disease pathophysiology. To the best of our knowledge, only in one case this has been done with advanced dMRI signal reconstruction models (different from 3D-SHORE) [86]. Of note, these studies usually employed classical univariate statistical approaches, although they suffer from several disadvantages. e.g., i) they usually need mass-univariate testing of several multivariate features, ii) the correlation between features must be tested individually for each pair of them, iii) the aforementioned limitations introduce the multiple comparison problems, iv) the significance of the results is severely compromised by the employment of high number of features, etc. Since many dMRI indices can be extracted from a single subject, and they can be collected as features of a values distribution within set of Region Of Interest (ROI)s, or even as voxel-level metrics, there is an inherent limitation when facing the problem with a univariate statistical technique. In this context, the introduction of multivariate approaches to data analysis allows to use of the entire information derivable from neuroimaging and clinical information and may reveal hidden characteristics of the link between the two. One of these methods is the Partial Least Squares regression, which grounds on simultaneous regression and dimensionality reduction of both the independent and dependent variables [87].

We chose to investigate the application of the PLS regression in our study to enable a joint description of the correlation patterns between neuroimaging and cognitive assessment with a relatively simple implementation. In particular, we aimed at highlighting this variation in an MS cohort of patients divided in Cognitive Impaired (CI) and Cognitive Preserved (CP) individuals.

4.2 Materials and Methods

The study population included thirty-six MS patients with a suspected cognitive impairment, who gave their written informed consent prior to undergoing MRI data acquisition and neuropsychological cognitive assessment.

MRI data were acquired using a 3T Philips Achieva scanner (Philips Medical Systems, the Netherlands) with an 8-channel head receiver coil. The MRI scan consisted of a two-shells dMRI (Repetition Time (TR)/Echo Time (TE) = 9300/109 ms, flip angle = 90°, Field of View (FOV) = 112 × 112 mm^2, 2-mm isotropic resolution, 62 slices, *b*-values = 700/2000 s/mm^2 with 32/64 gradient directions respectively and 7 *b0* volumes). In addition, T1-w (TR/TE = 8.1/3 ms, 180 slices, 1-mm isotropic resolution), T2-weighted (T2-w) (TR/TE = 2500/228 ms, 180 slices, 1-mm isotropic resolution), and 3D Fluid-Attenuated Inversion Recovery (FLAIR) (TR/TE = 8000/290 ms, 180 slices, 0.9 × 0.9 × 0.5 mm^2 resolution) images were acquired for anatomical information.

The neuropsychological evaluation relied on the Brief Repeatable Battery (BRB) of neuropsychological tests, along with the Stroop Test comprising the Effect Interference Time (ST-EIT) and the Stroop Test comprising the Effect Interference Error (ST-EIE). The BRB is composed of the Selective Reminding Test including the long-term storage (SRT-LTS), the Selective Reminding Test including the consistent retrieval (SRT-CLTS), the Selective Reminding Test including the delayed recall (SRT-D), the Spatial Recall Test (SPART), the Spatial Recall Test Delayed (SPART-D), the Symbol Digit Modalities Test (SDMT), the Paced Auditory Serial Addition Task (PASAT) (with a rate of number presentation of 3 s [PASAT-3], and 2 s [PASAT-2]), and the Word List Generation (WLG). All the assessments were done within 2 years of distance from the MRI acquisition.

Grounding on this information, an expert neuropsychologist (M. P.) classified the patients in CI and CP according to the results of a previous study [88]. The division led to the first group of 25 CI (15 females, age: 48.0 ± 7.2 years [y], disease duration: 9.2 ± 7.6 y) and the second group of 11 CP (9 females, age: 41.9 ± 8.7 y, disease duration: 8.6 ± 8.3 y).

dMRI data were corrected for motion, eddy-current and Echo Planar Imaging (EPI) distortions. Brain extraction and masking were then performed, and the transformation matrix registering T1-w image to the subject's mean *b0* volume was calculated. Preprocessed data were fitted with DTI and 3D-SHORE models, extracting Fractional Anisotropy (FA) and Mean Diffusivity (MD) for the former, and Generalized Fractional Anisotropy (GFA), Propagator Anisotropy (PA), Mean Squared Diffusivity (MSD), and Return to the Origin Probability (RTOP), Return to the Axis Probability (RTAP), Return to the Plane Probability (RTPP) for the latter. A rigid co-registration between FLAIR and T1-w images of each subject was estimated since they were both acquired on the

same person. Thus, MS lesions were automatically segmented on the first and subsequently filled on the second image. This was done using the Lesion Prediction Algorithm (LPA) available in the Least Segmentation Toolbox (LST) for SPM12 (www.statistical-modelling.de/lst.html). The so-filled T1w image was then parcellated according to the Desikan-Killiany atlas in Freesurfer software (http://surfer.nmr.mgh.harvard.ed\u/). Only a set of regions including thalamus, caudate, putamen, hippocampus, posterior cingulate cortex, superior-frontal gyrus, insula, lateral occipital cortex, lingual cortex, pericalcarine and precuneus were retained for further analyses [89]. These ROIs were used as masks to calculate the median value for each dMRI microstructural index, after the application of the previously calculated transformation matrix.

Both dMRI features and cognitive scores were individually standardized by the population's mean and standard deviation, and subsequently deconfounded by covarying for gender, age and disease duration. A PLS regression with Least Absolute Shrinkage and Selection Operator (LASSO) regularization was then performed using dMRI features as X (matrix of dimensions 36×88, where 88 is given by 8 indices for 11 ROIs) and the neuropsychological scores as Y (matrix of dimensions 36×11). More in detail, the Nonlinear Iterative PArtial Least Squares (NIPALS) algorithm [90] was used to decompose both X and Y and find the principal components.

The PLS model's eigen-vectors and eigen-values were then analyzed, and a permutation test was performed to evaluate the regression's significance. In detail, the rows of the matrix Y were permuted with 10e4 permutations, and the resulting p was calculated as the number of times that the sum of the obtained eigenvalues outperformed the one obtained by the tested model divided by the number of permutations [91].

4.3 Results

The PLS regression applied to our data showed the first eigen-component as the most relevant latent component retrieved by the model since alone accounted for 45% of the total variation of the data versus the others 10. In Figure 4.1 is shown this eigen-component, describing the covariance between dMRI microstructural features and cognitive assessment. In detail, it highlighted for the dMRI features an anti-correlation between anisotropy and restriction indices (FA, GFA, PA, RTOP, RTAP, and RTPP), and diffusivity ones (MD and MSD). Moreover, it can be observed that cortical ROIs usually had a higher impact on the eigen-component compared to subcortical ones (except the hippocampus). This landscape of dMRI derived features correlations is completed by the description of the cognitive assessment variation. In particular, the BRB derived

Chapter4/Figures/PLS_canonical_LASSO_Ydeconfound.png

Fig. 4.1: PLS component's weights for the dMRI and the cognitive assessment features. dMRI derived weights are divided by index (colors). The ROIs' order is the same for each index and it is: thalamus, caudate, putamen, hippocampus, posterior cingulate cortex, superior-frontal gyrus, insula, lateral occipital cortex, lingual cortex, pericalcarine and precuneus. The opacity differentiates the subcortical (light) from the cortical ROIs (dark).

scores showed a correlation in agreement with anisotropy and restriction indices, while both ST-EIT and ST-EIE seems to be not relevant in the data variation description (low PLS weighths).

The projections of X and Y features to the latent space represented by the mentioned PLS component is shown in Figure 4.2. As it can be seen from the scatterplot, despite the relatively low number of subjects, CI and CP patients can be easily distinguished by a visual inspection.

Chapter4/Figures/scatter_first_comp_lasso_Ydeconfound.png

Fig. 4.2: Latent representation of the study cohort. The two classes of patients are represented in different colors: CI in red and CP in blue.

Finally, the permutation test confirmed the significance of our model resulting in $p = 0.018$.

4.4 Discussion

The PLS model performed in this work enabled the representation of different types of features, such as dMRI and neuropsychology, to a common space describing their covariance. Moreover, the data projection of the data to the latent space appeared optimal to classify CI and CP MS subjects of the study.

As aforementioned, to the best of our knowledge, few studies investigated the link between the cognitive assessment of MS patients and the possible GM microstructural alterations as derived by dMRI. In the cases found, classical univariate statistics were used, which did not allow the employment of the whole possible information for a joint description of their link.

However, despite the different statistical approach and the low use of advanced dMRI signal models observed in the literature, our results derived from DTI seem in line with the findings of Preziosa et al. [92]. In particular, they considered a population of relapse-onset MS patients divided in CI and CP like us. They found an opposite trend of cortical MD and FA as they significantly increased and decreased, respectively, in CI compared to CP. Besides this, Planche et al. [93] focused on microstructural modifications of the hippocampus in MS patients compared to healthy controls, looking for a

correlation with memory impairment. They also found an opposite trend between indices defined by a decreased FA together with an increased MD, characterizing pathological subjects. Interestingly, they found that MD was able to distinguish between patients with and without memory impairment.

Concerning the additional indices used in this study, we observed coherence between FA and the other anisotropy indices (GFA and PA) in agreement with the similar tissue property mapped. The same can be said for the MD compared to MSD, while restriction indices confirmed the expectation of coherence with anisotropy indices. In fact, they grow when the structure in which the represented diffusion process is more restricted, e.g. in corpus callosum, as well as anisotropy [94].

In addition, our model showed a usually higher impact of the cortical ROIs on the eigen-component compared to subcortical ones. This observation was in agreement with the higher involvement of the cortex in high-order cognitive abilities [95]. However, despite the executive functions are usually associated with the frontal lobe, recently, also posterior and subcortical ROIs have been found to play a role in cognitive processing.

Finally, we acknowledge two main limitations of our study consisting of the small size of our cohort and the absence of a group of neurologically healthy subjects. Overcoming these two issues would be helpful for the improvement of our PLS model generalization potential, and to reveal more specific differences between disease stages.

4.5 Conclusions

The joint variation of dMRI derived indices and neuropsychological test scores could significantly model the variation of the data. Moreover, different forms of cognitive impairment were qualitatively distinguishable in the latent space created by the PLS regression. Thus, multivariate approaches to statistical analysis combining neuroimaging and clinical study may have the potential of depicting subtle differences in different forms of the MS pathology.

The work presented in this Chapter was published in [96].

4.6 Compliance with Ethical Standards

This study was performed in line with the principles of the Declaration of Helsinki. Approval was granted by the local Ethics Committee (MSBioB Biological bank - A.O.U.I. Verona, protocol no. 66418, 25/11/2019).

5

Interpretable Deep Learning as a means for decrypting disease signature in multiple sclerosis

Acknowledged the relevance of diffusion Magnetic Resonance Imaging (dMRI) indices in Grey Matter (GM) for Multiple Sclerosis analysis, and its relation with cognitive impairment resulted from the study described in the previous Chapter, here, our aim is to decrypt the microstructural signatures of the Primary Progressive Multiple Sclerosis (PPMS) versus the Relapsing-Remitting Multiple Sclerosis (RRMS) state of disease based on diffusion and structural Magnetic Resonance Imaging (MRI) data.
Firstly, as a benchmark, we relied on structural T1-weighted (T1-w) MRI and a 3D Convolutional Neural Networks (CNN) trained to stratify PPMS and RRMS subjects. Within this task, the application of feature visualization methods, based on relevance decomposition such as Layer-wise Relevance Propagation, allowed detecting the voxels of the input data mostly involved in the classification decision, potentially bringing to light brain regions that might reveal the disease state. In particular, the so-called 'winning class' Layerwise Relevance Propagation (LRP) was adopted in this study.
In the second step, a selection of microstructural descriptors, based on the 3D Simple Harmonics Oscillator based Reconstruction and Estimation (3D-SHORE) and the set of new algebraically independent Rotation Invariant Features (RIF), was considered and used to feed as well 3D CNN models. Classical Diffusion Tensor Imaging (DTI), which are Fractional Anisotropy (FA) and Mean Diffusivity (MD) were used as benchmark for dMRI. Finally, T1-w images were also considered for the sake of comparison with the State-Of-the-Art (SOA). A CNN model was fit to each feature map and LRP heatmaps were generated for each model, target class and subject in the test set.
Average heatmaps were calculated across correctly classified patients and size-corrected metrics were derived on a set of Region Of Interest (ROI)s to assess the LRP contrast between the two classes.
Our results demonstrated that dMRI features extracted in GM tissues can help in disambiguating PPMS from RRMS patients and, moreover, that LRP heatmaps highlight

areas of high relevance which relate well with what is known from literature for Multiple Sclerosis (MS) disease. Within a patient stratification task, LRP allows detecting the input voxels that mostly contribute to the classification of the patients in either of the two classes for each feature, potentially bringing to light hidden data properties which might reveal peculiar disease-state factors.

5.1 Introduction

MS is a chronic, inflammatory, neurodegenerative disease of the Central Nervous System (CNS), characterized by the accumulation of White Matter (WM) and GM damage. It affects the brain and the spinal cord, resulting in physical and cognitive disability due to the damage of the myelin sheath wrapping WM axons as well as neurodegeneration and axonal loss [76]. Four principal clinical phenotypes of MS have been described, among which RRMS and PPMS MS are the most common [97, 98]. While demyelination and atrophy characterize both forms, their patterns and distribution vary across the brain, suggesting that different driving mechanisms might underpin these two main clinical manifestations [99]. Therefore, there is a growing clinical need to find specific fingerprints to distinguish between them in order to enable precision medicine, that is patient-specific treatments with clear clinical impact on treatment decision-making [98]. However, the mechanisms driving MS are still largely unknown, calling for new methods allowing to detect and characterize tissue degeneration since the early stages of the disease.

dMRI is increasingly exploited for assessing microstructural alterations occurring in MS [100, 101]. This technique allows to define numerical indices that describe the brain tissue microstructure based on the measurements of signal decay along a predefined set of directions, providing an *in-vivo* indirect measure of the geometry of the diffusion pores [102, 85]. In particular, novel acquisitions based on multi-shell schemes have opened the way to the definition of a wider set of indices capturing microstructure degeneration and informing on the underlying disease process [100].

Diffusion signal models are generally tailored on WM [103] and are well suitable for modeling WM damage and structural connectivity alterations due to the disease. However, their exploitability for deriving neuroanatomically plausible microstructural descriptors from GM is far from trivial and has still to be proven. In recent years, several studies have attempted the characterization of GM modulations through dMRI acquisitions in different pathologies such as Alzheimer's Disease (AD) [104] and migraine [105]. In MS both classical and advanced diffusion models were employed to investigate the disease patterns in different phenotypes [78, 106] and to longitudinally moni-

tor patients over time [107]. Their findings strengthened the hypothesis of a GM modulation in MS and highlighted the dMRI sensitivity in detecting those changes. For the specific task of disambiguating PPMS from RRMS subjects, microstructural indices derived from the 3D-SHORE model [102, 85] were used to demonstrate that the probability density function of the Return to the Plane Probability (RTPP) was significantly different between the two groups in *Hippocampus* relying on histogram features [108].

In the context of patient classification from neuroimaging data, CNNs have recently gained popularity thanks to their ability in solving complex classification tasks, though in general require large amounts of data for training due to the high number of parameters that need to be calculated. Besides the availability of big data, one of the main bottlenecks for the use of CNNs for clinical purposes is that they are notoriously hard to interpret in retrospect. For this reason, Deep Learning (DL) methods, including CNNs, are often criticized to be non-transparent and still considered “black boxes". Therefore, the availability of a means allowing to interpret the network decisions becomes the key element for their exploitability.

In the last years, a number of solutions have been proposed for visualizing what is actually learned by a CNN. Besides straightforward methods such as the extraction of activations and weights of the different layers, among the most widespread methods are: (i) sensitivity analysis or BackPropagation (BP) [109], in which the relevance score is calculated as the gradient of the output probability given the input, computed through the backpropagation algorithm; (ii) Guided BackPropagation (GBP) [22], which modifies BP by setting to zero the negative gradients; (iii) Deconvolution and occlusion [110], where recursively a part of the input image is covered by a black patch and the network output recalculated in order to assess the changes in the classification probability under the assumption that the covered region was relevant for the classification; and (iv) LRP [32], which allows to detect and visualize in a relevance *heatmap* the voxels of the input data that mostly contributed to the classification decision. To this end, the LRP algorithm uses the network weights and the neural activations resulting from the forward pass to propagate the output back through the network up until the input layer, in a backward pass.

5.2 XAI: application to Multiple Sclerosis

Focusing on eXplainable Artificial Intelligence (XAI) applications, this precise task was previously approached by [33], which employed 3D-CNNs for the classification between MS subjects and healthy controls based on structural MRI data. They initially pre-trained a 3D-CNN consisting of four convolutional layers followed by exponential

linear units and four max-pooling layers on a large data sample (921 subjects) from the Alzheimer's Disease Neuroimaging Initiative. Afterward, they specialized the CNN to discriminate between MS patients and controls on a smaller dataset of 147 subjects, reaching a classification accuracy of 87.04%. As the final analysis, they used the LRP heatmaps to assess the most relevant regions for the classification, analyzing both positive and negative relevance given their patients *versus* controls classification task. Feature visualization was also employed in [111] to distinguish 66 control subjects from 66 MS patients. They relied on Susceptibility-Weighted Imaging (SWI) and a 2D-CNN, since for each SWI volume they considered only one single 2D projection in a transverse orientation. The CNN was composed of five convolutional layers with ReLU activation functions followed by max-pooling layers and two final fully-connected layers. To interpret the classification decisions they investigated three different feature visualization methods, namely LRP, Deep Learning Important FeaTures (DeepLIFT) [112] and BP as reference. The resulting maps were analyzed with perturbation analysis. In perturbation analysis, information from the image is perturbed region-wise from most to least relevant according to the attribution map. The target output score of the classifier is affected by this perturbation and quickly drops if highly relevant information is removed. The faster the classification score drops, the better an interpretability method is capable to identify the input features responsible for correct classification. Their results highlighted the outstanding performance of LRP maps and DeepLIFT over simpler methods, strengthening the suitability of such methods to address clinically relevant questions.

However, the specific problem of stratifying MS patients according to their phenotype is still unexplored in literature. Only a few works were found addressing this task. [113] combined graph-based CNN with structural connectivity information from dMRI, relying in particular on a network-based representation of the structural connectome. They aimed at distinguishing between 90 MS patients divided into four clinical profiles, namely clinically isolated syndrome, RRMS, Secondary-Progressive MS (SPMS) and PPMS, and 24 healthy controls. The combination of different local graph features, such as node degree, clustering coefficient, local efficiency, and betweenness centrality allowed to achieve accuracy scores higher than the 80%. Zhang and colleagues [30] moved a step further by introducing feature visualization methods to investigate the MS patients' stratification. They relied on structural MRI and compared six different 2D-CNN architectures for classification into three classes, namely RRMS, SPMS and controls. Furthermore, they applied three different feature visualization techniques (Class Activation Mapping (CAM), Gradient (Grad)-CAM, and Grad-CAM++ [24, 25]) to achieve increased generalizability for CNN interpretation. Their results showed that Grad-CAM

had the best localization ability in finding differences between RRMS and SPMS for discriminating brain regions.

Among these, we consider the LRP to be the most promising tool for two main reasons. First, it provides an individual heatmap for each subject lying in the same space as the input image, indicating the weight of each voxel for the final (individual) classification decision. Second, LRP heatmaps have proven to be more eloquent than those provided by GBP [22] in that they reflect image-specific relevance, whereas GBP, relying on gradients, tends to emphasize the areas that are more susceptible to changes that might not coincide with the areas on which the CNN bases the decision [26].

In [33], LRP was employed on CNNs for classifying between MS subjects and healthy controls based on structural MRI data while [111] compared multiple visualizations methods applied to the same task relying on SWI. The specific problem of MS patients stratification was addressed from a DL point of view in [113] by combining CNNs and graph metrics derived from structural connectivity. Finally, [114] and [115] used a 2D-CNN model on structural MRI data for classifying physiological versus pathological subjects without stratification.

To the best of our knowledge, no attempts have still been made for exploiting LRP in the PPMS versus RRMS patients stratification task. Therefore, the objective of this work is twofold: disambiguating the considered groups of MS patients relying firstly on T1-w MRI and secondly on advanced dMRI models and DL techniques focusing on GM, and decrypting CNNs decisions through the adoption of LRP.

For the first aim, we relied solely on T1-w MRI input in order to obtain a benchmark for the subsequent analysis. This preliminary work will be presented in Section 5.3. Then in Section 5.4, considering the increased interest in assessing the role of GM in the MS disease fingerprinting, we will present our work relying on three dMRI signal models to derive microstructural indices with the specific goal of assessing their sensitivity to the microstructural contract between the PPMS and RRMS phenotypes. Being aware of the potentials and limitations of the considered models, we specifically aimed at capturing possible feature modulations in GM leaving the biophysical interpretation of the results to further investigations relying on multimodal acquisitions.

For both the second objectives we build on the claim in [26], that LRP has the potential to answer the question “What *speaks for* AD in this particular patient?" providing guidance to understanding the mechanisms ruling the disease. In our work, such a question can be reformulated as “What *speaks for* PPMS in this particular patient?", which is the core question that was addressed. This is a key issue to be solved and a very challenging problem because of the subtle intra-pathology tissue modulations differentiating the two stages of the disease. In this respect, the goal of this work was to investigate whether CNNs were able to blindly capture such subtle differences while

providing insightful information about the underlying mechanisms through the observation of LRP maps. Restraining to these two categories represents the worst case from the classifier perspective, because of the subtle microstructural differences across the two phenotypes. However, we consider this as an important task because it is close to clinical practice conditions where a matched cohort of control subjects could not be available. In particular, as LRP allows to map the value of the network decision function onto the input voxels shedding light on the reasons behind the classification decisions, it can potentially provide hints for the interpretation of the mechanisms at the basis of the MS disease course besides the primary classification task, opening new perspectives for diagnosis, prognosis and treatment.

5.3 Preliminary analysis based on T1-w MRI and winning class LRP feature visualization

This Section will present the preliminary results for the stratification of RRMS and PPMS patients based on T1-w MRI. The LRP visualization, in particular the so called 'winning class' LRP, was used to emphasize the critical brain regions for appropriately identifying the two patient populations in a 3D-CNN that was proposed to achieve this goal. A Spearman's association between the mean relevance for each ROI and the individual Expanded Disability Status Scale (EDSS) scores strengthened the results in the end.

5.3.1 Materials and Methods

Population, Data Acquisition and Image Processing

The population consisted of 91 subjects, including 46 RRMS (35 females, 52.5 ± 10.4 years old) and 45 PPMS (25 females, 47.2 ± 9.5 years old) patients. The EDSS score was 2.8 ± 1.4 and 4.8 ± 1.3 for the two groups, respectively. Group differences in age and EDSS score were tested through t-test, while differences in gender numerosity were evaluated through χ^2 test.

MRI acquisitions were performed on a 3T Philips Achieva scanner (Philips Medical Systems, the Netherlands) equipped with an 8-channel head coil. The following sequences were used for all patients: 1) 3D T1-w Fast Field Echo (Repetition Time (TR)/Echo Time (TE) = 8.1/3 ms, FA = 8°, Field of View (FOV) = 240 × 240 mm^2, 1-mm isotropic resolution, 180 slices); 2) 3D Fluid-Attenuated Inversion Recovery (FLAIR) image (TR/TE = 8000/290 ms, Inversion Time (TI) = 2356 ms, flip angle = 90°, FOV = 256

Chapter5/Figures_EDL_AI/Fig (1).png

Fig. 5.1: 3D CNN architecture with single channel T1-w input.

$\times$ 256 mm^2, 0.9 $\times$ 0.9 $\times$ 0.5 mm^3 resolution, 180 slices). All patients were recruited in our center according to their diagnosis based on the McDonald 2010 diagnostic criteria. The study was approved by the local ethics committee, and informed consent was obtained from all patients. All procedures were performed in accordance with the Declaration of Helsinki (2008).

For each subject, the FLAIR was linearly registered to the T1-w (FSL flirt tool) and the Lesion Prediction Algorithm (LPA) [116] was used to automatically segment and fill the WM lesions in the native T1-w image. Each filled T1-w image was then imported in the FreeSurfer software [117] to perform a complete brain parcellation with 112 anatomical ROIs. The binary mask representing the GM tissue probability thresholded at 95% was derived for each subject (FSL fast tool) and applied to all the filled T1-w.

Network Architecture

A 3D-CNN Visual Geometry Group (VGG) net architecture [21] was used. This architecture has been well assessed in combination with MRI data in few recent studies [118, 26, 119], and it has been shown to achieve comparable performance with respect to the Residual Neural Networks (ResNet) model [120] in distinguishing AD patients from controls [118]. In addition, the VGG model allows for a straightforward application of visualization techniques and is thus particularly suitable for this work aiming at interpretability.

The network structure consists of four volumetric convolutional blocks for feature extraction, two fully connected layers with batch normalization, and one output layer with `softmax` nonlinearity. Each convolutional block consists of a convolutional layer followed by a `ReLU`, batch normalization, and 3D pooling. A graphical representation of the 3D-CNN structure highlighting the main parameters for each layer is provided in Figure 5.1.

Training, Validation and Testing

Data augmentation was performed during the training/validation phase in order to improve the generalization capabilities of our models. In detail, the data augmentation consisted of: addition of random Gaussian noise ($\mu = 0$, $\sigma = 0.1$); random affine transformation from -5 to $+5$ degrees in the Z axis, and from -3 to $+3$ degrees in the X axis; random volume translation from -3 to $+3$ voxels along each of the three axis; flipping across the X axis. In addition, clipping of the values to the 99th percentile was performed.

The CNN was trained using a 5-fold Cross Validation (CV) strategy over a training/validation set of 71 subjects. On each fold, the 71 subjects were randomly split in five groups, each of 14 subjects (except one of 15 subjects). The experiment was repeated five times and, for each repetition, four groups were considered as training and the remaining one was kept unseen for validation. The cross-entropy loss was optimized by means of the Adam optimizer [121] during the training phase. Twenty subjects were kept unseen and considered as a testing set. In detail, the five models derived from each fold of the 5-fold CV, were used to perform the prediction over the test set, and the performance metrics were computed for each of them. The test subjects did not undergo the data augmentation transformations.

In this work, True Positive (TP)s and True Negative (TN)s represent the number of correctly classified PPMS and RRMS subjects, respectively. The CNN performance was reported, averaged over the five models, in terms of accuracy, sensitivity and specificity, while precision for each class was defined as $\text{precision}_{PPMS} = \text{TP}/(\text{TP}+\text{FP})$ and $\text{precision}_{RRMS} = \text{TN}/(\text{TN}+\text{FN})$. The whole deep learning analysis was carried out using Pytorch [122]. The computation was performed on a laptop (Ubuntu 18.6, Nvidia Geforce GTX 1050, Intel Core i7, 16 GB RAM). Torchsample wrapper was used as high-level interface.

CNN Visualization

LRP visualization was employed to identify which voxels in the input volume contributed most to the classification output. This technique is based on a backward procedure which is a conservative relevance redistribution of the output prediction probability through the CNN layers till the input volume.

The core rule of the LRP backward procedure is the relevance conservation per layer. Let s and $s+1$ be two successive layers of the network and j and k two “neurons" of those layers, respectively. The relevance of the neuron k for the prediction $f(x)$, where x is the input, can be written as R_k^{s+1}. This relevance is redistributed to the connected “neurons" in layer s through the following equation:

$$\sum_j R^s_{j\leftarrow k} = R^{s+1}_k \tag{5.1}$$

The iteration of Eq. 5.1 through all the CNN layers allows the decomposition of the relevance of prediction function $f(x)$, R_f, in terms of the input in the first layer.

Multiple rules can be applied for the distribution of the relevance [46]. In this work we used the β-rule [32, 123], setting $\beta = 0$, hence allowing only positive contribution to the relevance score, following [26] where they demonstrated the LRP robustness relatively to the β-value. Higher β-values decompose the relevance in positive and negative contribution, the latter usually considered when dealing with patient control classification task, not the object of this study. In this work, the classification aim of differentiating two groups of patients led to the computation not only of the TPs (PPMS) related LRPs, but also the TNs (RRMS) related ones by computing the so-called *winning class LRP*. In fact, to obtain LRP maps, a target class has to be defined and the resulting maps are strongly related to such class. In this work, since the two classes share the same importance, the backward procedure starts, for each subject, from the highest relative prediction probability present in the prediction function $f(x)$ (hence, there is not a fixed target class). In this way, widening the definition of LRPs given in [26] for AD, the resulting winning class LRPs will answer two questions: (i) '*What speaks for PPMS in this subject?*', for the subjects predicted as PPMS, (ii) '*What speaks for RRMS in this subject?*' for the subjects predicted as RRMS. To compute LRP maps the iNNvestigate library [124] was used.

LRP heatmaps analysis

The LRP heatmaps were generated for each subject of the test set, based on the best model (in terms of accuracy) among the five derived from the 5-fold CV. After, they were registered to the standard MNI space (voxel size = 1 mm) and averaged over the two groups of patients, for visualization purposes.

Fifteen brain ROIs were selected based on MSliterature [125, 126, 127, 128]: thalamus (Thal), caudate (Cau), putamen (Put), hippocampus (Hipp), insular cortex (Ins), temporal gyrus (TpG), superior frontal gyrus (SFG), cingulate gyrus (CnG), lateral occipital cortex (LOC), pericalcarine (PCN), lingual gyrus (LgG), cerebellum (Cer), temporal pole (TP), pallidum (Pall) and parahippocampal gyrus (PHG). The reference atlas was the Desikan-Killany available in FreeSurfer tool.

The average of the LRP derived relevance values for the 15 ROIs was computed across the correctly classified subjects of the test set.

Chapter5/Figures_EDL_AI/F3.png

Fig. 5.2: LRP heatmaps obtained from the T1-CNN model. The heatmaps are shown for both RRMS and PPMS patients, and are overlaid to the MNI152 template in coronal, sagittal and axial views (columns). Each LRP map is averaged across the correctly classified RRMS and correctly classified PPMS subjects of the test set, respectively. The reported values are clipped to the range 60th–99.5^{th} percentile, calculated over the RRMS and the PPMS class group mean heatmaps.

Finally, as explorative analysis, we investigated the LRP neurological plausibility, following [33] and [26]. The Spearman's correlation between the average LRP relevance for each ROI and the EDSS score was calculated together with the corresponding p-value.

5.3.2 Results

A preliminary analysis revealed that the EDSS score and the age were significantly different between RRMS and PPMS subjects ($p < 0.05$). The same held with gender numerosity ($p < 0.05$), this last observation reflecting the epidemiology of the disease.

The proposed T1-CNN achieved an average accuracy equal to 0.84 ± 0.10 over the five models derived from the 5-fold CV, one for each fold. The sensitivity and specificity were 0.74 ± 0.24 and 0.94 ± 0.08, showing that the CNN minimized the FPs, which are the wrongly classified RRMS subjects. The trend was confirmed by the precision$_{RRMS}$ which was 0.82 ± 0.14 while the precision$_{PPMS}$ was 0.94 ± 0.14.

Figure 5.2 shows the group LRP heatmaps, averaged over the correctly classified subjects of the test set for each class. For ease of visualization, the maps are clipped between the 60th and the 99.5^{th} percentile calculated over the respective LRP target group

Chapter5/Figures_EDL_AI/F4.png

Fig. 5.3: Size-normalized importance metric extracted from the LRP maps. The mean relevance value for each ROI is reported for all the correctly classified PPMS and RRMS subjects in the test set. The median relevance for PPMS (orange circle) and RRMS (blue circle) groups are also shown.

heatmap. As expected, considering how winning class LRP maps were calculated, high relevance was found in both PPMS and RRMS classes. Even if widespread relevance values were present in both classes, the pattern was slightly different. In fact, the RRMS derived LRP map showed high activation in the temporal cortex and cerebellum, particularly evident in the coronal and sagittal views, respectively. On the contrary, the PPMS derived LRP map showed low relevance in the temporal lobe, while high relevance was assigned to the frontal lobe as can be observed in the sagittal view.

ROI-based analysis was performed to quantitatively assess the relevant areas for the classification task, as a first step toward the clinical validation of the outcomes. Figure 5.3 illustrates the size-normalized importance metrics for the correctly classified patients of the test set, separately for the two classes. As previously stated for the qualitative analysis of the LRP maps, the temporal pole showed the highest relevance for both classes, as well as the highest distance between the medians, followed by the hippocampus, which moreover presents the highest gap between the two distributions. Other relevant ROIs were the insula and the cerebellum, the former showing a higher distance between the medians of the distributions. The other considered ROIs showed

Table 5.1: Spearman correlation results between the mean relevance for each ROI and the EDSS score.

	Ins	PCN	SFG	CnG	PHG
ρ	-0.42	0.16	**0.47**	-0.16	**-0.59**
p-value	0.08	0.51	**0.05**	0.53	**0.01**
	TP	LOC	LgG	TpG	Thal
ρ	-0.44	0.42	-0.40	**-0.52**	0.08
p-value	0.07	0.08	0.10	**0.03**	0.74
	Cau	Put	Hipp	Pall	Cer
ρ	-0.25	-0.43	**-0.50**	-0.13	-0.22
p-value	0.38	0.08	**0.03**	0.61	0.38

The ρ score and relative p-values (rows) are reported for each ROI (columns). The significant correlations (p-value < 0.05) are highlighted in bold.

lower relevance values and overlapped distributions, particularly evident for the superior frontal gyrus.

Finally, the results of the Spearman correlation analysis between the ROIs mean relevance values and the EDSS scores are reported in Table 5.1. Significant negative correlations were detected for the parahippocampal gyrus, the temporal gyrus, and the hippocampus (p-value < 0.05), showing a ρ value of -0.59, -0.52, and -0.50, respectively. On the contrary, a slightly positive correlation with $\rho = 0.47$ could be found for the superior frontal gyrus.

5.3.3 Discussion

In this work, we addressed the stratification problem between RRMS and PPMS subjects based on T1-w data. A 3D-CNN was proposed to this aim, and the LRP visualization technique was applied in order to highlight which are the key brain regions for correctly classifying the two patient populations. Finally, the outcomes were strengthened through a Spearman correlation between the mean relevance for each ROI and the individual EDSS scores. Distinguishing PPMS from RRMS based on GM features is one of the current challenges in MSresearch [129], and the identification of a biomarker allowing to capture the differences between PPMS and RRMS patients is hence one of the major challenges of personalized medicine [130].

The obtained accuracy of 0.84 ± 0.10 suggests that the combination of T1-w and CNNs can help in the classification task between MSsubtypes. Performance is com-

parable with that presented in [113] (average precision of 0.84 and an average recall of 0.8 on a dataset of 604 acquisitions) although achieved with different methods and data acquisitions. The other classification metrics, as formalized in this work, can be strictly related to the ease of classification of each class of patients. In our results, the precision$_{PPMS}$ and the specificity were respectively close to the precision$_{RRMS}$ and the sensitivity, indicating that the CNN better minimized the FPs (the wrongly classified RRMS subjects). This highlighted a probable better characterization of RRMS subjects with respect to PPMS.

The differentiation between healthy and pathological subjects is much more common in literature. In this respect, a 3D CNN based approach was proposed by [33], showing an accuracy of 87.04% on a set of 147 fully volumetric structural MRI acquisitions. Moreover, to better interpret the CNN performance, Eitel and colleagues adopted LRP visualization. The substantial difference in the research question makes these works not directly comparable to ours.

Through LRP visualization it was possible to identify the regions based on which the CNN model performed the classification between the two MS subtypes. The ROIs deemed as more relevant, which were also significantly correlated with the EDSS score, are generally involved in MS pathology. The parahippocampal gyrus and the hippocampus have been shown to have a high probability of focal GM demyelination in MS pathology [131, 132], and the temporal gyrus has been demonstrated to be correlated with cognitive performance in MS [133], while the superior frontal gyrus has been shown to be associated with fatigue, particularly in RRMS [79]. Interestingly, the superior frontal gyrus resulted significantly correlated with EDSS despite it showing overlapping relevance between the two classes. This reasonably calls for clinical validation of the outcomes. In fact, the relevance values allowed to understand how the voxels of certain ROIs contributed to the classification, but still did not allow to identify the underlying reasons (e.g. lesion load, atrophy, etc.) [26].

Despite the promising results obtained in this study, we acknowledge that our study can be improved especially for what concerns the robustness of the outcomes, which depends on the numerosity of the sample. This limitation also affected the hyperparameters optimization, which was performed on separate sets. A comparison with different classification techniques will be the object of future works.

Nevertheless, we consider these outcomes as the valuable first evidence of the potential of the proposed method in splitting apart the two MR phenotypes and providing hints on the possible subserving mechanisms of disease progression, and we leave the open issues mentioned above for future investigation.

5.3.4 Conclusions

This work corroborated the capability of T1-w combined with a 3D CNN classifier of distinguishing the different typologies of MS disease. In addition, we could highlight, through the application of LRP visualization, that the CNN classification was based on clinically relevant ROIs that significantly correlated with EDSS score. From a clinical perspective, our results strengthen the hypothesis of the suitability of GM features as biomarkers for MS pathological brain tissues. Moreover, this work has the potential to address clinically important problems in MS, like the early identification of the clinical course for diagnosis, personalized treatment and treatment decisions.

5.4 The contribution of dMRI and class specific LRP feature visualization

In what follows we will move a step forward proposing various CNN models for the purpose of detecting PPMS patients and capturing the microstructural features and key ROIs that influence the classifier decision. As dMRI derived indices, DTI (FA and MD), 3D-SHORE (Propagator Anisotropy (PA), Return to the Axis Probability (RTAP), Return to the Origin Probability (RTOP), and RTPP) and the recently proposed RIF (RIF1, RIF2) were considered. Only GM tissues were given as input to the CNN, and a T1-w-based CNN model was additionally trained for benchmarking. Confounds influence was deeply investigated proposing a *post-hoc* analysis after CNN training. LRP was then applied to retrieve the regions leading the classification. Two heatmaps were created for each CNN model, one for each target class in the test set, showing the importance of each voxel. Following that, the relevance of the 15 chosen brain regions was assessed and validated region-wise using three different importance metrics. Finally the LRP neurological plausibility was also verified through a Spearman's correlation between the relevant maps and a diffusivity index known for its role in MS stratification.

5.4.1 Materials and Methods

An overview of the complete process proposed in this work is presented in Figure 5.4.

Dataset

The population consisted of 91 subjects, including 46 RRMS (35 females, 42.5 ± 10.4 years old) and 45 PPMS (25 females, 47.2 ± 9.5 years old) patients. EDSS score was 2.8 ± 1.4 and 4.8 ± 1.3 respectively for the two groups. A significant group difference in

Chapter5/Figures/F1.png

Fig. 5.4: Schematic overview of the proposed pipeline. DTI, 3D-SHORE, RIF and T1-w MRI are considered separately as input to different 3D CNNs models, resulting in one CNN model for each index. For each CNN, the best model, derived from a 5-fold CV is retained and LRP maps are extracted for both target classes (RRMS-LRP and PPMS-LRP).

age, EDSS score ($p < 0.05$, obtained via a t-test) and gender numerosity ($p < 0.05$, obtained via a chi-squared test) were recorded, this last reflecting the epidemiology of the disease.

MRI acquisitions were performed on a 3T Philips Achieva scanner (Philips Medical Systems, Best, The Netherlands) equipped with an 8-channel head coil. The following sequences were used for all patients: 1) two-shell dMRI (TR/TE = 9300/109 ms, flip angle = 90°,FOV = 112 x 112 mm^2, 2-mm isotropic resolution, 62 slices, b-values = 700/2000 s/mm^2 with 32/64 gradient directions respectively and 7 $b0$ volumes); 2) 3D T1-w Fast Field Echo (TR/TE = 8.1/3 ms, FA = 8°, FOV = 240 x 240 mm^2, 1-mm isotropic resolution, 180 slices); 3) 3D FLAIR image (TR/TE = 8000/290 ms, TI = 2356 ms, flip angle = 90°, FOV = 256 x 256 mm^2, 0.9 x 0.9 x 0.5 mm^3 resolution, 180 slices). The study was approved by the local ethics committee, and informed consent was obtained from all patients. All procedures were performed in accordance with the Declaration of Helsinki (2008).

Data preprocessing

Diffusion MRI data denoising, Gibbs ringing removal, motion and Eddy currents distortion correction were performed using the DIFFPREP module of Tortoise software (https://tortoise.nibib.nih.gov/tortoise).
These steps led to preprocessed dMRI data with a size of $90 \times 125 \times 125$ voxels. The Brain Extraction Tool in FSL (https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/[134]) was used for skull stripping and for deriving a binary mask for each subject. In addition, the

individual $b0$-weighted image were averaged across all the $b0$ volumes and the resulting image was spatially normalised to the MNI152 standard space using first the FSL `epi_reg` tool, to register the $b0$ image to the respective T1-w one, and then ANTs (`http://stnava.github.io/ANTs/`), to normalise the T1-w to the standard MNI152 space.

For each subject, the FLAIR image was rigidly registered to the T1-w using the FSL `flirt` tool. The LPA [116] available in the Least Segmentation Toolbox (LST) for SPM12 (`www.statistical-modelling.de/lst.html`) was used to automatically segment and fill the WM lesions in the native T1-w image. Each filled T1-w image was then imported in the FreeSurfer software (`http://surfer.nmr.mgh.harvard.edu/`, Harvard University, Boston, MA, USA) to perform a complete brain parcellation with 112 anatomical ROIs. Filled images were also skull stripped (FSL `bet` tool) and projected back to the dMRI native space by inverting the previously estimated transformation matrix. The segmentation of the registered T1-w images into GM, WM and cerebrospinal fluid was finally performed (FSL `fast` tool) and a binary mask representing the GM tissue probability thresholded at 95% was derived for each subject. This was applied to all the indices maps and to the T1-w, as only GM tissue was considered throughout the subsequent analyses.

Diffusion signal modeling

Microstructural indices were derived from three analytical signal models: DTI [84] (FA and MD); 3D-SHORE [102] (PA, RTAP, RTOP and RTPP); and the recently proposed RIF [135]. DTI and 3D-SHORE indices were calculated using DIPY [136] while in-house software was used for RIF.

As opposed to DTI and 3D-SHORE indices, which are well known in the current literature [137, 138, 139], RIF have been only recently introduced. Basically, RIF are calculated on the Laplace-series expansion of a given spherical function and are high order rotation invariants related to the spherical mean, power-spectrum and bispectrum invariants if calculated on the diffusion signal. Moreover, they can be linked to statistical and geometrical measures of spherical functions, including the mean, the variance and the volume of the function.

In this work, we used 4th order Spherical Harmonics (SH) to fit the diffusion signal, but only the first two RIF, I_0 and I_{22}, were considered. Indeed, due to the characteristics of the diffusion process in GM, the corresponding spherical signal is almost flat, such that all the high order invariants vanish. The only two non negligible RIF are I_0 and I_{22} and are given by (5.2) and (5.3), respectively. In this work we will refer to them as RIF1 and RIF2 for convenience of notations:

$$RIF1(f) = c_{00}\sqrt{4\pi} \tag{5.2}$$

$$RIF2(f) = \sum_{m=-l}^{l} |c_{lm}|^2 \tag{5.3}$$

where, in this work, $l = 2$. If the RIF are calculated on the diffusion signal, as it was the case here, RIF1 corresponds to the mean of the diffusion signal across one shell, while RIF2 is related to the variance [135]. More details about RIF computation can be found in Appendix A. Since RIF were calculated separately on each shell, two maps were obtained for each RIF in our two-shells dMRI scheme. Overall, eleven features resulting from the DTI, 3D-SHORE, RIF and T1-w models were handcrafted and used for each patient.

Network architecture

Chapter5/Figures/F2.pdf

Fig. 5.5: 3D Convolutional Neural Network architecture with single channel diffusion Magnetic Resonance Imaging index input.

A 3D-CNN VGG net architecture [21] as presented in Section 5.3.1 and shown in Figure 5.5 was used also for this study. A different CNN was used separately for each input.

Of note, throughout the manuscript, we will refer to the eleven derived CNNs models as FA-CNN, MD-CNN, PA-CNN, RTAP-CNN, RTOP-CNN, RTPP-CNN, RIF1-CNN, RIF2-CNN and T1-CNN, respectively, based on the input feature.

Training, Validation and Testing

The feature maps resulting from the DTI, 3D-SHORE, RIF models and T1-w, separately masked to retain GM voxels only, were split in subsets to be used for training/validation (78% of the total, 71 subjects) and testing (22% of the total, 20 subjects). For the RIF, the two different maps for each index (one per shell) were considered together as separate channels of input data. Therefore, the whole input of the network was a four dimensional tensor of size 1×90×125×125 for DTI and 3D-SHORE indices, 2×90×125×125 for the RIF and 1×180×240×240 for T1, respectively. The optimal weights were learned during training by minimizing the cross-entropy loss by means of the Adam optimizer [121]. Training and validation were performed in all cases on batches of size four.

Data augmentation was performed keeping the same parameters described in Section 5.3.1. Multiple tests were performed to fit the hyperparameter values over the training/validation phases. Their values were varied across the respective feasible and empirical range and the ones leading to the best accuracy and lower loss were retained. Validation was performed following a 5-fold CV strategy. The 71 subjects used for training/validation were randomly split in five groups, resulting in folds of 14 subjects each (except one consisting of 15 subjects). The experiments were repeated five times and, for each run, four folds were used for training and the remaining one for validation. The best model, for each fold, was chosen as the one corresponding to the lowest loss and highest accuracy values obtained over the validation sets. The remaining 20 subjects were kept unseen and used for testing using the best model resulting from each fold of the 5-fold CV procedure. No data augmentation was performed on the test set.

The whole DL analysis was carried out using `Pytorch`. The computation was performed on a laptop (Ubuntu 18.6, Nvidia Geforce GTX 1050, Intel Core i7, 16 GB RAM). `Torchsample` wrapper was used as high-level interface.

Performance Assessment

Performance was assessed following the objective of avoiding the misclassification of PPMS patients. Accordingly, we called TP and TN the number of correctly classified PPMS and RRMS subjects, respectively.

Performance metrics

The following measures were calculated to assess the performance of each CNN model:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5.4}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{5.5}$$

$$Specificity = \frac{TN}{TN+FP} \tag{5.6}$$

$$Precision_{PPMS} = \frac{TP}{TP+FP} \tag{5.7}$$

$$Precision_{RRMS} = \frac{TN}{TN+FN} \tag{5.8}$$

Results were reported for the testing set in terms of mean and standard deviation of the classification measures over the five best models resulting from the 5-fold CV.

Controlling for confounding variables

In the biomedical field, great importance is attributed to the role of confounds, that could bias the results hiding, contrasting or annihilating other factors that could hold important clinical information. In general, when linear regression is used, deconfounding is applied before modeling, regressing out the confounds directly from the data. However, this is not a common practice when deep networks are used, relying on their ability of capturing all the discriminating features. Though, this does not provide any guarantee with respect to possible biases in the outcomes neither on the prevalence (or not) of confounding variables in shaping the results. In this work the issue was faced by following the *post-hoc* method described in [140] leaving the input data unchanged. In particular, logistic classification models were used to assess the role of age, sex and EDSS in the differentiation between PPMS and RRMS phenotypes. To this end, two logistic models were contrasted for each index through the Likelihood Ratio (LR) test, that is predicting the outcomes using either the confounding variables or the confounding variables and the CNN model predictions. The statistical significance of the LR, assessed through a χ-squared test, would reveal that the role of the confounds in shaping the classification outcomes is not prevalent.

Layer-Wise Relevance Propagation (LRP)

LRP visualization was employed to determine which features and voxels in the input volume contributed most to the classification output. This technique relies on a backward pass ruled by the conservative relevance redistribution procedure, proceeding backwards from the CNN output values (i.e., the classification probabilities) to the input layer. In this approach, each neuron of a layer receives a relevance score from the next layer and redistributes it to its predecessors in equal amounts until the input layer is reached. In this way, neurons that contribute the most to the deepest layer receive more relevance. More details about the analytic formulation can be found in Section 5.3.1.

In a multi-class classification task, $f(x)$ consists of multiple values indicating the probability for the input x to belong to each of the classes c_i, e.g. $f(x)$={$f_{C_1(x)}$,$f_{C_2(x)}$, ..., $f_{C_N(x)}$} where N is the total number of classes. In order to calculate the LRP map, the target class must be specified. Let n be the class index, then $C_n - LRP(x)$ is obtained by backpropagating $f_{C_n(x)}$ through the network. Following this notation, in this work the prediction $f(x)$ is defined as $f(x)$={$f_{C_{PPMS}(x)}$, $f_{C_{RRMS}(x)}$}. Differently from the winning class LRPs described in Section 5.3.1 and adopted in the previous work, here two targets class-driven LRP heatmaps were derived for each subject, providing complementary information about the significance of each voxel in the classification process.

It might be useful to point out here that the network behavior is not symmetric across the two classes. To get the flavor of this let's consider a toy example. Let $PPMS_i$ be a Primary Progressive patient and let us assume that the patient is correctly classified by the network. Then, let $f(PPMS_i)$ be the corresponding value of the decision function. To get the corresponding LRP map (that we call $PPMS_i$-LRPmap), such value is backprojected through the network. Though this same patient is a TN for the other class (e.g. the RRMS one) and it will contribute with the value $1 - f(PPMS_i)$ to the RRMS-LRPmap. Then, since the TPs of one class coincide with the TNs of the other and contribute with backprojected values that sum to one, in the computation of the respective LRP maps, that is PPMS-LPR and RRMS-LRP, these two will in general be different.

LRP heatmaps analysis

For each CNN model, PPMS-LRPand RRMS-LRPheatmaps were generated for each subject of the test set, based on the best model among the five derived from the 5-fold CV. This led to twenty-two LRP maps per subject, representing the performance of the six diffusion indices, RIF1 ($RIF1_{700}$ and $RIF1_{2000}$), RIF2 ($RIF2_{700}$ and $RIF2_{2000}$) and T1-w (thus eleven maps per target class).

Both qualitative and quantitative analyses were performed relying on a ROI-based approach. To this end, 15 brain ROIs, which were previously demonstrated to be highly relevant in the MS disease [126, 127], were selected. The chosen ROIs were: *Thalamus*, *Caudate*, *Putamen*, *Hippocampus*, *Insula*, *Precuneous*, *Superior Frontal Gyrus* and *Cingulate Gyrus*, *Lateral Occipital Cortex*, *Pericalcarine* and *Lingual Gyrus*, *Cerebellum*, *Temporal Pole*, *Pallidum* and *Parahippocampal Gyrus*.

Inspired to [26], size-normalized *importance* metrics were derived for both the PPMS- and RRMS-LRPheatmaps. In particular, the median of the relevance values of both the target and non-target classes were extracted and averaged for each ROI across the correctly classified subjects (TPs and TNs) of the test set. Then, two additional measures, which we call gain and differential gain were calculated. The first is inspired to the gain

metric used in [26], which was given by the ratio between the LRP median values of the two categories. In this work, we propose to use the difference between such values to avoid the divergence of the measure that could occur in sites with vanishing LRP. Following the definition of the LRP, we consider the difference in relevance to be more representative of the actual contribution of a given ROI in forming the classification decision. All these steps resulted in two values per ROI, named as PPMS-LRPgain and RRMS-LRPgain, respectively. The differential gain was calculated as the difference between RRMS-LRP gain and PPMS-LRP gain. This was computed to measure the difference in gain brought by a given ROI when switching across the target class.

As a final explorative analysis, we aimed at investigating the LRP neuroanatomical plausibility, following [33] and [26]. In particular, in [26] the hippocampal volume was used as a biomarker for AD, while in [33] the WM lesion load was considered for MS and the correlation between the LRP relevance sum and lesion sum was assumed to provide evidence in favor of the informative potential of the LRP. In a joint work [108], we have shown that the RTPP mean value in the *Hippocampus* was significantly different for PPMS and RRMS in a subset of the cohort of patients considered here. In this work, four LRP maps were available for each index (including RTPP), one pair for each target class. Relying on that, our working hypothesis was that the statistical significance of the Spearman correlation coefficient between the mean LRP value of either of the four maps in the *Hippocampus* and the RTPP mean value in the same region would provide the first evidence in favor of the neuroanatomical plausibility of LRP as a potential staging signature. Accordingly, four correlation coefficients were assessed.

5.4.2 Results

Qualitative microstructural assessment

In Figure 5.6, RIF, DTI and 3D-SHORE index maps are shown in a coronal slice for two representative subjects, one per phenotype. For ease of visualization, RIF values out of the range between the 5th and the 95th percentile were clipped, while the square root of the RTAP and the cubic root of the RTOP were computed to report the values in the same units as RTPP.

The contrast of DTI and 3D-SHORE indices is in agreement with previously reported results [141, 138]. In detail, the maps of anisotropy indices (FA and PA) revealed hyperintensity in regions where the diffusion orientation profile has one main direction (e.g. *Corpus Callosum*) and low contrast in regions with diffusion spread in many directions (e.g. *Ventricles*). RTAP, RTOP and RTPP had similar behavior, reporting high values in tissues featuring restricted diffusion, such as WM, that decrease in GM and

Chapter5/Figures/S1.pdf

Fig. 5.6: dMRI based indices for one representative RRMS patient and one PPMS patient (columns). Coronal slices are reported for each index (rows). Images are displayed in radiological convention.

reach the minimum in the cerebrospinal fluid. As expected, MD presented a complementary pattern being higher in tissues where diffusion is unconstrained. Finally, RIF maps appeared inline with the results in [135], showing reduced intensity at increasing b-value, RIF degree and order. Both RIF1 maps were brighter in WM compared to GM, while RIF2 patterns were similar to FA/PA. In the two classes of patients, the maps had very similar contrast.

3D-CNN classification performance

The classification performance for each CNN model is reported in Table 5.2 for the test set in terms of accuracy, precision (for each class), sensitivity and specificity. Average

values and standard deviations were calculated from the five best models resulting from the 5-fold CV.

CNNs models for PA, RTAP, RTOP, RTPP, RIF1 and T1-w reached an accuracy ≥ 0.75. In particular, T1-CNN was the most accurate with an average score of 0.84, while for the dMRI-based models RTPP-CNN reached an accuracy of 0.81, followed by PA-CNN and RTAP-CNN achieving a mean accuracy score of 0.80 and 0.76, respectively. RIF2-CNN showed the worst performance with an average score of 0.58. Considering the two different classes, the highest precision for RRMS was reached by PA-CNN with a mean score of 0.96, followed by T1-CNN and RTPP-CNN showing a mean score ≥ 0.80. For the PPMS group, while T1-CNN provided the highest precision (average precision 0.94), good performance was also reached by RIF1-, RTAP- and RTPP-CNN, all providing a precision ≥ 0.80. The highest sensitivity was reached by PA-CNN with a score of 0.96, followed by RIF2-CNN and RTPP-CNN (having mean sensitivity of 0.82 and 0.80, respectively). Finally, T1-CNN achieved the highest specificity of 0.94. Classification performance measures for the validation set were inline with those obtained on the test set, providing evidence of the absence of overfitting despite the relatively low cardinality of the sample with respect to the number of the network parameters.

Concerning the influence of the three confounds on the CNNs classification outcomes, the LRtest highlighted that all the logistic classification models to which the CNNs outcomes were added as predictor were significantly different (χ^2 test, $p < 0.05$) from the logistic classification model employing only the confounds as predictors, except for the RIF1-, RIF2- and RTOP-CNN.

Test	Accuracy	Precision RRMS	Precision PPMS	Sensitivity	Specificity
RIF1-CNN	0.75 ± 0.04	0.71 ± 0.04	0.82 ± 0.05	0.64 ± 0.08	0.86 ± 0.05
RIF2-CNN	0.58 ± 0.15	0.73 ± 0.26	0.58 ± 0.16	0.82 ± 0.16	0.34 ± 0.28
FA-CNN	0.70 ± 0.08	0.76 ± 0.15	0.77 ± 0.13	0.68 ± 0.31	0.72 ± 0.20
MD-CNN	0.70 ± 0.06	0.67 ± 0.07	0.74 ± 0.04	0.60 ± 0.13	0.80 ± 0.05
PA-CNN	0.80 ± 0.04	$\mathbf{0.96 \pm 0.09}$	0.73 ± 0.04	$\mathbf{0.96 \pm 0.08}$	0.64 ± 0.08
RTAP-CNN	0.76 ± 0.09	0.75 ± 0.11	0.80 ± 0.04	0.68 ± 0.20	0.84 ± 0.05
RTOP-CNN	0.75 ± 0.03	0.77 ± 0.08	0.77 ± 0.05	0.74 ± 0.15	0.76 ± 0.10
RTPP-CNN	0.81 ± 0.05	0.81 ± 0.10	0.82 ± 0.03	0.80 ± 0.11	0.82 ± 0.04
T1-CNN	$\mathbf{0.84 \pm 0.10}$	0.82 ± 0.14	$\mathbf{0.94 \pm 0.07}$	0.74 ± 0.24	$\mathbf{0.94 \pm 0.08}$

Table 5.2: Classification performance metrics calculated in the test set for each CNN model. Values were calculated by averaging the results of the five best models derived from 5-fold CV and reported together with the respective standard deviation values.

LRP heatmaps

Chapter5/Figures/F4.pdf

Fig. 5.7: LRP heatmaps obtained from CNNs models based on the first RIF1 and PA. The heatmaps are shown for both target classes RRMS-LRP and PPMS-LRP, columns), and are overlaid to the MNI152 template in coronal, axial and sagittal views (rows). Each LRP map is averaged across the correctly classified RRMS and correctly classified PPMS subjects of the test set, respectively. The reported values are clipped to the range 60th–99.5^{th} percentile, calculated over the RRMS and the PPMS class group mean heatmaps.

Chapter5/Figures/F5.pdf

Fig. 5.8: LRP heatmaps obtained from CNNs models based on RTAP, RTPP and T1-w. The heatmaps are shown for both target classes (RRMS-LRP and PPMS-LRP, columns), overlaid to the MNI152 template in coronal, axial and sagittal views (rows). Each LRP map is averaged across the correctly classified RRMS and correctly classified PPMS subjects of the test set, respectively. The reported values are clipped to the range 60th–99.5^{th} percentile, calculated over the RRMS and the PPMS class group mean heatmap.

Figure 5.7 and Figure 5.8 show the group LRP heatmaps (RRMS-LRP and PPMS-LRP), averaged over the correctly classified subjects of the test set for both RRMS and PPMS

classes. Only the maps derived from the best indices in terms of mean accuracy are shown (see Subsection 5.4.1 for details). For ease of visualization, the maps are clipped to the range 60th–99.5^{th} percentile calculated over the respective LRP target group mean heatmap.

As expected, higher contrast, reflecting higher relevance, characterizes the average LRP map of the target class (i.e., PPMS in PPMS-LRP and RRMS in RRMS-LRP) in all CNNs models. This follows from the LRP relevance propagation algorithm, which starts from a larger number in the output layer for the target class and the correctly classified subjects. Even if a widespread activation of the GM regions was present in all the heatmaps, the patterns of the two families of LRP maps were not overlapped and PPMS-LRP resulted in more scattered activations compared to the others. The LRP maps corresponding to the different indices revealed different activations, highlighting that these could mirror specific microstructural properties. In particular, for both the target classes and both LRP heatmaps, RIF1-CNN ($RIF1_{700}$ and $RIF1_{2000}$) showed a widespread activation over the GM regions, involving both cortical and subcortical structures. A similar pattern appeared also in RTAP-CNN, RTPP-CNN and T1-CNN LRP maps, showing higher frontal activation in PPMS-LRP maps of T1-CNN. A different pattern can be observed in both PA-CNN derived LRP heatmaps, revealing higher activation values in deep GM structures and considerably lower values in cortical structures.

Moving to the non-target class, high relevance values were present in the PPMS average heatmap for the RRMS-LRP, which were not found in the counterpart group for the PPMS-LRP maps. This is particularly evident for RIF1 and T1-CNN derived heatmaps. Of note, these maps showed lower relevance scores compared to the others. However, this depends on the higher number of voxels on which the relevance had to be redistributed for these two inputs (two maps for the RIF1-CNN and a larger map for the T1-w).

LRP ROI-based analysis

ROI-based analyses were performed to assess the relevant areas for the classification task. Figure5.9 illustrates the size-normalized importance metrics for the correctly classified patients of the test set, separately for the two classes and for the two LRP types. The ROIs mean relevance values for the wrongly classified subjects revealed an always positive and the non target class (TNs for the PPMS-LRP and TPs for RRMS-LRP following the notations) featured relevance values following the same trend of the correctly predicted ones across ROIs. It is important to highlight that the non-target class still owed some relevance for all the CNNs models in both cases, which was particularly high for PPMS-LRP.

Chapter5/Figures/F6.pdf

Fig. 5.9: Size-normalized importance metrics extracted from the LRP maps derived for the two classes, PPMS-LRP (top) and RRMS-LRP (bottom), from the CNNs models based on the first RIF1, PA, RTAP, and RTPP. For each LRP type, the mean relevance value for each ROI is reported for all the correctly classified PPMS and RRMS subjects in the test set. The median relevance for PPMS (orange circle) and RRMS (blue circle) groups are also shown.

Chapter5/Figures/F7.pdf

Fig. 5.10: Relevance gain measures. The gain score for each LRP type is shown for different regions. The gain per area for each class derived LRP type, respectively RRMS-LRP (blue) gain and PPMS-LRP (orange) gain, is defined as the difference between the median relevance of the target and the non-target classes in a given area, calculated over all the correctly classified subjects of the test set.

Considering the different models, the same behavior was reported in PPMS-LRP for RTAP-CNN and RTPP-CNN heatmaps. In addition, RIF1-CNN maps showed similar values between the two shells and, together with T1-CNN LRP maps, presented sensibly lower LRP values compared to the other indices. All the CNNs models highlighted almost the same regions leading the classification between RRMS and PPMS, for both RRMS-LRP and PPMS-LRP. Among the cortical ROIs, *Parahippocampal Gyrus* appeared among the first five most relevant ROIs in all cases except T1-CNN, showing also the greater distance between the PPMS- and RRMS-LRP values in RRMS-LRP for all CNNs models except RIF1-CNN. *Temporal Pole* appeared highly relevant especially for PA-, RTAP- and RTPP- and T1-CNN, for both LRP types. Moreover, it reached high relevance values also for the non-target group in PPMS-LRP. *Superior Frontal Gyrus* showed a large LRP value for all the CNNs models, particularly for PPMS-LRP but also for the non-target class in RRMS-LRP for RTAP- and RTPP-CNN. Finally, *Lateral Occipital Cortex* was highly relevant for RTPP in the PPMS group of PPMS-LRP heatmaps.

Concerning deep GM ROIs, *Insula* was among the most relevant ROIs for all CNNs except RTPP-CNN, for both RRMS-LRP and PPMS-LRP, while *Cingulate Gyrus* was highly relevant for RRMS-LRP of RIF1-CNN. Of note, those deep GM ROIs showed non-overlapping sets of relevance values between groups in RRMS-LRP of all the CNNs mod-

Chapter5/Figures/S5.png

Fig. 5.11: Relevance gain measures. The differential gain for each LRP type is shown for different regions. The differential gain per area for each class derived LRP type, respectively RRMS-LRP (blue) gain and PPMS-LRP (orange) is defined as the difference between the RRMS-LRP gain and PPMS-LRP gain in each ROI.

els. *Cerebellum*, lastly, was a remarkably relevant ROI for all the CNNs models and both LRP types, showing also disjoint distributions of LRP values in RRMS-LRP.

In Figure 5.10, the results for the gain values are reported. This metric revealed that the highest difference in ROI relevance between the two LRP types was found in *Parahippocampal Gyrus*, *Temporal Pole*, *Superior Frontal Gyrus*, *Cerebellum*, *Cingulate Gyrus* and *Insula*, confirming the previously presented qualitative results. This quantitative analysis demonstrated also a different sensitivity of the considered indices to tissue modulations across separate ROIs, in particular of T1-w compared to the other indices. Among the dMRI indices, PA was the most different compared to the others. The results for the differential gain are reported in Figure 5.11.

The final Spearman correlation analysis revealed a significant and positive correlation ($\rho = 0.77, p = 0.016$) for the non-target class (PPMS) between the mean RTPP RRMS-LRP value in the *Hippocampus* and the RTPP mean value in the same ROI. No other significant correlations could be detected.

5.4.3 Discussion

In this work, we introduced LRP as a forceful method for explaining individual CNNs decisions in MS patients stratification. We trained different CNNs models to detect

PPMS patients and capture the microstructural features as well as the main ROIs leading to the classifier decision. The dMRI considered indices were derived from DTI (FA and MD), 3D-SHORE (PA, RTAP, RTOP and RTPP) and from a novel framework for the extraction of rotation invariant features from dMRI signal (RIF1 and RIF2). Only the GM tissues were fed to the CNN and a CNN model based on T1-w was also trained for benchmarking. For each CNN model, two heatmaps indicating the relevance of each voxel were derived, one for each target class in the test set. The relevance of 15 selected brain regions were then evaluated region-wise using three different importance metrics: (i) the size-normalized PPMS (or RRMS) importance, which is the median value of the LRP map for the target and non-target class, respectively; (ii) the PPMS-LRP and RRMS-LRP gain, measured as the difference between size-normalized target and non-target importance measures; and (iii) the differential gain, which combined both RRMS-LRP and PPMS-LRP by measuring the difference between the RRMS-LRP gain and the PPMS-LRP gain.

Our results demonstrated that: 1) dMRI features extracted in GM tissues can help disambiguate PPMS from RRMS patients; 2) LRP heatmaps highlight areas of high relevance which relate well with what is known from literature for MS disease.

Starting from the classification performance, 3D-SHORE derived indices, as well as RIF1, outperformed the DTI-based ones, reaching comparable results with T1-CNN. Moreover, while literature generally reports WM features as a signature of the MS disease, our study highlighted the GM potential role in identifying PPMS from RRMS patients, opening the way to a new type of potential numerical biomarkers focusing on GM.

Moreover, the LRtest between the two models associated with each index in the *post-hoc* assessment of the prevalence of the confounding variable revealed that the classification results were not dominated by the confounds for the DTI-, 3D-SHORE- (except for the RTOP) and T1-CNN models.

These results provide evidence of the potential improvement brought by dMRI features other than DTI for MS staging, as well as of the eloquence of microstructural information in GM. In particular, the optimal performance reached by PA was in agreement with the results reported in [142], suggesting the sensitivity of anisotropy measures to MS modulation of GM tissues, although using classical DTI indices.

Precision, sensitivity and specificity values as defined in this work were tailored on the ease of classification of PPMS patients. More specifically, the precision was calculated separately for PPMS and RRMS classes and measured, respectively, how well the CNNs models could characterize the PPMS (RRMS) cohorts by minimizing the number of RRMS (PPMS) wrongly classified subjects. PA provided the best results for precision for RRMS, that is in minimizing the number of wrongly classified PPMS subjects,

which was the main objective of this work. These results prove that dMRI is highly relevant for detecting the first signs of the PPMS stage of the disease. Conversely, the T1-CNN reached the highest precision for PPMS, demonstrating its ability to minimize the number of wrongly classified RRMS. This behavior was further confirmed by the sensitivity and specificity values. Regarding sensitivity, PA reached the best performance, highlighting the index ability to distinguish the PPMS patients by minimizing the number of misclassifications. The maximum specificity was provided by T1-w that allows to characterize the RRMS subject class.

Although these two models showed outstanding performance related to the correct classification of one of the two classes, the performance metrics scores were relatively low. The index showing the highest stability across all the proposed measures was RTPP, which achieved an average value ≥ 0.80 in all the classification performance metrics meaning that it minimized at the same time both the PPMS and RRMS wrong classifications.

Regarding the recently proposed RIF, even though they did not achieve significance in the control for confounds analysis, are interesting to be analyzed. RIF1 outperformed RIF2 in differentiating PPMS and RRMS. This was expected because in GM diffusion signal tends to be mostly isotropic and thus poorly described by high order SH models. High order RIF would be more suitable for WM, where the signal is highly anisotropic, especially in regions having complex topology (crossing, fanning). In fact, since RIF were calculated on the diffusion signal, RIF1 represents the mean of the signal across one shell and thus it is proportional to the inverse of the diffusivity, while the second is related to the variance of the signal across one shell and thus it is more sensitive to the complexity of the tissue [135].

Comparing our results to the current literature, as pointed out in Section 5.1 the stratification of MS patients is still largely underinvestigated and few studies addressed this specific problem so far. Among these, [113] achieved an average precision of 0.84 and an average sensitivity of 0.80 on a dataset of 604 acquisitions (b-value 1000 mm^2/s). Although they relied on dMRI data, their focus was on connectivity while the methods proposed in this work availed of microstructural information. Moreover, our RTPP-CNN and PA-CNN achieved comparable accuracy values on a smaller dataset.

The differentiation between healthy and pathological subjects is much more common in literature than intra-pathology stratification. In [115] and [114], two different 2D-CNN architectures were combined with conventional structural MRI data to this end reaching high accuracy scores (98.77% and 98.23%, respectively). Using the different slices of each subject as a separate input led to a much larger sample size (e.g., [115] amounting to 1357 images in total for the 64 subjects) which brings a clear advantage for training.

A 3D-CNN based approach was proposed in [33], reaching an accuracy of 87.04% on a set of 147 fully volumetric structural MRI acquisitions. Despite the lower accuracy compared to the 2D-CNN based approaches, the use of a 3D-CNN architecture facilitated the interpretation of the CNN performance through the use of feature visualization techniques. However, the difference in the research question makes these works not directly comparable to ours. Concerning the feature visualization, [111] compared different techniques applied to a 2D-CNN trained on 66 healthy controls and 66 MS patients SWI data. Their results highlighted the outstanding performance, based on the quantitative image perturbation method, of LRP maps and DeepLIFT [112] over simpler methods, strengthening the exploitability of such methods to address clinically relevant questions.

Regarding neural network visualization, the application of specific techniques, such as the LRP here adopted, provides a mean for CNNs interpretability and, when used in combination with other clinical and imaging data could support diagnosis and treatment decisions. To the best of our knowledge, this is the first work showing an application of LRP visualization on a dMRI-based classification problem. By relying on this technique, it was possible to identify the regions playing a prominent role in the classification between the two MS phenotypes, which were the regions of higher LRP gain across groups. From the analysis of these maps, it was clear that the different indices showed a selective pattern, being sensitive to modifications related to the disease in different ROIs. In particular, for dMRI-derived CNNs models and for the target class, both PPMS-LRP and RRMS-LRP group average heatmaps for RIF1-, RTPP- and RTAP-CNN showed a complementary relevance pattern compared to PA-CNN ones, suggesting that different regions encountered specific microstructural alterations. This result pushes towards their integration within a unified model accounting for all the relevant information at a time. Though, this would require a larger sample to compensate for the input size and we leave it for future investigation.

RIF1-CNN mean LRP heatmaps showed a redundancy in the information provided by the two different shells. For clinical purposes and applications, further investigation should be carried on measuring the discriminative power of RIF1 as calculated on a single shell. Determining whether one single b-value acquisition would be sufficient would allow reducing the MRI scan time with clear clinical advantages.

Concerning ROI-based analysis, the regions playing a leading role were in agreement with the literature. Indeed, the *Parahippocampal Gyrus* and *Insular Cortex* have been shown to report high probability of focal GM demyelination in MS pathology [126], while the *Cerebellum* has been demonstrated to be a major site for demyelination, especially in PPMS patients [143]. The *Superior Frontal Gyrus* has been associated with

fatigue, particularly in RRMS [79], and the *Temporal Pole* appeared to be present in clinically relevant MS cortical atrophy patterns [144].

Finally, it is important to note that the focus in interpreting LRP maps was not on the absolute values of the relevance, but on the differences and overlaps between the violin plots of the considered ROIs in the two classes of patients. This means that the relevance values allowed to understand how the voxels of certain ROIs contributed to the classification, but still did not allow to identify the underlying reasons (lesion load, atrophy, etc.) [26].

In order to investigate whether higher importance scores could correspond to relevant microstructural modulations, the association between RTPP average values in the ROI and the mean value of the RRMS-LRP heatmap for the non-target PPMS class *Hippocampus* was assessed. The *Hippocampus* was chosen in light of previous results [108] showing RTPP sensitivity to GM differences between PPMS and RRMS. Though the interpretation of this correlation is far from trivial, the straight meaning is that the PPMS subjects showed some tissue modulations typical of RRMS subjects to which RTPP-CNN was sensible and which were significantly linked to a biomarker for MS patients stratification. Although a deeper investigation is needed, in a broader view, this result, together with [26] and [33] provides evidence of CNNs ability to learn to identify important disease biomarkers as relevant for the classification. Of note, a significant feature for MS is a lower diffusion restriction and massive neuronal loss and demyelination in *Hippocampus* [145] which is indeed well captured by RTPP in other studies [141, 146]. In the future, it could be valuable indeed to perform an assessment of the performance of the dMRI indices in the tissues of the analyzed MS groups. This would enable a better understanding of the pathophysiology beyond the microstructural changes induced by the disease.

Limitations and Future Works

In what follows, some of the main limitations of this study are briefly summarized, and some among the many possible future steps that will be taken to fully exploit the potential of the proposed method are presented. First, the low numerosity of the sample could cast shadows on the statistical significance of the outcomes and also affect the hyperparameters optimization that, in an optimal setting, should be performed on separate sets. In this work data augmentation was used in both the training and the validation sets of the 5-fold CV for simulating a larger sample of subjects.

The lack of healthy controls is a reason for concern as it impedes benchmarking the performance of the proposed architecture in the patients versus controls classification task. However, as mentioned in the discussion, the CNN architecture we chose was previously employed in [26] to differentiate AD patients from controls based on T1-w MRI,

and a slightly modified version in [33] to classify healthy controls and MS patients based on FLAIR and T1-w MRI. We acknowledge that their datasets were different from our local cohort, however, their work can be considered as a literature benchmark for the CNN in use. Our aim was to disentangle the relevance linked to the two groups of disease which is an important open issue *per-se*. In consequence, we focused on this and investigated whether relying on sophisticated methods, such as advanced dMRI-based indices and 3D-CNNs models, could help in differentiating the two MS groups.

The lack of a control class also impeded to assess the relevance of the voxels and ROIs in distinguishing each of the phenotypes (PPMS, RRMS) from healthy matched controls. This is an interesting issue because it could reveal shared features of the two manifestations of the pathology that could not be captured by the proposed analysis yet potentially being insightful for understanding the mechanisms of the disease. We leave this for future investigation.

Finally, as a future work, including WM in the analysis, would widen the spectrum of the microstructural features potentially distinguishing the two disease phenotypes as well as unraveling the link with GM tissue modulations.

5.4.4 Conclusions

This work provides evidence in favor of the capability not only of T1-w but also of dMRI indices of distinguishing the PPMS from the RRMS state of disease in MS. We proved that 3D-SHORE based indices and RIF1 outperformed FA and MD, pushing to shift the attention on dMRI features other than DTI ones. In addition, thanks to the use of a 3D-CNN and LRP visualization, we could observe that the CNNs classification was based on clinically relevant ROIs and that different indices were sensitive to GM modulation in different brain regions. Our results support the hypothesis of dMRI based indices suitability as numerical biomarkers for the characterization of pathological brain tissues. Moreover, this work has the potential to address clinically important problems in MS, like the early identification of the clinical course for diagnosis and provides evidence in favor of the feasibility of precision medicine.

The work presented in this Chapter was published in [147] and [148].

Part III

XAI for Imaging Genetics in Alzheimer's Disease continuum

6

Benchmarking the link between Polygenic Risk Scores and structural MRI

In this work we exploited Partial Least Squares (PLS) model for analyzing the genetic underpinning of grey matter atrophy in Alzheimer's Disease (AD). To this end, 42 features derived from T1-weighted (T1-w) Magnetic Resonance Imaging (MRI), including cortical thicknesses and subcortical volumes were considered to describe the imaging phenotype, while the genotype information consisted of 14 recently proposed AD related Polygenic Risk Score (PRS), calculated by including Single Nucleotide Polymorphism passing different significance thresholds. The PLS model was applied on a large study cohort obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database including both healthy individuals and AD patients, and validated on an independent ADNI Mild Cognitive Impairment (MCI) cohort, including Early Mild Cognitive Impairment (EMCI) and Late Mild Cognitive Impairment (LMCI). The experimental results confirm the existence of joint dynamics between brain atrophy and genotype data in AD while providing important generalization results when tested on a clinically heterogeneous cohort. In particular, less AD specific PRS scores were negatively correlated with cortical thicknesses, while highly AD specific PRSs showed a peculiar correlation pattern among specific subcortical volumes and cortical thicknesses. While the first outcome is in line with the well known neurodegeneration process in AD, the second could be revealing of different AD subtypes.

6.1 Introduction

Dementia is rapidly increasing around the world, with AD being its most common cause, accounting for around 60–80% of the total cases [149]. A recent survey estimated that over 50 million people are living with dementia worldwide and such a figure is expected to triple by 2050, largely driven by the increases in life expectancy (https://www.healthdata.org/gbd/2019). Given the projected trends in population

ageing and population growth, AD is thus becoming one of the most burdensome diseases, increasingly calling for a next generation framework of early diagnosis and biomarker-guided targeted therapies. In recent years, advances with biomarkers have sustained a shift in how the disease is considered, with AD being now conceptualized as a biological and clinical continuum covering three well known phases (preclinical, MCI, and dementia stages of AD) rather than as being part of the simple succession of clinically defined entities [149, 150]. While the primary pathological hallmark of AD is the accumulation of abnormal proteins (mainly amyloid-β and hyperphosphorylated tau) in the brain, leading to a progressive synaptic, neuronal and axonal damage [151, 152], its etiology is complex and much remains to be fully elucidated. For these reasons, an increasing number of studies have focused on exploring its biological/genetic drivers and brain imaging correlates, and on shading lights on their possible interplay. structural Magnetic Resonance Imaging (sMRI) is currently a key part of the diagnostic criteria for the differential diagnosis and longitudinal monitoring of patients with dementia. Several studies have consistently observed both global and local atrophic changes in AD, lying along the hippocampal pathway (entorhinal cortex, hippocampus, parahippocampal gyrus and posterior cingulate cortex) in the early stages of the disease, while atrophy in temporal, parietal and frontal neocortices emerges at later stages being associated with neuronal loss as well as with language, visuospatial and behavioral impairments [151, 153].

On the genetic side, PRS are gaining popularity since they represent a single (or few) score(s) combining the effects of multiple independent genetic variants in a subject's genome derived from a large Genome-Wide Association Study (GWAS) study. The PRS are informative about the individual overall genetic disease risk enabling the associations between genetic profiles and imaging features on smaller cohorts. This in particular is the target of Imaging Genetics (IG) which aims at investigating the effects of genetic variations on brain function and structure and in which our work is framed. Such methods, applied in particular to AD onset, allowed a better understanding of the genetic underpinnings on brain modulations [154]. PRS for AD have been shown to be associated with clinical diagnosis and disease progression [155], cognitive decline [156] and imaging biomarkers [157, 156, 158] both on healthy and cognitive impaired patients. Previous studies have generally focused on the hippocampal volume solely to evaluate its association with PRS for AD in cognitive impaired cohorts, considering its central role in AD pathophysiology [156, 155]. A wider range of brain morphometric features was investigated in association with PRS for AD in clinically normal cohorts [158].

To the best of our knowledge, the interaction between PRS for AD and a complete set of brain structural imaging phenotypes, such as cortical thickness and subcortical

volumes, has not been deeply investigated in a cognitive impaired cohort. Typically, univariate models have been applied to characterize IG associations, however such methods do not account for potential cross features interactions and are highly prone to multiple comparison problems leading to underpowered discoveries of significant associations [42]. Multivariate methods, on the other hand, can address such limitations. Latent variable and multi-view models, for example, aim at finding a latent low dimensional space by the optimization of a target function such that the projections of the features hold some maximized joint properties. PLS maximizes the covariance between the latent projections, further addressing features collinearity which generally affects both imaging and genetics derived features. PLS is increasingly being exploited in IG studies, particularly in imaging transcriptomics aiming at investigating the association between imaging phenotypes and gene expression values in brain disorders [159]. Moreover, relying on different genetic features, such as Single Nucleotide Polymorphism (SNP)s, Lorenzi *et al.* [42] exploited PLS to uncover the genetic underpinnings of brain atrophy in AD. Despite these promising results, the potentialities of a classical statistical model as PLS in the AD domain are still under investigated, though could help to disambiguate the associations between different feature sets considering its inherent ability to provide a straightforward explanation of the outcomes, which is not always the case for complex deep models.

The objective of our work was the characterization of the different stages of AD in the PLS latent space representation, which is indeed generated by meaningful associations found between brain morphometric features and PRS in AD. Moreover, in order to assess the generalization capability of our model computed on AD and healthy controls, an unseen cohort of subjects affected by MCI was used for testing.

6.2 Materials and Methods

Phenotypes and genotypes used in this study were derived from the ADNI database (`adni.loni.usc.edu`). The full cohort comprehended 826 subjects from the ADNI-1,ADNI-2 and ADNI-GO phases including 243 Controls (CN), 289 EMCI, 179 LMCI and 115 AD patients (age: 72.9 ± 6.2, 71.2 ± 7.2, 71.9 ± 7.7, and 74.8 ± 7.9; females/males: 131/112, 126/163, 79/100, and 47/68). AD and CN subjects were considered as the discovery cohort, while EMCI and LMCI were kept for testing. The considered imaging features were region-based morphometric descriptors derived from T1-w MRI images extracted by UCSF using FreeSurfer version 5.1 and accessed through the ADNI website (date accessed 18/02/2022). 84 anatomical Region Of Interest (ROI)s were included. The average thickness and the volume were considered for cortical and subcortical ROIs, respectively. The subcortical volumes were normalized by the intracranial volume of the

respective subject. 42 features were finally obtained by averaging left and right hemispheres and were considered as phenotype. The genetic information was represented by the 14 PRS proposed in [155]. Briefly, each PRS was calculated by including all independent SNPs passing a p-value threshold in the most recent GWAS [160]. The thresholds adopted were $1e-08$, $1e-07$, $1e-06$, $1e-05$, $1e-04$, 0.001, 0.01, 0.05, 0.1, 0.2, 0.4, 0.5, 0.75, 1. The related PRS will be named as PRS_*threshold_value*. SNPs in the extended APOE locus were excluded from the PRS construction to enable investigations of risk independent from APOE. We refer to [155] for further details on PRS computation. A standardization to reach zero mean and unitary standard deviation was applied to the feature sets. Age was then regressed out from the image-derived features, while the first two principal components, describing the genetic information of the whole population on which the PRS were calculated, were regressed out from the genetic features, following [155].

6.2.1 Partial Least Squares

Among the available methods adopted in imaging genetics research, the most straightforward are multivariate approaches.

Considering the subject k, the respective imaging and genetics derived features can be expressed as x_k and y_k respectively, where $dim(x_k) = D_i$ and $dim(y_k) = D_g$ represent the number of features for the two sets. We can define then the matrices X as the imaging feature matrix and Y as the genetics feature matrix, which for N subjects will have $dim(X) = N \times D_i$ and $dim(Y) = N \times D_g$, respectively. The goal of classical multivariate approaches is the identification of the linear transformation of X and Y into a lower dimensional hidden subspace where the projected data exhibits statistical similarity.

In this work we will focus on PLS [90], thus on the identification of linear transformation parameterized by the vectors $\mathbf{w}_x$ and $\mathbf{w}_y$ such that the covariance between the projection $\mathbf{X}\mathbf{w}_x$ and $\mathbf{Y}\mathbf{w}_y$ is maximized.

$$\mathbf{w}_x, \mathbf{w}_y = argmax_{w_x, w_y}\left(\frac{\mathbf{w}^T \mathbf{S}_{\mathbf{XY}} \mathbf{w}_y}{\sqrt{\mathbf{w}_x^T \mathbf{w}_x}\sqrt{\mathbf{w}_y^T \mathbf{w}_y}}\right) \tag{6.1}$$

where $\mathbf{S}_{XY}$ is the cross-covariance matrix between the feature matrices X and Y.

The PLS problem, as well as the Canonical Correlation Analysis (CCA), can be straightforwardly optimized by the solution of an eigenvalue problem, but to avoid the instabilities related to the decomposition of potentially large sample covariance matrices, it can be computed by leveraging on stable numerical schemes. The Nonlinear Iterative PArtial Least Squares (NIPALS), is a classical algorithm [90] to solve this problem.

PLS was hence applied in order to model the joint variation between phenotype and genotype observed in healthy and AD individuals, following [42, 157]. Then, the generalization capability of our model was assessed on an unseen cohort of MCI subjects.

6.2.2 Model significance and validation

The data variability explained by each component was calculated, and the number of components was chosen in order to allow to represent at least the 60% of it. A permutation test based on the obtained singular values was finally performed to assess the significance of the model In brief, the test checked whether the singular values obtained by the model were higher than the ones obtained by randomly permuting all rows of the phenotype matrix ($10e4$ permutations were used). The Mann Whitney non-parametric U-test was performed to assess the significance of the latent space projection difference across groups. Finally, the generalization of the PLS model was tested on the MCI group by statistically assessing the ability of the estimated PLS components in splitting EMCI and LMCI subjects, through group-wise comparison of the projections in the latent space.

6.3 Results

Two latent components were needed to explain at least the 60% of data variability, accounting for 54% and 18% of data variability, respectively. The PLS weights of phenotype and genotype in the first and second latent components of the model are shown in Figure 6.1. The PLS model associates a weight to each input feature reflecting its relevance in shaping the latent space, that is in the association between genotype and phenotype.

Chapter6/Figures/PLS_weigths_small.png

Fig. 6.1: First and second PLS components weights (rows) for the phenotype and the genotype features (columns).

The first component revealed a widespread negative correlation between phenotype and genotype. The five most relevant brain regions were postcentral gyrus, caudalanterior cingulate, insular cortex, lingual gyrus and cuneus. On the genetics side, the less AD specific PRS, hence the ones having a less stringent p-value threshold for SNPs inclusion, showed the highest weights. Moving to the second component, pallidum, hippocampus, caudate, entorhinal and inferiortemporal appeared as the most relevant regions. More in detail, pallidum was anticorrelated to the hippocampus volume, and entorhinal and inferiortemporal thicknesses, while it appeared to be correlated with the caudate volume. On the genetic side, this component highlighted the most AD specific PRS, hence the ones including SNPs peculiar for AD. These were positively correlated with pallidum and caudate volumes, while a negative correlation was found with hippocampus. Moreover, the permutation test confirmed the significance of our model resulting in $p = 0.0428$. The latent space representation of AD and CN groups is shown

Chapter6/Figures/Latent_space_training_validation.png

Fig. 6.2: Latent representation of the discovery set and MCI cohort validation set (rows) on the first two PLS components (columns) (AD: blue, CN: grey, EMCI: green, LMCI: violet).

in Figure 6.2 for the two PLS components. Both showed a separation between the two classes, particularly evident in the second one. The projection in the latent space led to significant group-wise differences for the phenotype on both PLS components, reach-

ing $p < 1e{-}12$ on the first and $p < 1e{-}17$ for the second one. Conversely, a trend towards significance was found for the AD vs CN difference in the genotype latent space projection, with $p = 0.086$ and $p = 0.121$ for the first and second components respectively.

Figure 6.2 proves also the model generalization capability by showing the projection of the MCI independent set on the latent space generated by AD and CN subjects. While the first component showed a major overlapping between EMCI and LMCI, the second one allowed a clearer separation, with the LMCI being distributed in the same latent space region as the AD and the EMCI being more central.

Finally, Figure 6.3 summarizes the PLS latent space projections scores for the MCI group on both components, separately for genotype and phenotype. A significant difference was found for the phenotype in both components, $p = 0.042$ and $p = 0.007$, respectively. The genotype differences did not reach significance, though a moderate trend toward significance was present in the second component ($p = 0.130$).

Chapter6/Figures/Score_projections.png

Fig. 6.3: Latent space projection scores of the MCI cohort on the first two PLS components. Significant differences between EMCI (blue) and LMCI (orange), as revealed by the Mann Whitney non-parametric U-test, are highlighted in red for both phenotype and genotype features.

6.4 Discussion

In this work we modeled the relation between gray matter atrophy and PRS via joint multivariate statistical modeling in AD, showing a good generalization of the results by testing the model on an unseen cohort of MCI subjects. Results showed that two PLS components explained a sufficient amount of data variability (> 60%). Both components showed significant separation between AD and CN in the latent space, confirmed also in the MCI projection. Moreover, the latent spatial distribution observed between AD and CN was replicated by the distribution of EMCI and LMCI in the same space.

The association between PRS and brain atrophy has been mainly addressed in literature via general linear model regression. In Scelsi *et al.* [155], for example, the authors focused on the hippocampus volume and found a significant negative correlation between such measure and AD specific PRS in cognitively impaired subjects, in line with our findings. The PRS association with a series of cortical features was explored by Sabunco *et al.* [158] on a healthy cohort. They calculate PRS involving up to 26 independent common sequence variants associated with AD and showed a correlation between late-onset AD PRS and cortical thickness in several AD-specific regions such as entorhinal cortex, temporopolar cortex, lateral temporal cortex, inferior parietal cortex, inferior parietal sulcus, posterior cingulate cortex, and inferior frontal cortex.

The PLS model, on the other side, is a well established method for multivariate analysis and has been widely employed in IG studies. In the work by Lorenzi and colleagues [42], it was used to link brain atrophy to the complete set of SNPs from AD patients, uncovering a significant link between the TRIB3 gene and the stereotypical pattern of grey matter loss in AD. They relied on few structural MRI features for collecting IDPs, while on the full set of SNPs for the genotype. A similar approach was followed in [161], where they were able to stratify the early stages of AD in the PLS latent space by exploiting T1-w features and cerebrospinal fluid levels of t-tau, p-tau and amyloid-beta biomarkers.

Thanks to the straightforward PLS explainability, we were able to recover the features leading the correlation between imaging and genetic features. The analysis of the weights associated with each feature can indeed allow to compare their relative importance and directly evaluate the genotype/phenotype association, highlighting those having a higher impact on the latent space derivation. In our model, the first component represented the great majority of data variability (54%) revealing an anticorrelation between less specific PRS scores and cortical thicknesses, that is in line with the well-known neurodegeneration process in AD. Indeed, the PRS included in this study were associated with disease progression and diagnosis, with an increasing score being correlated with the worsening of the disease. The negative correlation with the pheno-

type hence could be associated to a decrease in cortical thickness, typical of AD progression [153].

The second component, even if it explained a smaller fraction of the full data variability (18%), showed the most significant separation ($p < 1e-12$) between AD and CN, for the phenotype, that was well preserved in the independent MCI cohort ($p = 0.007$). The PRS having the highest associated weights were the ones showing low p-value cutoffs, namely PRS_1e-07, PRS_1e-06 and PRS_1e-08, indeed scores that include established AD risk variants. Such PRS showed an anticorrelation with hippocampus volume and entorhinal cortex thickness among the others, being among the well known most affected regions in AD [153]. Of interest, they were also correlated with pallidum and caudate volumes, with the former showing the highest associated weight. Singleton and colleagues [162] have shown that a significant difference in pallidum volume was present between two AD subtypes, namely typical AD and behavioral AD, with the former featuring an increased pallidum volume compared to the latter. Moreover, Chen *et al.* [163] found a difference between AD subtypes related to the starting site of atrophy, and were able to identify three AD subtypes: (i) typical, for which atrophy begins in hippocampus and amygdala, (ii) cortical, where atrophy starts in the temporal lobe, followed by cingulate and insula and (iii) subcortical with atrophy beginning in pallidum, putamen and caudate. Therefore, we hypothesize that the second component obtained by our model could explain particular differences found across AD groups. In fact, it appears to explain data variability highly specific for AD, due to the high weights associated with the most conservative PRS. On the phenotype, at the same time, high importance was assigned to regions that have been demonstrated to play a role in AD subtypes identification. Further investigation is however needed to strengthen our hypothesis.

6.5 Conclusions

The presented PLS model confirms that there exists a joint variation between grey matter atrophy and PRS in AD, spreading over all the regions considered in the study. Moreover, we were able to capture volumetric modulations that possibly relate to different AD subtypes.

The work presented in this Chapter was published in [164].

7

Assessing the link between diffusion and functional MRI and Polygenic Risk Scores

In this work we exploit Partial Least Squares (PLS) regression to firstly analyze the joint variation between genotype and White Matter (WM) phenotype indices and secondly between the genotype and the functional connectivity in Mild Cognitive Impairment (MCI). Differently from the previous Chapter, we will adopt two separate feature sets for the phenotype. The first set is represented by 192 *WM features derived from diffusion Magnetic Resonance Imaging (dMRI) and extracted through a tract-based spatial statistics (TBSS) analysis on four diffusion tensor based indices, while the second set is composed by within/between network connectivity derived from functional Magnetic Resonance Imaging (fMRI). The genotype information consists of two recently proposed, Alzheimer's Disease (AD) related, Polygenic Risk Score (PRS), namely PRS1 and PRS2. The study cohort is based on the Alzheimer's Disease Neuroimaging Initiative (ADNI) database and comprehends healthy subjects and MCI individuals, including subsets featuring patients showing early and late conditions. Different subjects were considered for the dMRI and fMRI studies due to data availability. The experimental results show that, for the dMRI, in the latent space found by the PLS model, the phenotype revealed an anti-correlation between diffusivity and anisotropy in the WM tracts typical of neurodegeneration, to which both the PRS features are correlated in the first PLS component, and only the PRS2 in the second PLS component. Concerning fMRI, In the first component, all Functional Connectivity (FC) coefficients had the same sign and were correlated with PRS2. Connectivities involving the dorsal attention (DAN) and frontoparietal control (CON) networks reached the highest weights, while within/between network FC for the limbic (LIM) were less represented. Overall, the within-network FC values were less pronounced compared to the between-network ones. In the second component, most of the FC features had zero weights. Visual (VIS) and somatomotory (SMN) showed a correlated trend while being anti-correlated with LIM, CON and default mode network as well as with PRS1. Our findings suggest that the two PRSs correlated with a possible pattern of aberrant*

within/between-network FC changes occurring in Resting State Network (RSN)s devoted to higher cognitive functions and more vulnerable in this pathology.

7.1 Introduction

MCI is a syndrome showing cognitive decline greater than expected for an individual's age and education level. However, this does not notably interfere with daily life activities. Patients showing memory complaints and deficits have a high risk of progression to dementia, in particular to the AD type [165].

While the interaction between genetic and environmental risk factors for AD with brain degeneration have been recently investigated [166], their role for MCI subjects is still less understood. Imaging Genetics (IG) methods can be applied to this aim, exploiting imaging techniques to investigate the effects of genetic variations on brain function or structure in order to better understand their impact on behavior and disease phenotypes [154].

On the genetic side, according to the available data, two types of genetic feature extraction can be performed: Genome-Wide Association Study (GWAS), and polygenic approaches. GWAS consist of observational studies of a genome-wide set of genetic variants in large populations of individuals, targeting the association of genetic variants with Imaging Derived Phenotype (IDP)s. Polygenic studies rely on genetic features derived from available large scale GWAS, and aim at assessing the associations between genetic profiles and IDPs, enabling focused studies on a smaller cohort.

PRS, in particular, are gaining popularity since they represent a single score combining the effects of multiple independent genetic variants derived from GWAS analysis in a subject's genome, capturing the individual overall genetic disease risk [167]. For AD, these scores have been shown to be associated with relevant phenotypes such as disease progression and cognitive decline [155].

On the imaging side, among the available techniques, dMRI is an *in-vivo* technique that allows defining numerical indices that well describe the brain tissue microstructure based on water molecules movement [84]. dMRI has been widely studied in the IG field in relation to Schizophrenia disease, showing promising results [168] and recently it started to be considered also to detect the potential correlation between brain microstructure and genomics in AD. For example, Horguslouglu et al. [169] recently showed that a single nucleotide variation in the gene CELF1 was significantly associated with WM microstructural changes in the hippocampus, while Elsheikh et al. [14], performed a longitudinal study relying on structural brain connectivity defined by tractography and genes.

In order to evaluate also neuronal activity in the brain while performing a given task or at rest it is possible to rely on the non-invasive Blood Oxygenation Level Dependent contrast (BOLD) fMRI. Several authors have demonstrated the functional significance of the spontaneous, low-frequency fluctuations (<0.1 Hz) occurring in the BOLD signal at rest and have proved the existence of spatially distinct brain areas sharing a synchronous BOLD activity, the so-called RSNs [170]. Different FC measures have been devised so far focusing either on the coherence or on Pearson's temporal correlation between time-series measured at different locations in the brain [171]. These features have been scarcely investigated in the IG framework, though could represent important biomarkers for a timely characterization of the underlying functional modulations in neurodegenerative disorders.

Multivariate methods can be used to uncover the interaction between IDPs and genetics features. Among the multivariate approaches, latent variable and multi-view models aim at finding a latent low dimensional space in which the projections of the features show some maximized characteristics. In particular, PLS, which aims at maximizing the covariance between the latent projections, has been started to be exploited in IG studies. Lorenzi et al. [42] exploited this method to uncover the genetic underpinnings of brain atrophy in AD.

In this work, we aim at building on top of [42] to the twofold aim of: (i) Investigating the genetic influence on WM microstructure modulation in MCI. To this end, microstructural indices were derived from dMRI and tract-based features were extracted; (ii) Investigating the genetic influence on FC patterns in MCI under the working hypothesis that FC measures in different RSNs could reveal subtle changes induced by the onset of the disease, allowing to disentangle age-related from pathological functional degeneration and thus potentially enabling early detection of the disease fingerprints. PLS was then applied to investigate the relationship between such features and two recently proposed PRS [172]. The Chapter will separately describe the works reflecting the two aims, in detail Section 7.2 will be devoted to the association between dMRI derived IDPs and PRS while Section 7.3 will describe the pipeline for the analysis of the association between the same PRS and fMRI derived IDPs.

7.2 Association between dMRI derived IDPs and PRS

7.2.1 Materials and Methods

Phenotypes and genotypes used in this study were obtained from the ADNI database (`adni.loni.usc.edu`). The selected cohort comprehended 86 subjects from the ADNI-3 phase including 37 Healty Controls (HC), 5 MCI, 31 Early Mild Cognitive Impairment

(EMCI), 13 Late Mild Cognitive Impairment (LMCI), aged 73.85 ± 5.74, 69.72 ± 9.27, 71.74 ± 6.85, and 67.27 ± 6.34, and with ratios of males/females equal to 17/20, 1/4, 15/16, and 8/5, respectively.

The considered phenotypes were tract-based features derived from dMRI images (Repetition Time (TR)/Echo Time (TE) = 56/7200, 2-mm isotropic voxel, b=1000s/mm^2). The diffusion volumes were preprocessed using FSL software (version 6.0, https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/) applying an initial step of brain extraction (`bet` tool) followed by Eddy currents correction (`eddy` tool). The Diffusion Tensor Imaging (DTI) [84] model was fitted to the corrected images and Fractional Anisotropy (FA), Mean Diffusivity (MD), Radial Diffusivity (RD) and Axial Diffusivity (AxD) indices were extracted. The TBSS pipeline from FSL was applied to FA to derive a group WM skeleton (FA threshold of 0.2) to which all subjects were linearly registered, and the same transformations were subsequently applied for all other indices in order to obtain skeletonized values for each subject. For all subjects, the average value of each index was extracted from 48 Region Of Interest (ROI)s derived from the JHU-DTI atlas available in FSL. The resulting phenotype matrix, of dimensions 86×192, was considered as the independent variable X of the PLS model.

The genotype was represented by PRS1 and PRS2 scores proposed in [172]. Briefly, the PRS1 included all independent Single Nucleotide Polymorphism (SNP)s passing the genome-wide suggestive (p=$1.0e-05$) threshold in the most recent GWAS [160], resulting in 55 SNPs. The PRS2 cutoff was p=0.5, thus including 101.450 SNPs. No correlation was found between these two scores. We refer to [172] for further details on PRS computation. Of note, the SNPs encoding APOE e4 and e2 were included in both scores, and APOE dominated the PRS1 but had negligible effect on PRS2 [172]. The resulting genotype matrix, of dimension 86×2, is referred as the dependent variable Y of PLS model.

A standardization to reach zero mean and unitary standard deviation was applied to both X and Y. Age and sex were then regressed out from the X matrix. PRS2 was standardized, and the first five principal components of the genetic information of the whole population on which the PRS were calculated, were regressed out. In fact, these represented the genetic population structure to which PRS2 was highly correlated.

The PLS regression was finally applied for modeling the joint variation between phenotype and PRS. Ten-fold Cross Validation (CV) was used to identify the number of PLS components minimizing the Predicted Residual Error Sum of Squares (press) error on non-overlapping datasets.

The data variation explained by each component was calculated, and a permutation test was performed to assess the significance of the model. In detail, the test checked

whether the eigenvalues obtained by the model were higher than the ones obtained by randomly permuting all rows of the Y matrix, with $10e4$ permutations.

7.2.2 Results

Chapter7/Figures/fig1.png

Fig. 7.1: First and second PLS components weights for the phenotype and the genotype features. The phenotype weights are grouped by index type (colors). The ROI order is the same for each index and is shown under the barplots.

10-fold CV revealed that both components were needed for minimizing the press error, accounting for the 45% and the 55% of variability of the data, respectively.

The PLS weights of phenotype and genotype in the first and second eigen-component of the model are shown in Figure 7.1. The first component revealed an opposite trend between FA and all other microstructural indices across all ROIs. Moreover, the coefficient of FA usually had lower values compared to the diffusivity indices. This trend

was also observed on the second component, though with opposite sign. Conversely, in the second component, the magnitude of FA coefficients was generally higher compared to the first, suggesting a stronger impact of this component on such anisotropy measure. AxD showed the highest differences, being anti-correlated with FA for all ROIs in the first component, while lemniscus and peduncle were correlated with FA in the second, though with a low PLS weight. RD and MD showed correlated trends in both components, being anti-correlated with the anisotropy index in all ROIs.

Concerning ROIs, other differences can be detected. In the first PLS component, the corona radiata featured the largest PLS weight for all indices, while the RD and MD showed relatively high PLS weights also in peduncle and thalamic radiation. The anterior part of peduncle and thalamic radiation reached relatively high weight also in AxD. For the second component, the external capsule showed the highest PLS weight, coherently with FA, MD and RD. In general, an hemispheric symmetry was found for ROIs showing the highest PLS weights.

Regarding genotype variation, the selected PRS resulted as anti-correlated with all diffusivity indices in the first component, while in the second component only PRS2 showed this anti-correlation. In particular, the PRS1 showed the highest absolute weight in the first component, while the PRS2 was the highest in the second one.

The p-value retrieved by the permutation test did not reach the significance.

7.2.3 Discussion

In this work we modeled the intrinsic relation between brain microstructure and PRS via joint multivariate statistical modeling, identifying a link between PRS and the WM damage seen in subjects with MCI as characterized by the dMRI-based indices.

The obtained results showed that the two PLS components explained a comparable amount of variability, confirmed by the 10-fold CV applied to the model.

The anti-correlation between anisotropy and diffusivity indices detected by both PLS components was coherent with the expectation for the MCI condition. In fact, MCI patients are likely affected by a progressive WM degeneration due to axonal loss and demyelination, resulting in increased water diffusivity in the tracts [173]. The ROIs showing the highest PLS score, hence the most relevant for the model were in agreement with previous studies on the modulation of dMRI derived indices in MCI patients [173], where the external capsule, the corona radiata and the tracts connecting the limbic regions it showed an increased AxD and RD in MCI compared to healthy controls.

In this work, the absence of significance of the permutation test could indicate a low specificity of the PRS scores to the MCI patients selected in this study cohort. However, their link with the expected microstructural modulation for MCI subjects, is promising and deserves further investigations.

To the best of our knowledge, this is the first attempt to exploit PLS model in order to link the PRS for AD with brain microstructure in people experiencing MCI. In the work of Lorenzi et al. [42], PLS was used to link brain atrophy to the complete set of SNPs from AD patients, uncovering significant link between the TRIB3 gene and the stereotypical pattern of grey matter loss in AD. They relied on structural Magnetic Resonance Imaging (sMRI) for the imaging side and, more importantly, on a much larger dataset. A similar approach was followed in [161] where, exploiting sMRI based features and cerebrospinal fluid levels of t-tau, p-tau and amyloid-beta biomarkers, they were able to stratify the early stages of AD in the PLS latent space.

Despite the promising preliminary results, the limitations of this study consisted of the small cohort of subjects which did not allow a strong validation of the model on an unseen cohort, and did not allow to completely exploit the PRS. In the future, we plan to add more subjects to the model, including also AD patients, aiming at better depicting the microstructural link with the genetic component of AD disease progression.

7.3 Association between fMRI derived IDPs and PRS

7.3.1 Materials and Methods

The data analyzed in the current study were collected from the ADNI database as part of ADNI-3 phase (http://adni.loni.usc.edu/). The selected cohort comprehended 177 subjects, including 95 HC (among which 52 were classified as Cognitively Normal and 43 with Significant Memory Concern) and 82 MCI (comprising 52 EMCI, 4 MCI and 26 LMCI, according to the ADNI database).

Rs-fMRI acquisitions were performed on a 3T scanner with the following sequence parameters: TR/TE = 3000/∼ 30 ms, FA = (90°), Field of View (FOV) = (220 × 220 × 163) mm, 3.4-mm isotropic voxel size. 200 fMRI volumes were acquired in almost all subjects, with minimal variations in a small subset (e.g, 197 or 195 volumes). T1-weighted images were also available (main parameters: TR = 2300 ms, FOV = 208 × 240 × 256 mm, 1-mm isotropic voxel size). Data were preprocessed using the FMRIB Software Library (FSL version 6.0) (https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/).

As minimal preprocessing, removal of the first five volumes, motion realignment (MCFLIRT), 4D mean intensity normalization, spatial smoothing with a 6-mm FWHM kernel and interleaved slice-timing correction were performed. A nuisance regression pipeline was then applied to regress out from the minimally preprocessed data the six motion parameters (plus their derivatives), the mean white matter/cerebro-spinal fluid (WM/CerebroSpinal Fluid (CSF)) signals and a linear trend component [174]. WM

and CSF signals were extracted from the corresponding partial volume maps after erosion and binarisation with a threshold of 0.8. The residuals resulting from this analysis were subsequently band-pass filtered (0.01-0.08 Hz). To further eliminate motion artifacts, scrubbing was applied to remove high-motion frames as defined by exceeding 0.5 mm framewise displacement, zero-padding these volumes together with one preceding and two subsequent volumes to keep the number of data points consistent across subjects. Lastly, the preprocessed Resting State functional Magnetic Resonance Imaging (rs-fMRI) images were spatially normalized to the 2-mm MNI space (non-linear registration). The FC matrices were generated using the Schaefer functional atlas [175] with 100 parcels and 7 RSNs, namely visual (VIS), somatomotory (SMN), dorsal attention (DAN), ventral attention (VAN), limbic (LIM), frontoparietal control (CON), and default mode networks (DMN). The symmetric connectivity matrices were calculated using Pearson's correlation coefficient.

In order to exploit the FC patterns in these well-known networks, summary measures representing the mean connectivity value inside a given network (within-network FC) and across edges connecting regions belonging to different networks (between-network FC) were derived from the full matrices. Within-network FC was calculated as the mean value of all the region-to-region connectivities within a specific network (e.g., DMN), while between-network FC was derived by averaging across the edges connecting a node in a network with the other nodes in the remaining networks (e.g, DMNSMN or DMN-VIS) [176, 177]. These operations led to 28 single FC features per subject to be used as IDPs. Regarding the genetics, two PRSs namely PRS1 and PRS2, proposed in [172], were chosen. These scores were based on a recent GWAS study [160] and were calculated according to SNPs passing the genome-wide suggestive threshold ($p = 1.0e{-}05$) and the $p = 0.5$ cutoff, respectively. For further information regarding the PRS calculation please refer to [172].

Finally, a PLS model was applied to the data to capture the association between the IDPs and genetics. The imaging features extracted represented the matrix X in the PLS model (177×28), while the genetic features composed by the two PRSs were presented as matrix Y (177×2). Before applying the model, genetic features and IDPs matrices were standardized by subtracting the mean and dividing by the standard deviation. Deconfounding was also applied to remove the bias from age and gender in the X matrix. Conversely, the first five principal components of the genetic information of the whole population on which the PRS were calculated were regressed out from PRS2 only, as these represented the genetic population structures to which such PRS was highly correlated [172]. The PLS model was applied with Least Absolute Shrinkage and Selection Operator (LASSO) regularization using a penalty value of 0.15. A permutation test, based on

Chapter7/Figures/meanmatrix.png

Fig. 7.2: Mean FC matrices, averaged across all subjects belonging to a given group. The 100 parcels were grouped according to the 7 resting-state networks in the Schaefer atlas, keeping the two hemispheres separately.

1000 permutations of the rows of the Y matrix, was used to test the significance of the obtained eigenvalues ($p < 0.05$).

7.3.2 Results

FC matrices were derived from all the subjects, leading to the final average matrices reported in Figure 7.2 for the two groups. The 28 summary FC measures per subject were included in the PLS model along with the two individual PRSs, and the resulting two components accounted for 54% and 46% of variability of the data, respectively.

The PLS weights of phenotype and genotype in both eigen-components are reported in Figure 7.3 In the first component, while all the coefficients had the same sign, differences could be appreciated in terms of weights across the FC features. The within/between-network connections involving DAN generally featured the highest values, along with those encompassing the CON network. In particular, the between-network FC related to CON appeared as having the highest weights in all cases. Conversely, the connections with LIM had lower weights, especially for the intra-network one which reached the lowest value. Of note, in all cases the within-network FC values had generally lower weights compared to the between-network ones. In the second

Chapter7/Figures/withlasso28feat.png

Fig. 7.3: PLS component's weights for imaging and genetic features. Darker colors represent the within-network connectivity while the lighter shades show the between-network connectivity.

component, most of the FC features had zero weights (or close to zero, as for LIM VIS). The weights for the within-network FC were prominent in all cases except for DAN and VAN, and these features presented an opposite trend across networks, differently than before. Indeed, VIS and SMN showed a correlated trend, while being anti-correlated with LIM, CON and DMN. The magnitude of LIM coefficients was generally higher compared to the others, suggesting a stronger impact of this component on such FC measures. Conversely, connectivities related to the SMN, DAN and VAN networks, which reached high coefficient values in the first component, appeared to have a negligible contribution in this second one. Finally, DMN resulted to be the network with the most surviving features either from within- or between-network connectivity (six out of the thirteen). Regarding genotype variation, the PRS2 showed the highest absolute weight in the first component, while the opposite pattern was found in the second one. PRS2 was correlated with all the FC features in the first component, while in the second one PRS1 presented these correlated patterns only for a subset of connectivities involving LIM, CON and DMN. Conversely, the three features related to the SMN network with

high weights (SMN SMN, VAN SMN, DMN SMN) were anti-correlated with the PRS1. Finally, the permutation test proved the significance of the model with p-value = 0.044.

7.3.3 Discussion

In this study, we investigated the associations between neuroimaging phenotypes and genetics via joint multivariate statistical modeling in patients with MCI. The phenotypic features were presented in terms of within/between-network FC derived from rs-fMRI scans, while two PRSs were used as genetic features. These combine the effects of multiple independent risk variants into single scores, being able to capture an individual's overall genetic disease risk [172].

PLS model was applied to maximize the covariance between the two sets of data with LASSO regularization retaining the most relevant features. Analysis of the PLS weights showed associations between specific imaging features and one of the PRSs. In particular, all FC features were correlated with PRS2, while only LIM, CON and DMN were correlated with PRS1 in the second component. These two PRSs have been demonstrated to be associated with clinical diagnosis, CSFtau levels and with progressive atrophy in AD, although the second one has a poorer association with traits [172].

Our results showed that the 28 FC features were differently represented in the two components and had a differential association with the two PRSs, as visible in Figure 7.3, suggesting these PRSs for AD might shape the FC fingerprints in a selective way and deserving further investigations. Interestingly, the connectivities involving DAN, VAN, CON and DMN featured the highest weights in either the first or second component. These are all RSNs involved in higher cognitive functions, they comprise highly connected regions, and are characterized by an increased vulnerability compared to other networks, such as VIS or SMN, in MCI and AD patients. Previous authors argued this might depend on their particular vulnerability to amyloid deposition since the preclinical stages of dementia [178]. DMN in particular has been largely investigated in the current literature, and both within- and between network changes have been reported between HC, MCI and AD patients [179, 180]. However, several studies are going beyond DMN and have recently demonstrated aberrant internetwork changes involving those brain systems that are closely correlated and play a crucial role in higher cognitive function, underlying the central role of the interactions between RSNs in understanding MCI and AD pathology [181]. To the best of our knowledge, this was the first attempt to exploit a multivariate PLS model for linking the PRS for AD with brain FC measures in MCI subjects.

In a previous work by Lorenzi and colleagues [42], PLS was used to associate brain atrophy to the complete set of SNPs from AD patients, uncovering a significant link between the TRIB3 gene and the stereotypical pattern of grey matter loss in AD. A similar

approach was followed in [161], where they were able to stratify the early stages of AD in the PLS latent space by exploiting classical structural features and disease-specific biomarkers such as cerebrospinal fluid levels of t-tau, p-tau and amyloid-beta.

Despite these promising preliminary results, we are aware of the small sample size which represents the main limitation of the current study and the lack of AD subjects. However, we consider the results as relevant as the within/between-network FC modulations that could be detected are in agreement with previously reported patterns in MCI and AD, providing evidence in favor of the suitability of PRS scores for explaining even in a selective way the genetic underpinning of such changes.

7.4 Conclusions

Concerning dMRI, the presented PLS model suggested that there exists a joint variation between brain microstructure and PRS in MCI subjects. Moreover, the PRS correlated with the expected pattern of WM degeneration identified by the dMRI indices. Moving to fMRI, the proposed PLS model with LASSO suggested a joint variation between FC and PRSs in MCI subjects. Moreover, the two PRSs correlated with a possible pattern of aberrant within/between-network FC changes occurring in RSNs devoted to higher cognitive functions and more vulnerable in this pathology.

The dMRI based work presented in this Chapter was presented in [182], while the fMRI based study was published in [157].

8

Describing genetics in a more informative way: investigating the link between gene variant scores and structural MRI

The joint modeling of genetic data and brain imaging information allows determining the pathophysiological pathways of neurodegenerative pathologies such as AD. Gray matter atrophy is a well established biomarker and genetic variants play a prominent role in disease development. Their joint modeling allows decoding their associations while pointing towards new potential genetic determinants. In this work, the Partial Least Squares (PLS) model was used for analyzing the genetic underpinnings of grey matter modulations relying on the Alzheimer's Disease Neuroimaging Initiative (ADNI) phase 3 dataset.
Cortical thicknesses and subcortical volumes derived from T1-weighted Magnetic Resonance were considered to describe the imaging phenotypes and gene variant scores were extracted by the Sequence Kernel Association Test (SKAT) filtering model. Moreover, a transcriptomic analysis was carried on to assess the expression of the resulting genes in cortical and subcortical structures as a form of post-hoc validation. Results highlighted meaningful genotype-phenotype interactions in each significant latent component. Two genetic variants relevant for Alzheimer's Disease (AD) were highlighted that are EPHX1 and BCAS1, respectively involved in neurodegenerative and myelination processes. In particular, the first was associated to the decrease of subcortical volumes and the second to a decrease in the temporal lobe thickness. Noteworthy, BCAS1 is particularly expressed in the dentate gyrus. Overall, the PLS model allowed capturing genotype-phenotype interactions in a restricted study cohort, inline with previous findings and suggesting new potential disease biomarkers.

This Chapter proposes new ways to investigate the genotype phenotype interactions in a restricted study cohort highlighting associations that are descriptive of the underlying mechanisms of neurodegeneration in AD continuum.

8.1 Introduction

Imaging Genetics (IG) has rapidly grown in the last decades, offering the possibility to detect associations between genotype and neuroimaging data and opening new avenues to understand the genetic impact on individual's phenotypes, traits or risk of developing a disease. Indeed, the primary aim of IG is to assess the genetic architecture of brain structure and function, providing new insights into the brain mechanisms and into their role in shaping complex neurological, psychiatric and neurodegenerative disorders such as AD [13, 183]. AD is the most common cause of dementia, affecting 46.8 million people worldwide [184]. The pathophysiology of AD and its genetic drivers have been widely studied in recent years as presented in previous Chapters (Chapter 6 and Chapter 7).

On the imaging side, structural Magnetic Resonance Imaging (sMRI) represents a key element of the diagnostic criteria for the differential diagnosis and longitudinal monitoring of patients with dementia. Several studies have consistently observed both global and local atrophic changes in AD, lying along the hippocampal pathway (entorhinal cortex, hippocampus, parahippocampal gyrus and posterior cingulate cortex) in the early stages of the disease, while atrophy in temporal, parietal and frontal neocortices emerge at later stages, being associated with neuronal loss leading to language, visuospatial and behavioral impairments [151, 185].

On the other hand, AD has a strong genetic component with more than 40 AD-associated genes/loci that have been identified by Genome-Wide Association Study (GWAS) and sequencing projects over the last ten years, supported by large international GWAS consortia such as the International Genomics of Alzheimer's Project (IGAP) [186, 160]. Segregation analyses have linked several genes to early-onset familial cases that are often explained by rare variants with a strong effect, including APP5, PSEN1, and PSEN2 [187]. Conversely, common risk variants for the more complex late-onset type of AD have been identified thanks to the analyses of massive GWAS data, with strongest genetic risk loci represented by TOMM40, APOE, CLU, PICALM and ADAM10 among the others [186, 187]. Therefore, combining the genetic information with quantitative traits of neuroimaging data to unravel the genetic causes of AD nicely fits within the IG framework and is increasingly pursued in recent years, as demonstrated in several reviews on the topic [13, 188]. Advances in this respect have been fostered by well-know large-scale projects such as the ADNI [189, 190], the UK Biobank [191] and the Enhancing Neuro Imaging Genetics through Meta Analysis (ENIGMA) consortium [192]. ADNI in particular represents the landmark AD biomarker study, being a large and rich repository of open-source genomics, neuroimaging (Magnetic Resonance Imaging (MRI) and positron emission tomography), cognitive, behavioral, and clinical data.

In particular, the last phase ADNI-3, started in 2016 and still ongoing, has introduced updated MRI technologies [190] which are still partly investigated in the current literature.

As such, these initiatives have facilitated the availability of large databases coupling imaging and genetics data acquired on the same subjects, greatly promoting the development of novel methodologies and applications in IG. The earliest IG studies focused on analyzing the influence of candidate genes and/or specific genetic variants on a series of brain Imaging Derived Phenotype (IDP)s, usually modelled as separate outcome variables in univariate / mass-univariate approaches [183]. These studies with candidate genes and candidate IDPs have proven the validity of the IG approach, allowing to test biologically plausible hypotheses and to cast light on the ways in which genetic variants shape brain morphology and functionality in different disorders including AD. However, such methods do not account for potential cross-feature interactions, in particular, they might ignore the genetic correlation among multiple phenotypes (pleiotropy) and are highly prone to multiple comparison problems leading to underpowered discoveries of significant associations [42, 13, 193]. Moreover, focusing on a single variable at a time might misattribute the nature of genetic effects on the brain and bias the interpretation of the results considering the complex relationships between genetics and IDPs, especially when effects are spatially distributed and encompass the whole brain [194].

As such, multivariate analysis methods are being increasingly exploited in this domain, in order to improve the discovery of multiple genotype-phenotype associations while circumventing the limitations inherent to univariate approaches. Methods able to capture the integrated genetic effects of a set of genetic variants rather than considering each single Single Nucleotide Polymorphism (SNP) might be of help for performing whole-brain association studies, for example relying on Polygenic Risk Score (PRS) [172] or SNP set approaches. For the latter, recent strategies [195, 196, 197] have proposed grouping SNPs together into SNP sets based on their location in a gene, haplotype blocks given by linkage disequilibrium (LD) or according to a given pathway. The SKAT represents in particular one of the most widely used SNP set approaches, being a flexible and computationally efficient logistic kernel-machine regression method to test for association between genetic variants in a region and a given trait while adjusting for covariates [195]. SKAT has been successfully used to study variants in AD [198, 199], but its potentialities in the IG framework have been only partially investigated so far. In this scenario, of note is the study by Lu et al. [197] where SKAT along with group Least Absolute Shrinkage and Selection Operator (LASSO) and Bayesian latent variable selection were tested on a cohort of AD subjects to identify associations between genes and nine imaging phenotypes, represented by regional volume measures. The authors

demonstrated the added value of such approaches which allow accounting for the correlation among SNPs and detecting causal SNP sets, limiting the burden of multiple comparison correction. However, these promising results were achieved by analyzing each regional imaging volume separately, though, as the author recognised, these are usually correlated and their joint modeling may hold an increased power borrowing additional information.

Therefore, multivariate methods represent the key to address such limitations, allowing to leverage the multiscale phenotype-genotype fingerprints while reducing the multiple testing burden, resulting in higher statistical power to identify significant associations [13, 188]. Latent variable and multi-view models, for example, aim at finding a latent low dimensional space by the optimization of a target function such that the projections of the features hold some maximized joint properties. Canonical Correlation Analysis (CCA)-based methods have been largely applied in the IG framework in the past years, resulting into linear combinations of the two sets of variables which have maximum correlation with each other. Such approaches demonstrated high precision in assessing correlation patterns between the given features [193], for example when considering SNPs and functional MRI features [200], or in its sparse and multi-view version to establish associations between SNPs, sMRI IDPs and cognitive outcomes [201]. PLS analysis, which aims at maximizing at each step the covariance rather than the correlation between the latent variables, has been less frequently applied for detecting the multivariate genotype-phenotype associations. Although CCA and PLS are mathematically related, studies demonstrated that PLS may be more suitable and have improved predictive power when dealing with high-dimensional datasets, especially those with highly collinear variables common in IG experimental settings [202, 203]. Interestingly, Lorenzi *et al.* [42] exploited PLS to uncover the genetic underpinnings of brain atrophy in AD by relying on SNPs and T1-w sMRI, demonstrating the presence of a significant link between TRIB3 and the stereotypical pattern of gray matter loss in AD. Despite these promising results, the potentialities of a classical statistical model as PLS are still under investigated, though could help to disambiguate the associations between different feature sets and provide a straightforward explanation of the outcomes.

Therefore, in this work, we aimed at investigating the genetic mechanisms underlying brain atrophy in the AD continuum relying on a data-driven PLS multivariate approach to model their joint covariation with a twofold goal: (i) exploring the association between imaging (sMRI IDPs) and genetics (SNP set) features identified in a study cohort of healthy controls (CN) and patients on the AD continuum (PAT) from ADNI-3 phase; (ii) analysing the common latent representation obtained from the model, focusing on the components that clearly distinguish between the classes. A post-hoc analysis was performed with dual objectives: to identify the input features driving the

	Discovery set		Validation set	
Diagnosis	Controls (CN)	Patients (PAT)	CN	PAT
Count	181	62	39	15
Age, y	71.19 (6.12)	72.21 (8.88)	70.45 (6.11)	72.83 (9.56)
Education, y	17.05 (2.11)	16.11 (2.56)	16.49 (2.29)	15.67 (2.74)
Sex, %females	60	39	64	27

Table 8.1: Sociodemographic characteristics of the study cohort. Age and education are reported as years mean and standard deviation [Mean (SD)] while sex as the percentage of female individuals.

genotype-phenotype associations using the model's weights and to assess the expression of the detected genes through a transcriptomic analysis. The resulting model was then furtherly validated on an independent cohort, representing the same class distribution as the discovery set.

8.2 Materials and Methods

Figure 8.1 shows an overview of pipeline proposed in this work. In what follows all the steps will be fully detailed.

8.2.1 Study cohort

Data used in this study were derived from the ADNI database (adni.loni.usc.edu), in particular from the ongoing ADNI-3 phase. The ADNI was launched in 2003 as a public–private partnership led by Principal Investigator Michael W. Weiner. Up-to-date information is available at www.adni-info.org.

Summary sociodemographic, clinical, and genetic information is available in Table 8.1. Participants selection was based on the availability of MRI and genetic data and ethnicity, restricting the analyses to participants with European ancestry. The final cohort comprehended 297 subjects divided into 220 healthy controls CN and 79 patients PAT, 19 of which were AD while the remaining were Mild Cognitive Impairment (MCI) subjects. 80% of subjects was considered as the discovery cohort, while the remaining 20% was kept for validation. A comparable proportion between PAT and CN was kept in the discovery and validation sets.

Chapter8/Figures/Pipeline.png

Fig. 8.1: Overview of the proposed pipeline. Phenotype and genotype, representing region based cortical thicknesses and subcortical volumes respectively, were given as input to Partial Least Square (PLS) modeling to model the underlying joint covariance. For each obtained Latent Component (LC) the latent space as well as the separation between patients (PAT) and controls (CN) was evaluated. Model explanations were extracted through the analysis of the PLS weights which allowed retrieving positive and negative associations between the genotype and phenotype. The model was validated through a permutation test as well as the projection of an independent validation set on the obtained latent space which allowed to verify model generalizability. Finally, a transcriptomic analysis was perform to investigate the brain expression of the most relevant genes.

3D T1-weighted (T1-w) MRI volumes were considered for IDPs extraction (sagittal accelerated MPRAGE, Repetition Time (TR)/Echo Time (TE) = shortest, Inversion Time (TI)=900 ms, flip angle = $9o$, Field of View (FOV) = $256\times256\text{mm}^2$, spatial resolution = $1\times1\times1\text{mm}^3$, slices = 176-211). More details about the data acquisition can be found in [190]. ADNI-3 participants were genotyped using the Illumina Infinium Global Screening Array v2.

8.2.2 Image processing and phenotype feature extraction

The T1-w volumes were minimally preprocessed for bias-field correction (*fsl_anat* tool [204]). Subsequently, 84 anatomical Region Of Interest (ROI)s were extracted using FreeSurfer version 7.0 [205]. The average thickness and volume were considered for cortical and subcortical ROIs, respectively. The subcortical volumes were further normalized by the estimated total intracranial volume of the respective subject. The ROIs were averaged over hemispheres resulting in 42 features to be used in the subsequent analyses.

Moreover, as preliminary analysis, a Mann Whitney non-parametric U-test was performed to assess the group-wise differences between PAT and CN, separately for each brain feature. This allows to gain a clear insight into relations already present in the input features. false discovery rate (FDR) correction ($p_{fdr} < 0.05$) was applied considering 42 comparisons.

8.2.3 Genetic processing and genotype feature extraction

Quality Control (QC) procedures were conducted on genotype data using the whole-genome association analysis toolset PLINK 1.9 [206]. SNPs and subjects were filtered out based on missingness ($geno > 0.2$, $mind > 0.1$), minor allele frequency ($MAF > 0.05$) and deviations from Hardy–Weinberg equilibrium ($hwe > 1\text{e}{-06}$). QC kept 303150 SNPs out of the 759993 SNPs collected in ADNI-3. No subjects were filtered out.

GWAS analysis was performed as benchmark, including the top ten principal components from a Principal Component Analysis (PCA) over genotype data, age and gender as covariates.

SNP set analysis was then performed using the SKAT model [195]. SNP sets were defined in order to correspond to different genes and will be referred as "genes" throughout the paper. Of note, only SNPs located in the gene's exon regions were included. This led to 17295 genes containing a total of 132312 SNPs. SKAT was hence used to test the association between each gene and the disease status (PAT or CN) using logistic kernel-machine-based test adjusted by covariates. The R package *SKAT* was used to perform the analysis, specifying a linear weighted kernel and the same set of covariates as for the GWAS analysis.

More in details of the model, a SNP data vector, representing in this case a gene can be defined as $\mathbf{G}_i = \{g_{i1}, \ldots, g_{ip}\}$, where p is the number of SNPs in the selected gene, for a subject i, where $i = 1, \ldots, n$. The relationship between the $\mathbf{G}_i$ and the subject's disease status y_i, is given by $y_i = \alpha_0 + \boldsymbol{\alpha}'\mathbf{C}_i + \boldsymbol{\beta}'\mathbf{G}_i + \epsilon$, where α_0 is an intercept term, $\mathbf{C}_i = \{c_{i1}, \ldots, c_{im}\}$ is the vector of the m covariates, $\boldsymbol{\alpha} = \{\alpha_1, \ldots, \alpha_m\}$ is the vector of regression

coefficients for covariates, $\boldsymbol{\beta} = \{\beta_1, \ldots, \beta_p\}$ is the vector of regression coefficients for the p SNPs, and ϵ is the error term (mean $\mu = 0$ and σ^2 variance).

Evaluating whether the gene variants influence the disease state, adjusting for covariates, corresponds to testing the null hypothesis $H_0 : \boldsymbol{\beta} = 0$, hence $\beta_{i1} = 0, \ldots, \beta_{ip} = 0$. SKAT tests H_0 by assuming that each β_v follows an arbitrary distribution with mean 0 and variance $w_v \tau$, where τ is a variance component and w_v is a predefined weight for the SNP g_{iv}. The null hypothesis can be rephrased as $H_0 : \tau = 0$, which can be tested through a variance-component score test, which it only requires fitting the null model $y_i = \alpha_0 + \boldsymbol{\alpha}_1' \mathbf{X}_i + \epsilon_i$.

The variance-component score statistics is given by $Q = \frac{(\mathbf{y}-\hat{\boldsymbol{\mu}})' K (\mathbf{y}-\hat{\boldsymbol{\mu}})}{2}$, where K is the weighting linear kernel, $\hat{\boldsymbol{\mu}} = \hat{\alpha}_0 + \mathbf{C}\hat{\boldsymbol{\alpha}}$ is the predicted mean of $\mathbf{y}$ under H_0 and $\hat{\alpha}_0$ and $\hat{\boldsymbol{\alpha}}$ are estimated under the null model by regressing $\mathbf{y}$ on only the covariates $\mathbf{C}$. More in detail of K, it is a $n \times n$ matrix where each cell measures the genetic similarity between two subjects i and i' in the gene given the p SNPs.

To derive a p-value for the considered gene, SKAT tests if Q follows a mixture of chi-squared distributions. In our analysis a gene is considered as significant if its associated p-value is $\leqslant 0.05$.

Once the significant genes were obtained through SKAT, a function to map the population significant genes to a subject specific measure was proposed. In detail, for a subject i and for a significant gene G resulted from SKAT application, a gene based variant score $\eta_i(G)$ was extracted representing a measure of how mutated are the entire SNPs located in G.

More in detail, the state of a SNP g_{iv} is 0, if no genetic variation between the specific subject i and the reference genome is present, 1 otherwise. No distinction for diploid variations at the same locus was considered. The gene variant score of G for the subject i is defined as

$$\eta_i(G) = \frac{\sum_{v=1}^{p} g_{iv}}{p}$$

and was calculated for each subject and each significant gene resulting from SKAT application. Of note, as it has been done and described for the phenotype, a Mann Whitney non-parametric U-test was performed to assess the group-wise differences between PAT and CN, followed by FDR correction ($p_{fdr} < 0.05$).

In order to better analyze the genes resulting from SKAT analysis, their association with the disease was furtherly assessed using Hetionet [207] and REACTOME pathway analysis (R package ReactomePA, [208]). In detail, Hetionet is an open-source biomedical graph database that combines the information from 29 public databases into a single resource. It contains 47031 nodes of 11 types (e.g. genes, diseases, pathways, compounds) and 2250197 edges of 24 types (e.g. upregulates/downregulates, interacts). All

the significant SKAT genes were searched inside Hetionet in order to retrieve their eventual link with AD. REACTOME was additionally used to perform enrichment analysis starting from the full set of significant SKAT genes. Significant pathways were selected based on the associated FDR-adjusted p-value ($p_{fdr} < 0.2$). Moreover, pathways associated with AD in Hetionet were furtherly selected.

8.2.4 Partial Least Squares analysis

Phenotype and genotype were organized in two separate data matrices, X and Y, respectively, subsequently divided in discovery and validation. To ensure that the differences in thickness and volumes magnitudes as well as with SKAT scores do not dominate the statistical model, the X and Y data matrices were z-scored column-wise, by subtracting the mean from each column and dividing by the standard deviation of that column. Moreover, the influence of age was regressed out only from the phenotype.

Aiming at the modeling of the joint variation between the morphometric IDPs and the gene variant scores observed in our cohort, the PLS model was applied following [42, 157, 164]. Our focus is on the symmetric PLS formulation computed using the NIPALS algorithm [90], among the numerous versions of PLS proposed in the literature. Within this setting, PLS is intended to estimate the Latent Component (LC)s that maximize the global covariance between the two input modalities. More in detail of the formulation, this model aims at the identification of linear transformation parameterized by the vectors $\mathbf{w}_x$ and $\mathbf{w}_y$ such that the covariance between the projection $\mathbf{X}\mathbf{w}_x$ and $\mathbf{Y}\mathbf{w}_y$ is maximized, namely:

$$\mathbf{w}_x, \mathbf{w}_y = argmax_{w_x, w_y}\left(\frac{\mathbf{w}^T \mathbf{S}_{\mathbf{XY}} \mathbf{w}_y}{\sqrt{\mathbf{w}_x^T \mathbf{w}_x}\sqrt{\mathbf{w}_y^T \mathbf{w}_y}}\right) \tag{8.1}$$

where $\mathbf{S}_{XY}$ is the cross-covariance matrix between the feature matrices X and Y. NIPALS algorithm optimizes this function iteratively. In each LC, each input feature is hence given a weight according to its relative importance for describing the global multimodal relationship of the input features. More in detail, the magnitude of the associated weight directly reflects the importance of each feature in the common latent space definition, the sign, instead, relates to the direction of the association between different features, direct or inverse, also referred to as correlation or anticorrelation. Importantly, this sign does not necessarily entail an effective increase/decrease of that feature's value in a group of subjects compared to the other, but simply describes the associations between features found by PLS. The preliminary analysis on the input described in Subsections 8.2.2 and 8.2.3 will aid the interpretation of the obtained associations

The optimal number of PLS LCs was chosen by calculating the data variability explained by each of them based on model singular values, which reflects the data variability captured by the LC. The ratio of the singular value to the sum of singular values from the decomposition was used to threshold the components in order to retain the 60% of explained data variability. A permutation test based on the obtained singular values was finally performed to assess the significance of the model [91]. In brief, the test checked whether the singular values associated to each LC were higher than the ones obtained by randomly permuting all rows of the phenotype matrix ($1e4$ permutations were used).

LC related projection scores were derived separately for genotype and phenotype by multiplying the inputs by the LC associated weights. The group-wise Mann Whitney non-parametric U-test was then performed on the projection scores to assess group-wise differences between PAT and CN. Only the associations generated by LCs showing significant separation between the two groups on both the phenotype and genotype projection scores were deeply investigated.

The generalization of the PLS model was tested on the unseen group by statistically assessing the ability of the estimated PLS components in splitting patients and controls through group-wise comparison of the projections in the latent space (Mann Whitney non-parametric U-test). All the PLS analysis and validation was performed using Python, relying in particular on scikit-learn library [209].

8.2.5 Transcriptomic analysis

Finally, a transcriptomic analysis was performed based on the Human Protein Atlas (HPA) database (`proteinatlas.org`,[210]). The HPA provides normalized transcript per million (nTPM) expression values within 13 brain regions based on RNAseq analysis of 1324 samples from several donors. Each of the most relevant genes resulting from the pathway analysis and the PLS model was checked for expression in brain tissues. Of note, the HPA does not have a reference template for brain regions definition. Their transcriptomic measures were hence aggregated in order to match the Desikan-Killiany Atlas considered for the IDPs definition. A full matching was not however possible due to missing infromation in HPA database, resulting in 39 regions out of 42 having transcriptomic data.

Chapter8/Figures/venn_up.jpg

Fig. 8.2: Overview of the gene's grouping. The Venn diagran is based on four main sets representing genes resulting from SKAT analysis (light blue), genes in Hetionet database (green), genes belonging to the significant pathways returned from the enrichment analysis (pink) and the subset of the latter representing the genes belonging to AD related pathways (yellow). Darker colors are used to represent the intersections between such principal clusters.

8.3 Results

8.3.1 SKAT results

GWAS [186] and SKAT [195] analysis was conducted on the study cohort after a QC preprocessing in order to identify genotype association from case-control data (PAT and CN subjects) as described in Section 8.2.3.
In our analysis GWAS approach failed to discover significant association between individual SNPs and AD due to the involvement of small populations and low prevalence disease [211] in the considered cohort.
The SKAT SNP set analysis highlighted 408 significant genes (p-value $\leqslant 0.05$). These were almost equally distributed in all chromosomes, though a higher predominance could be noted in chromosome 6 (48 genes, 12% of significant genes) and chromosome

ID	Pathway	Genes	*p*-value	*p*-fdr
R-HSA-211999	CYP2E1 reactions	**CYP2D6** / CYP2E1 / CYP2C9	0.001	0.08
R-HSA-211897	Cytochrome P450	**CYP2D6** / CYP2E1 / CYP2C9 / CYP4V2 / CYP11B1 / CYP4F12	0.005	0.12
R-HSA-211859	Biological oxidations	FMO2 / **CYP2D6** / CYP2E1 / GSTM5 / CYP2C9 / EPHX1 / CYP4V2 / MAT1A / UGT2B4 / CYP11B1 / CYP4F12 / MTARC1	0.007	0.14
R-HSA-112316	Neuronal System	**CHAT** / RPS6KA2 / KCNA7 / SLC1A2 / GABRG2 / KCNH5 / CACNA1A / KCNMB1 / CASK / SLC6A1 / KCNJ6 / KCNAB1 CHRNA5 / KCNMB3 / NRXN1 / ABAT / GRIN3A / GABBR1	0.010	0.17

Table 8.2: Significant pathway associated with AD in Hetionet. REACTOME was used to conduct enrichment analysis. For each pathway is reported the Reactome ID, the pathway name, SKAT genes included in the pathway (Hetionet genes are highlighted in bold), *p*-value and false discovery rate adjusted *p*-value.

1 (33 genes, 8% of significant genes).
12 SKAT significant genes had been found associated with AD in Hetionet, and we refer to these as "Hetionet genes", in details: PTGS2 and DPYD (Chr1), TF (chr3), PPARGC1A (Chr4), CDH12 (Chr5), VEGFA (Chr6), LPL (Chr8), CHAT and ABCC2 (Chr10), BDNF (Chr11), AKAP13 (Chr15), CYP2D6 (Chr22).
Enrichment analysis identified 53 significant pathways (FDR adjusted p-value < 0.2). Among these, four pathways were associated with AD in Hetionet, reported in Table 8.2 together with Reactome ID, SKAT genes included in each pathway, *p*-values and adjusted *p*-values. The Hetionet genes included in these pathways (CUP2D6 and CHAT) are highlighted in bold. Moreover, a schematic overview of the different gene subsets adopted in this study can be found in Figure 8.2. While the association between the *Neuronal System* pathway and AD is clear, for the other three pathways we highlight the relationship below. *Biological oxidation* has been demonstrated to be associated with cell toxicity in various neurodegenerative disorders such as AD or Parkinson's Disease. An accumulation of nucleic acid oxidation indicates a decreased capacity to repair the nucleic acid damage [212]. Furthermore, *CYP2E1* reactions pathway is closely associated with *Biological oxidation*. CYP2E1 gene is involved in oxidative stress and can cause cell death [213]. Finally, *Cytochromes P450* in the corresponding pathway constitute a superfamily of enzymes that catalyze the metabolism of drugs. Polymorphisms in cytochrome P450 genes may affect the enzyme catalytic activity and have been associated with AD in several studies [214, 215].

8.3.2 Phenotype and Genotype preliminary analysis

The preliminary analysis on the phenotype, aiming at assessing whether any between-group significant difference was present in the original space, revealed selective alterations surviving the FDR correction. Among the cortical IDPs, temporal regions were the most significant (p-value $\leqslant$ 1e-05 for enthorinal cortex, middle temporal gyrus and temporal pole; p-value $\leqslant$ 1e-04 for Fusiform gyrus, inferior temporal cortex, parahippocampal gyrus, superior temporal gyrus) followed by few parietal regions (precuneus and insular cortex, p-value $\leqslant$ 1e-04) and by bankssts and inferior/superior parietal gyri (p-value $\leqslant$ 1e-03). Moving to the subcortical IDPs, amygdala was the most significant one (p-value = 3.56e-08), followed by hippocampus and accumbens recording p-values of 2.15e-07 and 1e-04, respectively. All the statistics revealed a decrease of the measured features in PAT compared to CN. No significant differences were recorded for the remaining phenotype features. Moving to the genotype, 60 gene variant scores revealed significant differences between PAT and CN. ChrX had the highest percentage of significant genes (42%) compared to its total number of SKAT genes. It was followed by Chr4, Chr22, Chr1, Chr17 and Chr2 which showed a percentage of significant genes above the 15%. However, any comparison survived the FDR correction.

8.3.3 Partial Least Squares analysis

The X and Y matrices for PLS computation had dimension number of subjects (243 for the discovery and 54 for the validation) times the number of respective features (42 IDPs for X and 408 gene variant scores for Y). The PLS model calculation on the discovery set returned a total of 14 LCs to explain at least the 60% of data variability, the first accounting for the 12%, with the others monotonically decreasing till the 3%. In what follows, out of the 14 LCs, we will focus on the components defining a latent space in which a significant difference between groups was found, being the most informative for our scope. Of note, the permutation test confirmed the significance of our model resulting in a p-value = 0.001.

Latent space and projection scores

Among the 14 LCs, the 1st (LC1), the 2nd (LC2) and the 5th (LC5) where the ones generating a latent space showing significant differences between the projection scores of PAT and CN groups for both the genotype and phenotype in the discovery set. Such LCs accounted for the 12%, 7% and 4% of data variability, respectively. Figure 8.3 shows the latent space generated by such LCs, as well as the distribution of the related projection scores, separately for imaging and genetics and for both the discovery and the

Chapter8/Figures/latent_space_up.png

Fig. 8.3: Latent space and projection scores boxplots. The latent space generated by the PLS components showing a significant difference for the projection between PAT (blue) and CN (yellow) are shown in rows. The projection scores for phenotype and genotype separately are then reported for both the discovery set and the validation set (columns). Significant differences between CN and PAT projections, as derived from Mann Whitney non-parametric U-test, are highlighted with red asterisks (*, **, ***, **** refers to p-values $\leqslant 0.05, 1e-02, 1e-03, 1e-04$ respectively).

validation set. Focusing on the discovery set and particularly on the generated latent space, a high correlation was present between genotype and phenotype projections for all the considered LC (Pearson correlation coefficient equals to 0.82, 0.80 and 0.77 for LC1, LC2 and LC5, respectively), while a clearer separation between classes was present in LC2 and LC5, compared to LC1. Moving to the differences in projection scores between PAT and CN, LC1 showed a p-value $\leqslant$ 1e-02, namely p-value of 0.002 and 0.001 for both phenotype and genotype projection scores. LC2 and LC5, despite accounting for a minor data variability, appeared to be the most interesting regarding the differentiation between PAT and CN. LC2 showed strong significant differences for both phenotype and genotype, with a p-value of 1e-04 and 4e-05, respectively. A similar trend was found for LC5 which showed a p-value of 2e-08 and 2e-04 for the genotype and phenotype, respectively.

The projection of validation data on the generated latent space showed a similar distribution of PAT and CN subjects as the discovery set and confirmed some of the significant differences recorded for the discovery set. In detail, for LC1, a significance (p-value = 0.014) was found also on the validation set for the phenotype projection. The validation set on LC2 showed a significant difference for the genotype (p-value = 0.008) found also on LC5 (p-value = 0.010). For both LC2 and LC5 the phenotype in the validation set showed a trend towards the significance (p-values of 0.072 and 0.061, respectively).

The remaining LCs of the model did not show any significant difference between PAT and CN for both the phenotype and the genotype in the discovery or validation set.

Genotype - Phenotype relevance and associations

Figure 8.4 shows the phenotype (imaging) weights from our PLS model, separately for LC1, LC2 and LC5. Overall, a different pattern can be appreciated in the considered LCs, highlighting that each of them explains different brain structure modulations in the considered study cohort. Beginning from LC1, cortical areas featured high positive weights, suggesting their predominant role in shaping this component especially over frontal (parsopercularis, rostral middle and superior frontal gyri, precentral gyrus), temporal (inferior, middle and superior temporal gyri, fusiform gyrus, bankssts) and parietal areas (supramarginal gyrus, inferior parietal gyrus). Lower importance was instead given to all the subcortical regions. Conversely, LC2 assigned high (positive) weights to the subcortical features, showing an anticorrelation between them and the cortical ones. Of note, the most important volumes were hippocampus, amygdala, basal ganglia (putamen, globus pallidus, caudate, and accumbens), and thalamus. These resulted to be positively correlated with the entorhinal cortex and temporal pole,

Chapter8/Figures/Fig_brain_weights_seismic_up.png

Fig. 8.4: Significant PLS components' weights for the phenotype. Positive weights are shown in red, while negative ones are in blue. Drawings generated using BrainPainter [216].

Chapter8/Figures/Heatmap_up_up.png

Fig. 8.5: Association between phenotype and genotype. Heatmap of the PLS weights for the most significant components (rows), thresholded over the 75th percentile of the respective distribution. Background shades highlight cortical (Cort) and subcortical (Subc) features for phenotype, and different chromosomes (e.g. Chr1) for genotype. The corresponding feature name lists can be found in Supplementary Figure 8.10. Positive and negative PLS weights are shown in red and blue respectively.

similarly featuring high positive weights, and anticorrelated particularly with the cerebellum and rostral middle frontal gyrus. Finally, LC5 highlighted a strong separation between frontal regions (frontal pole, parsorbitalis, parstriangularis, cingulate gyri, in

particular caudal anterior, rostral anterior and posterior), medial temporal lobe (entorhinal cortex, parahippocampal gyrus), subcortical volumes (negative weights) with temporal lobe regions (temporal pole, middle temporal gyrus).

Chapter8/Figures/Heatmap_SUPP.png

Fig. 8.6: Heatmap of the genotype PLS weights for the most significant components (rows). Background shades highlight different chromosomes (e.g. Chr1) for genotype. The corresponding feature name lists can be found in Supplementary Figure 8.10. Positive and negative PLS weights are shown in red and blue respectively.

Moving to the genotype (gene variant scores), Figure 8.6 shows the associated PLS weights, separately for LC1, LC2 and LC5. Despite the relevance pattern resulted quite uniform across chromosomes, some differences emerged. More in detail, in LC1 Chr2, Chr3, Chr11, Chr21 and ChrX showed the genes featuring, on average, the highest negative weights in anticorrelation with Chr18 which had instead the highest positive ones. In LC2, the chromosomes featuring the highest positive weights were Chr7, Chr18, opposed to Chr17, Chr19, ChrX. Finally, for LC5, Chr12 showed the highest positive and negative weights. Chr4 and Chr11 (positive weights) were in negative correlation with several chromosomes (Chr9, Chr18, Chr20, Chr21, Chr22, ChrX).

In order to better emphasize the association between phenotype and genotype, Figure 8.5 reports an heatmap illustrating the relative PLS weights for each feature and component. For ease and clarity, imaging features were grouped by cortical (Cort) and subcortical (Subc) regions, while genes were grouped by their position on chromosomes. A threshold to retain only the weights higher than the 75th percentile of the respective distribution was applied. Moreover, Table 8.3 highlights the most relevant associations commented below.

A first macro-analysis was performed on the genetic side, by analyzing the global importance of each chromosome. In particular, the percentage of SKAT genes above the threshold normalized by the total number of SKAT genes in a given chromosome was computed. Results showed that the chromosomes featuring the highest percentage

of relevant genes were Chr18, Chr21 and ChrX for LC1, including the 40% (Chr18 and Chr21) and 42.8% (ChrX) of relevant SKAT genes. Chr7, Chr17, Chr18 and Chr19 had the 44.4%, 45%, 40%, 42.8% and 40% respectively covered by the most relevant SKAT genes in the LC2. Finally, for the LC5, the Chr4, Chr9, Chr12 and ChrX resulted in percentages of 41.2, 42.8, 40.9, 42.8 of SKAT genes with the associated weights above the threshold.

More in depth of the relevant genes in each component, the top 5 genes showing the highest importance for LC1 were the RIF1, ATP6V1G2, NFASC and FBXO403 (negative weights, Chr2, Chr6, Chr1, Chr8 respectively), in anticorrelation with MFSD6L (Chr17). The first four genes were also found to be anticorrelated with the most relevant cortical features on the phenotype, namely the frontal and temporal regions described in the previous paragraph. No weights higher than the 75th percentile threshold were recorded for the Hetionet genes in LC1. However, among the SKAT genes belonging to the four AD pathways, KCNA7 and KCNJ6 (negative weights, Chr19 and Chr21) were correlated with RIF1, ATP6V1G2, NFASC and FBXO43, inheriting the related relation with the phenotypic counterpart detailed above. On the other end, CYP11B1 (pathway R-HSA-211859, Chr8) was correlated with MFSD6L, hence in anticorrelation with the relevant LC1 cortical thickness features.

Moving to LC2, the top 5 most important genes were HNMT (negative weight, Chr2), in anticorrelation with PHF14, RFWD3, MORN1 and TEP1 (Chr7, Chr16, Chr1, Chr14) on the genetic side as well as anticorrelated with the subcortical volumes for the imaging features. Moreover, such PHF14, RFWD3, MORN1 and TEP1 in opposition, showed a correlation with the thickness of cerebellum cortex and rostral middle frontal gyrus among the others. The Hetionet genes relevant in LC2 were the BDNF and CYP2D6 (positive weights, Chr11 and Chr22). These were further correlated with the relevant subcortical regions in LC2. Analysing the genes belonging to the AD pathways, CACNA1A, FMO2, EPHX1 and CYP2D6 (negative weights, pathways R-HSA-112316 and R-HSA-211859, Chr19, Chr1, Chr1, Chr22) had associated weights over the threshold in LC2. More in details, CACNA1A, FMO2 and EPHX1 were correlated with HNMT and hence inheriting its relation with the phenotype. CYP2D6 was instead found in correlation with such aforementioned imaging features, as well with the genes in the top positions in terms of weights (PHF14, RFWD3, MORN1, and TEP1) and BDNF.

Finally, the top 5 genes for the LC5 were BCAS1, GLT6D1, TMPRSS15, COL6A3 (negative weights, Chr20, Chr9, Chr21, Chr2) and were anticorrelated CEP164 (Chr11). The latter was correlated with entorhinal cortex, fusiform gyrus and temporal pole, while the others were in correlation mainly with cingulate gyri and frontal pole. Among the Hetionet genes found in the SKAT set, CHAT, TF and BDNF (positive weights, Chr10, Chr3, Chr11) were found in correlation between each other in this LC, as well as with entorhinal, fusiform gyrus and temporal pole in the phenotype. An anticorrelation

LC1		LC2		LC5	
Imaging	Genetics	Imaging	Genetics	Imaging	Genetics
supramarginal gyrus	MFSD6L	hippocampus	PHF14	entorhinal cortex	BCAS1
middle temporal gyrus	RIF1	putamen	RFWD3	caudal anterior cingulate	GLT6D1
inferior parietal gyrus	ATP6V1G2	globus pallidus	MORN1	rostral anterior cingulate	TMPRSS15
superior frontal gyrus	NFASC	caudate	TEP1	fusiform gyrus	COL6A3
inferior temporal gyrus	FBDXD43	accumbens	HNMT	frontal pole	CEP164
precentral gyrus	KCNA7	amygdala	BDNF	parahippocampal gyrus	TF
rostral middle frontal gyrus	KCNJ6	thalamus	CACNA1A	parsorbitalis	CHAT
parsopercularis	CYP11B1	cerebellum	FMO2	temporal pole	MAT1A
bankssts		entorhinal cortex	EPHX1	posterior cingulate	CYP2C9
fusiform gyrus		rostral middle frontal gyrus	CYP2D6	parstriangularis	SLC1A2
superior temporal gyrus		temporal pole		middle temporal gyrus	KCNH5
					CYP4F12

Table 8.3: Most relevant association found though the PLS model in each significant component. The features are ordered in terms of associated weight, positive and negative weights are highlighted in red and blue respectively.

was found instead with cingulate gyri and frontal pole. Of note, LC5 was the component in which most of genes belonging to ADNI-related pathways were found with the highest weights, namely CHAT, MAT1A, CYP2C9, SLC1A2, KCNH5 and CYP4F12 (positive weights, pathways R-HSA-211999, R-HSA-211897, R-HSA-211859, R-HSA-112316, Chr10 for the first three, Chr11, Chr14, Chr19). All of them, with the exception of KCNH5, were correlated with TF, BDNF and CEP164, inheriting their association with phenotype, as well as with cingulate gyri and frontal pole.

Transcriptomic analysis

The transcriptomic analysis revealed that part of the relevant genes discussed in the previous Sections was also expressed in brain tissues. Figure 8.7, 8.8 and 8.9 represent the normalized expression profiles for the top 5 genes, hetionet genes and genes belonging to AD pathways in each significant component, respectively. More in detail, among the top 5 genes in LC1, RIF1, ATP6V1G2 and NFASC were generally expressed in brain. The NFASC and RIF1 showed also a peculiar expression pattern, the former featuring markedly higher levels in frontal lobe, cingulus and globus pallidus while the latter showing higher expression in the hippocampus. All the top 5 genes of LC2 were expressed in brain, with PHF14 and HNMT showing the highest values overall the brain tissues. Of note, the MORN1 was found to be highly expressed in cerebellar cortex compared to the other regions. Finally, BCAS1 and CEP164, prominent in LC5, showed a

Chapter8/Figures/top5_brain_expr.png

Fig. 8.7: Gene expression profiles for the top five genes in each significant PLS LC. Expression values are normalized in the range [0,10] and grouped in the same regions as T1-w parcellation for the available regions.

Chapter8/Figures/Hetionet_brain_expr.png

Fig. 8.8: Gene expression profiles for SKAT genes belonging to Hetionet. Expression values are normalized in the range [0,10] and grouped in the same regions as T1-w parcellation for the available regions.

Chapter8/Figures/AD_pathways_brain_expr.png

Fig. 8.9: Gene expression profiles for the genes belonging to the four AD pathways in each significant PLS LC. Epression values are normalized in the range [0,10] and grouped in the same regions as T1-w parcellation for the available regions.

brain-wide expression, while COL6A3 was found highly expressed especially in superior and middle temporal gyri. Among the genes involved in the AD pathways and showing an associated PLS weight above the threshold of the $75^{t}h$ percentile, high expression levels were shown by KCNJ6, relevant in LC1, the EPHX1 and CACNA1A (LC2), and the SLC1A2, CYP4V2, KCNH5 (LC5). The latter in particular was expressed in entorhinal gyrus as well as in thalamus and middle frontal gyrus. Finally, all the Hetionet genes found above the threshold were found to be expressed in brain with the exception of CYP2D6. A particularly high expression was found for TF (LC5).

8.4 Discussion

In this work, we addressed the twofold objective of modeling the associations between brain imaging and genetic in patients on the AD continuum, highlighting the genes and the brain regions leading them. This was achieved by modeling the joint covariation between the region-based cortical and subcortical atrophy, represented by 42 features, and 408 gene variant scores, calculated for the significant genes derived from the SKAT SNP set approach which allowed to exploit gene-based genetic information while reducing the number of genetics features considered. Compared with previous approaches, this study hence proposes a method to summarize the genetic information, overcoming the limitations of GWAS analysis. This is achieved by exploiting SKAT

as SNP set approach (considering only SNPs located in its exon regions) and then projecting back the results to a subject level by computing a subject and gene specific variant score representing how varied is the gene for the specific subject compared to a reference genome. In this way, gene variant score allows to characterize each significant gene highlighted by SKAT with a single value which can be then used along with imaging variables in multivariate IG models. To the best of our knowledge, the interaction between gene variant scores and a complete set of brain structural imaging phenotypes has not been yet deeply investigated in a cognitive impaired cohort, though can convey more meaningful information compared to considering each single SNP or summary risk scores at a time. Moreover we focused on ADNI-3 to investigate the potentialities of this dataset, as this is still under-investigated when considering IG associations mainly due to sample size limitations inherent to the available genetic data for this study cohort. Indeed we considered 297 individuals, divided into healthy CN and PAT (either MCI or AD) and further split into discovery and validation cohorts (80% and 20%, respectively). We finally proposed few validation techniques as a transcriptomic analysis on the obtained candidate genes and a preliminary statistical analysis on the input feature distribution to better interpret and validate the obtained results.

Summary of main findings

In terms of imaging and genetic variables, while the phenotype features are here represented by well-known morphometric measures for regions that have been proven to be involved at different levels in the neurodegeneration process typical of the AD continuum [151, 152], the significant genes resulting from SKAT method revealed twelve genes belonging to the Hetionet database. Hence, this method, though applied in a somehow limited cohort where conventional GWAS failed, was able to retrieve well known genes known for their association with AD. Moreover, four out of the significant pathways obtained through the enrichment analysis on the significant genes were as well associated with AD in Hetionet. The joint multivariate modeling between imaging and genetics relied on the PLS, an explainable model which allowed to derive significant genotype-phenotype associations as verified by permutation testing and returning LCs in which a clear and significant difference between the PAT and CN projections scores was recorded. In the LCs encompassing the most significant differences between PAT and CN, the relevant genotype-phenotype associations can be summarised as follows: (i) The correlation between the EPHX1 variant score (*Biological Oxidation* pathway), whose role in neurodegeneration is highly investigated and strongly supported by previous findings, and a decrease in subcortical volumes, typical of neurodegeneration. This result was also confirmed by the expression analysis which highlighted the EPHX1 to be widely expressed in brain; (ii) The correlation between the BCAS1 variant score

and a significant decrease in temporal lobe thickness (PAT < CN). This gene is indeed involved in the process of myelination, particularly investigated in the dentate gyrus, part of the temporal lobe; (iii) Multiple associations, for which further exploration is needed, between the decrease in cortical thickness or volume of well known brain regions involved in AD continuum with genes whose function is still questioned, though preliminary related to neurodegeneration and which will be further detailed below.

IDPs and genes preliminary analyses

Our investigation started from a preliminary analysis of the input features, where we tested whether significant between-group differences were present, considering each feature separately from the others and relying on Mann Whitney non-parametric U-test. On the phenotype, the test highlighted differences in thickness or volumes for well established brain regions affected by AD. It is a matter of fact that, while regions along the hippocampal pathway were found to be affected by atrophy in the early stages of the disease, temporal, parietal and frontal neocortices emerge at later stages [151, 153]. Moreover a very recent systematic review on prospective biomarkers of AD [217] performed meta-analyses based on random-effect models on 84 articles, concluding that 20 biomarkers were globally associated with AD progression. Among them, hippocampal volume, entorhinal cortex volume and middle temporal lobe volume resulted as promising prospective sMRI biomarkers for AD progression. All the aforementioned regions resulted as significantly different between PAT and CN also in our cohort, with reduced volumes in patients as expected and in line with the neurodegeneration atrophy pattern. Moreover, interestingly, such regions were also among the most relevant in the association with genetics computed through our PLS model. On the genetic side, no genes survived the FDR correction, probably due to the high number of comparisons to be considered (408). However, when no corrections were applied, 60 genes resulted significantly different between PAT and CN. Four of them, namely RIF1, PHF14, KCNH5 and HNMT were then found among the most relevant ones in the PLS LCs, hence holding an important role in the latent space definition. Of note, differences in both directions (PAT < CN and PAT > CN) were found for such gene variant scores, suggesting that variants in some genes could lead to increased resistance to the disease however they could also represent biases in considering all SNPs in the genes.

PLS model significance, validation and generalizability

Aiming at analysing the multivariate association between the complete set of genetic and imaging features typical of IG studies, latent view methods such as CCA or PLS have gained increased popularity. An extensive review on the models applied to this aim can be found in [13, 188]. Focusing on PLS, which is a tried-and-true technique

for multivariate analysis, this method has been used with promising results to establish a connection between brain atrophy and individual SNPs from AD patients in a recent paper by Lorenzi and colleagues [42], revealing a strong link between the TRIB3 gene and the characteristic pattern of grey matter loss in such disease. They used the entire set of SNPs for the genotype and sMRI characteristics as IDPs, demonstrating the generalizability of their model in a separate cohort. The same strategy was used by Casamitjiana and colleagues [161], who were able to stratify the early stages of AD in the PLS latent space by utilizing T1-w features and the amounts of the biomarkers t-tau, p-tau, and amyloid-beta in the cerebrospinal fluid. Finally, in a previous preliminary work this approach allowed us to uncover significant associations between brain atrophy and 14 ADNI-related PRSs, possibly revealing different association for different AD subtypes [164]. Interestingly, thanks to its advantages in scalability and its ability to facing collinearity, PLS is starting to be applied also in the imaging transcriptomics field [159], opening new opportunities to investigate how the spatial patterns of gene expression relate to anatomical variations in brain structure and function in both health and disease. When applying PLS model, a solution maximising the covariance between latent space projection is always found. The validation of the obtained results hence becomes stringent. The strategy that we followed to deal with the low number of subjects in our cohort was firstly the analysis of the singular values defining the LCs, for which the permutation test was implemented. The rows of the X matrix, representing the phenotype, were randomly permuted in order to break any existing connection between the IDPs and the gene variant scores. The singular values obtained from the permuted inputs were then compared with the true ones and it was hence possible to assess that a significant distance was recorded between the random singular values distribution and the true ones. This confirmed the importance of the genotype-phenotype association described by the LCs associated with such singular values. Secondly, the observation of the latent space, where, separately for each LCs, the IDPs projection was plotted against the gene variant score one, allowed to assess whether the solution provided by the PLS effectively found a covariance between features. Finally, the generalization of the model was investigated through the projection on the obtained latent space of set of subjects set aside from the full cohort.

Role of the input feature definition

Besides the study of the significance and generalizability, one main limitation of these approaches is that, in order to handle a large number of input features, which is always the case when considering the full set of SNPs or the total number of voxels for imaging, a large number of observations is needed, hence they are poorly applicable when dealing with small cohorts. To overcome this limitation, on the imaging side, features

computed over brain regions, rather than single voxels, were applied to check the association with genetics. These metrics for on sMRI data, could represent Grey Matter (GM) volumes for a set of regions of interests [218], or local gray matter density extracted through voxel based morphometry and then grouped by target regions [219]. A complete overview on the commonly used IDPs in IG studies can be found in [13]. Region based volumes and thicknesses were indeed considered in the present study. On the genetic side, features based on PRS [220, 221] or Polygenic Hazard Score (PHS) [222] have been proposed, rather than using a series of individual SNPs. These are based on the presence/absence of significant individual SNPs and allow to collapse all the genetic information into a single score per subject. PRS is a statistical index to estimate a subject's genetic liability to a trait or disease involving the most significant SNPs according to previous analyses, typically GWAS. Moreover, approaches based on genetic feature reduction have started to be investigated in IG studies. In particular, by relying on a mass univariate approach, Hibar and colleagues [223], employed PCA to summarize the SNPs for each gene and associated it with each brain voxel in T1-w MRI data from the ADNI first phase. While no associations survived multiple comparison adjustment, several genes known for their association with AD or brain functions were identified before correction. To further overcome the issues arising when the spatial information in images or the effect of multiple genetic variants are not taken into account in the models, [224] developed a novel method based on the random field theory and multi-locus least square kernel machines to evaluate the joint effect of multiple SNPs within each gene on more than 30,000 brain voxels. The authors applied this approach to the same ADNI cohort, demonstrating this was more sensitive compared with voxel-wise single-locus approaches and identifying a number of genes as having significant associations with volumetric changes, among which GRIN2B had a prominent role. Along the same line, Le Floch et al. [202] demonstrated the importance of a pre-filtering step on individual SNPs before any multivariate analysis which can improve performance for both PLS and CCA-based methods. In [225], Wang et al. proposed a sparse multivariate multiple regression model, where SNPs were grouped by genes and the estimation of the regression coefficients was based on penalized least squares and grouping structure. More recently, Greenlaw et al. [226] extended this approach by proposing a Bayesian group sparse regression which takes into account the sparsity at the gene level. The above methods were applied on the first ADNI cohorts (sMRI data) and using a preselection of around 40 genes (and related SNPs) associated with AD in the literature.

An additional important approach in this framework is represented by the SNP set analyses which allow to improve the detection power w.r.t. individual SNP analysis, combining the effects of multiple variants together and identifying multi-locus mechanisms for complex disease. SNP sets are defined by LD blocks, genes, pathways or other

criteria, which may offer biological insights for interpreting results. Different strategies have been proposed in this respect. Some methods select individual SNPs from different genomic regions associated with a given disease from literature [227] or resulting from independent analyses [220, 221]. Other methods select a single gene or SNP set based on a priori knowledge and examine the joint effects of multiple SNPs within this gene or set [228]. Finally, some methods exploit data-driven strategies to identify multiple SNP sets from the entire genome [197, 195, 196].

In IG, SNP set methods are used to select all SNPs belonging to significant SNP sets and then create regression models for associations with IDPs using these individual SNPs as a genetic feature. In this way, although the SNPs are selected according to a SNP set approach, the association with IDPs is made at individual SNP level. To overcome this limitation, global scores for each significant SNP set can be used as genetic features in regression models to probe the association with IDPs at SNP set level. In this work we firstly extracted genes from the selected ADNI-3 cohort using SKAT, then we introduced a gene variant score that gives for each subject and each SKAT significant gene a measure of the extent to which all the SNPs in a given gene are mutated. Such gene variant score allows to reduce the number of variables in the multivariate model switching to the gene-wise level approach which, starting from a set of SKAT genes, takes into account the absence or presence of all SNPs in the respective gene. The gene variant score computed on the 408 significant SKAT genes were considered as genetic input for the PLS model.

Significant latent components

More in detail of the proposed PLS model, results highlighted that three PLS components, namely LC1, LC2 and LC5, defined a latent space encompassing a significant separation between PAT and CN for both genotype and phenotype. Such significance was confirmed by the projection on the obtained latent space of the unseen validation cohort, holding the same class distribution of the discovery set. In fact, the separation between PAT and CN remained significant also for the validation set on either the phenotype or the genotype, the latter being particularly evident ($p < 1e^{-02}$) in the LC2. The global model significance was finally confirmed through the permutation test attaining a p-value of 0.001. The great advantage of PLS model relies on its straightforward explainability. In fact, by analysing the fitted feature weights it allows to retrieve the features driving the association between imaging and genetic features, separately for each component. A twofold result can hence be extracted: i) The intra-genotype and intra-phenotype relationships, that is observing how the different features belonging to the same data source are related to each other; ii) The analysis of the association between genotype and phenotype, highlighting those features that have a greater influ-

ence on the latent space derivation. Framing these concepts on our model allowed to explore the association between brain morphometric measures, and the variant score of genes selected from SKAT model. Multiple association patterns existed between the input features in the latent space, however particular attention will be given to the anticorrelations between IDPs and gene variant scores, which were indeed predominant in the LCs. The statistical analysis on the input features was instrumental to further elucidate the link found between genotype and phenotype.

Relevant IDPs-gene interactions in the first latent component

Analysing each component, in the LC1 the major role was played by the anticorrelation between the cortical thickness features (in particular supramarginal gyrus, middle/inferior/superior temporal gyri and rostral middle, superior frontal gyri) and the variant scores associated to genes RIF1, ATP6V1G2, NFASC, FBXO43, KCNA7 and KCNJ6. RIF1 resulted as significantly higher in PAT compared to CN, and this trend is extended to all its correlated genes in this LC. Interestingly, RIF1, ATP6V1G2, NFASC and KCNJ6 were generally expressed in brain, as a result of our transcriptomic analysis, with NFASC being particularly expressed in frontal lobe among the other regions and KCNJ6 in the hippocampus. NFASC gene is highly investigated in association with AD, transcripts were shown to be involved in synapse formation and stabilization, and were found as elevated in the subjects with MCI converting to AD compared with stable MCI as well as significantly correlated with p-tau [229]. Moreover, Duits *et al.* concluded that, together with other peptides, NFASC transcripts could have a role in early events in the AD pathophysiological cascade. The other mentioned genes appeared also to have a role in neurodegeneration, even if their involvement in brain is still under investigation. Of interest, the RIF1 was found to protect telomeres and chromosome breaks, which is in turn a process involved in brain ageing [230]. Concerning ATP6V1G2, its main role appeared to be related to neurons energy metabolism, in particular lysosome acidification. Noori et al. [231], found ATP6V1G2 was among the genes being downregulated in neurodegenerative diseases. This downregulation may result in short ATP supply in neurons due to the failure of energy metabolism, which is however highly needed for protein clearance mechanisms. Finally, KCNA7 and KCNJ6 are part of the *Neuronal System* pathway (R-HSA-112316) and belong to the voltage-gated potassium channel gene family. In particular, KCNJ6 is associated with Down's syndrome [232], which has a well-established increased risk for AD [233]. No direct interaction between KCNA7 and AD was found, however potassium channel are becoming a target for the treatment of neurological disorders and autoimmune diseases [234].

Relevant IDPs-gene interactions in the second latent component

Moving to LC2, this showed the most significant differences between the latent projections of PAT and CN subjects. It was mainly defined by an anticorrelation between subcortical volumes, among which hippocampus, putamen end pallidum had the most relevant weights, and HNMT, CACNA1A, FMO2 and EPHX1 gene variant scores. Of interest, HNMT was also among the genes showing a significant increase in the variant score for PAT compared to CN (Mann Whitney U-test, uncorrected). In literature, it was found to be correlated with intellectual disability [235] in a cohort of patients affected by non-syndromic autosomal recessive intellectual disability. CACNA1A, FMO2 and EPHX1 instead belonged to the four ADNI- related pathways highlighted in Section 8.3.1. More in detail, CACNA1A (*Neuronal System* pathway, R-HSA-112316) was demonstrated to be linked with familial AD in a cohort of patients presenting cerebellar damage with amyloid plaques [236]. EPHX1 (*Biological Oxidation* pathway, R-HSA-211859)has been highly investigated in literature so far. Transcripts have been detected in various areas of the brain such as cerebellum, frontal, occipital, pons, red nucleus, and substantia nigra regions. Indeed, EPHX1 was recorded as highly expressed brain-wide by our transcriptomic analysis and its role in pathogenesis of neurodegeneration was supported by previous findings demonstrating a differential expression in patients with AD [237]. This finding well relates with our results according to which the EPHX1 variant score was anticorrelated subcortical volumes, which also showed a significant dicrease in PAT compared to CN, while it was correlated with cerebellar thickness modulations. Finally, FMO2 belongs to the *Biological Oxidation* pathway (R-HSA-211859), however its direct role in neurodegeneration has not been yet demonstrated.

Relevant IDPs-gene interactions in the fifth latent component

Finally, LC5 was mainly defined by the anticorrelation of BCAS1, GLT6D1, TMPRSS15, COL6A3 and KCNH5 with enthorinal cortex, fusiform gyrus and temporal pole thicknesses. Of note, KCNH5 showed a significant increase in its variant score for PAT compared to CN (Mann Whitney U-test, uncorrected), hence strengthtening the assumption that the increase in such gene variant score, as well as the one of its correlated genes in LC5, is linked to a decrease in the thickness values of the mentioned brain regions. More in detail of the genes, BCAS1 is involved in the process of myelination. In fact, an explorative proteomic study of the dentate terminal zone showed that, in that region, its transcripts were among the top 10 decreased proteins showing the largest changes in AD [238]. Of interest, the dentate gyrus is part of the temporal lobe, which is indeed among the most important regions for this LC. Moreover, as a result of our transcriptomic analysis, BCAS1 also showed a brain-wide high expression. GLT6D1 was

found to be associated with periodontitis. Despite its apparent distance from AD, recent experimental studies indicated that a periodontitis-causing bacterium might be a causal factor for AD since it was identified in the brain of AD patients, while in mice it provoked brain colonization and increased production of amyloid-β [239]. However, a recent bidirectional Mendelian randomization study to examine the potential causal relationship between chronic periodontitis and AD did not result in significant evidence [240]. Further studies are hence needed to deeply investigate such association. Concerning TMPRSS15, in some early-onset patients with AD induced by APP duplication (due to down syndrome), the duplicated region also contains TMPRSS15, which is hypothesized to participate in neurogenesis and/or APP metabolism [241], as detailed for gene KCNJ6, relevant in LC1. Interestingly, COL6A3 was found expressed in particular in superior temporal gyrus, which is among the significant brain regions whose thickness was decreased in PAT compared to CN. In literature, this gene has been associated with the Collagen VI protein whose lack was demonstrated to have a role in neurodegeneration [242]. Moreover, variants in this gene were found in patients affected by recessive isolated dystonia, a human brain disorder [243]. Finally, KCNH5 (*Neuronal System* pathway, R-HSA-112316) was found to be particularly expressed in entorhinal cortex, which in turn is among the most relevant regions in LC5. This gene encodes a member of voltage-gated potassium channels. Members of this family have diverse functions, including regulating neurotransmitter. It also appeared to have a role in neurodegeneration [244] even if an extensive analysis is not yet present in literature.

Overall, through the PLS model we obtained a latent representation of the input features dominated by significant genotype-phenotype associations. Most of the variant scores associated to the genes standing in the highest positions in the LCs weights were correlated with what is well known for the phenotype in AD and moreover they were found to be related to neurodegeneration as well as expressed in brain. The most relevant findings were the correlation between the EPHX1 variant score and a decrease in subcortical volumes and the correlation between the BCAS1 variant score and a significant decrease in temporal lobe thickness, as discussed in the previous paragraphs. Besides these well assessed associations, with the PLS model we retrieved multiple additional genotype-phenotype associations which are still underinvestigated in literature. Among the others the NFASC and the ATP6V1G2, were among the most relevant gene variant scores for the LC1. They are highly studied in AD but still not related with brain modulations. Our model, instead, found a significant association between the increase of the associated variant score and a decrease in temporal and frontal gyri cortical thicknesses which deserves further analysis. Moreover, in the LC5 we found multiple genes, namely GLT6D1, TMPRSS15 and COL6A3, whose involvement in AD has still not been fully proven but which resulted as significantly associated with decreased morphome-

tric values in well known AD affected regions as the entorhinal cortex, fusiform gyrus and temporal lobe. Therefore, in summary, the PLS model allowed on one side to retrieve well assessed genotype-phenotype association for which their role in the disease was already established in the current literature, and on the other side to unveil highly relevant associations between still not AD-related genes and decreased morphometric values in brain regions with a prominent role in AD, opening the way to further exploration directions.

Limitations and future directions

We have to acknowledge some limitations in the current study. First of all we recognise the small sample size of our cohort, especially concerning patients data. This was due to the still limited number of subjects available in ADNI-3 phase (ongoing). However, this did not impact on the significance of the results thanks to the adoption of the gene variant scores, even if, at the same time did not allow to use different validation techniques such as bootstrap analysis. Meanwhile ADNI-3 cohort includes the most complete set of imaging acquisition in ADNI database hence it will allow the inclusion of different IDPs providing different views on the brain modulations. This approach would be of high interest being the AD an intrinsically multiview disease, however the present study based on T1-w MRI could be considered as benchmark and strting point for future analysis. Diffusion MRI and functional MRI derived IDPs, such as tract based measures or connectivity features could be included in order to investigate how both the microstructure and function are affected by AD continuum and contemporary associated with gene variants. The gene variant score introduced in this work is computed on all SNPs located in the same gene with each SNP being equally weighted. It could be interesting to modify the gene variant score in order to weigh the SNPs differently based, for example, on the associations of individual SNPs with the disease or imaging phenotype (i.e. p-value from GWAS). In this direction more sophisticated, still explainable or interpretable models could be introduced in order to account for the multi-channel information which can't be successfully addressed through the classical definition of PLS model which allows only the inclusion of two channels, while keeping a clear interpretation of the results, in the input matrices X and Y.

8.5 Conclusions

The presented PLS model confirms that there exists a joint variation between grey matter atrophy and gene variant scores in AD. Such associations described a latent representation of the input features dominated by significant genotype-phenotype associations validated also through the transcriptomic analysis. This work hence proposed

new ways to investigate the genotype phenotype interactions in a restricted study cohort highlighting that simple yet explainable models can still allow uncovering associations that are descriptive of the underlying mechanisms of neurodegeneration in AD continuum.

The work presented in this Chapter is in preparation for journal submission [245].

8.6 Supplementary Figures

Chapter8/Figures/genes_table.png

Fig. 8.10: List of SKAT genes names organized by Chromosome and relative starting position.

9

An interpretability framework for a multi-channel variational autoencoder

In this work we proposed an interpretability framework for the Multi-channel Variational Autoencoder (MCVAE) for analyzing the genetic underpinnings of Grey Matter (GM) and White Matter (WM) modulations in the Alzheimer's Disease (AD) continuum.Three Channels were considered as the input. Cortical thicknesses and subcortical volumes derived from T1-weighted (T1-w) MRI represented the structural Magnetic Resonance Imaging (sMRI) channel, WM features derived from diffusion Magnetic Resonance Imaging (dMRI) and extracted through a tract-based spatial statistics analysis on four diffusion tensor based indices represented instead the dMRI channel while the gene variant score calculated for all the significant genes resulting from Sequence Kernel Association Test (SKAT) filtering as presented in Chapter 8 was considered for the genetic channel. We put forth a metric to compare two alternative MCVAEs that account for a different amount of Latent Variable (LV)s. Finally, we suggested a modification of the MCVAE to apply post-hoc interpretability approaches. Our results showed a common latent space well aligned across different channels and displayed a clustering across the various stages of AD despite the poor reconstruction ability. We were able to retrieve the most pertinent features for the decoding of the various channels using the SHapley Additive exPlanations (SHAP) application were extremely comparable for each distinct decoded feature.

This Chapter provides new approaches for using eXplainable Artificial Intelligence (XAI) methods in a generative framework to examine genotype-phenotype interactions, highlighting connections between GM, WM and genetics that are descriptive of the neurodegeneration mechanisms in the AD continuum.

9.1 Introduction

There is still much debate about the pathophysiology underpinning brain modulations in AD. For example, while sMRI derived Imaging Derived Phenotype (IDP)s are well assessed biomarkers for AD detection [151] it is not clear whether white matter alterations are related to, or independent of, gray matter degeneration in AD. For this reason, the role of microstructure was recently being considered in AD studies [169, 14]. Hence, the joint analysis of biomedical data is fundamental for a deep understanding of the relationship between biomarkers.

Moreover, the link between brain IDPs and genetics is an hot topic to such an extent that an entire research field, IG, focuses on this aim as already shown in Chapters 6, 7 and 8. On the genetic side, numerous methodologies have been employed to extract genotype traits linked with AD, with the majority of studies still concentrating on the detection of Single Nucleotide Polymorphism (SNP)s. However, recent strategies moved forward grouping SNPs into SNP sets [195, 196, 197]. A natural grouping technique consists of taking all SNPs that are found within a gene. The SKAT [195], a logistic kernel-machine regression model for testing the association between SNP sets and disease, is one of the most popular methods for the SNP set approach. SKAT has been utilized in several AD studies. In [198], for instance, the scientists discovered a gene (as an SNP set) around APOE that is highly connected with the condition.

For the genotype-phenotype association the most straightforward method is the massive univariate correlation analysis [246]. Unfortunately, the modeling capacity of this method is very limited, and it is prone to false positives when the data dimension is large. To overcome the limitations of mass-univariate analysis, more advanced latent space based methods were proposed aiming at reaching a low-dimensional space representation where desired statistical features, such as maximum correlation (Canonical Correlation Analysis (CCA)) or maximum covariance (Partial Least Squares (PLS)) are enforced. However, because they are not generative, these approaches provide limited information regarding how this latent representation is reflected in the data as they do not explicitly provide a mean to sample observations when the distribution of latent variables and parameters is known.

To this aim the key work on the Variational AutoEncoder (VAE) [121], proposes a powerful generative model for high-dimensional single-modality data. VAEs are models that couple a recognition function, or encoder, to infer a lower dimensional representation of the data, with a generative function, or decoder, which transforms the latent representation back to the original observation space. The VAE is a Bayesian model where the LVs are inferred by estimating the associated posterior distributions. Inference is efficiently performed through VI which is a well-known method for computing poste-

rior distributions when conventional integrations are infeasible. Moreover, the posterior moments can be parameterized through neural networks. For this reason, VAEs are adaptable and capable of accounting for any type of data. Indeed, by combining various data sources, the combined analysis of heterogeneous channels would be also possible in this environment. However, representing concatenated multi-channel data with a VAE may raise interpretability challenges because it is difficult to isolate the contribution of a single channel from the description of the latent representation. In addition, at the time of testing, the model can often only be applied to data containing complete channel information.

To address this issue, Antelmi and colleagues [247] proposed the multichannel VAE proposed which generalizes the single-channel VAE by assuming that, in a multi-channel scenario, the latent representation associated with each channel must match a shared target distribution. This was achieved by imposing a constraint on the latent representations, where each latent representation is required to correspond to a shared target prior. This technique was successfully adopted in multiple scenarios, the most outstanding was the very recent work of Diaz Pinto and colleagues in which they achieved the task of predicting myocardial infarction through retinal scans and minimal personal information [248]. In relation to AD an extension of MCVAE was recently exploited in [249] aiming to the modeling of the spatio-temporal dynamics governing the joint evolution of longitudinal imaging and clinical biomarkers along the history of the disease in order to simulate the effect of intervention time and drug dosage on the biomarkers' progression.

The main drawback of these models is the lack of interpretability due to their increased complexity. In detail, when considering a linear encoder and decoder architectures it would possible to check the relationship between each input feature by analyzing the weights associated to each LV, however, it would be impossible to analyze which specific feature contributed the most to the reconstruction or generation of another one. This step could be achieved by relying on interpretability methods, as presented in Chapter 2 and Chapter 3 and would allow shedding light on a different level of features associations.

The aim of this work is then to propose an extended interpretability framework for the MCVAE in which perturbation based methods, such as SHAP, are employed to uncover the specific relationship between features. The clinical outcome is data fusion for IG in AD continuum including genetics and different brain imaging techniques such as the diffusion MRI other than the common structural MRI as different channels in the MCVAE model.

9.2 Materials and methods

9.2.1 Study Cohort

Data used in this study were derived from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu).

Phenotype and genotype were selected in particular from the ongoing ADNI3 dataset, and were the same considered for the work presented in Chapter 8. Summary sociodemographic, clinical, and genetic information is available in Table 8.1. Participants selection was carried on April 2022 and was based on the availability of MRI and genetic data and ethnicity, all participants had European ancestry. The final cohort comprehended 297 subjects divided into 220 CN and 79 PAT, 19 of which were AD while the remaining were MCI subjects.

3D T1-w MRI and dMRI volumes were considered for IDPs extraction (3D T1-w: sagittal accelerated MPRAGE, Repetition Time (TR)/Echo Time (TE) = shortest, Inversion Time (TI)=900 ms, flip angle = $9o$, Field of View (FOV) = $256\times256\text{mm}^2$, spatial resolution = $1\times1\times1\text{mm}^3$, slices = 176-211; dMRI:(TR/TE = 56/7200, $2mm$ isotropic voxel, b=$1000s/mm^2$).

9.2.2 Phenotype and Genotype processing

The T1-w volumes were minimally preprocessed for bias-field correction (*fsl_anat* tool [204]). Subsequently, 84 anatomical Region Of Interest (ROI)s were extracted using FreeSurfer version 7.0 [205]. The average thickness and volume were considered for cortical and subcortical ROIs, respectively. The subcortical volumes were further normalized by the estimated total intracranial volume of the respective subject. The ROIs were averaged over hemispheres resulting in 42 features to be used in the subsequent analyses.

The dMRI volumes were preprocessed using FSL software (version 6.0, https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/) applying an initial step of brain extraction (`bet` tool) followed by Eddy currents correction (`eddy` tool). The data was then denoised using Local PCA via empirical thresholds relying on Python dipy library. Subsequently non linear registration to the MNI space was applied in order to correct for Echo Planar Imaging (EPI) induced distortions (ANTs toolbox [250]). The Diffusion Tensor Imaging (DTI) [84] model was fitted to the corrected images and Fractional Anisotropy (FA), Mean Diffusivity (MD) indices were extracted. The tract-based spatial statistics (TBSS) pipeline from FSL was applied to FA to derive a group WM skeleton (FA threshold of 0.2) to which all subjects were linearly registered. The same transformations were subsequently applied to all the other indices in order to obtain skeletonized values for each

subject. For all the subjects, the average value of each index was extracted from 48 ROIs derived from the JHU-DTI atlas available in FSL.

Genetic processing was conducted as described in Chapter 8, Section 8.2.3. In brief, SKAT [195] model was used to filter out genes relevant for the differentiation between AD patients at various stages and healthy subjects. Then a gene variant score was calculated for each subject and each gene in order to be associated with IDPs.

9.2.3 Multi channel Variational Autoencoder

MCVAEs are the multi channel extension of the Bayesian generative models VAEs. In what follows, a brief overview of VAEs will be firstly given, being instrumental to the subsequent MCVAEs description. Finally, details about the experimental setting of this work will be presented.

Variational Autoencoder

VAEs are bayesian generative models composed of two main parts: i) the encoder, which is used to obtain a lower dimension representation of the input data, ii) The decoder, which decodes the latent representation to obtain back data in the original, higher-dimension space [251]. VAEs are Bayesian model in the sense that they infer latent variables by estimating the associated posterior distribution.

Let $X = \{x^{(\mathrm{i})}\}^{\mathrm{d}}{}_{\mathrm{i}=1}$ be and observation set consisting of d i.i.d. samples of a continuous or discrete random variable **x**. The following generative random process is assumed for the observation set:

$$\begin{aligned} z &\sim p(z) \\ x &\sim p_\theta(x \mid z) \end{aligned} \tag{9.1}$$

where p(z) is the prior distribution of the latent and unobserved random variable and $p_\theta(x \mid z)$ is the likelihood distribution for the observations conditioned on the latent variables. It is assumed that the likelihood functions of these two families are differentiable w.r.t. θ, z. In this context, solving the inference problem enables the identification of the shared latent space that generates the observed data. The inference problem is solved by determining the posterior $p_\theta(z \mid x)$, which is not always analytically quantifiable. Variational inference can be employed as presented in [247] by introducing the distribution $q_\phi(z \mid x)$, to approximate the true posterior distribution. The distribution $q_\phi(z \mid x)$ then encodes a data point $x^{(i)}$ the latent space distribution of z, while $p_\theta(z \mid x,)$ decodes the latent representation **z** over a possible distribution **x** values.

Multi channel extension

MCVAEs are the multi channel extension of VAEs presented in [247]. Let now be $x = x_1, ..., x_C$ an observation set over C channels, which could represent the number of modalities of the data, and x_c is a $d-$dimensional vector, representing the measurements for a specific channel. In the multi channel framework we assume that $\mathbf{z}$ is the $l-$dimensional latent variable commonly shared by x_c. The following generative process hence holds for the observation set

$$\begin{aligned} &\mathbf{z} \sim p(\mathbf{z}) \\ &x_c \sim p_{\theta_c}(x_c \mid \mathbf{z}), c \in 1 \dots C \end{aligned} \tag{9.2}$$

where $p(\mathbf{z})$ is the prior distribution of the latent space and $p_{\theta_c}(x_c \mid \mathbf{z})$ is the likelihood distribution for the observations conditioned on the latent variables $\mathbf{z}$. As above for VAEs, the solution of the inference problem is given by deriving the posterior $p_{\theta_c}(\mathbf{z} \mid x_c)$. Of note, every likelihood distribution belongs to the same family $\mathbf{P}$, but has a different set of parameters θ_c. In a similar manner, as detailed for the VAE, due to the true posterior intractability, every channel's likelihood distribution is approximated by a function $(q_{\phi_c}(\mathbf{z} \mid x_\mathrm{c})$, belonging to the same family $\mathbf{Q}$, parameterized by ϕ_c.

Every channel hence, brings some information about the latent variable distribution, however providing a different approximation. In this framework, Antelmi and colleagues [247] proposed to impose a constraint enforcing each $q_{\phi_c}(\mathbf{z} \mid x_\mathrm{c})$ to be as close as possible to the target posterior distribution, in terms of the Kullback-Leibler (KL) divergence. This constraint was specified as

$$\underset{q \in \mathbf{Q}}{\arg\min} \mathbb{E}_c \big[D_{KL}(q_{\phi_c}(\mathbf{z} \mid x_\mathrm{c}) \mid\mid p_\theta(\mathbf{z} \mid x_1, ..., x_\mathrm{C})) \big] \tag{9.3}$$

where $\mathbb{E}_\mathrm{c}$ represents the average over channels computed empirically. Practically, solving the objective in the above equation allows minimizing the discrepancy between the variational approximations and the target posterior. Optimizing equation 9.3 equals to optimizing the following lower bound:

$$L(\theta, \phi, x) = \mathbb{E}_\mathrm{c} \big[L_\mathrm{c} - D_\mathrm{KL}(q_{\phi_\mathrm{c}}(\mathbf{z} \mid x_\mathrm{c}) \mid\mid p_\theta(\mathbf{z})) \big] \tag{9.4}$$

where $L_\mathrm{c} = \mathbb{E}_{q(\mathbf{z} \mid x_\mathrm{c})} \big[\sum_{i=1}^{C} \log p_{\theta_\mathrm{i}}(x_\mathrm{i} \mid \mathbf{z}) \big]$ is the expected log-likelihood of decoding each channel from the latent representation of the channel x_c only. Moreover, the L_c term imposes to each channel c to reconstruct (decode) itself alongside every other channel allowing to reconstruct a missing channel x_miss from the available ones $\{x_\mathrm{av}\}$. More details about the derivation and optimization of the lower bound can be found in [247].

Antelmi et al. [247] highlighted that using a sparse version of the mcVAE ensures the evidence lower bound generally reaches the maximum value at convergence when the number of latent dimensions coincides with the true one used to generate the data. Consequently, the sparse version of the MCVAE was taken into account in this work.

Experimental Settings

In this work a three channels MCVAE was applied considering 42 sMRI derived features representing cortical thicnkesses and subcortical volumes as sMRI channel, 54 features representing tract based average values for FA and MD as dMRI channel and 408 gene mutation scores as the genetic channel. A gaussian distribution was assumed for both q and p, hence the encoder aimed at finding the parameters ϕ_c representing the mean and variance for each q_c. A single layer architecture (linear) was employed after multiple empirical tests by varying the number of layers and introducing non linarities, as well as dropout layers. The single layer linear architecture for the encoder and the decoder was also the best performing in the experimental tests on a similar setting as the one we propose, in [247]. Two models were compared in order to include a different number of LVs. 20 LVs were firstly selected for the sparse MCVAE, after the dropout probability analysis 3 LVs were then chosen for the second model. We will refer to such models as 3LVs-sparse and 20LVs-sparse. For each model, validation was performed following a 5-fold Cross Validation (CV) strategy. The 297 subjects used for training/validation were randomly split in five groups, keeping the same class ratio in each fold, resulting in folds of 60 subjects each (except one consisting of 59 subjects). The experiments were repeated five times and, for each run, four folds were used for training and the remaining one for validation. Performance was assessed through the computation of Mean Square Error (MSE) and R^2 averaged over each decoded feature separately for each channel. Moreover, for the best reconstructed channel, the feature specific MSE was also derived.

Comparisons across models were performed by analyzing the encoding matrix of each fold for each MCVAE. This is particularly interesting for sparse models since each LV has an associated dropout probability, and through this metric it is possible to confirm whether the LVs holding the least dropout probability in the 20LV-sparse are similar to the ones obtained with the 3LV-sparse In fact, the last layer for each encoding is a matrix in which each row vector represents the encoding weight defining each LV. A measure of similarity between two vectors representing LVs across different models can be hence calculated by computing the cosine similarity as

$$cos\theta = \frac{|\vec{r}_i| \cdot |\vec{r}_j|}{\|\vec{r}_i\| \left\|\vec{r}_j\right\|} \tag{9.5}$$

where r_i and r_j represent the i-th and j-th encoding row of two different models. The resulting similarity ranges from -1 meaning exactly opposite, to 1 meaning exactly the same, with 0 indicating orthogonality or decorrelation, while in-between values indicate intermediate similarity or dissimilarity.

For the subsequent analysis the best split in terms of both reconstruction performance and cosine similarity was considered.

9.2.4 Interpretability analyses

Among all the available XAI methods, in order to extract the importance of a set of tabular features, the SHAP model [252] was chosen. SHAP is starting to be widely applied being a model-agnostic explanation method, hence applicable to any kind of Deep Learning (DL) model. It belongs to the class of additive feature attribution methods which builds on the game theory concept of Shapley values. In a generative task, it assigns a quantitative value to each feature depending on its contribution to a specific feature reconstruction. The SHAP method represents Shapley values as a linear model of feature coalitions. It requires retraining the model on all feature subsets $S \subseteq F$, where F is the set of all input features. It assigns an importance value to each input feature that represents the effect on the selected output feature reconstruction given by the inclusion of that input feature. To do so, a model $f_{S\cup\{i\}}$ is trained with that feature present, and another model f_S is trained with the feature withheld. Reconstructions from the two models are compared on the current input, and this is performed for all possible subsets. Shapley values are then computed as the weighted average of all possible differences and used as feature attributions. SHAP has a solid theoretical foundation, making this approach quite robust, and it allows to derive contrastive explanations comparing the individual prediction vs the average one. Moreover, SHAP values present properties of local accuracy, missingness, and consistency, which are not simultaneously found in other methods. An important drawback of SHAP is that it provides additive contributions of the different variables. However, if the model is not additive, the Shapley values might be misleading.

The SHAP method was applied to the best model obtained during CV in order to uncover which input features contributed the most to the generation of the best reconstructed output features. To this aim a subclass of VAE was created, with the only difference being the forward method returning only the mean value of the decoded distribution, instead of a dictionary of values, as in the original model. In this way, SHAP values are computed for an input feature vector (such as genetic), according to the contribution of each component in the reconstruction of a desired output feature vector (such as sMRI), with the decoded value being the mean of the decoded distribution.

3LV-sparse

	sMRI			dMRI			Genetics	
from	MSE	R^2	from	MSE	R^2	from	MSE	R^2
dMRI	0.923(0.189)	-0.002(0.106)	sMRI	0.812(0.246)	-0.032(0.119)	dMRI	1.009(0.154)	-0.028(0.044)
Genetics	1.182(0.229)	-0.299(0.186)	Genetics	1.253(0.289)	-0.655(0.303)	sMRI	1.008(0.154)	-0.027(0.044)
Genetics+dMRI	0.955(0.191)	-0.036(0.086)	Genetics+sMRI	0.882(0.252)	-0.129(0.12)	sMRI+dMRI	1.006(0.154)	-0.026(0.042)

20LV-sparse

	sMRI			dMRI			Genetics	
from	MSE	R^2	from	MSE	R^2	from	MSE	R^2
dMRI	0.886(0.182)	0.039(0.087)	sMRI	0.785(0.243)	0.007(0.102)	dMRI	1.009(0.154)	-0.028(0.043)
Genetics	1.026(0.206)	-0.117(0.119)	Genetics	1.112(0.28)	-0.457(0.242)	sMRI	1.008(0.153)	-0.027(0.043)
Genetics+dMRI	0.902(0.188)	0.024(0.075)	Genetics+sMRI	0.843(0.251)	-0.074(0.105)	sMRI+dMRI	1.006(0.153)	-0.026(0.041)

Table 9.1: Reconstruction performance of the 3LV-sparse and 20LV-sparse models (rows) in terms of MSE and R^2 [mean (SD)]

9.3 Results

The performance for the models 3LV-sparse and 20LV-sparse averaged over the 5 folds of CV procedure, and over the features of each channel, are reported in Table 9.1. The performance is expressed in terms of MSE and R^2 of the reconstruction and its standard deviation. The reconstruction of each channel is presented considering the other two channels both separately and together as input. In order to better understand the obtained results in terms of MSE, the feature range for the standardized input feature was [-5.355, 4.847] for sMRI, [-5.611, 4.237] for dMRI, and [-2.989, 4.658] for genetics. There is not an evident difference between the two models 3LV-sparse and 20LV-sparse, indeed the reconstruction performance is very similar with imaging channels being better reconstructed one from the other than from genetics. The MSE of the reconstruction of sMRI from dMRI was 0.923 ± 0.182 and 0.886 ± 0.87 respectively for the 3LV-sparse and the 20LV-sparse. Adding the genetics to the reconstruction input sloped the reconstruction MSE to 0.955 ± 0.191 for 3LV-sparse and 0.902 ± 0.188 for the 20LV-sparse. Considering solely the genetic as input yields higher MSE values. Moving to the reconstruction of dMRI features, considering as input the sMRI allows to obtain an average MSE of 0.812 ± 0.246 and 0.785 ± 0.243 respectively for the 3LV-sparse and 20LV-sparse. Also in this case, adding the genetics as reconstruction input resulted in an increase of the MSE with the 3LV-sparse reaching a value of 0.822 ± 252 and the 20LV-sparse of 0.843 ± 0.251. As it was for the sMRI feature reconstruction, also for the dMRI, reconstructing only from genetic features yielded the worst performance. Finally, the reconstruction of genetic features was not successfully achieved either considering sMRI or dMRI as input as it is possible to verify in Table 9.1.

Chapter9/Figures/LV_comparison.png

Fig. 9.1: Dropout probabilities for each LV of the 3LV-sparse and 20LV-sparse models (rows). Colors indicate the mapping between 3LV-sparse LVs and 20LV-sparse extracted through cosine similarity.

9.3.1 Model comparison

We recall that this and the following analyses were performed considering the best model obtained through the 5-fold CV process. During the training phase of sparse models, a dropout probability is computed and associated with each LV. This information is essential in defining the optimal number of LVs, which are deemed as necessary when their dropout probability is lower than an empirical threshold. This threshold was set as 0.5, after a few empirical tests as suggested also in [247]. Of interest, as shown in Figure 9.1 the model 20LV-sparse has one LVs under the threshold, namely the LV 11 which is mapped to the LV 1 of the model 3LV-sparse. More in detail, the cosine distance between those LVs was equal to -0.79. The other two LVs of the model 3LV-sparse did not map to any of the LVs of the 20LV sparse, however considering an MCVAE with only one LV would have led to poorer feature reconstruction performance, hence the 3LVs were chosen for the reference MCVAE. This result was consistent among the five different splits of the 5folds-CV procedure.

9.3.2 Latent space

Figure 9.2 shows the latent space generated by the first and the second LV separately for each input channel (columns) for both the training and the validation set (rows). The LVs were chosen based on the associated dropout probability, choosing the most important ones. In general, the three latent spaces were well aligned, with the dMRI and the genetic ones being more similar compared with the sMRI one. For the training

Chapter9/Figures/Latent_space.png

Fig. 9.2: Latent representation of the training and validation sets (rows) considering the two most significant LVs, separately for each input channel, namely dMRI, sMRI and genetics (columns, AD: blue, CN: orange, MCI: green)

set, a clear class clustering was present and coherent among the input channels, with the AD subjects being located in the right-down part, and the MCI being located in the center and overlapped with CN mostly located in the up-left region. This clustering was not clearly present for the testing set.

9.3.3 Best reconstructed features

Figure 9.3 shows the scatter plots of the ten best decoded sMRI features(above) and dMRI features (below) for the validation set, starting separately from the genetic and dMRI channels for sMRI reconstruction and from genetics and sMRI for dMRI features. The reconstructed features are plotted against the respective ground truth. The acronym tables for dMRI and sMRI features can be found in Section 9.6.

Table 9.2 quantifies the reconstruction performance of such features.

From the scattering plots it is evident a better reconstruction of dMRI features, compared with sMRI ones, that was confirmed by the respective MSE values. In particular, focusing on dMRI reconstruction, the best reconstructed features were the MD-derived either considering genetics or sMRI as input. Considering the decoding performance of genetic features, the best reconstructed dMRI metric was the pontine crossing tract (PCT) with an MSE of 0.594, followed by some tracts of the internal and external capsule (PLIC, ALIC, EC, MSE = 0.617, 0.686, 0.752), the corona radiata (PCR, SCR; MSE

Chapter9/Figures/Recosntruction.png

Fig. 9.3: Scatter plots of the best reconstructed sMRI and dMRI features (above and below, respectively) decoded from the other two respective channels.

= 0.817, 0.741), the corticospinal tract (CST; MSE = 0.643), the Superior Longitudinal fasciculus (SLF; MSE = 0.784) and the cerebral peduncle (CP; MSE = 0.841). Some of these features were also among the best ones decoded from sMRI achieving generally lower MSE values. In fact, decoding from sMRI, the tracts part of the external and internal capsule were among the best reconstructed with MSE values of 0.260, 0.322, 0.389 and 0.395 for RLIC, PLIC, ALIC and EC. The SCR reached an MSE of 0.296, much lower compared to the 0.741 reached by decoding from genetics. The other well dMRI decoded features starting from sMRI were the superior longitudinal fasciculus (SLF; MSE=0.313), the sagittal stratum (SS; MSE = 0.344), the fornix (FXST; MSE=0.365), the PCT with MSE=0.393 and the superior fronto-occipital tract (SFO; MSE = 0.398).

Concerning the reconstruction of sMRI features, better reconstruction performance was achieved decoding from dMRI compared to genetics. Starting from the best reconstructed features decoded from dMRI, the Accumbens (AC) was the feature with the lowest MSE of 0.790 followed by the bankssts (BSTS) with an MSE equal to 0.826, and the posterior cingulate gyrus (PCG), MSE = 0.844. All the other top ten best reconstructed features had an associated MSE higher than 0.9. Considering the genetic as decoding input, all the sMRI reconstructed features had an MSE greater than 0.9, with the AC being the best reconstructed feature with an MSE of 0.926. The other features were mainly

sMRI feature reconstruction

from Genetics	MSE	from dMRI	MSE
AC	0.926	AC	0.790
PCG	1.056	BSTS	0.826
CU	1.059	PCG	0.844
PaCG	1.085	PCU	0.923
RACG	1.099	FG	0.943
PCU	1.122	PoCG	0.953
ICG	1.144	STG	0.971
PA	1.203	CU	0.987
TTG	1.210	PCAL	0.988
FP	1.248	PaCG	1.001

dMRI feature reconstruction

from Genetics	MSE	from sMRI	MSE
PCT_MD	0.594	RLIC_MD	0.260
PLIC_MD	0.617	SCR_MD	0.296
CST_MD	0.643	SLF_MD	0.313
ALIC_MD	0.686	PLIC_MD	0.322
SCR_MD	0.741	SS_MD	0.344
EC_MD	0.752	FXST_MD	0.365
CST_FA	0.758	ALIC_MD	0.389
SLF_MD	0.784	PCT_MD	0.393
PCR_MD	0.817	EC_MD	0.395
CP_MD	0.841	SFO_MD	0.398

Table 9.2: Best feature reconstruction MSE for the decoding of sMRI and dMRI channels.

part of the cingulus (PCG, RACG, ICG) and the cuneus (CU, PCU), each having an MSE value higher than 1.

9.3.4 SHAP feature importance

Figures 9.4 and 9.5 show the SHAP values associated with each input feature for the reconstruction of the features in a different channel.

In detail Figure 9.5 shows the contribution of genetic (above) and sMRI features (below) to the decoding of the top 10 best reconstructed dMRI features. Starting with the decoding from genetics, the gene ranking was highly stable across reconstructed fea-

Chapter9/Figures/SHAP_sMRI.png

Fig. 9.4: SHAP values associated to the most contributing features of channels dMRI and genetics (rows) for the decoding of the best reconstructed sMRI features (columns).

tures, with the top 10 relevant genes being the same for all the dMRI IDPs with the genes HDAC7, TRHDE and BCAS1 always in the top 3. The same behavior was recorded for the decoding of dMRI from sMRI IDPs, where the most relevant features were TH, RACG and SFG, being among the top three contributing features for all the top ten best reconstructed dMRI IDPs.

Figure 9.4 instead shows the contribution of genetic (above) and dMRI features (below) to the decoding of the top 10 best reconstructed sMRI features. The most relevant genes were generally relevant for all the sMRI features but in different rankings based

Chapter9/Figures/SHAP_dMRI.png

Fig. 9.5: SHAP values associated to the most contributing features of channels sMRI and genetics (rows) for the decoding of the best reconstructed dMRI features (columns).

on the reconstructed feature. The most frequent gene was the PTPN13, being relevant for PCG, PsCG, PCU, ICG, TTG and FP. It was followed by RIF1 and HABP4 which were relevant for AC, PCG, PsCG, PCU, PA and TTG for RIF1 and PCG, CU, PsCG, PCU, TTG and FP for HABP4. The decoding of the same channel from dMRI instead saw the same dMRI IDPs being relevant for all the top 10 reconstructed sMRI features, with the MD features being the most frequent and RLIC, SFO and ACR the most relevant tract. Only the AC had different dMRI contributions having the RLIC and FX from FA and the EC from MD as the most relevant dMRI features for its reconstruction.

9.4 Discussion

In this work we proposed an interpretability framework for the MCVAE aiming at unraveling the underlying association between brain structural and microstructural modulation with genetics. We hence developed a three channels MCAVE considering ROI based volumes and thicknesses as the sMRI channel, tract-based FA and MD values for dMRI, and gene variant scores for genetics. We compared a different number of LVs, namely 20 and 3, as well as proposed a metric to compare the different LVs among different MCVAEs. We also analyzed the obtained latent space and retrieved the best reconstructed features separately for each channel decoded from the other two. Finally, we proposed a modification of the MCVAE which allowed to apply the SHAP model to retrieve which feature of the input channel mostly contributed to the reconstruction of the features in the output.

In literature, the MCVAE model recently proposed by [247] was applied in [248] allowing to predict myocardial infarction through retinal scans. More in detail they trained an MCVAE and a deep regressor model to estimate left ventricular mass and left ventricular end-diastolic volume and predict the risk of myocardial infarction (AUC=0.80± 0.02, sensitivity=0.74 ± 0.02, specificity=0.71 ± 0.03) using just the retinal images and demographic data. Their results indicated that it is possible to identify patients at high risk of future myocardial infarction from retinal imaging available in every optician and eye clinic. In relation to AD an extension of MCVAE was recently exploited in [249] and [253] to model the dynamics governing the joint evolution of longitudinal imaging and clinical biomarkers along disease progression. In their first work [249], they accomplished to simulate the effect of the intervention time and drug dosage on the biomarkers' progression in an ADNI study cohort. Their results were compatible with the outcomes observed in past clinical trials, and suggest that anti-amyloid treatments should be administered at least 7 years earlier than what is currently being done in order to obtain a statistically powered improvement of clinical endpoints. In their second work [253] they assessed the generalization of the model by testing it on an independent study cohort of the Geneva Memory Center (GMC). They showed that the difference between the temporal evolution of similar biomarkers simulated on the ADNI and GMC cohorts remained below 10%, confirming model robustness and good generalization and highlighting its potential for clinical and pharmaceutical studies.

To the best of our knowledge, our work is the first attempt of considering genetics and dMRI as input to the MCVAE to investigate their association in AD proposing also a metric allowing to compare models with different numbers of LVs. Moreover, the application of SHAP to a generative framework is far from trivial and allowed to deeply uncover feature associations.

With this work we firstly demonstrated that the 3LV-sparse and the 20LV-sparse had similar reconstruction performance, with the LV having the lowest associated dropout probability in the models mapping one to the other. This confirmed that both models reached a similar combination of the input channels leading to the definition of the latent space, at least in the most significant LV. Moreover, the reconstruction performance was comparable across the 20LV-sparse and the 3LV-sparse with the dMRI being generally better reconstructed compared to sMRI and genetics. In particular, between the two imaging channels, it was evident a better contribution of sMRI in decoding dMRI than viceversa. The genetic channel did not achieve sufficient reconstruction performance either decoding from sMRI or dMRI. At the same time, the reconstruction of the imaging channels decoding from genetics achieved lower performance compared with the decoding from imaging channels themselves. This behavior could be due to the MCVAE assumption on the latent space. In fact, the posterior distribution estimated for the latent space of each channel is a multivariate gaussian, which well fits the distribution of imaging data but not the genetic ones which instead follow a categorical distribution. This represents the major limitation of this approach with the chosen input data.

Leaving the genetics aside, moving to the best reconstructed imaging features, as already stated for the general performance, the dMRI reached the best MSE score, in particular when decoded from sMRI. The best reconstructed brain tracts were part of the internal capsule and the corona radiata which were also among the dMRI derived indices having a high associated weight in the PLS model discussed in Chapter 7, associated with two PRS, in MCI patients. Importantly, these tracts were also found with increased diffusivity in MCI compared to healthy controls in [173]. In general, the MD derived IDPs were the best decoded, with the exception of the corticospinal tract, which was among the top ten tracts best decoded from genetics for both FA and MD indices. Moving to the sMRI decoded features, the reconstruction performance was very poor in decoding either from dMRI or genetics, highlighting that the model was not able to find relevant associations between those channels.

Being aware that this is not compliant with what is stated in Chapter 3 we decided to perform the additional step of *post-hoc* feature ranking following the aim of applying an interpretability method to a generative model in order to better investigate the feature-feature interactions. In fact, through the adoption of perturbation based methods such as SHAP, we were able to retrieve, for an input and a decoded channels, which features of the input contributed the most, hence obtaining higher Shapley values, to the reconstruction of each feature in the output channel. This was possible thanks to a slight modification of the forward method of the MCVAE, in which instead of returning both the mean and the variance of the latent space distribution, only the mean was

considered. This step has not affected the reconstruction performance since, by design, the original MCVAE only considers the mean of the distribution for the decoding phase. With this step, we were able to show that the features most contributing to the decoding were relevant for multiple decoded features, with some of them being highly frequent, especially for the reconstruction of dMRI IDPs. We will not go into much in detail about the feature contribution due to the generally poor performance achieved. However, of interest, there is a grain of truth since the genes that mostly contributed to the decoding of imaging channel were also found in our previous work relating SKAT genes and sMRI IDPs through the PLS model presented in Chapter 8.

9.4.1 Limitations and future works

Despite the presented results not being completely satisfactory in this preliminary phase, we achieved promising results in the generation of an interpretability framework for generative models such as the MCVAE. Future works will focus on the utilization of a different prior latent space distribution from the gaussian used in this study in order to better reflect the input feature distributions and hence allowing to reach also better reconstruction performance. In this direction, it was recently proposed the heterogeneous longitudinal VAE (HL-VAE) which extends the existing temporal and longitudinal VAEs to heterogeneous data providing efficient inference for high-dimensional datasets and including likelihood models for continuous, count, categorical, and ordinal data while accounting for missing observations [254] which seems promising to overcome the limitations of this work. Finally, The extension of the study cohort is also a future step that would allow having an independent test set from the validation sets of the 5-fold CV considered in this study.

9.5 Conclusions

In this work we presented an interpretability framework for MCVAE including two imaging channels, the sMRI and dMRI derived IDPs and one genetics channel composed by gene variant scores applied to the analysis of the AD continuum. Moreover, we proposed metric to assess the similarities between two different MCVAEs accounting for a different number of LVs. Finally, we proposed a modification of the MCVAE in order to apply of interpretability methods such as SHAP considered in this work. Despite the poor reconstruction performance the obtained latent space was well aligned across different channels and showed a clustering between the different stages of AD. Thanks to the application od SHAP we retrieved the most relevant features for the decoding of the different channels allowing us to verify that the most relevant features

were highly similar for each different decoded feature. This work is the first step towards the application of interpretability to the MCVAE allowing to deeply understand the most relevant feature contribution and associations leading to the generation of a common latent space across different input channels.

The preliminary work presented in this Chapter was partly submitted to the VIII Congress of the National Group of Bioengineering (GNB) [255].

9.6 Supplementary tables

Tract	Acronym
Middle cerebellar peduncle	MCP
Pontine crossing tract	PCT
Genu of corpus callosum	GCC
Body of corpus callosum	BCC
Splenium of corpus callosum	SCC
Fornix	FX
Corticospinal tract	CST
Medial lemniscus	ML
Inferior cerebellar peduncle	ICP
Superior cerebellar peduncle	SCP
Cerebral peduncle	CP
Anterior limb of internal capsule	ALIC
Posterior limb of internal capsule	PLIC
Retrolenticular part of internal capsule	RLIC
Anterior corona radiata	ACR
Superior corona radiata	SCR
Posterior corona radiata	PCR
Posterior thalamic radiation	PTR
Sagittal stratum	SS
External capsule	EC
Cingulum	CgC
Cingulum	CgH
Fornix/Stria terminalis	FXST
Superior longitudinal fasciculus	SLF
Superior fronto-occipital fasciculus	SFO
Uncinate fasciculus	UNC
Tapetum	TAP

Table 9.3: dMRI tracts acronyms table.

Brain region	Acronym	Brain Region	Acronym
bankssts	BSTS	posteriorcingulate	PCG
caudalanteriorcingulate	CACG	precentral	PrCG
caudalmiddlefrontal	CMFG	precuneus	PCU
cuneus	CU	rostralanteriorcingulate	RACG
entorhinal	EC	rostralmiddlefrontal	RMFG
fusiform	FG	superiorfrontal	SFG
inferiorparietal	IPG	superiorparietal	SPG
inferiortemporal	ITG	superiortemporal	STG
isthmuscingulate	ICG	supramarginal	SMG
lateraloccipital	LOG	frontalpole	FP
lateralorbitofrontal	LOFG	temporalpole	TP
lingual	LG	transversetemporal	TTG
medialorbitofrontal	MOFG	insula	IN
middletemporal	MTG	Cerebellum-Cortex	CER
parahippocampal	PHIG	Thalamus	TH
paracentral	PaCG	Caudate	CA
parsopercularis	POP	Putamen	PU
parsorbitalis	POR	Pallidum	PA
parstriangularis	PTR	Hippocampus	HI
pericalcarine	PCAL	Amygdala	AM
postcentral	PoCG	Accumbens-area	AC

Table 9.4: sMRI regions acronyms table.

Part IV

Open challenges in XAI: validation strategies

10

A new stability criterion for XAI methods: Application to AD classification

This Chapter will present a new method to assess the robustness of feature rankings provided by eXplainable Artificial Intelligence (XAI) methods, especially in presence of multicollinear feature. The framework was tested to solve Alzheimer's Disease (AD) classification problem while interpreting the outcome. Our findings indicate that our method was able to disentangle the list of the informative features underlying dementia, with important implications for aiding personalized monitoring plans.

10.1 Introduction

The utilization of Artificial Intelligence (AI) systems in a growing number of sensitive domains with ramifications for the social, ethical, medical, and safety sectors inevitably raises problems of trust, bias, and interpretability due to the fact that the majority of AI technologies in use today are effectively "black boxes".

Local Interpretable Model-Agnostic Explanations (LIME) [8] and SHapley Additive exPlanations (SHAP) [252] are the two most popular and well-performing model-agnostic methods for categorical variables. Particularly, the latter has a number of desired theoretical qualities (i.e., local correctness, missingness, consistency) and generates explanations that are more consistent with human explanations [252], making it more applicable than the others for a number of applications. There are now applications of the SHAP approach in several fields of study, including neuroimaging and brain age [19], coronary heart disease [256], chemistry [257]. However, the list of the most informative characteristics may be impacted if the predictor variables are highly collinear, which may lead to erroneous explanations [258, 259]. This is a prevalent difficulty in real-world issues, when a large number of variables are considered and correlations are frequently unavoidable. Moreover, the inability of SHAP to appropriately handle multicollinearity between features might significantly hinder the understanding of relative

explanations. Therefore, it is necessary to evaluate the results of SHAP and, more generally, XAI methodologies in order to create confidence in the chosen approach. Several qualitative and quantitative methodologies and proxies have been developed to evaluate the success of XAI methods and were already presented in Chapter 2. In this work, we provide a novel heuristic for assessing the robustness of a given XAI method's list of useful predictors. Our idea is that the consistency of the explanations may rely on a) the explanation technique, b) the selected classifier, and c) the data provided. All of these criteria play a vital part in generating the list of model-determining attributes. In specifically, we evaluated the suggested technique on a well-known biological topic, namely the binary categorization of controls and dementia patients and the selection of the most informative characteristics. To do this, data from a publically accessible database (Alzheimer's Disease Neuroimaging Initiative (ADNI) [260]) were analyzed to create a set of characteristics to be utilized in conjunction with several classifiers, while SHAP was selected as the XAI approach. This would also make it possible to validate the congruence between SHAP explanations from various models and their behavior in facing multicollinearity.

10.2 Materials and methods

The proposed pipeline examines how a given classifier and SHAP deal with multicollinearity between features to provide an overall robust and stable list of significant features. For SHAP specification please refer to the Appendix A.

10.2.1 Proposed analysis pipeline

Given a selected classifier and SHAP values, the objective is to determine the stability of the feature ranking supplied by SHAP, in relation to features context. In other words, if the SHAP value assigned to a specific feature depends only on that feature, deleting the others should not affect it. Based on this working hypothesis the proposed proxy is an iterative procedure in which in each step the most significant features, according to the associated SHAP values, were sequentially eliminated from the classifier input. SHAP values and feature ranking were then recalculated for this new model and so on.

In order to evaluate the stability of the SHAP values the Normalized Movement Rate (NMR) was calculated. The NMR is a metric ranging between 0 and 1, describing the agreement between feature rankings. NMR = 1 indicates a drastic shift in the order of predictors when features are eliminated repeatedly from the model, while NMR = 0 shows that the list of predictors does not change when the top features are eliminated

at each iteration. The closer the NMR value is to 1, the more unstable the ordered list, whereas a number near to 0 indicates a stable feature ordering.

More in detail, the steps of the proposed method are briefly outlined in the pseudocode (Algorithm 1). After training, testing, and obtaining the ordered list of informative features for the classifier from SHAP, the NMR measure was calculated to determine the method's stability. Specifically, in the iterative technique, the top feature in terms of associated SHAP values from the list is removed. At each iteration, the SHAP values and feature ranking were recomputed, while evaluating how many predictors changed their positions compared to the previous step as well as the movements they did across the list. These steps were repeated until two features remained in the predictor list. NMR is finally calculated as the ratio between the movement rate, given by the total sum of the movements the features did across each list over the number of possible movements, and the total Number of Features (NF) (minus 2). Therefore, the NMR value is calculated by comparing sub-lists of informative features that have different lengths by removing the top feature at each iteration.

Algorithm 1 Calculating NMR value

Ensure: NF=number of features
Require: Train the model for classification/regression etc.
Require: Test the model
Require: Apply SHAP and rank the feature importance in descending order
  while $NF \neq 2$ **do**
    Remove the predictor with associated highest SHAP value
    Apply SHAP and rank feature importance in descending order
    Calculate M as the count of the predictors which changed their position compared to the list at the previous step
    Calculate the maximum possible movements as: $MPS = 2 \times \sum_{i=1}^{i+2}(NF - i)$ for $i < NF$
    Calculate the movement rate as: $MR = \frac{M}{MPM}$
  end while
  Calculate NMR as: $NMR = \sum(MR)/NF_{total} - 2$

10.2.2 Data

To exemplify the presented strategy, it was applied on data collected from ADNI3 database (http://adni3.loni.usc.edu/). The selected cohort consisted of a total of 475 individuals, including 300 Healty Controls (HC) (254 Controls (CN), and 46 Significant Memory Concern (SMC)) and 175 patients with dementia (comprising 70 Early Mild Cognitive Impairment (EMCI), 55 Mild Cognitive Impairment (MCI), 34 Late Mild Cognitive Impairment (LMCI) and 16 AD). Freesurfer v.7.1 was used to process T1-weighted (T1-w) images (https://surfer.nmr.mgh.harvard.edu). The thickness/volume

values corresponding to 25 regions of interest were considered as input variables (predictors). Age, gender, level of education, and APOE were considered as confounding variables and regressed out.

10.2.3 Implementation settings

The first question concerned the relationship between the proposed stability proxy and the chosen classifier. In order to forecast the class of each subject, five different classifiers were investigated, namely Decision Tree (DT), Light Gradient Boosting Machine (LGBM), Logical Regression (LR), Random Forest (RF), and Support Vector Classifier (SVC). To determine the appropriate model parameters, the study cohort was divided in training (80%) and test (20%) subjects. Hyperparameters tuning was then carried out through 10-fold cross validation on training data, optimizing the following parameters: max_depth, min_samples_leaf, min_samples_split, splitter and criterion for DT; max_bin, boosting_type, metric, num_leaves and min_child_samples for LGBM; penalty and inverse of regularization strength for LR; bootstrap, min_samples_split, min_samples_leaf, max_features, max_depth and n_estimators for RF; cache_size, kernel type gamma and regularization parameter for SVC. The parameters leading to the maximum accuracy value were selected, and the matching optimum model was then applied to the test data. Finally, the SHAP model was globally applied to each of the five models to determine the ranking (descending order) of informative predictors in the test data. SHAP TreeExplainer was used for DT and RF, whereas KernelExplainer was utilized for LGBM and SVC, and LinearExplainer was utilized for LR.

The second question instead concerned the effect of feature collinearity on the model output. Consequently, we utilized Principal Component Analysis (PCA) to generate a list of independent characteristics. The number of components was determined automatically so that 90% of the variation could be explained, resulting in 15 Principal Components (PC)s. To derive the NMR, the same procedures as in Algorithm 1 were performed using a fresh set of uncorrelated PCs and applying the SHAP technique globally to these new features. The original ranks (before/after PCA) were then compared across techniques using Spearman's rank correlation to determine the inter-method agreement.

10.3 Results

We used PCA on the initial set of 25 features and computed the correlation between the 15 PCs and the raw features, as well as between the PCs themselves. To demonstrate that the PCA successfully eliminated the association between features, we provide their

correlation in Figure 10.1. As predicted, the correlation between the features was extremely low, with values 1e-15 for all pairings. If we do not apply PCA to the feature space (Figure 10.2), Spearman's correlation coefficients between each pair of variables indicate substantial correlations in the majority of cases, with values ranging from 0.01 to 0.79. This indicates the presence of a multicollinearity problem among the chosen features and warrants the application of PCA.

Chapter10/Figures/Salih2.png

Fig. 10.1: Spearman's correlation after applying PCA.

Consequently, we use the PCA, generate the SHAP ranking lists, and execute the iterative approach, deleting successively the most relevant feature, before computing the NMR according to Algorithm 1. The NMR values for each classifier using the original features (NMRorig) and after PCA (NMRPCA) are shown in Figure 10.3. Our findings demonstrated a drop in the NMR following PCA, indicating that the important rankings supplied by SHAP are more stable and reliable following the elimination of feature collinearity. Particularly, the value decreased for four out of five models, when SVC reached NMR= 0 and LR and RF values were near to 0. Only for DT did the NMR not change after PCA was applied. The Figure also displays the accuracy of each model, with ACCorig referring to the model's accuracy when using the raw set of features and ACCPCA equating to the model's accuracy while employing PCs. For each model, two values are reported: one when all predictors (original, 25 or after PCA, 15 features) are used, and the other when only the final two variables are considered (i.e., the ones deemed as having the lowest importance). When just the less significant characteris-

Chapter10/Figures/Salih1.png

Fig. 10.2: Spearman's correlation between the original features.

tics were maintained, accuracy decreased in both techniques, although the decline was more pronounced when working with PCs. After PCA, the accuracy values of all models improved overall. To determine how much the results depend on the chosen classifier, we compare the SHAP lists from various models by computing their correlation before/after PCA (Figure 10.4). The upper (blue) triangle highlights the Spearman's correlation between the ordered SHAP ranks. To determine how much the results depend on the chosen classifier, we compare the SHAP lists from various models by computing their correlation before/after PCA (Figure 10.4). The upper (blue) triangle displays the Spearman's correlation between the ordered SHAP ranks, showing low values (near to 0) in the majority of cases. In contrast, when examining the SHAP-provided sorted lists for the 15 PCs (lower triangle, yellow), higher values were discernible. This implies that reducing the dependence between predictors using PCA can result in feature ranking that is independent of the classifier, as expected.

10.4 Discussion and Conclusions

In this study, we developed a straightforward yet practical way for evaluating any XAI method and quantifying the predictor list's stability. The suggested measure may be applied easily to estimate the degree of confidence in the ordered list of informative predictors provided by a certain collection of features, classifier, and XAI technique. Our findings on SHAP indicate that if the original set of features is used and a signifi-

Chapter10/Figures/Salih_tab1.png

Fig. 10.3: NMR and accuracy (ACC) values for each model combined with the raw features or PCs. ACCorig: accuracy using the original features; ACCPCA: accuracy using 15 PCs.

Chapter10/Figures/Salih_tab2.png

Fig. 10.4: Spearman's correlation for the ranked lists using the original 25 features (blue) and 15 PCs (yellow).

cant correlation is present, there will be a degree of instability in the rankings from all the models, but SVC is the classifier with the lowest NMR and hence the most stable lists. In contrast, when PCA is applied, the new collection of uncorrelated variables results in stable ranks for the majority of classifiers. As PCs are combinations of traits, one may argue that we might lose the semantic meaning of interpretability in this manner. Consequently, it would be required in this case to understand each PC by assessing the magnitude of the original variables in the eigenvector defining each component, so providing information linked with the original space. XAI is commonly utilized in sensitive areas, such as the prediction of long-term mortality [261], admission to the critical care unit, and extubation failure [262, 263]. In such fields, it is crucial to select the most informative predictors. Our pipeline may be applied in any domain using a wide range of models and XAI techniques to evaluate the feature rankings' dependability.

This work was led by my colleague Ahmed Salih and was published in [264].

11

Consistency assessment of eXplainable Artificial Intelligence (XAI) methods on tabular data: application to upper limb rehabilitation outcome after stroke

In this Chapter we propose a comparison between XAI permutation and perturbation based methods applied to tabular data to assess their consistency on the same problem. The clinical outcome is the prediction of Upper Limb (UL) functional recovery following rehabilitation after stroke.

While stroke is one of the leading causes of disability, the prediction of UL functional recovery following rehabilitation is still unsatisfactory, hampered by the clinical complexity of post-stroke impairment. Predictive models leading to accurate estimates while revealing which features contribute most to the predictions are the key to unveil the mechanisms subserving the post-intervention recovery, prompting a new focus on individualized treatments and precision medicine in stroke. In this study, we had the twofold goal of evaluating whether Machine Learning (ML) can allow to derive accurate predictions of UL recovery in sub-acute patients, and disentangling the contribution of the variables shaping the outcomes. To do so, Random Forest equipped with four XAI methods was applied to interpret the results and assess the feature relevance and their consensus. Our results revealed increased performance when using ML compared to conventional statistical approaches. Moreover, the features deemed as the most relevant were concordant across the XAI methods, suggesting a good consistency of the results. In particular, the baseline motor impairment as measured by simple clinical scales had the largest impact, as expected. Our findings highlight the core role of ML not only for accurately predicting the individual follow-up outcome scores after rehabilitation, but also for making ML results interpretable when associated to XAI methods. This provides clinicians with robust predictions and reliable explanations that are key factors in therapeutic planning/monitoring of stroke patients.

11.1 Introduction

While ML methods have been proven to be highly promising in different domains [265, 266, 267] quantitative methods to reliably assess the variables that are important in this prediction are needed to disentangle the contributions of the different features shaping the final estimates. XAI is still largely unexplored, especially in post-stroke UL recovery framework. Indeed, providing details on the predictors contributing most to a given outcome and on their relevance would provide meaningful information and make the data-driven solution worth of the clinicians' trust [268]. Some initial insights in this respect have been given by [269], using conventional approaches to define a variable importance such as the magnitude of the regression coefficients for Elastic Net (EN) or the permutation feature importance for Random Forest (RF), that is the decrease in a model score when a single feature value is randomly shuffled. However, a wider range of interpretability methods could be applied to explain any black-box model, where the information is hidden inside the model structure and should be jointly applied to assess the consistency and reproducibility of the results given by a single method. Information about feature importance could be gained by perturbation-based methods as SHapley Additive exPlanations (SHAP) [270], recently explored in a predictive (classification) model for functional recovery after post-stroke rehabilitation [268], and Local Interpretable Model-Agnostic Explanations (LIME) [8] that are, by far, the most comprehensive and dominant across literature methods for visualizing feature interactions and importance [271].

Stroke is one of the leading causes of disability worldwide [272]. The World Health Organization (WHO) estimates that European countries' stroke events are likely to increase by 30% between 2000 and 2025, with an expected increase in persistent disability [273]. At the rehabilitation discharge, around 64% of patients regain walking functions [274], while upper limb (UL) impairments chronically affect the functional independence and satisfaction in $50-70\%$ of all stroke patients [275]. Within this framework, rehabilitation still plays a crucial role in promoting the highest functional recovery and lowering the level of disability starting from the acute and sub-acute stages [275].

The last five years have witnessed a growing interest in ML for decision support and functional outcome prediction in stroke [276, 277, 278]. ML methods have been applied to identify the factors affecting home discharge after stroke inpatient rehabilitation [279, 280], to predict overall motor outcome after acute hemorrhagic or ischemic stroke [281, 282, 283] and to predict post-stroke activities of daily living in sub-acute stroke patients [284]. Almost all these studies in the current literature focus either on classification tasks, for example to classify stroke patients into classes based on significant changes or those likely to have favorable outcomes after a given time (e.g.,

good and poor outcome), or on predicting longitudinal score changes [278]. However, as stated by Bonkhoff & Grefkes [277], directly predicting the final follow-up scores of functional recovery rather than focusing on coarse-grained classifications or on the prediction of score changes is a desirable next step to enhance current scenarios. To the best of our knowledge, only two studies have used classical statistical and ML methods to predict individual UL scores following rehabilitation so far [285, 269]. Tozlu et al. [269] in particular assessed the performance of five ML methods in predicting UL motor function in chronic stroke patients after six weeks of intervention (task-oriented rehabilitation combined with TMS protocol stimulation) based on demographic, clinical, neurophysiological, and imaging features. Their results showed that EN followed by RF performed best when predicting post-intervention upper-extremity Fugl-Meyer Assessment (FMA-UE) using demographic and baseline clinical data, and that pre-intervention FMA-UE and hemispheric difference in motor threshold were the most important predictors.

Therefore, in this study we aimed at extending the results obtained by classical statistical analyses relying on a ML-based model to predict UL recovery in sub-acute stroke patients admitted to a neurorehabilitation unit, with a special focus on model interpretability with different XAI approaches. In particular, we aimed at directly predicting the final follow-up scores of functional recovery rather than focusing on patient classifications or on the prediction of longitudinal score changes, as generally done. Firstly, we evaluated the performance of Multiple Linear Regression (MLR) and a well-known ML approach (RF regression) in predicting UL motor function scores after six weeks of rehabilitation in sub-acute stroke patients based on demographic, clinical, and neuropsychological measures. RF has previously demonstrated good predictive performance [269, 268], even with small sample sizes and high dimensional data, and requires only limited hyperparameter tuning (e.g., the number of trees in the forest, the maximum number of features considered for splitting a node, or the maximum number of levels). Then, we focused on interpreting the results provided by the ML black-box model in order to elucidate what are the features having a stronger influence on the prediction task and what are their relative importance. To do so, four different XAI approaches, including permutation and perturbation-based methods, were used to verify the consistency and robustness of the results, in light of the different nature of these interpretability methods. To the best of our knowledge, this represents the first attempt of using such methods for dealing with this crucial clinical issue, paving the way for the implementation of solutions to aid the clinical support and treatment of sub-acute stroke patients.

11.2 Materials and Methods

The data were extracted from a local dataset, comprising all patients admitted to the Neurorehabilitation Unit (AOUI Verona) between January 2018 and October 2020. The initial cohort included 192 subjects. Inclusion criteria were confirmed diagnosis of first-ever stroke (ischemic or haemorrhagic) verified by Computed Tomography (CT) or Magnetic Resonance Imaging (MRI), age ≥ 18 years old, no orthopedic limitation and pre-existing neurological disorders affecting the UL, time from stroke < 6 weeks. Exclusion criteria were severe comprehension (NIHSS score= 2) and cognitive deficits (Mini-Mental State Evaluation < 18/30), premature discharged or admission to other medical/surgery facilities for complications, having neurosurgery in the acute and subacute stages, having a maximum score in all UL sensorimotor scales (FMA, Motricity Index [MI] and FMA Sensory).

Starting from the initial cohort of 192 subjects, records that were incomplete at admission (T0) for one of the main clinical/cognitive scales or lacking the FMA-UE at discharge (T1) were excluded. This procedure led to the final dataset including 95 records. The mean time (± standard deviation) elapsed between stroke event and dismission was 37.71 (±15.43) days. This study was approved by the local Ethics Committee for Clinical Sperimentation (CESC) of Verona and Rovigo (no. 2320CESC). Patients gave their consent to participate in the study as part of consent to usual care. The study was conducted following the declaration of Helsinki.

Demographic and clinical information was retrieved from the patient medical chart and rehabilitative log. All patients underwent the same assessment protocol at admission and discharge from the Neurorehabilitation Unit.

We selected a comprehensive set of assessment measures based on the International Classification of Functioning, Disability and Health (ICF) framework, outlined in Table 11.1. In particular, they represent features that are relevant to the patient in the subacute phases of recovery and that are routinely collected in neurorehabilitation departments.

All patients underwent intensive, multidisciplinary UL rehabilitation treatment consisting of 2 hours/day for six days/week for all the length of stay [286]. This was focused on passive mobilization and stretching, exercises based on active motility tasks and normal daily living activities selected by the physiotherapist according to the residual UL movements and the patient preference for activities. Tasks were tailored to the patient functioning and progressively increased in difficulty as the patient improved in performance by tuning the speed or accuracy, repetition, or creating performance-sensitive adaptations.

Table 11.1: OVERVIEW OF THE EXTRACTED FEATURES. Abbreviations: L = left; H = hemorrhagic; I = ischemic; TACS = Total anterior circulation stroke; PACS = Partial anterior circulation stroke; POCS = Posterior circulation syndrome; FMA-UE = Fugl-Meyer upper-extremity; FMA-UE-S = Fugl-Meyer upper-extremity Sensory; MI-UE = Motricity Index upper-extremity; MI-LE = Motricity Index lower-extremity; TCT = Trunk Control Test; BI = Barthel Index.

Domain	Type	Feature
Demographical & Clinical	Continuous	Age
	Dichotomous	Diabetes
		Hypertension
		Dysphagia
		Aphasia (Mild-moderate)
		Thrombolysis
		Somatosensory deficits
		Affected hemisphere (L)
		Bilateral involvement
	Categorical	Sex
		Stroke Type (H,I)
		TACS, PACS, POCS
Cognitive	Continuous	Picture naming
		Broken Heart Cancellation
		Space asymmetry
		Object asymmetry
Body function	Continuous	FMA-UE
		FMA-UE-S
		MI-UE
		MI-LE
		TCT
Disability	Continuous	BI

Based on all the collected data, 24 relevant variables were derived for each patient. These included 14 features on demographic and clinical domains that are age, gender, stroke type, affected hemisphere, presence of a bilateral hemisphere involvement, Bamford classification, thrombolysis treatment in the acute stage, presence of hypertension, diabetes, dysphagia, presence of mild-moderate aphasia and somatosensory deficits [287, 288, 289, 290, 291, 292]. For the latter, a baseline score ≤ 11 points was chosen as threshold to indicate somatosensory deficits, accounting for a measurement error of 10% [293].

Within the International Classification of Functioning, Disability and Health framework (ICF), we selected features exploring the cognitive, body function and disability domains in the sub-acute stage [288, 289, 290, 292], as also detailed in Section II.B. Four variables related to the cognitive domain included the Oxford Cognitive Screen (OCS) sub-items "picture naming" and "Broken Heart cancellation task", space and objective

asymmetry. Body function and disability measures (6 features) included the FMA-UE, FMA-UE-S, TCT, MI-UE, MI-LE, and BI.

11.2.1 Data modeling

Chapter11/Figures/Fig1.png

Fig. 11.1: Overview of the proposed pipeline. Twenty-four pre-treatment features, grouped in demographical and clinical, cognition, body function and disability metrics, were given as input to two separate data modeling techniques, namely forward and backward stepwise MLR and RF regression, with the aim of predicting the individual post-intervention FMA-UE score at follow-up, after inpatient stroke rehabilitation. 5-folds Cross Validation (CV) was employed for RF hyperparameters optimization. Finally, aiming at a clear interpretation of the outcome, MLR coefficients were analysed to retrieve the most important features for the prediction, while four XAI methods (Permutation Feature Importance (PFI), Random forest Feature Importance (RFI), LIME and SHAP) were employed and compared to explain RF outcome.

The general workflow of our study is displayed in Figure11.1. Descriptive statistics were used to assess patients' demographic and clinical characteristics. Stepwise MLR based on sum of squared errors criterion was initially applied to predict the outcome variable, represented by the post-intervention FMA-UE score at T1. This approach used both forward and backward stepwise regression to determine the best prediction model, starting from a constant model (Penter=0.05, Premove=0.10). In terms of predictors, any possible issue related to multicollinearity was preliminary studied using the Belsley collinearity diagnostics [294]. Multicollinearity was considered to occur if a component associated with a condition index > 30 contributes strongly (variance proportion > 0.5) to the variance of two or more variables. This analysis confirmed that all the 24 variables derived at the baseline (T0) were non-collinear and were thus consid-

ered as possible independent variables in the stepwise. Of note, three out of the four Bamford classes were considered as variables as the fourth one could be derived from the others. All these analyses were performed using MATLAB R2020a (Mathworks Inc).

A RF regression analysis was then carried out with Python software (version 3.7) and scikit-learn library (version 0.24.1), using the 24 abovementioned features (Table I) to predict the individual post-intervention FMA-UE scores. RF is a supervised ML algorithm that uses ensemble learning method for regression (and classification). It is based on two concepts, that are regression trees and bagging, i.e., bootstrapping and aggregating. A regression tree divides in a recursive way the input data by finding combinations of thresholds associating the value ranges of the input variables to the outcome. Each sample can fall into a single leaf, according to the unique path that this sample cover across the tree, which depends on the different threshold combinations. These thresholds are optimized according to an impurity criterion, chosen to be the Mean Square Error (MSE) in our case. To limit the risk of overfitting, several regression trees are randomly generated by bootstrapping the original data, that is selecting a subset of variables for building each tree. Outputs from the forest are finally averaged to obtain the final prediction.

RF has a series of hyperparameters to be set, defining the main structure and associated characteristics of the model, which impact on model accuracy and performance. In our study, four main hyperparameters were tuned, relying in particular on a grid search 5-fold CV to define the best combination of parameters using the same scoring criterion as for the evaluation of the model performance (MSE). The simplest one is represented by number of trees (`n_estimators`), for which the range $200-800$ by increments of 20 was here tested. The total number of features to be randomly selected for defining each regression tree is given by the `max_features` parameter, and three possible values were considered $(4, 8, 24)$. Of note, the recommended default value for this parameter in a regression problem is $m = p/3$ with p = number of total features (thus $m = 8$ in our case) [295, 296], though other options as $m = p$ or $m = log2(p)$ are currently implemented in the different packages and are still debated. Therefore, we decided to include m as tuning parameter, selecting such possible levels to limit the computational time. The minimum number of samples required to split each node can be controlled by the `min_samples_split` parameter (range $2-16$), while the minimum number of samples for each leaf is known as `min_samples_leaf` (range $1-16$). For the RF model with the best set of hyperparameters, the Out-Of-Bag (OOB) score, Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE) were calculated to quantify the performance for the regression task.

11.2.2 XAI and RF model explanations

XAI recently emerged as one of the hottest topics aimed at overcoming the interpretability issue typical of ML (and Deep Learning (DL)) methods, proposing strategies for understanding the outcomes of ML algorithms. Multiple categorizations of interpretability methods can be found in literature. Referring to Holzinger and colleagues [10], such methods can be divided into *post-hoc* methods, which explain what the model predicts in term of what is readily interpretable, and ante-hoc, which instead are methods that incorporate explainability directly into their structure, such as linear regression. In our context, as RF is not intrinsically interpretable since its prediction results from averaging several hundreds of decision trees, *post-hoc* approaches have to be applied to shed lights on this black-box model. Moreover, such approaches are the most suitable in our case since they allow to explain models that are already trained without the need to modify the intrinsic model structure. Among them, classical impurity-based or permutation-based algorithms are generally used in combination with RF, such as RFI [295] and PFI [295, 297]. The latter in particular allows to determine the most important features by running the model on permuted versions of the input. However, more accurate techniques have been recently proposed having the same aim. Particularly relevant are the advanced local surrogate methods that aims at replacing the decision function with a directly interpretable local surrogate model (e.g., LIME [8] and SHAP [252]). These methods perturb the input slightly and test the changes in prediction. These four different approaches were applied to our RF model to interpret the model predictions and quantify the feature importance. We relied in particular on a Leave-One-out (LOO) training strategy, where at each run $n-1$ subjects were used to train the model and the left out for testing and evaluating the feature importance of the different variables.

11.3 Results

The stepwise regression led to the following results for the best predictive model: F-stat=127 (p-value=$2.27e-32$), RMSE = 6.9, MAPE = 17.2%, R^2 = 0.807, Adjusted R^2 = 0.8. The significant features for the final model are reported in Table 11.2 and were the FMA-UE, the MI-UE and the TCT, with the FMA-UE showing the lowest p-value ($p<0.001$). For MLR, the weights associated to each feature can be directly interpreted as the relative feature importance to the model.

Concerning RF analysis, the grid search CV defined a model with `n_estimators=260`, `max_features=8`, `min_sample_split=4`, and `min_sample_leaf=1` as the best estimator. In terms of performance in predicting the post-intervention FMA-UE, the OOB

Table 11.2: STEPWISE MLR RESULTS. Abbreviations: Estimate = Coefficient estimates for each corresponding term in the model; SE = Standard error of the coefficients; tSTat = t-statistic, p = p-value; FMA-UE = Fugl-Meyer upper-extremity; MI-UE = Motricity Index upper-extremity; TCT = Trunk Control Test.

Variable	Estimate	SE	tSTat	p
Intercept	17.794	2.243	7.931	<0.001
FMA-UE	0.365	0.091	4.001	<0.001
MI-UE	0.139	0.064	2.155	0.034
TCT	0.135	0.043	3.153	0.002

score was 0.84, while the RMSE was 6.17 and MAPE = 15.4%. Results for the model explanations with the four XAI methods are reported in Figure 11.2 as global variable importance. In particular, the mean values across all the left-out data in the LOO-strategy were derived for each variable and reported in a descending order, informing on the global feature importance. Of note, the mean of the absolute values was calculated for both SHAP and LIME.

Classical impurity-based and permutation-based importance results are reported in the top panels, showing the mean Mean Decreased Impurity (MDI) and mean Mean Decreased Accuracy (MDA) values associated to each feature across all the subjects, respectively. Importantly, in these two methods the evaluation of the feature importance is based on changes in model performance, with more important features having a sharper impact. Agreement was found concerning the most important variables ranked in the top four positions, that are the post-intervention FMA-UE, followed by MI-UE, TCT and MI-LE. Of note, the latter was not included in the best model from stepwise MLR since it was not deemed as significant in the prediction task, differently from the other three variables. A high similarity could also be noted for the remaining features, suggesting a good overlap between the results provided by these two simplest approaches in our scenario.

LIME results are here reported as global averages across all the subjects for each feature. In LIME, the impact of each feature on the prediction is defined by its weights, i.e., the regression coefficients, in the simple local model. In this way an estimate of the relative importance of the different features and their impact is directly provided, with higher weights suggesting a stronger influence on the prediction. The top four features (FMA-UE, MI-UL/LE and TCT) are in agreement with those identified by the two conventional approaches.

Similar results are also shown by the SHAP method, as visible in Figure 11.2-bottom right. As this approach is based on the magnitude of feature attributions, features with large absolute Shapley values are the most important for the prediction. Considering all the four XAI methods cases, agreement was globally found across the different rank-

Chapter11/Figures/F2.png

Fig. 11.2: Global feature importance, ranked in descending order, for each of the four XAI methods.

ings, as it can be qualitatively evaluated in Figure 11.2. This was further confirmed by the Spearman's rank correlation coefficient calculated for all the ranking pairs, revealing high correlation between RFI and PFI ($\rho = 0.97$), followed by PFI vs SHAP ($\rho = 0.96$) and RFI vs SHAP ($\rho = 0.94$). Conversely, the correlation of LIME ranking with all the others was lower, confirming what can be visually appreciated from the qualitative evaluation of the lists (LIME vs SHAP: $\rho = 0.64$; LIME vs PFI: $\rho = 0.56$; LIME vs RFI: $\rho = 0.51$). Across all XAI methods, a perfect match was visible for the top four most important features. Age, BI and FMA-UE-S also appeared as having a moderate impact in explaining the model predictions in this ML framework, while the four variables related to the cognitive domain are less stable across the XAI methods and result in diverse positions.

Of note, LIME showed the most different ranking among the least important features compared to the other XAI methods.

11.4 Discussion

Summary of main findings

In this observational study, statistical (stepwise MLR) and ML (RF) methods were applied to predict the UL rehabilitation outcomes, as represented by individual follow-up scores of functional recovery (FMA-UE at T1), in a cohort of sub-acute stroke patients admitted to a Neurorehabilitation Unit. Moreover, different XAI methods were used towards understanding and interpreting model predictions, highlighting the impact of the different features in shaping the final results. A comprehensive set of features investigating the motor, somatosensory and cognitive domains in addition to other demographic and clinical information suggestive of a complex health status (i.e., presence of dysphagia, diabetes, thrombolysis) were used, aiming at defining the optimal predictive model and assessing their relative importance in explaining the model outcome.

The main finding of this observational study is that, in the sub-acute phase of stroke recovery, the post-intervention UL recovery can be predicted with good accuracy by a ML-based model (RMSE=6.17, MAPE=15.2%), further improving the results from simple regression analyses. Moreover, XAI methods allow to open the black-box and to easily interpret the outcome results, revealing in our scenario a good concordance between the feature importance rankings provided by the different approaches. From a clinical perspective, our results demonstrated that the assessment at the admission unit using three clinical scales (FMA-UE, MI-UE, TCT), which are validated for stroke patients and easy to administer, can be used to predict UL rehabilitation outcomes at discharge with good results. This was confirmed using both MLR and four different XAI methods applied to the RF outcome results, underlying the main contribution given by these clinical scales in the prediction task. In addition, all the XAI approaches applied to RF suggested that also MI-LE is associated with the recovery of the UL, although this was not pointed out by the stepwise MLR. Interestingly, cognitive function assessed with the OCS and clinical comorbidities did not significantly influence the outcome at discharge.

Machine Learning and XAI methods

In this study, stepwise MLR and RF models were used to predict the continuous measure of post-intervention FMA-UE after rehabilitation and identify the critical clinical variables for recovery prediction. In the current literature, inconsistent conclusions

have been often found when comparing the performance of classical models to different ML algorithms for diagnostic or prognostic clinical prediction models [298], mainly due to the different sample sizes, analysed variables and predictive models. However, different researchers have proposed ML-based strategies for stroke prediction with excellent results. In our case, while the results of the two methods were broadly equivalent, RF allowed to further improve the prediction achieving better performance as confirmed by the different metrics applied to evaluate the results. This is in agreement with a previous study on chronic stroke patients [269], where the authors reported the best performance when using EN, neural networks and RF combined with demographic and clinical data for predicting post-intervention recovery and changes. Other studies have also evaluated the performance of RF models in predicting the rehabilitation outcomes in this population [268, 278], though most of times not specifically focusing on UL recovery [284, 299]. In all these cases, findings demonstrated that RF regression algorithms were able to estimate the outcome values with high accuracy and reached better accuracy than other ML methods. RF can be also applied to effectively estimate long-term outcome prediction of mortality and morbidity in stroke patients, as recently demonstrated in [300].

The two approaches we adopted also differed in estimating the importance attributed to some variables and in particular to the MI-LE scales, which is ranked fourth by all the XAI methods applied to RF but is not considered in the MLR as the p-value for its inclusion was not significant (p-value=0.415). Considering the four XAI methods here adopted, a perfect agreement for the top four features was present, despite the different principles at the basis of each method (impurity/permutation/perturbation-based) which preclude to take this agreement for granted. The overall concordance between the feature importance provided by the different approaches was further confirmed by the Spearman's correlation, which also highlighted a lower concordance of LIME ranking with all the others (ρ = 0.51-0.64). While XAI methods seem promising tools for making ML models more useable in practical clinical applications [10], examples in the biomedical field and particularly in the rehabilitation framework are still very scarce [301, 268], especially in terms of comparative studies. Indeed, to the best of our knowledge, systematic studies assessing the performance of different XAI methods applied on the same tabular (biomedical) data for interpreting the prediction/classification outcomes of ML methods are still lacking in the current literature, precluding a direct comparison of our findings. Weak correlations have been shown between LIME and SHAP scores when applied to interpreting brain age predictions from imaging variables and Deep Learning models [302, 303]. LIME has also demonstrated to be less stable and to identify a very different set of important biomarkers compared to SHAP and permutation-based methods in tumour data and survival prediction [304].

Many explainability methods are either built for a specific model or focus on specific data; as such, there is no “best-in-class” method and comparative analyses should be increasingly performed, also relying on objective measures [305]. Thus, multiple XAI methods could be used to examine the "under the hood" workings of AI models and examine their biological plausibility, allowing to foster confidence in model outputs to end-users.

In our scenario, even if a consensus for the first four features was present, the feature ranking for the less important features showed less consistency. This was particularly evident between LIME and the three other methods considered. Such differences are mainly seen for one-hot-encoded variables as Affected Hemisphere Bilateral, Affected Hemisphere Left, or the triplete TACS, POCS and PACS which are designed to be mutually exclusive. LIME does not allow to correctly treat such variables since, considering that each data point is perturbed to create the approximate model, perturbing a one-hot-encoded variable may result in unexpected meaningless features, hence the respective LIME values might be unreliable and should be cautiously interpreted.

Results implication for practice and research

The idea of predicting a functional recovery in stroke patients has been the focus of a relevant field of research in the past years. However, most of the existing studies have been focused on the acute phase (within three days after stroke) using bedside clinical tests (active finger extension [AFE] and Shoulder abduction) or on the chronic phase of recovery with expensive neurophysiological and neuroradiological biomarkers [290, 287, 291, 269]. Moreover, almost all studies focused on classification tasks, for example to classify stroke patients based on significant changes in upper-extremity motor function [269], or to binary classify those likely to have favorable outcomes after a given time (e.g., good and poor outcome) [306, 268, 282, 307, 308]. No evidence has been reported on predicting UL rehabilitation outcomes (individual scores) by ML methods in the sub-acute stages. These studies are essential to meet the emerging need for accurate and reliable prediction of the functional recovery of the UL in the rehabilitation phase improving the patient’s hospital stay management. Specifically, resources could be allocated in advance according to the patients’ needs and expectations of recovery. In keeping with previous findings on this topic, the present results suggested that a set of three simple outcomes might predict UL rehabilitation outcomes in sub-acute stroke patients through a relative simple ML approach, paving the way for an initial translation of these methods in the clinical context. Indeed, the most challenging clinical question in stroke rehabilitation is “What is this patient’s potential for recovery?” to make decisions regarding the content and focus of therapy. We believe the most-important fea-

tures we identified by the XAI methods are easy-to-use clinical biomarkers to predict the final FMA-UE score that can unravel factors important to the recovery process.

Study strength and limitations

The study strengths are the set of comprehensive assessment procedures including simple validated clinical scales feasible to be carried out in a real-world context of hospitalization, and measurements of somatosensory, motor, cognitive domains along with clinical features potentially relevant to the topic (i.e., thrombolysis) and comorbidities (i.e., diabetes, hypertension). Moreover, we included patients that well represent the real-life scenario in a neurorehabilitation context for what concerns the stroke type (ischemic and hemorrhagic) and the site of stroke (anterior and posterior circulation infarcts). Finally, all patients underwent the same amount of intensive rehabilitation within a multidisciplinary framework (2 hours/day, six days/week).

We also acknowledge some study limitations, first among the others the moderate sample size, which requires to confirm these preliminary findings in larger groups. As the reliability of the XAI values is closely related to the accuracy of the predictive models, future developments should focus on defining increasingly accurate models on larger cohorts. Moreover, the exclusion of patients with aphasia that hinders the assessment procedure execution might limit the generalization of the present findings, as well as the use of the FMA-UE-S as a measure of somatosensory impairment in stroke patients. More time-consuming assessments (i.e., Nottingham Sensory Assessment scale) might be more appropriate. Future studies could also incorporate additional information, for example from imaging or genetic data, to broaden the picture and further increase the prediction accuracy.

11.5 Conclusions

In summary, this study presents a ML approach based on a RF model for an accurate prediction of UL recovery scores at discharge after inpatient stroke rehabilitation. XAI methods allowed to interpret the ML outcomes and to identify those variables having a more prominent role in the prediction, reporting a consensus across the top features identified by all the chosen methods. From a clinical perspective, our findings may improve the management of sub-acute stroke patients in a rehabilitation unit. Moreover, the present study highlights the importance of an accurate clinical assessment to quantify sensorimotor impairments at admission objectively. Further research is needed to strengthen the general agreement on ML and XAI applications in clinical settings.

The work presented in this Chapter was published in [309].

12

Consistency assessment of eXplainable Artificial Intelligence (XAI) methods on volumetric data: application to Multiple Sclerosis (MS) patients stratification

In this Chapter we propose a comparison between XAI visualization methods applied to fully volumetric data data to assess their consistency on the same problem. The growing availability of novel interpretation techniques opened the way to the application of deep learning models in the clinical field where their use is still largely unexploited. In particular, this holds great potential for the analysis of neuroimaging data, that being multi-modal, multi-dimensional and heterogeneous pose a great challenge to classical statistical and machine learning approaches. In this framework, we focus on a case study that is the stratification of MS patients in the Primary Progressive (PP) versus the Relapsing Remitting (RR) state of the disease with the twofold goal of detecting the two disease phenotypes and identifying those factors that most influence the classification task. To this end, different feature visualization techniques were applied, namely BackPropagation (BP), Guided BackPropagation (GBP) and the Layerwise Relevance Propagation (LRP). Under the assumption that the agreement across these methods is an indication of the robustness of the results, the voxels of the input data mostly involved in the classification decision were identified and their association with clinical scores was assessed, potentially bringing to light brain regions which might reveal disease signatures. Indeed our results highlighted regions such as the Parahippocampal Gyrus, among the others, showing both high consistency across the three visualization methods and a significant correlation with the Expanded Disability Status Scale (EDSS) score, witnessing in favor of the neurophisilogical plausibility of these findings.

12.1 Introduction

In this work we aimed at analysing the consistency across interpretability methods. Three feature visualization methods were compared, namely BP [21], GBP [22] and LRP [32]. A consensus analysis across the feature visualization methods was also carried

out based on Normalized Mutual Information (NMI) [310], under the assumption that consistency across methods would be an indication of the neuroanatomical plausibility of the outcomes. The clinical application was providing hints for the interpretation of the mechanisms at the basis of the MS disease course, building on top of the work presented in Chapter 5 and exploiting XAI for opening new perspectives for diagnosis, prognosis and treatment.

To the best of our knowledge, only our preliminary work [148], attempted the exploitation of 3D-Convolutional Neural Networks (CNN)s to differentiate the Primary Progressive Multiple Sclerosis (PPMS) and the Relapsing-Remitting Multiple Sclerosis (RRMS) applying only the LRP method to detect the most impacting regions to the CNN outcome. In that study we exploited only Grey Matter (GM) features derived from both T1-weighted (T1-w) and diffusion Magnetic Resonance Imaging (dMRI) brain acquisitions for a total of 91 subjects equally split in PPMS and RRMS categories. Our results demonstrated that LRP heatmaps highlighted areas of high relevance which relate well with what is known from literature for the MS disease.

12.2 Materials and Methods

An overview of the whole process is provided in Figure 12.1.

The population is the same as the one employed in Chapter 5 and consisted of 91 subjects, including 46 RRMS (35 females, 52.5 ± 10.4 years old) and 45 PPMS (25 females, 47.2 ± 9.5 years old) patients. For major details about population description and Magnetic Resonance Imaging (MRI) acquisition parameters as well as preprocessing and processing please refer to Section 5.3.1 of Chapter 5. In this comparative study only T1-w MRI was considered.

A Visual Geometry Group (VGG) like 3D-CNN [21] was employed, reflecting the one T1-CNN described in Section 5.3.1 of Chapter 5.

Data augmentation was performed during the training/validation phase in order to improve the generalization capabilities of our model due to the scarcity of the data, following the same approaches of our previous work (sec 5.3.1, Chap 5).

The CNN was trained using a 5-fold CV strategy over a training/validation set of 91 subjects. On each fold, the 91 subjects were randomly split in five groups of 18 subjects each (except one of 19 subjects). The experiment was repeated five times and, for each repetition, four groups were considered as training and the remaining one for validation. The cross-entropy loss was optimized by means of the Adam optimizer [121] during the training phase.

The CNN performance was reported as the average over the validation set of the five models, in terms of accuracy, sensitivity and specificity, while precision for each

Chapter12/Figures/Pipeline.png

Fig. 12.1: Overview of the interpretability and validation pipeline adopted for MS patients stratification. The MS masked T1-w volumes were given as input to a 3D CNN architecture trained and validated through a 5-folds Cross Validation (CV) procedure. The BP, GBP and the LRP maps were derived for 20 subjects, 10 for each class, namely PPMS and RRMS. For each visualization technique a quantitative analysis was carried out by computing region based violin plots across 14 Region Of Interest (ROI)s. In order to assess the robustness of the results, the ROI based Normalized Mutual Information between all the possible combinations of BP, GBP and LRP was computed. Finally, the neuroanatomical plausibility was investigated through the assessment of the Spearman correlation between the relevance value in each ROI and the EDSS, separately for BP, GBP and LRP heatmaps.

class was defined as detailed in what follows. True Positives (TrueP) and True Negatives (TrueN) represent the number of correctly classified PPMS and RRMS subjects, respectively, while False Positives (FalseP) and False Negatives (FalseN) count the wrongly classified RRMS and PPMS, respectively. The class-specific precision was defined as $\text{precision}_{PPMS} = \text{TrueP}/(\text{TrueP} + \text{FalseP})$ and $\text{precision}_{RRMS} = \text{TrueN}/(\text{TrueN} + \text{FalseN})$. The whole Deep Learning (DL) analysis was carried out using the software toolkit Pytorch [122]. The computation was performed on a laptop (Ubuntu 18.6, Nvidia Geforce GTX 1050, Intel Core i7, 16 GB RAM). Torchsample wrapper was used as high-level interface.

Confounding variables influence assessment

To assess the influence of confounding variables on the classification outcome we adopted a *post-hoc* analysis following the method proposed by Dinga and colleagues [140] already presented in Chapter 5 Section 5.4.1. In particular, [140] proposed to control for confounds at the level of Machine Learning (ML) predictions relying on logistic classification models. This allows to understand what information about the outcome can be explained using model predictions that is not already explained by confounding variables. Following [140], this information can be obtained by calculating the Likelihood Ratio (LR), or difference in log-likelihood, of two models: i) the model predicting the outcome using only confounding variables and ii) the model predicting the outcome using both confounding variables and the CNN predictions as calculated during the training phase, under the assumption that the statistical significance of the LR, assessed through a χ-squared test, would reveal that the role of the confounds in shaping the classification outcomes is not prevalent.

In our study this method was used to assess the role of age, sex and EDSS in the differentiation between PPMS and RRMS phenotypes.

12.2.1 Convolutional Neural Networks Visualization methods

To identify the regions on which the CNN model based the classification decision we employed three interpretability methods, that are BP, GBP and LRP (Appendix A). As the classification aims at differentiating two groups of patients, relevance maps were derived for both TrueP (PPMS) and TNs (RRMS) samples and were referred to as *winning class* heatmaps. More in detail, in a multi-class classification task, the CNN prediction function $f(x)$ consists of multiple values indicating the probability for the input x to belong to each of the classes c_i, e.g. $f(x) = \{f_{C_1(x)}, f_{C_2(x)}, \ldots, f_{C_N(x)}\}$ where N represents the total number of classes. Indeed, to obtain relevance maps, a target class has to be defined and the resulting maps are strongly dependent on the class. Let n be the class index and the BP the algorithms used to compute the saliency map. Then, $C_n - BP(x)$ is obtained by backpropagating $R_L = f_{C_n(x)}$ through the network. Following this notation, in this work the prediction $f(x)$ is defined as $f(x) = \{f_{C_{PPMS}(x)}, f_{C_{RRMS}(x)}\}$. In particular, since the two classes share the same importance, there is not a fixed target class. For the correctly classified PPMS subjects, the relative PPMS-BP, PPMS-GBP and PPMS-LRP were calculated starting from the respective $f_{C_{PPMS}(x)}$. On the contrary, for the correctly classified RRMS subjects the RRMS-BP, RRMS-GBP and RRMS-LRP were calculated starting from the relative $f_{C_{RRMS}(x)}$.

In this way, the resulting winning class relevance maps will answer to two questions: (i) "What speaks for PPMS in this subject?", for the subjects correctly classified as PPMS,

and (ii) "What speaks for RRMS in this subject?" for those correctly predicted as RRMS. To cope with the low numerosity of the dataset, the heatmaps were derived for 20 randomly sampled subjects, 10 per class, using the best model out of the 5-folds CV model set, resulting in three maps per subject that were subsequently analysed as detailed hereafter.

12.2.2 Relevance heatmaps analysis

The Captum library [311] was used to compute BP and GBP maps, while the iNNvestigate library [124] was employed for LRP. The relevance maps were registered to the standard MNI space (voxel size = 1 mm isotropic) and averaged over the two groups of patients separately, for visualization purposes.

Fourteen brain ROIs were selected based on MS literature [125, 126, 127, 128]: thalamus (Thal), caudate (Cau), putamen (Put), hippocampus (Hipp), insular cortex (Ins), temporal gyrus (TpG), superior frontal gyrus (SFG), cingulate gyrus (CnG), lateral occipital cortex (LOC), pericalcarine (PCN), lingual gyrus (LgG), cerebellum (Cer), temporal pole (TP) and parahippocampal gyrus (PHG). The reference atlas was the Desikan-Killiany available in FreeSurfer.

The mean relevance values for each of the 14 ROIs was computed, for each of the three heatmaps in the subjects' space per condition. Each heatmap was previously normalized by the respective L^2-norm for direct comparison, following [312].

In order to assess the robustness of the heatmaps as descriptors of the relevance of the different ROIs for the considered task, a consensus analysis was performed across the outcomes of the visualization methods. The underlying assumption is that the agreement across methods witnesses in favour of the robustness, or consistency, of the outcome. However, this does not guarantee the neuroanatomical plausibility of the so detected regions, which needs to be probed relying on additional criteria as will be discussed hereafter. Jointly, such two steps can be regarded as a cross-method validation of the relevance maps.

Assessing the consistency of the heatmaps

To this end, the NMI was used as metric, calculated as presented in [310]. More in detail, given two images I and K, the NMI is calculated as

$$NMI(I;K) = \frac{H(I) + H(K)}{H(I,K)} \tag{12.1}$$

where $H(I)$ and $H(K)$ represent the marginal entropy for the images I and K, respectively, while $H(I,K)$ is the joint entropy of I and K. In this study the entropy was estimated on the probability density function relying on the joint histogram of I and K.

Following this definition, the NMI ranges from 1 to 2, where $NMI = 1$ means independent variables while $NMI = 2$ corresponds to $I = K$. The NMI was calculated at the region level between all the possible combinations of the normalized BP, GBP and LRP.

Assessing the neuroanatomical plausibility of the heatmaps

As explorative analysis, we investigated the plausibility of the outcomes of the three considered feature visualization methods. Inspired by [33] and [26], the Spearman correlation between the average BP, GBP and LRP relevance values for each ROI and the EDSS score were calculated, together with the corresponding p-value, both uncorrected and adjusted with Bonferroni correction for multiple comparisons. A total of 42 comparisons were performed (equal to the number of the considered regions multiplied by the number of feature visualization methods).

12.3 Results

A preliminary analysis revealed that the EDSS score and the age were significantly different between RRMS and PPMS subjects ($p < 0.05$), and thus constituted confounding variables. The same held with gender numerosity ($p < 0.05$), this last observation reflecting the epidemiology of the disease.

The proposed CNN achieved an average accuracy on the validation sets equal to 0.81 ± 0.08 over the five models derived from the 5-fold CV, one for each fold. The sensitivity and specificity were 0.74 ± 0.22 and 0.80 ± 0.11, respectively, showing that the CNN minimized the FalsePs, that is the wrongly classified RRMS subjects. This trend was confirmed by the precision$_{RRMS}$ which was 0.80 ± 0.15 while the precision$_{PPMS}$ was 0.76 ± 0.15.

Concerning the influence of the three confounds on the CNNs classification outcomes, the LR test revealed that the logistic classification model to which the CNN outcomes were added as predictor was significantly different (χ^2 test, $p < 0.05$) from the logistic classification model employing only the confounds as predictors, confirming that the classification was not driven by the confounding variables.

12.3.1 Qualitative Assessment of the Relevance Heatmaps

Figure 12.2 shows the BP, GBP, and LRP heatmaps averaged over the correctly classified subjects for each class, respectively. For ease of visualization, the maps were clipped between the 50th and the 99.5^{th} percentile calculated over the respective target group heatmap. As expected, considering that winning class heatmaps were calculated for each method, high relevance was found in both PPMS and RRMS classes.

Chapter12/Figures/F3.png

Fig. 12.2: BP, GBP and LRP heatmaps obtained from the T1-w based CNN model. The heatmaps are shown for both RRMS and PPMS patients, and are overlaid to the MNI152 template in coronal, sagittal and axial views (columns). Each interpretability map is averaged across the correctly classified RRMS and correctly classified PPMS subjects, respectively. The reported values are clipped to the range 60th–99.5^{th} percentile, calculated over the RRMS and the PPMS class group mean heatmaps.

In general, a shared relevance pattern could be detected across ROIs, with the TrueP in the RRMS maps showing the highest similarity. Considering that the colormap is based on the percentile calculated for each class separately for each method, it is evident that BP shows widespread high voxel sensitivity values that do not correspond to regions of major interest, with the exception of the TrueP for the RRMS. The BP heatmap was more spread and noisier compared to ones resulting from the other feature visualization techniques. A similar pattern was found between GBP and LRP maps for both PPMS and RRMS maps, with the GBP showing overall a more widespread and scattered relevance compared to the LRP.

More in detail of the different techniques, starting from the BP maps, the noisy pattern was particularly evident for the PPMS-BP. Both RRMS-BP and PPMS-BP highlighted higher relevance in the temporal lobe, particularly evident in the coronal and temporal views. Moving to the GBP maps, even if widespread relevance values were present in both classes, the pattern was slightly different. In fact, the RRMS-GBP map showed high activation in the temporal lobe and Cer, as highlighted in both the coronal and the sagittal views. On the contrary, the PPMS-GBP maps showed low relevance in the temporal lobe, while high relevance was assigned to the frontal lobe as can be observed in the sagittal view. The LRP maps replicated the same trend described for GBP. However, a sharper and less scattered pattern was found for LRP maps better highlighting only the most relevant regions.

12.3.2 Quantitative Assessment of the Heatmaps

ROI-based analysis was performed to quantitatively assess the relevant areas for the classification task, as a first step towards the clinical validation of the outcomes. Figure 12.3 illustrates the average L^2-norm normalized relevance per ROI for the correctly classified patients, separately for the two classes and for the three visualization methods adopted in this study.

Starting from a general overview, a similar trend can be detected between the BP, GBP and LRP, all showing high relevance for both subject classes in regions such as TP, Ins, Cer and Hipp, with the LRP maps showing a generally higher relevance score. The RRMS mean relevance values were consistently higher compared to the PPMS ones for all the feature visualization methods, with the exception of the SFG, LOC and LgG where the PPMS relevance mean values were higher than the RRMS one. The BP maps median relevance values for the two classes were highly overlapped in almost all the considered regions. On the contrary, the GBP and LRP maps showed a distinct relevance distribution for the two classes, as it is particularly evident in the Hipp where the two distributions resulted completely disjoint, with higher difference for GBP maps.

Chapter12/Figures/F4.png

Fig. 12.3: Size-normalized importance metric extracted from the GBP, BP and LRP maps (columns). The mean relevance value for each region is reported for all the correctly classified RRMS and RRMS subjects. The median relevance for PPMS (orange circle) and RRMS (blue circle) groups are also shown. The relevance values are also normalized by the L2 norm for direct comparison.
Abbreviations: thalamus (Thal), caudate (Cau), putamen (Put), hippocampus (Hipp), insular cortex (Ins), temporal gyrus (TpG), superior frontal gyrus (SFG), cingulate gyrus (CnG), lateral occipital cortex (LOC), pericalcarine (PCN), lingual gyrus (LgG), cerebellum (Cer), temporal pole (TP) and parahippocampal gyrus (PHG).

Stability analysis

The consensus analysis was performed to assess differences and similarities across the three feature visualization methods. Figure 12.4 shows the NMI obtained for 14 brain regions. The NMI was calculated on the L^2-norm normalized relevance heatmaps, for the three combinations, namely BP versus GBP, BP versus LRP and GBP versus LRP. In general, a similar NMI trend can be observed across the methods, with the cortical regions showing a higher NMI compared to the subcortical ones. The highest NMI was found, as expected, between BP and GBP maps, which showed an NMI value greater than 1.2 for all the ROIs. More in detail, the PHG resulted as the region featuring the highest similarity between the heatmaps derived from the two methods, followed by PCN, Ins, CnG and TpG. Of note, the Cer showed the lowest variability in the NMI across subjects. Moving to the similarity between LRP and the two gradient-based methods, the NMI resulted generally lower for the comparison between BP and LRP compared to the GBP vs LRP, though sharing the same trend. Noteworthy, the RRMS showed higher similarity across methods compared to the PPMS class. The TrueP appeared as the most similar for both the comparisons, followed by the PHG and the Ins. The subcortical regions showed the lowest NMI, with the exception of the Hipp which instead showed high NMI values when comparing BP and LRP, and GBP and LRP, respectively.

Chapter12/Figures/F5.png

Fig. 12.4: Region based NMI metric extracted from the comparison between BP and GBP, BP and LRP, GBP and LRP (columns). The NMI value for each ROI is reported for all the correctly classified PPMS (orange dots) and RRMS subjects (blue dots).
Abbreviations: thalamus (Thal), caudate (Cau), putamen (Put), hippocampus (Hipp), insular cortex (Ins), temporal gyrus (TpG), superior frontal gyrus (SFG), cingulate gyrus (CnG), lateral occipital cortex (LOC), pericalcarine (PCN), lingual gyrus (LgG), cerebellum (Cer), temporal pole (TP) and parahippocampal gyrus (PHG).

Neuroanatomical plausibility

The Spearman correlation analysis between the ROI-wise mean relevance values for BP, GBP and LRP and the EDSS scores are reported in Table 12.1. Significant positive correlations (p-value < 0.05, uncorrected) were detected for the SFG, for the three interpretability methods. Among the other regions, PHG, TP, Ins, TpG, and Hipp showed a significant negative correlation with EDSS for both GBP and LRP heatmaps. In particular, the largest negative correlation with the EDSS, was found for LRP and GBP mean relevance in the PHG ($\rho = -0.81$ and -0.74, respectively), followed by the TpG which showed a negative ρ value of -0.65 and -0.64 for GBP and LRP, respectively. Finally, when applying Bonferroni correction for multiple comparisons, the PHG still showed a significant correlation with EDSS for both GBP and LRP (adjusted p-value of 0.011 and 0.001, respectively).

12.4 Discussion

In this work we addressed the stratification problem between RRMS and PPMS patients based on T1-w data. A 3D-CNN was used to this aim, and three different interpretability methods for feature visualization were applied and compares, namely BP, GBP and

Table 12.1: Spearman correlation results between the mean BP, GBP and LRP heatmap values for each ROI and the EDSS score (uncorrected p-values).
Abbreviations: thalamus (Thal), caudate (Cau), putamen (Put), hippocampus (Hipp), insular cortex (Ins), temporal gyrus (TpG), superior frontal gyrus (SFG), cingulate gyrus (CnG), lateral occipital cortex (LOC), pericalcarine (PCN), lingual gyrus (LgG), cerebellum (Cer), temporal pole (TP) and parahippocampal gyrus (PHG).

	BP		GBP		LRP	
	ρ	p-value	ρ	p-value	ρ	p-value
Ins	-0.06	0.782	**-0.56**	**0.015**	**-0.56**	**0.012**
PCN	**-0.61**	**0.005**	-0.19	0.43	0.13	0.596
SFG	**0.52**	**0.022**	**0.58**	**0.008**	**0.52**	**0.021**
CnG	0.11	0.664	-0.34	0.148	-0.26	0.285
PHG	-0.29	0.221	**-0.74**	**0.001**	**-0.81**	**3e-05**
TP	-0.32	0.185	**-0.57**	**0.011**	**-0.56**	**0.013**
LOC	0.12	0.630	0.40	0.088	0.30	0.214
LgC	-0.16	0.505	-0.08	0.726	**-0.54**	**0.016**
TpG	-0.18	0.459	**-0.65**	**0.003**	**-0.64**	**0.003**
Thal	0.35	0.133	-0.08	0.746	0.13	0.599
Cau	0.09	0.686	-0.31	0.196	-0.36	0.132
Put	-0.08	0.729	**-0.55**	**0.016**	**-0.54**	**0.017**
Hipp	-0.31	0.188	**-0.59**	**0.008**	**-0.59**	**0.007**
Cer	0.04	0.890	**-0.52**	**0.022**	-0.43	0.068

The ρ score and relative p-values (rows) are reported for each Region Of Interest (columns). The significant correlations (p-value < 0.05) are highlighted in bold. The correlations surviving the Bonferroni correction are shown underlined.

LRP, in order to assess their consistency across methods, strengths and weaknesses as well as highlighting the key brain regions involved in the classification of the two patients populations. Then, Spearman correlation was used to assess the concordance between the ROI-wise mean relevance and the individual EDSS scores for each of the 14 considered ROIs.

We already stated in our previous works the performance of the classification task between MS subtypes held by the combination of T1-w and CNNs, as well as its comparison with state-of-the-art performance (Chap 5)

A 3D-CNN based approach was proposed in [33], showing an accuracy of 87.04% on a set of 147 fully volumetric structural MRI acquisitions. Despite the lower accuracy compared to the 2D-CNN based approaches, the use of a 3D-CNN architecture facilitated the interpretation of the CNN performance through the use of feature visualization techniques. However, the essential difference in the research question makes these works not directly comparable to ours. Concerning the feature visualization, [111] compared different techniques applied to a 2D-CNN trained on 66 healthy controls and 66 MS patients Susceptibility-Weighted Imaging (SWI) data. Their results highlighted

the superiority of LRP and Deep Learning Important FeaTures (DeepLIFT) [112] over simpler methods, relying on the perturbation based analysis on the derived heatmaps, strengthening the exploitability of such methods to address clinically relevant questions.

The application of the BP, GBP, and LRP provides a means for CNNs interpretability and, when used in combination with other clinical and imaging data, could support diagnosis and treatment decisions. By relying on these techniques, it was possible to identify the regions playing a prominent role in the classification between the two MS phenotypes as the regions of highest difference in relevance across groups. In our results, the relevance maps of both RRMS and PPMS showed that BP was highly sensitive for both classes, highlighting a scarce class related relevance which was instead found for the other methods. Moreover, high relevance was found also in brain regions that were masked in the input, revealing the noisy pattern of BP maps and rising a warning on the interpretation of such results. We recall that only the GM-masked T1-w values were given as input to the CNN model. A more focused pattern was instead found for GBP and LRP which consistently showed relevance values only in the GM regions of the input volume. More in depth of the ROI-based analysis of the three feature visualization methods, the ROIs leading the CNN classification were coherent across methods and were in agreement with the clinical literature findings. Indeed, TP, which showed the highest relevance for BP, GBP, and LRP, as well as the highest NMI for all the comparisons, has been reported to be present in MS cortical atrophy patterns [144]. The Ins, which was the second region for mean relevance value for all methods and NMI for all the comparisons, have been shown to reveal high probability of focal GM demyelination in MS pathology [126]. The Cer has been demonstrated to be a major site for demyelination, especially in PPMS patients [143]. Finally, an important feature for MS is a lower diffusion restriction and massive neuronal loss and demyelination in Hipp [145, 51], which was a region holding high relevance for all the three feature visualization methods, as well as high NMI for all the considered comparisons.

It is important to note that the focus in interpreting feature visualization maps was not on the absolute values of the relevance, but on the differences and overlaps between the violin plots of the considered ROIs in the two classes of patients. This means that the relevance values allowed to understand how the voxels of certain ROIs contributed to the classification, but still did not allow identifying the subserving mechanism (lesion load, atrophy, etc.) [26].

In general the NMI between all the possible couplings of the three methods resulted, as expected, the highest between BP and GBP, being the two methods both based on gradients and having a similar computation with the exception for the ReLU layers. This was confirmed by the larger difference found instead between both BP and GBP, with

respect to LRP. In fact, the LRP is not directly based on gradients but on the backpropagation of the prediction values constrained by the relevance conservation rule. Of interest, higher NMI was found for the RRMS, particularly in the comparison between the gradient based methods and LRP, reflecting the clinically assessed higher variability in PPMS subjects compared to RRMS. However, the ROIs showing the highest NMI were the same for all the combinations, confirming their importance for the CNN outcome and providing evidence of the plausibility of the results.

In order to investigate whether high importance scores could correspond to clinically assessed differences across classes, the association between the mean relevance value in each ROI and the EDSS was also computed. All the regions featuring high relevance and NMI across methods showed also a significant Spearman correlation with EDSS. Of note, the PHG was the only region surviving the Bonferroni correction showing a significant and negative correlation with the EDSS, although it was not among the ROIs showing the highest relevance. For clinical assessment, this region has been associated with fatigue, particularly in RRMS [79].

Although deeper investigation would be needed to drive strong conclusions, these results, together with [26] and [33], provide evidence of the potential of the joint exploitation of CNNs and visualization methods for identifying relevant disease biomarkers for the considered disease phenotypes, as well as of the core role of visualization methods in pursuing the objective assessment of the plausibility and clinical relevance of the results.

12.4.1 Limitations and future works

One of the obvious possible improvements to our work would be the increase of the number of subjects, even though the classification performance and the consensus across the heatmaps obtained by different visualization methods witness in favor of the robustness and reliability of the results. Moreover, for the consistency analysis, different metrics are being proposed in literature focusing not only on assessing the similarity between heatmaps, but also on their consistency and bias, such as the mutual verification proposed in [312]. In a clinical context it is mandatory to obtain highly reliable and still understandable explanations in order to spread the use of ML and DL methods.

Another possible improvement would be single-subject analysis. Since the interpretability methods adopted provide a heatmap for each subject indicating the contribution of each voxel to the final classification decision, a subject specific analysis could be carried out, moving step forward the personalized precision medicine.

Then, additional feature visualization methods could be exploited, such as DeepLIFT analyzed also in [111], in order to further investigate the consensus across more advanced interpretability methods.

Overall, we consider these outcomes as a valuable evidence of the potential of the proposed method in splitting apart the two MS phenotypes and providing hints on the signatures of possible subserving mechanisms of disease progression. We leave the open issues mentioned above for future investigation.

12.5 Conclusions

This work corroborated the capability of T1-w combined with a 3D-CNN classifier in distinguishing the different typologies of MS disease. In addition, we could highlight, through the application and the consistency analysis across the three considered feature visualization techniques, that the CNN classification was based on ROIs holding clinical relevance whose heatmap NMI was high and which mean values significantly correlated with EDSS score. From a clinical perspective, our results strengthen the hypothesis of the suitability of GM features as biomarkers for MS pathological brain tissues. Moreover, this work has the potential to address clinically important problems in MS, like the early identification of the clinical course for diagnosis, personalized treatment and treatment decision.

The work presented in this Chapter will soon be published as book Chapter in [313].

Part V

Conclusions and future directions

13

Conclusions and future work

The works illustrated in this Thesis show how it is possible to enable Artificial Intelligence (AI) in medicine through the utilization of eXplainable Artificial Intelligence (XAI) resulting in a positive impact in solving multiple practical clinical applications, characterized by highly complex underlying mechanisms.

In the following Sections, we summarize the contributions of this Thesis proposing research perspectives in the field of the XAI application for joint modeling of heterogeneous data, as well as XAI validation strategies.

13.1 Summary of the main contributions

Proposition of XAI taxonomy

It is clear that there is still a lack of consensus in the literature on the taxonomy of XAI methods. In Part I, we clarified the main concepts related to XAI, especially differentiating interpretability from explainability, each relating to clear and different XAI approaches and not being used as synonyms. It is indeed fundamental to differentiate between methods that are explainable by design (*explainable*) and methods that can be interpreted *post-hoc* (*interpretable*). This is crucial for the choice of the right method for a given problem helping to better understand when to use one or the other approach. We also moved a step further defining four validation attributes that encompass the different proxies present in literature namely *stability*, *consistency*, *plausibility*, and *understandability* which could allow the evaluation of the quality of the explanations provided by existing XAI methods. Importantly, our contribution is also the addition of *plausibility* to the list which is particularly relevant in the medical field. Stated that explainability methods are vital to gain a deep understanding of Machine Learning (ML) and Deep Learning (DL) model predictions, their application must be faced

with caution, assessing the four characteristics presented above in order to be practically used and earn reliable and meaningful insights on the approached problem. The work presented in this part was published in [51]

XAI for subject stratification in Multiple Sclerosis

In Part II of this thesis we presented multiple approaches to study Multiple Sclerosis (MS) for uncovering different MS stages based either on cognitive impairment or MS type, namely Relapsing-Remitting Multiple Sclerosis (RRMS) and Primary Progressive Multiple Sclerosis (PPMS).

In Chapter 4, we proposed a multivariate model to estimate the joint variation of diffusion Magnetic Resonance Imaging (dMRI) derived indices and neuropsychological tests scores, demonstrating that different forms of cognitive impairment were qualitatively distinguishable in the latent space created by the Partial Least Squares (PLS) regression. Thus, multivariate approaches to statistical analysis combining neuroimaging and clinical studies may have a potential in depicting subtle differences in different forms of the MS pathology.

Moving a step forward, in Chapter 5, exploiting all the volumetric information of both T1-weighted (T1-w) and dMRI derived indices thanks to the use of a 3D Convolutional Neural Networks (CNN) and Layerwise Relevance Propagation (LRP) visualization, we have observed that the CNNs based disease state detection relied on clinically relevant Region Of Interest (ROI)s and that different indices were sensitive to Grey Matter (GM) modulation in different brain regions. We hence provided evidence in favor of the capability of dMRI indices of distinguishing different stages of the disease in MS. In particular, we were able to prove that 3D Simple Harmonics Oscillator based Reconstruction and Estimation (3D-SHORE) based indices and RIF1 outperformed Fractional Anisotropy (FA) and Mean Diffusivity (MD), pushing to shift the attention on dMRI features other than Diffusion Tensor Imaging (DTI) ones. Our results, thanks to the use of XAI, support the hypothesis of dMRI based indices suitability as numerical biomarkers for the characterization of pathological brain tissues.

Importantly, in this works we devised validation strategies based not only on model significance or generalizability (eg. permutation test for PLS or 5-folds Cross Validation (CV) for 3D-CNN models) but also on confound influence assessment, far from trivial when dealing with volumetric data, as well as on neuroanatomical *plausibility* by assessing the explanations correlation with well established features relevant for MS.

The works presented in this part were published in [96, 148, 147] and have the potential to address clinically important problems in MS, like the early identification of the clinical course for diagnosis and provides evidence in favor of the feasibility of precision medicine, through the adoption of XAI techniques.

XAI for Imaging Genetics in Alzheimer's Disease (AD) continuum

In Part III of this thesis we presented multiple approaches for applying XAI to study Imaging Genetics (IG) in AD continuum. We firstly presented our studies based on PLS, an explainable multivariate model which allowed to uncover the association between imaging derived Imaging Derived Phenotype (IDP)s and genetics and for which explanations were directly retrieved through the observation of PLS weights. All the PLS models were validated through a permutation test which resulted significant for both structural Magnetic Resonance Imaging (sMRI) and functional Magnetic Resonance Imaging (fMRI) based studies, as well as CV strategies to assess the number of Latent Variable (LV) needed to explain at least the 60% of data variability.

More in detail, in Chapter 6 we exploited 14 different Polygenic Risk Score (PRS) for AD, calculated by including Single Nucleotide Polymorphism (SNP)s passing different significance thresholds, to check their association with brain ROI volumes and thicknesses. The PLS model was applied to a large study cohort obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database including both healthy individuals and AD patients, and validated on an independent ADNI Mild Cognitive Impairment (MCI) cohort, including Early and Late Mild Cognitive Impairment (LMCI) subjects. The experimental results confirmed the existence of a joint dynamics between brain atrophy and genotype data in AD, while providing important generalization results when tested on a clinically heterogeneous cohort. In particular, less AD specific PRS scores were negatively correlated with cortical thicknesses, while highly AD specific PRSs showed a peculiar correlation pattern among specific subcortical volumes and cortical thicknesses. While the first outcome is in line with the well known neurodegeneration process in AD, the second could be revealing of different AD subtypes.

In Chapter 7 we relied on only two PRSs to check their association separately with dMRI and fMRI derived IDPs. dMRI IDPs were composed by tract based features calculated over four DTI based indices, namely FA, MD, Radial Diffusivity (RD) and Axial Diffusivity (AxD), while fMRI IDPs were represented by within/between network connectivities. The study cohort was as well based on the ADNI database, focusing on the ADNI-3 phase, and comprehended healthy subjects and MCI individuals, including subsets featuring patients showing early and late condition. Different subjects were used for the dMRI and fMRI studies due to data availability. The experimental results showed an anti-correlation between diffusivity and anisotropy in the WM tracts typical of neurodegeneration, to which both the PRS were correlated. Moreover, thanks to the fMRI based analysis, we retrieved the correlation between the PRS2 and connectivities involving the dorsal attention (DAN) and frontoparietal control (CON). Visual (VIS) and somatomotory (SMN) showed a correlated trend, while being anti-correlated with limbic (LIM), CON and default mode network as well as with PRS1.

Such findings suggest that the two PRSs correlated with a possible pattern of aberrant within/between-network Functional Connectivity (FC) changes occurring in Resting State Network (RSN)s devoted to higher cognitive functions and more vulnerable in this pathology.

It is evident that the presented PLS models confirmed the existence of a joint variation between grey matter atrophy, White Matter (WM) diffusivity, FC and PRS in AD, each resulting in a different view of the disease, confirming the necessity to look towards interpretable heterogeneous data integration models.

In order to better deal with a small study cohort while aiming at summarizing the genetic information in a more informative way, in Chapter 8 we introduced the gene based variant scores and we verified its associations with sMRI derived IDPs. The gene variant scores were calculated for all the significant genes resulting from Sequence Kernel Association Test (SKAT) filtering which allowed to obtain 408 scores for genes selected based on their potentiality in distinguishing control subjects from AD patients in various stages. The proposed gene variant score allowed to obtain informative genetic features which did not result from the classical genome-wide analysis. Relying on such scores and the sMRI IDPs we were able to obtain a significant PLS model, validated on an unseen cohort of subjects where the obtained IDPs-gene variant scores associations succeeded in differentiating the two groups of patients in their latent space projection. Moreover, a transcriptomic analysis was carried on to assess the brain expression of the resulting important genes. Results highlighted meaningful genotype-phenotype interaction in each significant latent component. Among the others, the correlation between the EPHX1 associated variant score, highly investigated for its role in neurodegeneration, and a decrease in subcortical volumes, as well as the association between the BCAS1 variant score, gene involved in the process of myelination and found particularly expressed in dentate gyrus, and a significant decrease in temporal lobe thickness (Patients (PAT) < Controls (CN)) were retrieved. With this Chapter we proposed new ways to investigate the genotype-phenotype interactions in a restricted study cohort highlighting associations that are descriptive of the underlying mechanisms of neurodegeneration in AD continuum. In addition, we proposed different validation strategies for the assessment of the neurobiological *plausibility* of the obtained explanations, such as the transcriptomic analysis on the most significant genes.

Finally, in Chapter 9 we aimed at fusing the multiple IDPs and genetics together in the Multi-channel Variational Autoencoder (MCVAE), while proposing an interpretable framework for generative models. We hence exploited a three channels MCVAE equipped with ROI based volumes and thicknesses as the sMRI channel, tract based FA and MD values for dMRI and the gene variant scores for genetics. We firstly proposed a metric based on the cosine distance to compare the different LVs among different MCVAEs. We

also analyzed the obtained latent space and retrieved the best decoded features separately for each channel starting from the other two. Finally, we proposed a modification of the MCVAE which allowed to apply the SHapley Additive exPlanations (SHAP) model to retrieve which features of the input channel mostly contributed to the reconstruction of the features in the output. Despite the reconstruction performance of the MCVAE were not satisfying, the obtained latent space was interesting: a clustering between different classes was present and the latent spaces generated from each channel were well aligned. Thanks to the application of the SHAP model we were able to verify that the most relevant genetic features for the decoding of the imaging channels well mapped with the results presented in Chapters 8 and 7.

The work presented in this part allowed to obtain a wide overview on AD continuum, including different imaging techniques and genetic features. We showed how the adoption of simple yet explainable methods such as PLS, even in presence of a small study cohort, has the potential to uncover meaningful genotype-phenotype associations, strengthened by the multiple model validations as well as explanation validation through the *plausibility* assessment. Moreover, we presented an interoperability framework based on the MCVAE equipped with SHAP which has the potential to open the way to interpretable data fusion.

Moreover, the Chapter presented in this Part were published in [182, 164, 157] while two journal papers [255, 245] and one conference proceeding [314] are in preparation and will be soon submitted.

Open challenges in XAI: validation strategies

Part IV was devoted to XAI methods validation. In detail, with the work presented in Chapter 10 we presented a simple yet useful approach to evaluate any XAI method and quantify the **stability** of the list of informative predictors. The proposed measure, the Normalized Movement Rate (NMR), can be easily employed to determine how much it is possible to trust the ordered list of informative predictors given by a specific set of features, classifier and explainability method. The method was applied to the task of classifying AD from control subjects based on different classifiers such as Likelihood Ratio (LR), Support Vector Classifier (SVC), and Decision Tree (DT). Our findings on SHAP demonstrated that if the original set of features presents high correlation, a certain degree of instability would be present in the rankings from all the models though SVC is the classifier leading to the lowest NMR and thus having the most stable lists. Conversely, when applying Principal Component Analysis (PCA) on the feature set, the new set of uncorrelated variables leads to stable rankings for most of the classifiers. The strength of the presented pipeline is that it can be implemented in any domain with different models and XAI methods to evaluate the reliability of the feature rankings.

Concerning XAI methods *consistency*, in Chapter 11 we proposed to compare four different XAI *post-hoc* methods on Random Forest (RF) model aiming at predicting post-stroke UL functional recovery following rehabilitation. Predictive models leading to accurate estimates while revealing which features contribute most to the predictions are the key to unveiling the mechanisms subserving the post-intervention recovery, prompting a new focus on individualized treatments and precision medicine in stroke. The proposed RF was equipped with four XAI methods, namely Random forest Feature Importance (RFI), Permutation Feature Importance (PFI), Local Interpretable Model-Agnostic Explanations (LIME) and SHAP were applied to interpret the results and assess the features' relevance and *consistency*. Our results revealed increased performance when using ML compared to conventional statistical approaches. Moreover, the features deemed as the most relevant were concordant across the XAI methods, suggesting good stability of the results. Our findings highlight the core role of ML not only for accurately predicting the individual follow-up outcome scores after rehabilitation but also for making ML results interpretable when associated to XAI methods. This provides clinicians with robust predictions and reliable explanations that are key factors in the therapeutic planning/monitoring of stroke patients. From a clinical perspective, our findings may improve the management of sub-acute stroke patients in a rehabilitation unit.

Finally, in Chapter 12 we proposed the XAI **consistency** analysis for visualization methods applied to fully volumetric data with the clinical outcome being the stratification of MS patients. Different feature visualization techniques were applied, namely BackPropagation (BP), Guided BackPropagation (GBP) and LRP to a 3D-CNN fed with T1-w Magnetic Resonance Imaging (MRI), masked in order to retain only GM. After model validation through 5-folds CV and *post-hoc* confound assessment, under the assumption that the agreement across explanations derived from different XAI methods is an indication of the robustness of the results, we calculated a consensus metric, the Normalized Mutual Information (NMI), across each couple of XAI methods. The voxels of the input data mostly involved in the classification decision were identified and their association with clinical scores was assessed, potentially bringing to light brain regions that might reveal disease signatures. Indeed our results highlighted regions such as the Parahippocampal Gyrus, among the others, showing both high *stability* across the three visualization methods and a significant correlation with the Expanded Disability Status Scale (EDSS) score, witnessing in favor of the neurophysiological *plausibility* of these findings.

The contribution of this part was supported by multiple publications [264, 309, 313]

13.2 Publications

This thesis led to the development of the following publications:

Book Chapters

[313] Federica Cruciani, Lorenza Brusini, Mauro Zucchelli, G Retuci Pinheiro, Francesco Setti, I Boscolo Galazzo, Rachid Deriche, Leticia Rittner, Massimiliano Calabrese, and Gloria Menegaz. "Explainable Deep Learning for decrypting disease signatures in Multiple Sclerosis". In: *Explainable Deep Learning AI Methods and Challenges*. Elsevier, 202.

International Journals

[51] Ilaria Boscolo Galazzo, Lorenza Brusini, Muge Akinci, Federica Cruciani, Marco Pitteri, Stefano Ziccardi, Albulena Bajrami, Marco Castellaro, Ahmed MA Salih, Francesca B Pizzini, et al. "Unraveling the MRI-Based Microstructural Signatures Behind Primary Progressive and Relapsing–Remitting Multiple Sclerosis Phenotypes". In: *Journal of Magnetic Resonance Imaging* 55.1 (2022), pp. 154–163.

[148] Federica Cruciani, Lorenza Brusini, Mauro Zucchelli, G Retuci Pinheiro, Francesco Setti, I Boscolo Galazzo, Rachid Deriche, Leticia Rittner, Massimiliano Calabrese, and Gloria Menegaz. "Interpretable deep learning as a means for decrypting disease signature in multiple sclerosis". In: *Journal of Neural Engineering* 18.4 (2021), 0460a6.

[309] Marialuisa Gandolfi, Ilaria Boscolo Galazzo, Rudy Gasparin Pavan, Federica Cruciani, Alessandro Picelli, Silvia Francesca Storti, Nicola Smania, and Gloria Menegaz. "eXplainable AI allows predicting upper limb rehabilitation outcomes in subacute stroke patients". In: *IEEE Journal of Biomedical and Health Informatics* (2022).

International Journals in preparation

[245] Federica Cruciani, Antonino Aparo, Lorenza Brusini, Carlo Combi, Silvia F. Storti, Rosalba Giugno, Gloria Menegaz, and Ilraia Boscolo Galazzo. "Identifying the joint signature of brain atrophy and gene variant scores in the Alzheimer's Disease". In: *Journal of Biomedical Informatics* (In submission).

[255] Federica Cruciani, Ettore Cinquetti, Lorenza Brusini, Antonino Aparo, Carlo Combi, Ilaria Boscolo Galazzo, and Gloria Menegaz. "Exploring the potential of MCVAE for patients stratification and skewed data compensation across the AD continuum". In: (In preparation).

Conference Proceedings

[96] Lorenza Brusini, Federica Cruciani, Ilaria Boscolo Galazzo, Marco Pitteri, Silvia F Storti, Massimiliano Calabrese, Marco Lorenzi, and Gloria Menegaz. "Multivariate data analysis suggests the link between brain microstructure and cognitive impairment in multiple sclerosis". In: *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2021, pp. 685–688.

[147] Federica Cruciani, Lorenza Brusini, Mauro Zucchelli, Gustavo Retuci Pinheiro, Francesco Setti, Ilaria Boscolo Galazzo, Rachid Deriche, Leticia Rittner, Massimiliano Calabrese, and Gloria Menegaz. "Explainable 3D-CNN for multiple sclerosis patients stratification". In: *International Conference on Pattern Recognition*. Springer. 2021, pp. 103–114.

[157] Heba Elshatoury, Federica Cruciani, Francesco Zumerle, Silvia F Storti, André Altmann, Marco Lorenzi, Gholamreza Anbarjafari, Gloria Menegaz, and Ilaria Boscolo Galazzo. "Disentangling the association between genetics and functional connectivity in Mild Cognitive Impairment". In: *2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*. IEEE. 2021, pp. 1–4.

[164] Federica Cruciani, Andre Altmann, Marco Lorenzi, Gloria Menegaz, and Ilaria Boscolo Galazzo. "What PLS can still do for Imaging Genetics in Alzheimer's disease". In: *2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*. IEEE. 2022, pp. 1–4.

[264] Ahmed Salih, Ilaria Boscolo Galazzo, Federica Cruciani, Lorenza Brusini, and Petia Radeva. "Investigating Explainable Artificial Intelligence for MRI-based Classification of Dementia: a New Stability Criterion for Explainable Methods". In: *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2022, pp. 4003–4007.

Conference Proceedings in preparation

[314] Federica Cruciani et al. "An interpretability framework for MCVAE: preliminary results of the application to Imaging Genetics in Alzheimer's disease". In: *VIII*

Congress of the National Group of Bioengineering (GNB) 2023. GNB. In submission.

Conference Abstracts

[182] Federica Cruciani, Lorenza Brusini, Giorgio Dolci, Ilaria Boscolo Galazzo, Andre Altmann, Marco Lorenzi, and Gloria Menegaz. "A multivariate imaging study for decrypting the link between polygenic risk scores and microstructure in Mild Cognitive Impairment". In: *2021 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*. IEEE. 2021, pp. 1–4.

13.3 Future developments

This work opened the way to other important aspects and perspectives that may be worth exploring. We already presented possible future directions in each Chapter and we will summarize them hereafter.

- From Part I still some work is needed to achieve a generalized knowledge about XAI, as well as a critical application of validation methods as presented also in Part IV. It would be of high interest to generate an XAI framework for *post-hoc* analysis applicable to any model or input. The framework should include both the phases of explanation extraction, XAI validation and eventually the feedback loop for model improvement. There exist frameworks for the first two steps separately. For example, the python library *captum* [311] allows to obtain different explanations relying on the state-of-the-art interpretability methods, focusing on brain imaging. For the validation phase the library *quantus* [12] was recently proposed allowing the calculation of different state-of-the-art validation metrics. However, a complete framework accounting for both tasks and adding the possibility to exploit explanations to improve the model, as well as including explanations *plausibility* which is fundamental in healthcare, is still missing.
- Concerning the application of XAI to the specific task of MS patients stratification, following the works presented in Part II, future works will be focused on the development of an interpretable model accounting for the information derived from all the available input data which comprehends both dMRI and sMRI. Moreover, the lack of healthy controls is a reason of concern as it impedes benchmarking the performance of the proposed architecture in the patients versus controls classification task. This would allow also to assess the relevance of the voxels and ROIs in distinguishing

each of the phenotypes (PPMS, RRMS) from healthy matched controls which could reveal shared features of the two manifestations of the pathology that could not be captured by the proposed analysis yet potentially being insightful for understanding the mechanisms of the disease. Finally, including WM in the analysis would widen the spectrum of the microstructural features potentially distinguishing the two disease phenotypes as well as unravel the link with GM tissue modulations. Nevertheless, we consider these outcomes as the valuable first evidence of the potential of the proposed method in splitting apart the two MRI phenotypes and providing hints on the possible subserving mechanisms of disease progression, and we leave the open issues mentioned above for future investigation.

- Moving to XAI application in IG we already performed different analysis including different feature sets with the PLS model. However, given the data availability of ADNI dataset, in particular ADNI-3 phase including the most complete set of imaging acquisition in ADNI database, the fusion of different IDPs together would provide different views on the AD related brain modulations. This approach was attempted through the adoption of the MCVAE model which however could be improved through the utilization of an prior latent space distribution matching each input distribution, relaxing the gaussianity assumption for each channel that was imposed in our work. This would allow to better reflect the input feature distributions and hence allowing to reach also better reconstruction performance. In this respect, it is worth mentioning the recently proposed heterogeneous longitudinal Variational AutoEncoder (VAE) (HL-VAE) that extends the existing temporal and longitudinal VAEs to heterogeneous data providing efficient inference for high-dimensional datasets and including likelihood models for continuous, count, categorical, and ordinal data while accounting for missing observations [254], which could be an interesting path to explore.
- From a feature point of view, in particular for genetics, the gene variant score was computed on all SNPs located in the same gene with each SNP being equally weighted. It could be interesting to modify the gene variant score in order to weigh the SNPs differently based, for example, on the associations of individual SNPs with the disease or imaging phenotype (i.e. the p-value obtained from Genome-Wide Association Study (GWAS)).
- Other than the points expressed above, when dealing with diseases, the future direction is not only the inclusion of different modalities in an interpretable model but also the generation of a biomarker, accounting for all the different input features, expressing the disease status. Disease Progression Modelling (DPM) has existed for around ten years. In response to the then-emerging hypothetical models of AD progression [315, 151] that resemble the hypothetical cascade of dynamic biomarkers,

computer science researchers took on the challenge of producing quantitative and actionable disease signatures. Of high interest, recently Abi Nader and colleagues [249, 253] presented SimulAD, a computational DPM originally developed on the ADNI database to simulate the evolution of clinical and imaging markers characteristic of AD, and to quantify the disease severity of a subject. Their framework is based on the modeling of the spatio-temporal dynamics governing the joint evolution of imaging and clinical biomarkers along the history of the disease, and allows the simulation of the effect of intervention time and drug dosage on the biomarkers' progression. When applied to multi-modal imaging and clinical data from the ADNI cohort the method enables to generate hypothetical scenarios of amyloid lowering interventions. The biomarkers used in their studies included sMRI, fMRI, Positron Emission Tomography (PET) and clinical scores. The inclusion of microstructure measures derived from dMRI, being highly informative for the disease as shown in this thesis, could add precious information to the obtained DPM. We already started following this path during my three months visit to EPIONE research group based at INRIA Sophia Antipolis, however the low number of subjects having dMRI in multiple timepoints still not allowed to properly include them in the model. This is still a promising direction applicable not only to AD but also to other neurodegenerative diseases such as the MS object of this thesis and it would be of high interest to add also the XAI step to this framework to deeply uncover feature contributions and associations to the DPM generation.

13.4 Final remarks

For AI pervading medicine still some work has to be done, however, thanks to the continuous and tireless research into XAI an increasing awareness and critical use of AI methods is making its way into real-life applications. With this Thesis, we clarified XAI concepts and showed real-life XAI applications in medical framework, as well as proposed validation strategies in order to build trust on the obtained results. The advancement of mathematical and statistical techniques, as well as the adoption of XAI methods and validation strategies will enable the development of personalized medicine and treatment by taking a data-driven and objective approach to healthcare.

Background

A

Appendix - Background

A.1 Data and feature extraction

A.1.1 Brain imaging

The choice of the feature type is strongly linked to the approach and more than everything to the data availability. On the imaging side, the focus will mainly be on two techniques of Magnetic Resonance Imaging (MRI) acquisitions: structural Magnetic Resonance Imaging (sMRI) and diffusion Magnetic Resonance Imaging (dMRI) which will be detailed in the following subsections.

Structural MRI (sMRI)

With the term sMRI we refer to the classical MRI acquisitions which could be T1- or T2-weighted according to the acquisition parameters.

As the name suggests, MRI utilizes the magnetic properties of molecules, specifically hydrogen, and other naturally occurring elements to record the brain. Structural MRI (sMRI) is the main modality obtained during an MRI scan session and displays the contrast between different tissues in the brain. An MRI scanner essentially consists of a large magnet that aligns protons in your molecules with its magnetic field. Perturbing these protons and measuring how they react to those perturbations allows us to distinguish between different tissues. For example, tissues that contain fat will show up as light voxels in a T1-weighted (T1-w) MRI scan. T1-w is a method of measuring what happens after the perturbations. Tissues or areas that contain water, and thus more hydrogen atoms, will be dark on a T1-w sMRI scan.

Most studies dealing with sMRI on different data exploited Region Of Interest (ROI) based features, like volume and thickness [42, 158, 247].

Diffusion MRI (dMRI)

dMRI is a non-invasive imaging method that is able to provide information in-vivo on the cerebral tissue microstructure and cytoarchitecture. This method has been established over the last two decades as one of the most promising methods for the study of tissue microstructure due to its unique sensitivity to the displacement of water molecules at the scale of microns. It allows to define numerical indices that describe the brain tissue microstructure based on the measurements of signal decay along a predefined set of directions, providing an *in-vivo* indirect measure of the geometry of the diffusion pores [316, 85] while enabling the construction of the structural connectome through tractography [317].

The diffusion process and the Ensemble Average Propagator

The intrinsic thermal energy of the tissues induces a random motion on the water molecules. This process was first defined by Einstein as a *self diffusion* process [318]. If we consider an ensemble of water particles, the probability density function of a particle undergoing a displacement $r \in \mathbb{R}^3$ during diffusion time $\tau \in \mathbb{R}^+$ is

$$P(\mathbf{r};\tau) = \frac{1}{\sqrt{4\pi|\mathbf{D}|\tau}} e^{\frac{-||\mathbf{r}\mathbf{D}^{-1}||^2}{4\tau}} \tag{A.1}$$

where D is the *diffusion tensor*, namely the 3×3 symmetric matrix describing the magnitude of the diffusion in each direction. The quantity $P(r,t)$ is called Ensemble Average Propagator (EAP) [319].

Cellular structures can hinder or restrict the motion of water molecules at microscopic scale. This link between tissue microstructure and diffusion process is the reason why the dMRI signal has the potential to reveal information about the cellular-scale organization of the tissue.

Diffusion Weighted Imaging

To obtain the EAP it is necessary to acquire the dMRI signal, which can be obtained with the Pulsed Gradient Spin Echo (PSGE) sequence applied to a MRI protocol. Figure A.1 shows how the PSGE builds on top of the standard spin-echo MRI acquisition sequence by applying two additional gradients. These gradients are characterized by their strength G and their duration $\delta\, ms$ and are applied with a delay of $\Delta\, ms$.

The first diffusion gradient is applied after a 90 degree Radio Frequency Spin Echo (RFSE) pulse, which projects the spin on the plane perpendicular to main magnetization direction; the second diffusion gradient is applied after a 180 degree RFSE pulse which refocuses the spins.

Appendix/Figures/impulso.pdf

Fig. A.1: Typical PSGE sequence, where the two pulses having width δ ms, are Δ ms apart from each other. G represents the strength of the diffusion gradient.

The application of the two pulses will re-establish the original orientation of the spins that were processing along the applied gradient direction, which had caused a variation of the field intensity. The spins that have moved between the application of the two gradients will be subject to a different field strength during the second pulse, hence they will not return to their exact initial state. The resulting phase shift will be reflected in a decreased intensity of the measured MR signal. This means that the obtained diffusion weighted images will show low intensity where the diffusivity along the applied gradient is high.

To quantify this signal loss it is necessary to acquire the signal without any diffusion gradient (G=0) to obtain the reference signal which depends solely on the amount of spins in the voxel. The diffusion signal is weighted along the applied gradient direction with *b-value*

$$b = \gamma^2 G^2 \delta^2 (\Delta - \delta/3) \tag{A.2}$$

which is measured in s/mm^3, and where γ (MHz/T) is the nuclear gyromagnetic ratio of the water proton. The diffusion signal can be derived using the Stejskal-Tanner equation [320]

$$S(b) = S_0 e^{-bD} \tag{A.3}$$

where S_0 is the signal acquired at G=0 and D is the Apparent Diffusion Coefficient (Apparent Diffusion Coefficient (ADC)) which can be computed as $D(\mathbf{u}) = \mathbf{u}^T \mathbf{D} \mathbf{u}$, where $\mathbf{D}$ the diffusion tensor.

The estimation of the diffusion of water molecules in the brain requires the sensing of the dMRI signal along multiple spatial directions. Using PSGE sequence at a certain b-value it is possible to obtain a snapshot of the diffusion process in a given direction. The most common diffusion-weighted acquisition scheme is composed of multiple gradients acquired at the same b-value but in multiple directions which are spread uniformly on the surface of a sphere called shell. The estimation of the diffusion of water molecules in the brain requires the sensing of the dMRI signal along multiple spatial directions.

Finally, the diffusion signal can be related to the EAP via a Fourier relationship under the q-space formalism [321]

$$E(\mathbf{q},\tau) = \int_{\mathbb{R}^3} P(\mathbf{r},\tau) e^{j2\pi\mathbf{q}\mathbf{r}} d\mathbf{r} \tag{A.4}$$

where $\tau = \Delta - \delta/3$ is the diffusion time, $\mathbf{q}$ is expressed as $\mathbf{q} = \frac{\gamma\delta\mathbf{G}}{2\pi}$ and $\mathbf{r}$ is the spin displacement. In order to recover the full probability density function of the water molecules displacement it is necessary to calculate the inverse Fourier transform of the diffusion signal. However, this is not generally applicable since the number of samples acquired in the q-space is limited. To overcome this problem a number of reconstruction models have been proposed in the literature [84, 102]. The main goal of the reconstruction models considered in this thesis is to derive the EAP from the raw signal.

Diffusion Tensor Imaging

Diffusion Tensor Imaging (DTI) [84] models the diffusion signal as a single multivariate Gaussian function and its equation can be directly derived from the Fourier transform of Eq. A.1

$$E(\mathbf{q}) = exp(-4\pi^2\tau\mathbf{q}^T\mathbf{D}\mathbf{q}) \tag{A.5}$$

where $\mathbf{D}$ is the diffusion tensor, a 3×3 symmetric positive definite matrix.

The coefficients of the diffusion tensor can be estimated from the diffusion signal samples using linear least squares and taking $\ln E(\mathbf{q})$.

One of the most important advantages of the diffusion tensor is that its eigenvalues and eigenvectors decomposition can be directly linked to the biological properties of the tissues

$$\mathbf{D} = \lambda_1\mathbf{v}_1\mathbf{v}_1^T + \lambda_2\mathbf{v}_2\mathbf{v}_2^T + \lambda_3\mathbf{v}_3\mathbf{v}_3^T \tag{A.6}$$

with λ_i the i_{th} biggest eigenvalue associated with the eigenvector $\mathbf{v}_i$.

The biggest eigenvector corresponds to the main diffusion direction in the tissue. In White Matter (WM) voxel this direction corresponds to the average axons main axis. The associated eigenvalue λ_1 represents the ADC in the principal diffusion direction.

Several indices have been proposed in the literature, calculated from the tensor eigenvalues, such as Fractional Anisotropy (FA), Mean Diffusivity (MD), Radial Diffusivity (RD), Axial Diffusivity (AxD)

$$FA = \sqrt{\frac{1}{2}\frac{(\lambda_1-\lambda_2)^2+(\lambda_1-\lambda_3)^2+(\lambda_2-\lambda_3)^2}{\lambda_1^2+\lambda_2^2+\lambda_3^2}} \tag{A.7}$$

$$MD = \frac{\lambda_1+\lambda_2+\lambda_3}{3} \tag{A.8}$$

$$RD = \frac{\lambda_2+\lambda_3}{2} \tag{A.9}$$

$$AxD = \lambda_1 \tag{A.10}$$

Although this model has been proven to be suitable for describing weakly constrained diffusion and simple WM topologies, it can not cope with complex architectures like fiber crossing, fanning and kissing, and the non-Gaussianity of the diffusion process inside a restricted medium such as the WM [322].

Simple Harmonic Oscillator based reconstruction and estimation

To overcome the limitation induced by the simple signal reconstruction obtained by fitting the DTI model to the dMRI data, novel reconstruction models have been proposed in the literature. One of them is the 3D Simple Harmonics Oscillator based Reconstruction and Estimation (3D-SHORE) [102, 316]. This model is based on the solution of the 3D quantum harmonic oscillator using the orthonormalized basis

$$E(\mathbf{q}) = \sum_{l=0,even}^{N_{max}} \sum_{n=l}^{\frac{(N_{max}+l)}{2}} \sum_{m=-l}^{l} c_{nlm}\Phi_{nlm}(\mathbf{q}) \tag{A.11}$$

In this equation, N_{max} is the minimal order of functions, $\Phi_{nlm}(\mathbf{q})$ are the functions forming the 3D-SHORE orthonormal basis and are given by

$$\Phi_{nlm}(\mathbf{q}) = [\frac{2(n-l)!}{\zeta^{\frac{3}{2}}\Gamma(n+\frac{3}{2})}]^{\frac{1}{2}}(\frac{q^2}{\zeta})^{\frac{1}{2}}exp(\frac{-q^2}{2\zeta})L_{n-l}^{l+\frac{1}{2}}(\frac{q^2}{\zeta})Y_l^m(\mathbf{u}) \tag{A.12}$$

where Γ is the Gamma function and ζ is a scaling parameter determined by the diffusion time and the mean diffusivity [323], [141]. For the 3D-SHORE mode, the EAP is obtained by plugging Equation A.12 into Equation A.1 [102], [141]. Due to the linearity of the Fourier transform, the EAP basis is thus expressed in terms of the same set of coefficients c_{nlm} as the diffusion signal.

From the fitting of the signal with 3D-SHORE model multiple indices can be derived that better characterize the diffusion properties of the underlying microstructure. Return to the Origin Probability (RTOP), Return to the Axis Probability (RTAP) and Return to the Plane Probability (RTPP) [102] represent respectively the value of the EAP in zero, the integral of the EAP along the main diffusion direction and over the plane passing through the origin and perpendicular to the main diffusion direction

$$RTOP = P(\mathbf{0}) \tag{A.13}$$

$$RTAP = \int_R P(\mathbf{r}_\parallel) d\mathbf{r}_\parallel \tag{A.14}$$

$$RTPP = \int_R^2 P(\mathbf{r}_\perp) d^2\mathbf{r}_\perp \tag{A.15}$$

where $\mathbf{r}_\parallel$ is the main diffusion direction, and $\mathbf{r}_\perp$ indicates the plane orthogonal to the main diffusion direction and passing through the origin. It has been shown [102], [324] under the assumption of narrow pulses and long diffusion time, RTAP and RTPP are proportional to the inverse of the mean apparent cross-sectional area and length of the compartment where diffusion takes place, respectively. From the EAP it is possible to derive also a propagator anisotropy index, depending on the angular distance between the isotropic part of the EAP, that is encoded in the coefficient c_{n00}, and the full EAP as in [102]

$$PA = \sqrt{1 - \frac{\sum_{n=0}^{N_{max}} c_{n00}^2}{\sum_{n,l,m}^{N_{max}} c_{nlm}^2}} \tag{A.16}$$

Rotation Invariants Features

Novel Rotation Invariant Features (RIF) were proposed in [325]. Usually the most used dMRI RIF are MD and FA [84] which can be calculated from a rank-2 tensor representation of the ADC profile [102].

The full RIF set, instead, comprises indices calculated on the Laplace-series expansion of a spherical function and are the natural expansion of spherical mean, power-spectrum and bispectrum invariants. Moreover, they can be linked to statistical and geometrical measures of spherical functions, such as the mean, the variance and the volume of the spherical signal. They can be applied not only in dMRI [326] but also in computer vision [327] and pattern recognition [328].

The general form of the new RIF can be written as:

$$I_l[f] = \sum_{m_1=-l_1}^{l_1} \dots \sum_{m_d=-l_d}^{l_d} c_{l_1 m_1} \dots c_{l_d m_d} G(l_1, m_1 | \dots | l_d, m_d) \tag{A.17}$$

Where $\mathbf{l} = [l_1, l_2..., l_d]$ are the considered Spherical Harmonics (SH) order, f is the function from which the RIF will be calculated, $c_{l_i m_i}$ are the SH coefficients of order i and G represents the generalized Gaunt coefficient, namely the integral of l SH [329].

Essentially the equation (A.17) is the sum over all the m of the product of the SH coefficients times the integral of multiple SH. Given N the truncation degree of the SH, it is possible to calculate the number of SH coefficients as $n_c = \frac{(N+1)(N+2)}{2}$. In [325] they show that the number of algebraically independent invariants corresponds to n_c-3. Algebraic independence is a fundamental property ensuring that none of the elements of the considered set of RIF can be expressed as an algebraic function of one or more of the other elements of the set. Therefore, it allows obtaining a non-redundant representation of the rotation invariant properties of the considered spherical function.

Practically, for a given b-value, it is possible to express the signal as a spherical function by considering its truncated SH-series expansion

$$E(b, \mathbf{u}) = \sum_{l=0,even}^{N} \sum_{m=-l}^{l} c_{lm}(b) Y_l^m(\mathbf{u}) \tag{A.18}$$

The equation (A.18) was then replaced in the invariants equation (A.17) to obtain the invariants on the signal at a given b-value. In this work, a SH degree $N = 4$ was considered, therefore 12 algebraic independent invariants for each b-value were generated for each of the subjects.

A.2 EXplainable Artificial Intelligence (XAI) methods

In this Section we will present in detail the eXplainable Artificial Intelligence (XAI) methods used throughout the works presented in this thesis

A.2.1 Perturbation based methods

Local Interpretable Model-Agnostic Explanations (LIME)

This is a model-agnostic approach that can be applied to explain any black-box model, aiming at understanding how the predictions change when the data samples are perturbed, enforcing local fidelity and interpretability. Given a global complex model, LIME focuses on training local surrogate models to explain individual predictions [8]. It belongs to the family of additive feature attribution methods, which have an explanation model that is a linear function of binary variables:

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z_i{}' \tag{A.19}$$

where g is the explanation model, z' is a binary variable representing whether the feature is present or not, M is the number of simplified input features and ϕ_i is the feature attribution. To find ϕ, the vector that assigns a feature importance ϕ_i to each input feature, LIME minimises the following objective function:

$$\xi = \arg\min_{g \in G} L(f, g, \pi'_x) + \Omega(g) \tag{A.20}$$

with f representing the original prediction model, G is the family of possible explanations, L is the loss (e.g, Mean Square Error (MSE)) calculated over a set of samples in the simplified input space weighted by the local kernel π'_x which defines how large the neighborhood around the instance is, and $\Omega(g)$ penalizes the complexity of g. Going more into the details, the method starts from an individual instance to be explained and the associated prediction given by the target black-box model. It then perturbs the related input features within a neighborhood proximity to measure the changes in predictions. Based on such perturbations, a new artificial dataset is created and the corresponding predictions using the black-box model are derived. The new data points are also weighted based on their proximity to the original observation, with closer data having higher weights, in order to ensure local fidelity. An exponential kernel with an Euclidean distance is generally chosen for this purpose. An interpretable surrogate simple model (e.g., linear regression) is then trained on the generated perturbed dataset. This approximates the original model behavior in the neighborhood of the original samples and can be exploited to understand which variables are the most important for the ML prediction by comparing the respective coefficients. Therefore, the impact of each feature on the prediction is defined by its weights in the local model, which is assumed to be a good approximation of the more complex original model. This approach has several advantages, including the fact that the same local, interpretable model can be used for explanation even if the underlying ML model is changed, it works for different data types including tabular data, text and images, and it is easy to use. One main drawback of such methods is related to the correct definition of the neighborhood chosen to randomly perturbating feature values, and of the kernel function for assigning the distance-based weights. Moreover, instability of the explanations has been demonstrated in some scenarios. While LIME inherently provides local model interpretability, averaging the scores assigned to each feature across all the local explanations allows to produce a global explanation.

SHapley Additive exPlanations (SHAP)

This is a model-agnostic explanation method belonging to the class of additive feature attribution methods which uses a similar concept to LIME and builds on the game theory concept of Shapley values [330]. These values are used to determine contribution of each player in a coalition or a cooperative game. Indeed, originally, Shapley proposed a game theory method for assigning fair payouts to players depending on their contribution to the total gain. In a prediction task, this corresponds to assign a quantitative value to each feature depending on its contribution to a specific prediction. The SHAP method [252] computes the Shapley values and represents them as a linear model of feature coalitions. It requires retraining the model on all feature subsets $S \subseteq F$, where F is the set of all input features. It assigns an importance value to each feature that represents the effect on the model prediction given by the inclusion of this feature. To do so, a model $f_{S\cup\{i\}}$ is trained with that feature present, and another model f_S is trained with the feature withheld. Predictions from the two models are compared on the current input, and this is repeated for all possible subsets. Shapley values are then computed as the weighted average of all possible differences and used as feature attributions:

$$\phi_i = \sum_{S\subseteq F\setminus\{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!}[f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S)] \tag{A.21}$$

According to [252], SHAP values attribute to each feature the change in the expected model prediction when conditioning on that feature. In this framework, the difference between the prediction and the average prediction, considered as baseline reference, is perfectly distributed among all the features. Therefore, SHAP values of all the features sum up to explain the difference between the actual prediction and the baseline.

SHAP has a solid theoretical foundation, making this approach quite robust, and it allows to derive contrastive explanations comparing the individual prediction vs the average one. It also has a fast implementation for tree-based models (TreeSHAP [270], which was chosen in our study. Moreover, SHAP values present properties of local accuracy, missingness, and consistency, which are not simultaneously found in other methods. An important drawback of SHAP is that it provides additive contributions of the different variables. However, if the model is not additive, the Shapley values might be misleading. The variable importance at a global level is given by the mean absolute SHAP value across all observations.

A.2.2 Random Forest specific model explanations

Random forest Feature Importance (RFI)

This approach relies on one of the most widely used methods for assessing feature importance measures in the specific case of tree-based models that is the Mean Decreased Impurity (MDI). MDI is defined as the total decrease in impurity when a given variable is used to split a node, averaged over all trees of the ensemble [295].
Indeed, in the context of ensembles of randomized trees represented by a tree structure T, Breiman [295] proposed to evaluate the importance of a variable x for predicting Y by adding up the weighted impurity decreases for all nodes t where x is used, averaged over all N_T trees in the forest:

$$MDI(x) = \frac{1}{N_T} \sum_{T} \sum_{t \in T : v(s_t) = x} p(t) \Delta i(s_t, t) \tag{A.22}$$

where $p(t)$ is the proportion N_t/N of samples reaching the node t, $v(s_t)$ is the variable used in split s_t, and $\Delta i = i(t) - p_L i(t_L) - p_R i(t_R)$ with impurity measure $i(t)$ (variance for regression), t_L and t_R representing the partitions of the N_t node samples, $p_L = N_{tL}/N_t$ and $p_R = N_{tR}/N_t$. While this approach is widely used for RF, having the advantage of being easy and fast to compute as it does not require permuting the data, it has several drawbacks. In particular, this approach is known to be biased in favor of variables with many possible split points and with high category frequencies [331], and to be biased in presence of correlated features.

Permutation Feature Importance (PFI)

This approach, known as Mean Decreased Accuracy (MDA), can be used as an alternative to overcome the drawbacks of default feature importance computed with MDI. It measures the changes in the model's prediction error when a single feature value is randomly shuffled (MSE in our case). This modification of the model score is indicative of how much the model depends on that feature. Indeed, permuting an important feature would result in a large decrease in accuracy while permuting a less relevant feature would have a negligible effect [295, 297].
This measure is computed for each predictor variable x as follows: 1) the prediction accuracy A_t is calculated using the Out-Of-Bag (OOB) observations; 2) the observations are shuffled, and the prediction accuracy A_t^* is recomputed; 3) the difference $A_t - A_t^*$ is averaged over all the trees in the forest, returning the permutation importance of x:

$$PFI(x) = \frac{1}{N_T} \sum_{t=1}^{N_T} A_t - A_t^* \tag{A.23}$$

Of note, as stated above, the MSE metric is generally used for defining the A_t and A_t^* terms in regression problems. This technique benefits from being model-agnostic and can be calculated many times with different permutations, allowing a straightforward interpretation of the results. However, it is computationally expensive and might have problems with highly correlated features.

A.2.3 Saliency maps

Backpropagation

BP [21] relies on the visualization of the gradient of the network output probability with respect to the input image. For a given voxel, this gradient describes how much the output probability changes when the voxel value changes. In Convolutional Neural Networks (CNN)s, the gradient is straightforwardly available since it can be easily computed via the backpropagation algorithm used for training, and it is equivalent to the visualization of the partial derivatives of the network output with respect to each input feature scaled by its value. Given an input x and a function S_c that describes the model output for the class c, the BP can be expressed as

$$BP(x) = \frac{\partial S_c}{\partial x}. \tag{A.24}$$

Differently from the backpropagation algorithm used for training, to obtain the saliency map what is backpropagated is not the classification error, or loss, but directly the classification probability for a given class. The absolute value of the resulting coefficients is taken as the relevance score representing the feature importance.

For an intuition of this method, following the example in [22] and Figure A.2, we can exemplify what happens in the backward pass to obtain the saliency maps for the convolutional and, in particular, for the Rectified Linear Unit (Rectified Linear Unit (ReLU)) layer which holds the difference between Guided BackPropagation (GBP) and GBP. The ReLU layer is responsible for adding the non linearity to the neural network. For short, it is a piece-wise linear function that is defined to be zero for all negative values of the node input and one otherwise, thus keeping unchanged the positive input values while annihilating the negative ones.

Given the input image or volume $\mathbf{x}$, during the forward pass, each CNN layer l returns a feature activation map f^l till the last layer L. Then, starting from f^L it is possible to generate the backpropagation map R^L by zeroing all the neuron activations except the one to be backpropagated, that is the one related to the target class, and to start the backward pass to reconstruct the input image $\mathbf{x}$ showing the part of the input image that is most strongly activating this neuron. Each R^l represents an intermediate step

in the calculation of the BP, for the intermediate layer l. When reaching the CNN input layer, the reconstruction R^1 will have the same size of the input $\mathbf{x}$.

Starting from the convolutional layers, the respective activation in the forward pass can be expressed as $f_{l+1} = f_l \circledast K_l$, where K_l is a convolutional kernel. The gradient with respect to the output feature map R^L, for layer l is then

$$R^l = \frac{\partial R^L}{\partial f_l} = \frac{\partial R^L}{\partial f_{l+1}} \circledast \hat{K}_l \tag{A.25}$$

where $\hat{K}_l$ is the flipped version of the kernel K_l and f is the visualized neuron activity. The convolution with the flipped kernel exactly corresponds to computing the lth layer reconstruction R_l.

Moving to the ReLU layers, the activation during the forward pass can be defined as $f_{l+1}(x) = ReLU(f_l) = max(f_l, 0)$. The respective backpropagation computes the gradient of R^L with respect to the ReLu layer l as

$$R^l = \frac{\partial R^L}{\partial f_l} = \frac{\partial R^L}{\partial f_{l+1}} \cdot (f_l > 0) = R^{l+1} \cdot (f_l > 0) \tag{A.26}$$

where the $(\cdot)$ operator represents the element-wise multiplication. The element that defines which gradients are backpropagated, in this case $(f_l > 0)$, is also known as sign operator.

By iterating these steps backward through the network layers results in R^1, representing the image-specific class saliency map, highlighting the areas of the given input that is discriminative with respect to that class.

Guided Backpropagation

This method was presented by [22] and is a modified version of BP with respect to the backward pass through the ReLU layer. More in detail, GBP combines the approach used for the Deconvolution Network (DeconvNet) [110] with the one described for the BP, leading to more focused heatmaps.

Here, we briefly present the DeconvNet in order to introduce the GBP . The focus will be on the backpropagation through the ReLU layer which holds the main difference between BP and DeconvNet. Following the same mathematical framework as for the BP, the DeconvNet backward pass through the ReLu layer l can be described as

$$R^l = \frac{\partial R^L}{\partial f_l} = \frac{\partial R^L}{\partial f_{l+1}} \cdot (R^{l+1} > 0) = R^{l+1} \cdot (R^{l+1} > 0) \tag{A.27}$$

This differs from the BP since the sign indicator is based on the output reconstruction R^{l+1} of the precedent layer and not on the input activation f_{l+1} to the precedent layer

as for BP. This allows only positive error signals to be backpropagated through the net to obtain the final saliency map.

Moving to the GBP , this method defines the backpropagation through the ReLU layer l as

$$R^l = \frac{\partial R^L}{\partial f_l} = \frac{\partial R^L}{\partial f_{l+1}} \cdot (R^{l+1} > 0) \cdot (f_l > 0) = R^{l+1} \cdot (R^{l+1} > 0) \cdot (f_l > 0) \tag{A.28}$$

Like DeconvNets, in GBP only positive error signals are backpropagated setting the negative gradients to zero, which amounts to the application of the ReLU to the error signal itself during the backward pass. Moreover, like in BP, only positive inputs are considered. The advantage of retaining only positive gradients is to prevent a backward flow of negative contributions corresponding to neurons which inhibit the activation of the higher level neuron. As opposed to BP, this can act as an additional guidance signal when traversing the network. As above, the absolute value of the gradient is taken as the relevance score.

Appendix/Figures/F1.png

Fig. A.2: Differences between BackPropagation (BP), GBP and DeconvNet in the backpropagation through the ReLU layers, following the example proposed in [22]

Chapter5/Figures_EDL_AI/F2.png

Fig. A.3: Overview of LRP visualization procedure.

Layerwise Relevance Propagation (LRP)

LRP is slightly different from both BP and GBP since it is based on a backward procedure which is a conservative relevance redistribution of the output prediction probability through the CNN layers till the input volume, as shown in Figure A.3.

Briefly, let i and j be the indices for neurons at two successive layers r and $r+1$. Let R_j^{r+1} be the relevance of neuron j for the prediction $f(x)$ where x is the neural network input. The relevance R_j^{r+1} is redistributed to the connected neurons in layer r such that the relevance conservation holds:

$$\sum_i R_{i\leftarrow j}^r = R_j^{r+1}. \tag{A.29}$$

By iterating Eq. A.29 through all the layers, it is possible to decompose the relevance score of the prediction $f(x)$, R_f in terms of the input variables of the first layer. This allows to easily visualize the relevance values as heatmaps.

Different rules have been applied in literature for redistributing the relevance [46]. In this work we used the β-rule as in [32, 123]:

$$R_{i\leftarrow j}^{r,r+1} = \left((1+\beta)\frac{w_{ij}^{+}}{w_j^{+}} - \beta\frac{w_{ij}^{-}}{w_j^{-}} \right) R_j^{r+1} \tag{A.30}$$

In this equation, $w_{i,j}^{+/-}$ is the amount of positive/negative contribution that node j transfers to node i, divided by the sum over all positive/negative contributions of the nodes in layer r. In fact $w_j^{+/-} = \sum_i w_{i,j}^{+/-}$, such that the relevance is conserved from layer $r+1$ to layer r. This approach was adopted following [26], where multiple β values were tested to assess their impact on the resulting heatmaps. The authors proved the LRP robustness relatively to the corresponding β-value. The β-rule decomposes the relevance score in positive and negative contributions, and weights the relative importance according to the β parameter. Setting $\beta = 0$ adds only positive contributions to the relevance score. Negative contributions hold an inhibitory effect highlighting the voxels that are antagonist to those having strong positive impact on the classification function. Since in this work we aimed at detecting those voxels playing in favour of the correct classification of PPMS patients, we constrained β to be zero.

Bibliography

[1] Ryszard Stanislaw Michalski, Jaime Guillermo Carbonell, and Tom M Mitchell. *Machine learning: An artificial intelligence approach*. Springer Science & Business Media, 2013.

[2] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. "Deep learning with differential privacy". In: *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*. 2016, pp. 308–318.

[3] Jürgen Schmidhuber. "Deep learning in neural networks: An overview". In: *Neural networks* 61 (2015), pp. 85–117.

[4] Andreas Holzinger. "Interactive machine learning for health informatics: when do we need the human-in-the-loop?" In: *Brain Informatics* 3.2 (2016), pp. 119–131.

[5] Protima Khan, Md Fazlul Kader, SM Riazul Islam, Aisha B Rahman, Md Shahriar Kamal, Masbah Uddin Toha, and Kyung-Sup Kwak. "Machine learning and deep learning approaches for brain disease diagnosis: principles and recent advances". In: *IEEE Access* 9 (2021), pp. 37622–37655.

[6] Andreas Holzinger, Chris Biemann, Constantinos S Pattichis, and Douglas B Kell. "What do we need to build explainable AI systems for the medical domain?" In: *arXiv preprint arXiv:1712.09923* (2017).

[7] Bas HM van der Velden, Hugo J Kuijf, Kenneth GA Gilhuijs, and Max A Viergever. "Explainable artificial intelligence (XAI) in deep learning-based medical image analysis". In: *Medical Image Analysis* (2022), p. 102470.

[8] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "" Why should i trust you?" Explaining the predictions of any classifier". In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016, pp. 1135–1144.

[9] Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. "Explainable AI: A Review of Machine Learning Interpretability Methods". In: *Entropy* 23.1 (2021), p. 18.

[10] Andreas Holzinger, Georg Langs, Helmut Denk, Kurt Zatloukal, and Heimo Müller. "Causability and explainability of artificial intelligence in medicine". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 9.4 (2019), e1312.

[11] Rongtao Jiang, Choong-Wan Woo, Shile Qi, Jing Wu, and Jing Sui. "Interpreting Brain Biomarkers: Challenges and solutions in interpreting machine learning-

based predictive neuroimaging". In: *IEEE Signal Processing Magazine* 39.4 (2022), pp. 107–118.

[12] Anna Hedström, Leander Weber, Dilyara Bareeva, Franz Motzkus, Wojciech Samek, Sebastian Lapuschkin, and Marina M-C Höhne. "Quantus: an explainable AI toolkit for responsible evaluation of neural network explanations". In: *arXiv preprint arXiv:2202.06861* (2022).

[13] Li Shen and Paul M Thompson. "Brain imaging genomics: integrated analysis and machine learning". In: *Proceedings of the IEEE* 108.1 (2019), pp. 125–162.

[14] Samar SM Elsheikh, Emile R Chimusa, Nicola J Mulder, and Alessandro Crimi. "Genome-wide association study of brain connectivity changes for Alzheimer's disease". In: *Scientific reports* 10.1 (2020), pp. 1–16.

[15] Lada Kohoutová, Juyeon Heo, Sungmin Cha, Sungwoo Lee, Taesup Moon, Tor D Wager, and Choong-Wan Woo. "Toward a unified framework for interpreting machine-learning models in neuroimaging". In: *Nature protocols* 15.4 (2020), pp. 1399–1435.

[16] Zachary C Lipton. "The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery." In: *Queue* 16.3 (2018), pp. 31–57.

[17] Xinyang Feng, Zachary C Lipton, Jie Yang, Scott A Small, and Frank A Provenzano. "Estimating brain age based on a uniform healthy population with deep learning and structural magnetic resonance imaging". In: *Neurobiology of aging* 91 (2020), pp. 15–25.

[18] Loveleen Gaur, Mohan Bhandari, Tanvi Razdan, Saurav Mallik, and Zhongming Zhao. "Explanation-Driven Deep Learning Model for Prediction of Brain Tumour Status Using MRI Image Data". In: *Frontiers in Genetics* (2022), p. 448.

[19] Gareth Ball, Claire E Kelly, Richard Beare, and Marc L Seal. "Individual variation underlying brain age estimates in typical development". In: *NeuroImage* 235 (2021), p. 118036.

[20] Caroline Machado Dartora, Luıs Vinicius de Moura, Michel Koole, and Ana Maria Marques da Silva. "Discriminating Aging Cognitive Decline Spectrum Using PET and Magnetic Resonance Image Features". In: *Journal of Alzheimer's Disease* (2022), pp. 1–15.

[21] Karen Simonyan and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556* (2014).

[22] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. "Striving for simplicity: The all convolutional net". In: *arXiv preprint arXiv:1412.6806* (2014).

[23] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. "Learning deep features for discriminative localization". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 2921–2929.

[24] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. "Grad-cam: Visual explanations from deep networks via gradient-based localization". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 618–626.

[25] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks". In: *2018 IEEE winter conference on applications of computer vision (WACV)*. IEEE. 2018, pp. 839–847.

[26] Moritz Böhle, Fabian Eitel, Martin Weygandt, and Kerstin Ritter. "Layer-Wise Relevance Propagation for Explaining Deep Neural Network Decisions in MRI-Based Alzheimer's Disease Classification". In: *Frontiers in Aging Neuroscience* 11 (2019), p. 194. ISSN: 1663-4365. DOI: 10.3389/fnagi.2019.00194. URL: https://www.frontiersin.org/article/10.3389/fnagi.2019.00194.

[27] Florian Dubost, Pinar Yilmaz, Hieab Adams, Gerda Bortsova, M Arfan Ikram, Wiro Niessen, Meike Vernooij, and Marleen de Bruijne. "Enlarged perivascular spaces in brain MRI: Automated quantification in four regions". In: *Neuroimage* 185 (2019), pp. 534–544.

[28] Wen Wei, Emilie Poirion, Benedetta Bodini, Stanley Durrleman, Nicholas Ayache, Bruno Stankoff, and Olivier Colliot. "Predicting PET-derived demyelination from multimodal MRI using sketcher-refiner adversarial training for multiple sclerosis". In: *Medical image analysis* 58 (2019), p. 101546.

[29] Sabyasachi Chakraborty, Satyabrata Aich, and Hee-Cheol Kim. "Detection of Parkinson's disease from 3T T1 weighted MRI scans using 3D convolutional neural network". In: *Diagnostics* 10.6 (2020), p. 402.

[30] Yunyan Zhang, Daphne Hong, Daniel McClement, Olayinka Oladosu, Glen Pridham, and Garth Slaney. "Grad-CAM helps interpret the deep learning models trained to classify multiple sclerosis types using clinical brain magnetic resonance imaging". In: *Journal of Neuroscience Methods* 353 (2021), p. 109098.

[31] Sumeet Shinde, Shweta Prasad, Yash Saboo, Rishabh Kaushick, Jitender Saini, Pramod Kumar Pal, and Madhura Ingalhalikar. "Predictive markers for Parkinson's disease using deep neural nets on neuromelanin sensitive MRI". In: *NeuroImage: Clinical* 22 (2019), p. 101748.

[32] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. "On pixel-wise explanations for non-

linear classifier decisions by layer-wise relevance propagation". In: *PloS one* 10.7 (2015), e0130140.

[33] Fabian Eitel, Emily Soehler, Judith Bellmann-Strobl, Alexander U Brandt, Klemens Ruprecht, René M Giess, Joseph Kuchling, Susanna Asseyer, Martin Weygandt, John-Dylan Haynes, et al. "Uncovering convolutional neural network decisions for diagnosing multiple sclerosis on conventional MRI using layer-wise relevance propagation". In: *arXiv preprint arXiv:1904.08771* (2019).

[34] Alison Deatsch, Matej Perovnik, Mauro Namıas, Maja Trošt, Robert Jeraj, Alzheimer's Disease Neuroimaging Initiative, et al. "Development of a deep learning network for Alzheimer's disease classification with evaluation of imaging modality and longitudinal data". In: *Physics in Medicine & Biology* 67.19 (2022), p. 195014.

[35] Shilpa Dang and Santanu Chaudhury. "Novel relative relevance score for estimating brain connectivity from fMRI data using an explainable neural network approach". In: *Journal of Neuroscience Methods* 326 (2019), p. 108371.

[36] Saumya Jetley, Nicholas A Lord, Namhoon Lee, and Philip HS Torr. "Learn to pay attention". In: *arXiv preprint arXiv:1804.02391* (2018).

[37] Florian Dubost, Hieab Adams, Pinar Yilmaz, Gerda Bortsova, Gijs van Tulder, M Arfan Ikram, Wiro Niessen, Meike W Vernooij, and Marleen de Bruijne. "Weakly supervised object detection with 2D and 3D regression neural networks". In: *Medical Image Analysis* 65 (2020), p. 101767.

[38] Chunfeng Lian, Mingxia Liu, Li Wang, and Dinggang Shen. "End-to-end dementia status prediction from brain mri using multi-task weakly-supervised attention network". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2019, pp. 158–167.

[39] Luisa M Zintgraf, Taco S Cohen, Tameem Adel, and Max Welling. "Visualizing deep neural network decisions: Prediction difference analysis". In: *arXiv preprint arXiv:1702.04595* (2017).

[40] Marko Robnik-Šikonja and Igor Kononenko. "Explaining classifications for individual instances". In: *IEEE Transactions on Knowledge and Data Engineering* 20.5 (2008), pp. 589–600.

[41] Dasom Seo, Kanghan Oh, and Il-Seok Oh. "Regional multi-scale approach for visually pleasing explanations of deep neural networks". In: *IEEE Access* 8 (2019), pp. 8572–8582.

[42] Marco Lorenzi, Andre Altmann, Boris Gutman, et al. "Susceptibility of brain atrophy to TRIB3 in Alzheimer's disease, evidence from functional prioritization in imaging genetics". In: *Proceedings of the National Academy of Sciences* 115.12 (2018), pp. 3162–3167.

[43] Latarsha J Carithers, Kristin Ardlie, Mary Barcus, Philip A Branton, Angela Britton, Stephen A Buia, Carolyn C Compton, David S DeLuca, Joanne Peter-Demchok, Ellen T Gelfand, et al. "A novel approach to high-quality postmortem tissue procurement: the GTEx project". In: *Biopreservation and biobanking* 13.5 (2015), pp. 311–319.

[44] Kacper Sokol and Peter Flach. "Explainability fact sheets: a framework for systematic assessment of explainable approaches". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 2020, pp. 56–67.

[45] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. "Axiomatic attribution for deep networks". In: *International conference on machine learning*. PMLR. 2017, pp. 3319–3328.

[46] Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. "Methods for interpreting and understanding deep neural networks". In: *Digital Signal Processing* 73 (2018), pp. 1–15.

[47] Bhagya Nathali Silva, Murad Khan, and Kijun Han. "Towards sustainable smart cities: A review of trends, architectures, components, and open challenges in smart cities". In: *Sustainable Cities and Society* 38 (2018), pp. 697–713.

[48] Bernease Herman. "The promise and peril of human evaluation for model interpretability". In: *arXiv preprint arXiv:1711.07414* (2017).

[49] Jason Smucny, Ge Shi, and Ian Davidson. "Deep Learning in Neuroimaging: Overcoming Challenges With Emerging Approaches". In: *Frontiers in Psychiatry* 13 (2022).

[50] Carlo Combi, Beatrice Amico, Riccardo Bellazzi, Andreas Holzinger, Jason H Moore, Marinka Zitnik, and John H Holmes. "A manifesto on explainability for artificial intelligence in medicine". In: *Artificial Intelligence in Medicine* 133 (2022), p. 102423.

[51] Ilaria Boscolo Galazzo, Lorenza Brusini, Muge Akinci, Federica Cruciani, Marco Pitteri, Stefano Ziccardi, Albulena Bajrami, Marco Castellaro, Ahmed MA Salih, Francesca B Pizzini, et al. "Unraveling the MRI-Based Microstructural Signatures Behind Primary Progressive and Relapsing–Remitting Multiple Sclerosis Phenotypes". In: *Journal of Magnetic Resonance Imaging* 55.1 (2022), pp. 154–163.

[52] Katja Franke and Christian Gaser. "Ten years of BrainAGE as a neuroimaging biomarker of brain aging: what insights have we gained?" In: *Frontiers in neurology* 10 (2019), p. 789.

[53] James H Cole, Riccardo E Marioni, Sarah E Harris, et al. "Brain age and other bodily 'ages': implications for neuropsychiatry". In: *Molecular psychiatry* 24.2 (2019), pp. 266–281.

[54] S. M. Smith, D. Vidaurre, F. Alfaro-Almagro, et al. “Estimation of brain age delta from brain imaging”. In: *Neuroimage* 200 (Oct. 2019), pp. 528–539.

[55] James H Cole. “Multimodality neuroimaging brain-age in UK biobank: relationship to biomedical, lifestyle, and cognitive factors”. In: *Neurobiology of aging* 92 (2020), pp. 34–42.

[56] L. C. Lowe, C. Gaser, and K. Franke. “The Effect of the APOE Genotype on Individual BrainAGE in Normal Aging, Mild Cognitive Impairment, and Alzheimer’s Disease”. In: *PLoS One* 11.7 (2016), e0157514.

[57] Han Peng, Weikang Gong, Christian F Beckmann, Andrea Vedaldi, and Stephen M Smith. “Accurate brain age prediction with lightweight deep neural networks”. In: *Medical image analysis* 68 (2021), p. 101871.

[58] Ruey-Kai Sheu and Mayuresh Sunil Pardeshi. “A Survey on Medical Explainable AI (XAI): Recent Progress, Explainability Approach, Human Interaction and Scoring System”. In: *Sensors* 22.20 (2022), p. 8068.

[59] Ann-Marie G. de Lange, Melis Anatürk, Sana Suri, et al. “Multimodal brain-age prediction and cardiovascular risk: The Whitehall II MRI sub-study”. In: *NeuroImage* 222 (2020), p. 117292.

[60] Stephen M Smith, Lloyd T Elliott, Fidel Alfaro-Almagro, et al. “Brain aging comprises many modes of structural and functional change with distinct genetic and biophysical associations”. In: *Elife* 9 (2020), e52677.

[61] X. Niu, F. Zhang, J. Kounios, et al. “Improved prediction of brain age using multimodal neuroimaging data”. In: *Hum Brain Mapp* 41.6 (Apr. 2020), pp. 1626–1643.

[62] Chang-Le Chen, Yung-Chin Hsu, Li-Ying Yang, Yu-Hung Tung, Wen-Bin Luo, Chih-Min Liu, Tzung-Jeng Hwang, Hai-Gwo Hwu, and Wen-Yih Isaac Tseng. “Generalization of diffusion magnetic resonance imaging–based brain age prediction model through transfer learning”. In: *Neuroimage* 217 (2020), p. 116831.

[63] Fidel Alfaro-Almagro, Paul McCarthy, Soroosh Afyouni, Jesper LR Andersson, Matteo Bastiani, Karla L Miller, Thomas E Nichols, and Stephen M Smith. “Confound modelling in UK Biobank brain imaging”. In: *NeuroImage* 224 (2021), p. 117002.

[64] Benedikt Jónsson, Gyda Bjornsdottir, TE Thorgeirsson, Lotta Marıa Ellingsen, G Bragi Walters, DF Gudbjartsson, Hreinn Stefansson, Kari Stefansson, and MO Ulfarsson. “Brain age prediction using deep learning uncovers associated sequence variants”. In: *Nature communications* 10.1 (2019), pp. 1–10.

[65] Katja Franke, Gabriel Ziegler, Stefan Klöppel, et al. “Estimating the age of healthy subjects from T1-weighted MRI scans using kernel methods: Exploring the influence of various parameters”. In: *NeuroImage* 50.3 (2010), pp. 883–892.

[66] Alba Xifra-Porxas, Arna Ghosh, Georgios D Mitsis, and Marie-Hélène Boudrias. "Estimating brain age from structural MRI and MEG data: Insights from dimensionality reduction techniques". In: *NeuroImage* 231 (2021), p. 117822.

[67] Matthias S Treder, Jonathan P Shock, Dan J Stein, Stefan DuPlessis, Soraya Seedat, and Kamen A Tsvetanov. "Correlation constraints for regression models: controlling bias in brain age prediction". In: *Frontiers in psychiatry* 12 (2021), p. 25.

[68] Ricardo Pio Monti, Alex Gibberd, Sandipan Roy, Matthew Nunes, Romy Lorenz, Robert Leech, Takeshi Ogawa, Motoaki Kawanabe, and Aapo Hyvärinen. "Interpretable brain age prediction using linear latent variable models of functional connectivity". In: *PloS one* 15.6 (2020), e0232296.

[69] Marco Palma, Shahin Tavakoli, Julia Brettschneider, Thomas E Nichols, Alzheimer's Disease Neuroimaging Initiative, et al. "Quantifying uncertainty in brain-predicted age using scalar-on-image quantile regression". In: *NeuroImage* 219 (2020), p. 116938.

[70] Arinbjörn Kolbeinsson, Sarah Filippi, Yannis Panagakis, Paul M Matthews, Paul Elliott, Abbas Dehghan, and Ioanna Tzoulaki. "Accelerated MRI-predicted brain ageing and its associations with cardiometabolic and brain disorders". In: *Scientific Reports* 10.1 (2020), pp. 1–9.

[71] Jaroslav Rokicki, Thomas Wolfers, Wibeke Nordhøy, et al. "Multimodal imaging improves brain age prediction and reveals distinct abnormalities in patients with psychiatric and neurological disorders". In: *Human Brain Mapping* (2020).

[72] Denis Engemann, Oleh Kozynets, David Sabbagh, Guillaume Lemaıtre, Gael Varoquaux, Franziskus Liem, and Alexandre Gramfort. "Combining magnetoencephalography with magnetic resonance imaging enhances learning of surrogate-biomarkers". In: *Elife* 9 (2020), e54055.

[73] Tobias Kaufmann, Dennis van der Meer, Nhat Trung Doan, Emanuel Schwarz, Martina J Lund, Ingrid Agartz, Dag Alnæs, Deanna M Barch, Ramona Baur-Streubel, Alessandro Bertolino, et al. "Common brain disorders are associated with heritable patterns of apparent aging of the brain". In: *Nature neuroscience* 22.10 (2019), pp. 1617–1623.

[74] Gidon Levakov, Gideon Rosenthal, Ilan Shelef, et al. "From a deep learning model back to the brain—Identifying regional predictors and their relation to aging". In: *Human brain mapping* 41.12 (2020), pp. 3235–3252.

[75] Johnny Wang, Maria J Knol, Aleksei Tiulpin, et al. "Gray matter age prediction as a biomarker for risk of dementia". In: *Proceedings of the National Academy of Sciences* 116.42 (2019), pp. 21213–21218.

[76] Alastair Compston and Alasdair Coles. "Multiple sclerosis". In: *The Lancet* 372.9648 (2008), pp. 1502–1517. ISSN: 0140-6736. DOI: https://doi.org/10.1016/

S0140-6736(08)61620-7. URL: http://www.sciencedirect.com/science/article/pii/S0140673608616207.

[77] DH Miller, AJ Thompson, and M Filippi. "Magnetic resonance studies of abnormalities in the normal appearing white matter and grey matter in multiple sclerosis". In: *Journal of neurology* 250.12 (2003), pp. 1407–1419.

[78] Silvia De Santis, Matteo Bastiani, Amgad Droby, Pierre Kolber, Frauke Zipp, Eberhard Pracht, Tony Stoecker, Sergiu Groppa, and Alard Roebroeck. "Characterizing microstructural tissue properties in multiple sclerosis with diffusion MRI at 7 T and 3 T: the impact of the experimental design". In: *Neuroscience* 403 (2019), pp. 17–26.

[79] Massimiliano Calabrese, Francesca Rinaldi, Paola Grossi, Irene Mattisi, Valentina Bernardi, Alice Favaretto, Paola Perini, and Paolo Gallo. "Basal ganglia and frontal/parietal cortical atrophy is associated with fatigue in relapsing—remitting multiple sclerosis". In: *Multiple Sclerosis Journal* 16.10 (2010). PMID: 20670981, pp. 1220–1228. DOI: 10.1177/1352458510376405. eprint: https://doi.org/10.1177/1352458510376405. URL: https://doi.org/10.1177/1352458510376405.

[80] Massimiliano Calabrese and Marco Castellaro. "Cortical gray matter MR imaging in multiple sclerosis". In: *Neuroimaging Clinics* 27.2 (2017), pp. 301–312.

[81] Nancy D Chiaravalloti and John DeLuca. "Cognitive impairment in multiple sclerosis". In: *The Lancet Neurology* 7.12 (2008), pp. 1139–1151.

[82] Valentina Zipoli, Benedetta Goretti, Bahia Hakiki, et al. "Cognitive impairment predicts conversion to multiple sclerosis in clinically isolated syndromes". In: *Multiple Sclerosis Journal* 16.1 (2010), pp. 62–67.

[83] Michael Lanz, Horst K Hahn, and Helmut Hildebrandt. "Brain atrophy and cognitive impairment in multiple sclerosis: a review". In: *Journal of Neurology* 254.2 (2007), pp. II43–II48.

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

[85] Evren Özarslan, Cheng Guan Koay, Timothy M Shepherd, Michal E Komlosh, M Okan İrfanoğlu, Carlo Pierpaoli, and Peter J Basser. "Mean apparent propagator (MAP) MRI: a novel diffusion imaging method for mapping tissue microstructure". In: *NeuroImage* 78 (2013), pp. 16–32.

[86] M Bester, JH Jensen, JS Babb, et al. "Non-Gaussian diffusion MRI of gray matter is associated with cognitive impairment in multiple sclerosis". In: *Multiple Sclerosis Journal* 21.7 (2015), pp. 935–944.

[87] Marco Lorenzi, Andre Altmann, Boris Gutman, et al. "Susceptibility of brain atrophy to TRIB3 in Alzheimer's disease, evidence from functional prioritization

in imaging genetics". In: *Proceedings of the National Academy of Sciences* 115.12 (2018), pp. 3162–3167.

[88] Marco Pitteri, Chiara Romualdi, Roberta Magliozzi, et al. "Cognitive impairment predicts disability progression and cortical thinning in MS: An 8-year study". In: *Multiple Sclerosis Journal* 23.6 (2017), pp. 848–854.

[89] Massimiliano Calabrese, Richard Reynolds, Roberta Magliozzi, et al. "Regional distribution and evolution of gray matter damage in different populations of multiple sclerosis patients". In: *PloS one* 10.8 (2015), e0135428.

[90] Herman Wold. "Path models with latent variables: The NIPALS approach". In: *Quantitative sociology*. Elsevier, 1975, pp. 307–357.

[91] Anthony Randal McIntosh and Nancy J Lobaugh. "Partial least squares analysis of neuroimaging data: applications and advances". In: *Neuroimage* 23 (2004), S250–S263.

[92] Paolo Preziosa, Elisabetta Pagani, Maria E Morelli, et al. "DT MRI microstructural cortical lesion damage does not explain cognitive impairment in MS". In: *Multiple Sclerosis Journal* 23.14 (2017), pp. 1918–1928.

[93] Vincent Planche, Aurélie Ruet, Pierrick Coupé, et al. "Hippocampal microstructural damage correlates with memory impairment in clinically isolated syndrome suggestive of multiple sclerosis". In: *Multiple Sclerosis Journal* 23.9 (2017), pp. 1214–1224.

[94] Alexandru V Avram, Joelle E Sarlls, Alan S Barnett, et al. "Clinical feasibility of using mean apparent propagator (MAP) MRI to characterize brain tissue microstructure". In: *NeuroImage* 127 (2016), pp. 422–434.

[95] Irene Cristofori, Shira Cohen-Zimerman, and Jordan Grafman. "Executive functions". In: *Handbook of clinical neurology*. Vol. 163. Elsevier, 2019, pp. 197–219.

[96] Lorenza Brusini, Federica Cruciani, Ilaria Boscolo Galazzo, Marco Pitteri, Silvia F Storti, Massimiliano Calabrese, Marco Lorenzi, and Gloria Menegaz. "Multivariate data analysis suggests the link between brain microstructure and cognitive impairment in multiple sclerosis". In: *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2021, pp. 685–688.

[97] Barrie J Hurwitz. "The diagnosis of multiple sclerosis and the clinical subtypes". In: *Annals of Indian Academy of Neurology* 12.4 (2009), p. 226.

[98] Fred D Lublin, Stephen C Reingold, Jeffrey A Cohen, Gary R Cutter, Per Soelberg Sørensen, Alan J Thompson, Jerry S Wolinsky, Laura J Balcer, Brenda Banwell, Frederik Barkhof, et al. "Defining the clinical course of multiple sclerosis: the 2013 revisions". In: *Neurology* 83.3 (2014), pp. 278–286.

[99] Claudia Lucchinetti, Wolfgang Brück, Joseph Parisi, Bernd Scheithauer, Moses Rodriguez, and Hans Lassmann. "Heterogeneity of multiple sclerosis lesions:

implications for the pathogenesis of demyelination". In: *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society* 47.6 (2000), pp. 707–717.

[100] Silvia De Santis, Tobias Granberg, Russell Ouellette, Constantina A Treaba, Elena Herranz, Qiuyun Fan, Caterina Mainero, and Nicola Toschi. "Evidence of early microstructural white matter abnormalities in multiple sclerosis from multi-shell diffusion MRI". In: *NeuroImage: Clinical* 22 (2019), p. 101699.

[101] Massimo Filippi and Matilde Inglese. "Overview of diffusion-weighted magnetic resonance studies in multiple sclerosis". In: *Journal of the neurological sciences* 186 (2001), S37–S43.

[102] Evren Özarslan and Thomas H Mareci. "Generalized diffusion tensor imaging and analytical relationships between diffusion tensor imaging and high angular resolution diffusion imaging". In: *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* 50.5 (2003), pp. 955–965.

[103] Ileana O Jelescu and Matthew D Budde. "Design and validation of diffusion MRI models of white matter". In: *Frontiers in physics* 5 (2017), p. 61.

[104] Philip SJ Weston, Ivor JA Simpson, Natalie S Ryan, Sebastien Ourselin, and Nick C Fox. "Diffusion imaging changes in grey matter in Alzheimer's disease: a potential marker of early neurodegeneration". In: *Alzheimer's research & therapy* 7.1 (2015), p. 47.

[105] MA Rocca, A Ceccarelli, A Falini, P Tortorella, B Colombo, E Pagani, G Comi, G Scotti, and M Filippi. "Diffusion tensor magnetic resonance imaging at 3.0 tesla shows subtle cerebral grey matter abnormalities in patients with migraine". In: *Journal of Neurology, Neurosurgery & Psychiatry* 77.5 (2006), pp. 686–689.

[106] Celia Oreja-Guevara, Marco Rovaris, Giuseppe Iannucci, Paola Valsasina, Domenico Caputo, Rosella Cavarretta, Maria Pia Sormani, Pasquale Ferrante, Giancarlo Comi, and Massimo Filippi. "Progressive gray matter damage in patients with relapsing-remitting multiple sclerosis: a longitudinal diffusion tensor magnetic resonance imaging study". In: *Archives of Neurology* 62.4 (2005), pp. 578–584.

[107] Marco Bozzali, Mara Cercignani, Maria Pia Sormani, Giancarlo Comi, and Massimo Filippi. "Quantification of brain gray matter damage in different MS phenotypes by use of diffusion tensor MR imaging". In: *American Journal of Neuroradiology* 23.6 (2002), pp. 985–988.

[108] Ilaria Boscolo Galazzo, Lorenza Brusini, Muge Akinci, Federica Cruciani, Abulena Bajrami, Marco Castellaro, Ahmed Abdo, Francesca B. Pizzini, Jorge Jovicich, Massimiliano Calabrese, and Gloria Menegaz. "Microstructural modulations in the hippocampus allow to characterizing relapsing-remitting versus pri-

mary progressive multiple sclerosis". In: *MICCAI workshop BrainLesion*. Unpublished results.

[109] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. "Deep inside convolutional networks: Visualising image classification models and saliency maps". In: *arXiv preprint arXiv:1312.6034* (2013).

[110] Matthew D Zeiler and Rob Fergus. "Visualizing and understanding convolutional networks". In: *European conference on computer vision*. Springer. 2014, pp. 818–833.

[111] Alina Lopatina, Stefan Ropele, Renat Sibgatulin, Jürgen R Reichenbach, and Daniel Güllmar. "Investigation of deep-learning-driven identification of multiple sclerosis patients based on susceptibility-weighted images using relevance analysis". In: *Frontiers in neuroscience* 14 (2020), p. 609468.

[112] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. "Learning important features through propagating activation differences". In: *arXiv preprint arXiv:1704.02685* (2017).

[113] Aldo Marzullo, Gabriel Kocevar, Claudio Stamile, Françoise Durand-Dubief, Giorgio Terracina, Francesco Calimeri, and Dominique Sappey-Marinier. "Classification of Multiple Sclerosis Clinical Profiles via Graph Convolutional Neural Networks". In: *Frontiers in neuroscience* 13 (2019), p. 594.

[114] Yu-Dong Zhang, Chichun Pan, Junding Sun, and Chaosheng Tang. "Multiple sclerosis identification by convolutional neural network with dropout and parametric ReLU". In: *Journal of computational science* 28 (2018), pp. 1–10.

[115] Shui-Hua Wang, Chaosheng Tang, Junding Sun, Jingyuan Yang, Chenxi Huang, Preetha Phillips, and Yu-Dong Zhang. "Multiple sclerosis identification by 14-layer convolutional neural network with batch normalization, dropout, and stochastic pooling". In: *Frontiers in neuroscience* 12 (2018).

[116] Paul Schmidt, Christian Gaser, Milan Arsic, Dorothea Buck, Annette Förschler, Achim Berthele, Muna Hoshi, Rüdiger Ilg, Volker J Schmid, Claus Zimmer, et al. "An automated tool for detection of FLAIR-hyperintense white-matter lesions in multiple sclerosis". In: *Neuroimage* 59.4 (2012), pp. 3774–3783.

[117] Bruce Fischl. "FreeSurfer". In: *Neuroimage* 62.2 (2012), pp. 774–781.

[118] Sergey Korolev, Amir Safiullin, Mikhail Belyaev, and Yulia Dodonova. "Residual and plain convolutional neural networks for 3D brain MRI classification". In: *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*. IEEE. 2017, pp. 835–838.

[119] Johannes Rieke, Fabian Eitel, Martin Weygandt, John-Dylan Haynes, and Kerstin Ritter. "Visualizing convolutional networks for MRI-based diagnosis of Alzheimer's

disease". In: *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*. Springer, 2018, pp. 24–31.

[120] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep Residual Learning for Image Recognition". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2016, pp. 770–778.

[121] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[122] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. "Automatic differentiation in PyTorch". In: *Conference on Neural Information Processing Systems (NIPS)*. 2017.

[123] Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, Klaus-Robert Müller, and Wojciech Samek. "Layer-Wise Relevance Propagation for Neural Networks with Local Renormalization Layers". In: *Artificial Neural Networks and Machine Learning – ICANN 2016*. Ed. by Alessandro E.P. Villa, Paolo Masulli, and Antonio Javier Pons Rivero. Cham: Springer International Publishing, 2016, pp. 63–71. ISBN: 978-3-319-44781-0.

[124] Maximilian Alber, Sebastian Lapuschkin, Philipp Seegerer, Miriam Hägele, Kristof T Schütt, Grégoire Montavon, Wojciech Samek, Klaus-Robert Müller, Sven Dähne, and Pieter-Jan Kindermans. "iNNvestigate neural networks!" In: *J. Mach. Learn. Res.* 20.93 (2019), pp. 1–8.

[125] Hanneke E Hulst and Jeroen JG Geurts. "Gray matter imaging in multiple sclerosis: what have we learned?" In: *BMC neurology* 11.1 (2011), p. 153.

[126] Jeroen JG Geurts, Massimiliano Calabrese, Elizabeth Fisher, and Richard A Rudick. "Measurement and clinical effect of grey matter pathology in multiple sclerosis". In: *The Lancet Neurology* 11.12 (2012), pp. 1082–1092.

[127] Massimiliano Calabrese, Richard Reynolds, Roberta Magliozzi, Marco Castellaro, Aldo Morra, Antonio Scalfari, Gabriele Farina, Chiara Romualdi, Alberto Gajofatto, Marco Pitteri, et al. "Regional distribution and evolution of gray matter damage in different populations of multiple sclerosis patients". In: *PloS one* 10.8 (2015), e0135428.

[128] Arman Eshaghi, Razvan V Marinescu, Alexandra L Young, Nicholas C Firth, Ferran Prados, M Jorge Cardoso, Carmen Tur, Floriana De Angelis, Niamh Cawley, Wallace J Brownlee, et al. "Progression of regional grey matter atrophy in multiple sclerosis". In: *Brain* 141.6 (2018), pp. 1665–1677.

[129] Roberta Magliozzi, Owain W. Howell, Richard Nicholas, Carolina Cruciani, Marco Castellaro, Chiara Romualdi, Stefania Rossi, Marco Pitteri, Maria Donata Benedetti, Alberto Gajofatto, Francesca B. Pizzini, Stefania Montemezzi, Sarah Rasia, Rug-

gero Capra, Alessandra Bertoldo, Francesco Facchiano, Salvatore Monaco, Richard Reynolds, and Massimiliano Calabrese. "Inflammatory intrathecal profiles and cortical damage in multiple sclerosis". In: *Annals of Neurology* 83.4 (2018), pp. 739–755. DOI: 10.1002/ana.25197. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ana.25197. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/ana.25197.

[130] Hernan Inojosa, Undine Proschmann, Katja Akgün, and Tjalf Ziemssen. "A focus on secondary progressive multiple sclerosis (SPMS): challenges in diagnosis and definition". In: *Journal of neurology* (2019), pp. 1–12.

[131] Marco Vercellino, Silvia Masera, Marcella Lorenzatti, Cecilia Condello, Aristide Merola, Alessandra Mattioda, Antonella Tribolo, Elisabetta Capello, Giovanni Luigi Mancardi, Roberto Mutani, et al. "Demyelination, inflammation, and neurodegeneration in multiple sclerosis deep gray matter". In: *Journal of Neuropathology & Experimental Neurology* 68.5 (2009), pp. 489–502.

[132] Jeroen JG Geurts and Frederik Barkhof. "Grey matter pathology in multiple sclerosis". In: *The Lancet Neurology* 7.9 (2008), pp. 841–851.

[133] Asaf Achiron, Joab Chapman, Sigal Tal, Eran Bercovich, Hararai Gil, and Anat Achiron. "Superior temporal gyrus thickness correlates with cognitive performance in multiple sclerosis". In: *Brain Structure and Function* 218.4 (2013), pp. 943–950.

[134] Stephen M. Smith, Mark Jenkinson, Mark W. Woolrich, Christian F. Beckmann, Timothy E.J. Behrens, Heidi Johansen-Berg, Peter R. Bannister, Marilena De Luca, Ivana Drobnjak, David E. Flitney, Rami K. Niazy, James Saunders, John Vickers, Yongyue Zhang, Nicola De Stefano, J. Michael Brady, and Paul M. Matthews. "Advances in functional and structural MR image analysis and implementation as FSL". In: *NeuroImage* 23 (2004). Mathematics in Brain Imaging, S208–S219. ISSN: 1053-8119. DOI: https://doi.org/10.1016/j.neuroimage.2004.07.051. URL: http://www.sciencedirect.com/science/article/pii/S1053811904003933.

[135] Mauro Zucchelli, Samuel Deslauriers-Gauthier, and Rachid Deriche. "A computational Framework for generating rotation invariant features and its application in diffusion MRI". In: *Medical Image Analysis* 60 (2020), p. 101597.

[136] Eleftherios Garyfallidis, Matthew Brett, Bagrat Amirbekian, Ariel Rokem, Stefan Van Der Walt, Maxime Descoteaux, and Ian Nimmo-Smith. "Dipy, a library for the analysis of diffusion MRI data". In: *Frontiers in neuroinformatics* 8 (2014), p. 8.

[137] Alexandru V Avram, Joelle E Sarlls, Alan S Barnett, Evren Özarslan, Cibu Thomas, M Okan Irfanoglu, Elizabeth Hutchinson, Carlo Pierpaoli, and Peter J Basser.

"Clinical feasibility of using mean apparent propagator (MAP) MRI to characterize brain tissue microstructure". In: *NeuroImage* 127 (2016), pp. 422–434.

[138] Ilaria Boscolo Galazzo, Lorenza Brusini, Silvia Obertino, Mauro Zucchelli, Cristina Granziera, and Gloria Menegaz. "On the Viability of Diffusion MRI-Based Microstructural Biomarkers in Ischemic Stroke". In: *Frontiers in Neuroscience* 12 (2018), p. 92. ISSN: 1662-453X. DOI: 10.3389/fnins.2018.00092. URL: https://www.frontiersin.org/article/10.3389/fnins.2018.00092.

[139] Lorenza Brusini, Silvia Obertino, Ilaria Boscolo Galazzo, Mauro Zucchelli, Gunnar Krueger, Cristina Granziera, and Gloria Menegaz. "Ensemble average propagator-based detection of microstructural alterations after stroke". In: *International journal of computer assisted radiology and surgery* 11.9 (2016), pp. 1585–1597.

[140] Richard Dinga, Lianne Schmaal, Brenda WJH Penninx, Dick J Veltman, and Andre F Marquand. "Controlling for effects of confounding variables on machine learning predictions". In: *BioRxiv* (2020).

[141] Mauro Zucchelli, Lorenza Brusini, C Andrés Méndez, Alessandro Daducci, Cristina Granziera, and Gloria Menegaz. "What lies beneath? Diffusion EAP-based study of brain tissue microstructure". In: *Medical image analysis* 32 (2016), pp. 145–156.

[142] Massimiliano Calabrese, Francesca Rinaldi, Dario Seppi, Alice Favaretto, Letizia Squarcina, Irene Mattisi, Paola Perini, Alessandra Bertoldo, and Paolo Gallo. "Cortical diffusion-tensor imaging abnormalities in multiple sclerosis: a 3-year longitudinal study". In: *Radiology* 261.3 (2011), pp. 891–898.

[143] Alexandra Kutzelnigg, Jens C Faber-Rod, Jan Bauer, Claudia F Lucchinetti, Per S Sorensen, Henning Laursen, Christine Stadelmann, Wolfgang Brück, Helmut Rauschka, Manfred Schmidbauer, et al. "Widespread demyelination in the cerebellar cortex in multiple sclerosis". In: *Brain pathology* 17.1 (2007), pp. 38–44.

[144] Martijn D Steenwijk, Jeroen JG Geurts, Marita Daams, Betty M Tijms, Alle Meije Wink, Lisanne J Balk, Prejaas K Tewarie, Bernard MJ Uitdehaag, Frederik Barkhof, Hugo Vrenken, et al. "Cortical atrophy patterns in multiple sclerosis are non-random and clinically relevant". In: *Brain* 139.1 (2016), pp. 115–126.

[145] Jeroen JG Geurts, Lars Bö, Stefan D Roosendaal, Thierry Hazes, Richard Daniëls, Frederik Barkhof, Menno P Witter, Inge Huitinga, and Paul van der Valk. "Extensive hippocampal demyelination in multiple sclerosis". In: *Journal of Neuropathology & Experimental Neurology* 66.9 (2007), pp. 819–827.

[146] Ilaria Boscolo Galazzo, Lorenza Brusini, Silvia Obertino, Mauro Zucchelli, Cristina Granziera, and Gloria Menegaz. "On the viability of diffusion MRI-based microstructural biomarkers in ischemic stroke". In: *Frontiers in neuroscience* 12 (2018), p. 92.

[147] Federica Cruciani, Lorenza Brusini, Mauro Zucchelli, Gustavo Retuci Pinheiro, Francesco Setti, Ilaria Boscolo Galazzo, Rachid Deriche, Leticia Rittner, Massimiliano Calabrese, and Gloria Menegaz. "Explainable 3D-CNN for multiple sclerosis patients stratification". In: *International Conference on Pattern Recognition*. Springer. 2021, pp. 103–114.

[148] Federica Cruciani, Lorenza Brusini, Mauro Zucchelli, G Retuci Pinheiro, Francesco Setti, I Boscolo Galazzo, Rachid Deriche, Leticia Rittner, Massimiliano Calabrese, and Gloria Menegaz. "Interpretable deep learning as a means for decrypting disease signature in multiple sclerosis". In: *Journal of Neural Engineering* 18.4 (2021), 0460a6.

[149] Harald Hampel, Rhoda Au, Soeren Mattke, Wiesje M van der Flier, Paul Aisen, Liana Apostolova, Christopher Chen, Min Cho, Susan De Santi, Peng Gao, et al. "Designing the next-generation clinical care pathway for Alzheimer's disease". In: *Nature Aging* 2.8 (2022), pp. 692–703.

[150] Clifford R Jack Jr, David A Bennett, Kaj Blennow, Maria C Carrillo, Billy Dunn, Samantha Budd Haeberlein, David M Holtzman, William Jagust, Frank Jessen, Jason Karlawish, et al. "NIA-AA research framework: toward a biological definition of Alzheimer's disease". In: *Alzheimer's & Dementia* 14.4 (2018), pp. 535–562.

[151] Giovanni B Frisoni, Nick C Fox, Clifford R Jack, Philip Scheltens, and Paul M Thompson. "The clinical use of structural MRI in Alzheimer disease". In: *Nature Reviews Neurology* 6.2 (2010), pp. 67–77.

[152] Giovanni B Frisoni, Daniele Altomare, Dietmar Rudolf Thal, Federica Ribaldi, Rik van der Kant, Rik Ossenkoppele, Kaj Blennow, Jeffrey Cummings, Cornelia van Duijn, Peter M Nilsson, et al. "The probabilistic model of Alzheimer disease: the amyloid hypothesis revised". In: *Nature Reviews Neuroscience* 23.1 (2022), pp. 53–66.

[153] Meredith N Braskie, Arthur W Toga, and Paul M Thompson. "Recent advances in imaging Alzheimer's disease". In: *Journal of Alzheimer's Disease* 33.s1 (2013), S313–S327.

[154] Thomas Steckler and Giacomo Salvadore. "Neuroimaging as a Translational Tool in Animal and Human Models of Schizophrenia". In: *Translational Neuroimaging*. Elsevier, 2013, pp. 195–220.

[155] Marzia A Scelsi, Raiyan R Khan, Marco Lorenzi, Leigh Christopher, Michael D Greicius, Jonathan M Schott, Sebastien Ourselin, and Andre Altmann. "Genetic study of multimodal imaging Alzheimer's disease progression score implicates novel loci". In: *Brain* 141.7 (2018), pp. 2167–2180.

[156] Elizabeth C Mormino, Reisa A Sperling, Avram J Holmes, Randy L Buckner, Philip L De Jager, Jordan W Smoller, Mert R Sabuncu, Alzheimer's Disease Neuroimag-

ing Initiative, et al. “Polygenic risk of Alzheimer disease is associated with early- and late-life processes”. In: *Neurology* 87.5 (2016), pp. 481–488.

[157] Heba Elshatoury, Federica Cruciani, Francesco Zumerle, Silvia F Storti, André Altmann, Marco Lorenzi, Gholamreza Anbarjafari, Gloria Menegaz, and Ilaria Boscolo Galazzo. “Disentangling the association between genetics and functional connectivity in Mild Cognitive Impairment”. In: *2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*. IEEE. 2021, pp. 1–4.

[158] Mert R Sabuncu, Randy L Buckner, Jordan W Smoller, Phil Hyoun Lee, Bruce Fischl, Reisa A Sperling, and Alzheimer’s Disease Neuroimaging Initiative. “The association between a polygenic Alzheimer score and cortical thickness in clinically normal subjects”. In: *Cerebral cortex* 22.11 (2012), pp. 2653–2661.

[159] Aurina Arnatkeviciute, Ben D Fulcher, Mark A Bellgrove, and Alex Fornito. “Imaging transcriptomics of brain disorders”. In: *Biological Psychiatry Global Open Science* (2021).

[160] Brian W Kunkle, Benjamin Grenier-Boley, Rebecca Sims, Joshua C Bis, Vincent Damotte, Adam C Naj, Anne Boland, Maria Vronskaya, Sven J Van Der Lee, Alexandre Amlie-Wolf, et al. “Genetic meta-analysis of diagnosed Alzheimer’s disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing”. In: *Nature genetics* 51.3 (2019), pp. 414–430.

[161] Adrià Casamitjana, Paula Petrone, José Luis Molinuevo, et al. “Projection to Latent Spaces Disentangles Pathological Effects on Brain Morphology in the Asymptomatic Phase of Alzheimer’s Disease”. In: *Frontiers in neurology* 11 (2020), p. 648.

[162] Ellen H Singleton, Yolande AL Pijnenburg, Carole H Sudre, Colin Groot, Elena Kochova, Frederik Barkhof, Renaud La Joie, Howard J Rosen, William W Seeley, Bruce Miller, et al. “Investigating the clinico-anatomical dissociation in the behavioral variant of Alzheimer disease”. In: *Alzheimer’s research & therapy* 12.1 (2020), pp. 1–12.

[163] Hanyi Chen, Eric de Silva, Carole H Sudre, Jo Barnes, Alexandra L Young, Neil P Oxtoby, Frederik Barkhof, Daniel C Alexander, and Andre Altmann. “What do data-driven Alzheimer’s disease subtypes tell us about white matter pathology and clinical progression?” In: *Alzheimer’s & Dementia* 17 (2021), e054028.

[164] Federica Cruciani, Andre Altmann, Marco Lorenzi, Gloria Menegaz, and Ilaria Boscolo Galazzo. “What PLS can still do for Imaging Genetics in Alzheimer’s disease”. In: *2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*. IEEE. 2022, pp. 1–4.

[165] Serge Gauthier, Barry Reisberg, Michael Zaudig, et al. “Mild cognitive impairment”. In: *The lancet* 367.9518 (2006), pp. 1262–1270.

[166] Michael W Weiner, Dallas P Veitch, Paul S Aisen, et al. “2014 Update of the Alzheimer’s Disease Neuroimaging Initiative: a review of papers published since its inception”. In: *Alzheimer’s & dementia* 11.6 (2015), e1–e120.

[167] International Schizophrenia Consortium. “Common polygenic variation contributes to risk of schizophrenia that overlaps with bipolar disorder”. In: *Nature* 460.7256 (2009), p. 748.

[168] Christian Clemm von Hohenberg, Marlene C Wigand, Marek Kubicki, et al. “CNTNAP2 polymorphisms and structural brain connectivity: a diffusion-tensor imaging study”. In: *Journal of psychiatric research* 47.10 (2013), pp. 1349–1356.

[169] Emrin Horgusluoglu-Moloch, Gaoyu Xiao, Minghui Wang, et al. “Systems modeling of white matter microstructural abnormalities in Alzheimer’s disease”. In: *NeuroImage: Clinical* 26 (2020), p. 102203.

[170] Bharat B Biswal, Joel Van Kylen, and James S Hyde. “Simultaneous assessment of flow and BOLD signals in resting-state functional connectivity maps”. In: *NMR in Biomedicine* 10.4-5 (1997), pp. 165–170.

[171] David M Cole, Stephen M Smith, and Christian F Beckmann. “Advances and pitfalls in the analysis and interpretation of resting-state FMRI data”. In: *Frontiers in systems neuroscience* (2010), p. 8.

[172] Andre Altmann, Marzia A Scelsi, Maryam Shoai, Eric de Silva, Leon M Aksman, David M Cash, John Hardy, Jonathan M Schott, and Alzheimer’s Disease Neuroimaging Initiative. “A comprehensive analysis of methods for assessing polygenic burden on Alzheimer’s disease pathology and risk beyond APOE”. In: *Brain communications* 2.1 (2020), fcz047.

[173] Xiaoyan Sun, David Salat, Kristen Upchurch, et al. “Destruction of white matter integrity in patients with mild cognitive impairment and Alzheimer disease”. In: *Journal of investigative medicine* 62.7 (2014), pp. 927–933.

[174] Nicolai Franzmeier, Julia Neitzel, Anna Rubinski, Ruben Smith, Olof Strandberg, Rik Ossenkoppele, Oskar Hansson, and Michael Ewers. “Functional brain architecture is associated with the rate of tau accumulation in Alzheimer’s disease”. In: *Nature communications* 11.1 (2020), pp. 1–17.

[175] Alexander Schaefer, Ru Kong, Evan M Gordon, Timothy O Laumann, Xi-Nian Zuo, Avram J Holmes, Simon B Eickhoff, and BT Thomas Yeo. “Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI”. In: *Cerebral cortex* 28.9 (2018), pp. 3095–3114.

[176] Bianca De Blasi, Lorenzo Caciagli, Silvia Francesca Storti, Marian Galovic, Matthias Koepp, Gloria Menegaz, Anna Barnes, and Ilaria Boscolo Galazzo. “Noise re-

moval in resting-state and task fMRI: functional connectivity and activation maps". In: *Journal of Neural Engineering* 17.4 (2020), p. 046040.

[177] Jian Zhai and Ke Li. "Predicting brain age based on spatial and temporal features of human brain functional networks". In: *Frontiers in human neuroscience* 13 (2019), p. 62.

[178] Michel J Grothe, Stefan J Teipel, and Alzheimer's Disease Neuroimaging Initiative. "Spatial patterns of atrophy, hypometabolism, and amyloid deposition in Alzheimer's disease correspond to dissociable functional brain networks". In: *Human brain mapping* 37.1 (2016), pp. 35–53.

[179] Christian Sorg, Valentin Riedl, Mark Mühlau, Vince D Calhoun, Tom Eichele, Leonhard Läer, Alexander Drzezga, Hans Förstl, Alexander Kurz, Claus Zimmer, et al. "Selective changes of resting-state networks in individuals at risk for Alzheimer's disease". In: *Proceedings of the National Academy of Sciences* 104.47 (2007), pp. 18760–18765.

[180] Maja AA Binnewijzend, Menno M Schoonheim, Ernesto Sanz-Arigita, Alle Meije Wink, Wiesje M van der Flier, Nelleke Tolboom, Sofie M Adriaanse, Jessica S Damoiseaux, Philip Scheltens, Bart NM van Berckel, et al. "Resting-state fMRI changes in Alzheimer's disease and mild cognitive impairment". In: *Neurobiology of aging* 33.9 (2012), pp. 2018–2028.

[181] Pan Wang, Bo Zhou, Hongxiang Yao, Yafeng Zhan, Zengqiang Zhang, Yue Cui, Kaibin Xu, Jianhua Ma, Luning Wang, Ningyu An, et al. "Aberrant intra-and inter-network connectivity architectures in Alzheimer's disease and mild cognitive impairment". In: *Scientific reports* 5.1 (2015), pp. 1–12.

[182] Federica Cruciani, Lorenza Brusini, Giorgio Dolci, Ilaria Boscolo Galazzo, Andre Altmann, Marco Lorenzi, and Gloria Menegaz. "A multivariate imaging study for decrypting the link between polygenic risk scores and microstructure in Mild Cognitive Impairment". In: *2021 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*. IEEE. 2021, pp. 1–4.

[183] Jingyu Liu and Vince D Calhoun. "A review of multivariate analyses in imaging genetics". In: *Frontiers in neuroinformatics* 8 (2014), p. 29.

[184] Martin J Prince, Fan Wu, Yanfei Guo, Luis M Gutierrez Robledo, Martin O'Donnell, Richard Sullivan, and Salim Yusuf. "The burden of disease in older people and implications for health policy and practice". In: *The Lancet* 385.9967 (2015), pp. 549–562.

[185] Meredith N Braskie and Paul M Thompson. "Understanding cognitive deficits in Alzheimer's disease based on neuroimaging findings". In: *Trends in cognitive sciences* 17.10 (2013), pp. 510–516.

[186] Jean-Charles Lambert, Carla A Ibrahim-Verbaas, Denise Harold, Adam C Naj, Rebecca Sims, Céline Bellenguez, Gyungah Jun, Anita L DeStefano, Joshua C Bis, Gary W Beecham, et al. "Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease". In: *Nature genetics* 45.12 (2013), pp. 1452–1458.

[187] Iris E Jansen, Jeanne E Savage, Kyoko Watanabe, Julien Bryois, Dylan M Williams, Stacy Steinberg, Julia Sealock, Ida K Karlsson, Sara Hägg, Lavinia Athanasiu, et al. "Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk". In: *Nature genetics* 51.3 (2019), pp. 404–413.

[188] Yu Xin, Jinhua Sheng, Miao Miao, Luyun Wang, Ze Yang, and He Huang. "A review of imaging genetics in Alzheimer's disease". In: *Journal of Clinical Neuroscience* 100 (2022), pp. 155–163.

[189] Susanne G Mueller, Michael W Weiner, Leon J Thal, Ronald C Petersen, Clifford Jack, William Jagust, John Q Trojanowski, Arthur W Toga, and Laurel Beckett. "The Alzheimer's disease neuroimaging initiative". In: *Neuroimaging Clinics* 15.4 (2005), pp. 869–877.

[190] Michael W Weiner, Dallas P Veitch, Paul S Aisen, Laurel A Beckett, Nigel J Cairns, Robert C Green, Danielle Harvey, Clifford R Jack Jr, William Jagust, John C Morris, et al. "The Alzheimer's Disease Neuroimaging Initiative 3: Continued innovation for clinical trial improvement". In: *Alzheimer's & Dementia* 13.5 (2017), pp. 561–571.

[191] Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, et al. "UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age". In: *PLoS medicine* 12.3 (2015), e1001779.

[192] Paul M Thompson, Jason L Stein, Sarah E Medland, Derrek P Hibar, Alejandro Arias Vasquez, Miguel E Renteria, Roberto Toro, Neda Jahanshad, Gunter Schumann, Barbara Franke, et al. "The ENIGMA Consortium: large-scale collaborative analyses of neuroimaging and genetic data". In: *Brain imaging and behavior* 8.2 (2014), pp. 153–182.

[193] Natalia Vilor-Tejedor, Diego Garrido-Martın, Blanca Rodriguez-Fernandez, Sander Lamballais, Roderic Guigó, and Juan Domingo Gispert. "Multivariate Analysis and Modelling of multiple Brain endOphenotypes: Let's MAMBO!" In: *Computational and Structural Biotechnology Journal* 19 (2021), pp. 5800–5810.

[194] Chun Chieh Fan, Robert Loughnan, Carolina Makowski, Diliana Pecheva, Chi-Hua Chen, Donald J Hagler, Wesley K Thompson, Nadine Parker, Dennis van der Meer, Oleksandr Frei, et al. "Multivariate genome-wide association study

on tissue-sensitive diffusion metrics highlights pathways that shape the human brain". In: *Nature communications* 13.1 (2022), pp. 1–10.

[195] Michael C Wu, Seunggeun Lee, Tianxi Cai, Yun Li, Michael Boehnke, and Xihong Lin. "Rare-variant association testing for sequencing data with the sequence kernel association test". In: *The American Journal of Human Genetics* 89.1 (2011), pp. 82–93.

[196] Priyanka Nakka, Benjamin J Raphael, and Sohini Ramachandran. "Gene and network analysis of common variants reveals novel associations in multiple complex diseases". In: *Genetics* 204.2 (2016), pp. 783–798.

[197] Zhao-Hua Lu, Hongtu Zhu, Rebecca C Knickmeyer, Patrick F Sullivan, Stephanie N Williams, Fei Zou, and Alzheimer's Disease Neuroimaging Initiative. "Multiple SNP Set Analysis for Genome-Wide Association Studies Through Bayesian Latent Variable Selection". In: *Genetic epidemiology* 39.8 (2015), pp. 664–677.

[198] Kwangsik Nho, Sungeun Kim, Emrin Horgusluoglu, Shannon L Risacher, Li Shen, Dokyoon Kim, Seunggeun Lee, Tatiana Foroud, Leslie M Shaw, John Q Trojanowski, et al. "Association analysis of rare variants near the APOE region with CSF and neuroimaging biomarkers of Alzheimer's disease". In: *BMC medical genomics* 10.1 (2017), pp. 45–52.

[199] Joshua C Bis, Xueqiu Jian, Brian W Kunkle, Yuning Chen, Kara L Hamilton-Nelson, William S Bush, William J Salerno, Daniel Lancour, Yiyi Ma, Alan E Renton, et al. "Whole exome sequencing study identifies novel rare and common Alzheimer's-Associated variants involved in immune response and transcriptional regulation". In: *Molecular psychiatry* 25.8 (2020), pp. 1859–1875.

[200] Pascal Zille, Vince D Calhoun, and Yu-Ping Wang. "Enforcing co-expression within a brain-imaging genomics regression framework". In: *IEEE transactions on medical imaging* 37.12 (2017), pp. 2561–2571.

[201] Xiaoke Hao, Chanxiu Li, Lei Du, Xiaohui Yao, Jingwen Yan, Shannon L Risacher, Andrew J Saykin, Li Shen, and Daoqiang Zhang. "Mining outcome-relevant brain imaging genetic associations via three-way sparse canonical correlation analysis in Alzheimer's disease". In: *Scientific reports* 7.1 (2017), pp. 1–12.

[202] Édith Le Floch, Vincent Guillemot, Vincent Frouin, Philippe Pinel, Christophe Lalanne, Laura Trinchera, Arthur Tenenhaus, Antonio Moreno, Monica Zilbovicius, Thomas Bourgeron, et al. "Significant correlation between a set of genetic polymorphisms and a functional brain network revealed by feature selection and sparse Partial Least Squares". In: *Neuroimage* 63.1 (2012), pp. 11–24.

[203] Claudia Grellmann, Sebastian Bitzer, Jane Neumann, Lars T Westlye, Ole A Andreassen, Arno Villringer, and Annette Horstmann. "Comparison of variants of

canonical correlation analysis and partial least squares for combined analysis of MRI and genetic data". In: *Neuroimage* 107 (2015), pp. 289–310.

[204] Mark Jenkinson, Christian F Beckmann, Timothy EJ Behrens, Mark W Woolrich, and Stephen M Smith. "Fsl". In: *Neuroimage* 62.2 (2012), pp. 782–790.

[205] Bruce Fischl, David H Salat, Evelina Busa, Marilyn Albert, Megan Dieterich, Christian Haselgrove, Andre Van Der Kouwe, Ron Killiany, David Kennedy, Shuna Klaveness, et al. "Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain". In: *Neuron* 33.3 (2002), pp. 341–355.

[206] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel AR Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul IW De Bakker, Mark J Daly, et al. "PLINK: a tool set for whole-genome association and population-based linkage analyses". In: *The American journal of human genetics* 81.3 (2007), pp. 559–575.

[207] Daniel S Himmelstein and Sergio E Baranzini. "Heterogeneous network edge prediction: a data integration approach to prioritize disease-associated genes". In: *PLoS computational biology* 11.7 (2015), e1004259.

[208] Marc Gillespie, Bijay Jassal, Ralf Stephan, Marija Milacic, Karen Rothfels, Andrea Senff-Ribeiro, Johannes Griss, Cristoffer Sevilla, Lisa Matthews, Chuqiao Gong, et al. "The reactome pathway knowledgebase 2022". In: *Nucleic acids research* 50.D1 (2022), pp. D687–D692.

[209] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[210] Evelina Sjöstedt, Wen Zhong, Linn Fagerberg, Max Karlsson, Nicholas Mitsios, Csaba Adori, Per Oksvold, Fredrik Edfors, Agnieszka Limiszewska, Feria Hikmet, et al. "An atlas of the protein-coding genes in the human, pig, and mouse brain". In: *Science* 367.6482 (2020), eaay5947.

[211] Vivian Tam, Nikunj Patel, Michelle Turcotte, Yohan Bossé, Guillaume Paré, and David Meyre. "Benefits and limitations of genome-wide association studies". In: *Nature Reviews Genetics* 20.8 (2019), pp. 467–484.

[212] Paula I Moreira, Akihiko Nunomura, Masao Nakamura, Atsushi Takeda, Justin C Shenk, Gjumrakch Aliev, Mark A Smith, and George Perry. "Nucleic acid oxidation in Alzheimer disease". In: *Free Radical Biology and Medicine* 44.8 (2008), pp. 1493–1505.

[213] Eun-Joo Shin, Chu Xuan Duong, Xuan-Khanh Thi Nguyen, Zhengyi Li, Guoying Bing, Jae-Hyung Bach, Dae Hun Park, Keiichi Nakayama, Syed F Ali, Anumantha G Kanthasamy, et al. "Role of oxidative stress in methamphetamine-induced

dopaminergic toxicity mediated by protein kinase Cδ". In: *Behavioural brain research* 232.1 (2012), pp. 98–113.

[214] Constance Chace, Deborah Pang, Catherine Weng, Alexis Temkin, Simon Lax, Wayne Silverman, Warren Zigman, Michel Ferin, Joseph H Lee, Benjamin Tycko, et al. "Variants in CYP17 and CYP19 cytochrome P450 genes are associated with onset of Alzheimer's disease in women with down syndrome". In: *Journal of Alzheimer's Disease* 28.3 (2012), pp. 601–612.

[215] Natalia Mast, Aicha Saadane, Ana Valencia-Olvera, James Constans, Erin Maxfield, Hiroyuki Arakawa, Young Li, Gary Landreth, and Irina A Pikuleva. "Cholesterol-metabolizing enzyme cytochrome P450 46A1 as a pharmacologic target for Alzheimer's disease". In: *Neuropharmacology* 123 (2017), pp. 465–476.

[216] Razvan Marinescu, Arman Eshaghi, Daniel Alexander, and Polina Golland. "BrainPainter: A software for the visualisation of brain structures, biomarkers and associated pathological processes". In: *arXiv preprint arXiv:1905.08627* (2019).

[217] Rui-Xian Li, Ya-Hui Ma, Lan Tan, and Jin-Tai Yu. "Prospective Biomarkers of Alzheimer's Disease: A Systematic Review and Meta-analysis". In: *Ageing Research Reviews* (2022), p. 101699.

[218] Xiaofeng Zhu, Heung-Il Suk, Heng Huang, and Dinggang Shen. "Structured sparse low-rank regression model for brain-wide and genome-wide associations". In: *International conference on medical image computing and computer-assisted intervention*. Springer. 2016, pp. 344–352.

[219] Xiaoqian Wang, Hong Chen, Jingwen Yan, Kwangsik Nho, Shannon L Risacher, Andrew J Saykin, Li Shen, Heng Huang, and ADNI. "Quantitative trait loci identification for brain endophenotypes via new additive model with random networks". In: *Bioinformatics* 34.17 (2018), pp. i866–i874.

[220] International Schizophrenia Consortium Manuscript preparation Purcell Shaun M. shaun@ pngu. mgh. harvard. edu 1 2 b Wray Naomi R. 5 Stone Jennifer L. 1 2 3 4 Visscher Peter M. 5 O'Donovan Michael C. 6 Sullivan Patrick F. 7 Sklar Pamela 1 2 3 4, Polygene analyses subgroup Wray Naomi R. 5 Macgregor Stuart 5 Sklar Pamela sklar@ chgr. mgh. harvard. edu 1 2 3 4 d Sullivan Patrick F. 7 O'Donovan Michael C. 6 Visscher Peter M. 5, Hugh Gurling, Douglas HR Blackwood, Aiden Corvin, Nick J Craddock, Michael Gill, Christina M Hultman, George K Kirov, et al. "Common polygenic variation contributes to risk of schizophrenia and bipolar disorder". In: *Nature* 460.7256 (2009), pp. 748–752.

[221] Valentina Escott-Price, Rebecca Sims, Christian Bannister, Denise Harold, Maria Vronskaya, Elisa Majounie, Nandini Badarinarayan, Gerad/Perades, IGAP consortia, Kevin Morgan, et al. "Common polygenic variation enhances risk prediction for Alzheimer's disease". In: *Brain* 138.12 (2015), pp. 3673–3684.

[222] Rahul S Desikan, Chun Chieh Fan, Yunpeng Wang, Andrew J Schork, Howard J Cabral, L Adrienne Cupples, Wesley K Thompson, Lilah Besser, Walter A Kukull, Dominic Holland, et al. "Genetic assessment of age-associated Alzheimer disease risk: Development and validation of a polygenic hazard score". In: *PLoS medicine* 14.3 (2017), e1002258.

[223] Derrek P Hibar, Jason L Stein, Omid Kohannim, Neda Jahanshad, Andrew J Saykin, Li Shen, Sungeun Kim, Nathan Pankratz, Tatiana Foroud, Matthew J Huentelman, et al. "Voxelwise gene-wide association study (vGeneWAS): multivariate gene-based association testing in 731 elderly subjects". In: *Neuroimage* 56.4 (2011), pp. 1875–1891.

[224] Tian Ge, Jianfeng Feng, Derrek P Hibar, Paul M Thompson, and Thomas E Nichols. "Increasing power for voxel-wise genome-wide association studies: the random field theory, least square kernel machines and fast permutation procedures". In: *Neuroimage* 63.2 (2012), pp. 858–873.

[225] Hua Wang, Feiping Nie, Heng Huang, Sungeun Kim, Kwangsik Nho, Shannon L Risacher, Andrew J Saykin, Li Shen, and Alzheimer's Disease Neuroimaging Initiative. "Identifying quantitative trait loci via group-sparse multitask regression and feature selection: an imaging genetics study of the ADNI cohort". In: *Bioinformatics* 28.2 (2012), pp. 229–237.

[226] Keelin Greenlaw, Elena Szefer, Jinko Graham, Mary Lesperance, Farouk S Nathoo, and Alzheimer's Disease Neuroimaging Initiative. "A Bayesian group sparse multitask regression model for imaging genetics". In: *Bioinformatics* 33.16 (2017), pp. 2513–2522.

[227] Liana G Apostolova, Shannon L Risacher, Tugce Duran, Eddie C Stage, Naira Goukasian, John D West, Triet M Do, Jonathan Grotts, Holly Wilhalme, Kwangsik Nho, et al. "Associations of the top 20 Alzheimer disease risk variants with brain amyloidosis". In: *JAMA neurology* 75.3 (2018), pp. 328–341.

[228] Zhiyuan Xu, Chong Wu, Wei Pan, Alzheimer's Disease Neuroimaging Initiative, et al. "Imaging-wide association study: integrating imaging endophenotypes in GWAS". In: *Neuroimage* 159 (2017), pp. 159–169.

[229] Flora H Duits, Gunnar Brinkmalm, Charlotte E Teunissen, Ann Brinkmalm, Philip Scheltens, Wiesje M Van der Flier, Henrik Zetterberg, and Kaj Blennow. "Synaptic proteins in CSF as potential novel biomarkers for prognosis in prodromal Alzheimer's disease". In: *Alzheimer's research & therapy* 10.1 (2018), pp. 1–9.

[230] Gabriele A Fontana, Julia K Reinert, Nicolas H Thomä, and Ulrich Rass. "Shepherding DNA ends: Rif1 protects telomeres and chromosome breaks". In: *Microbial Cell* 5.7 (2018), p. 327.

[231] Ayush Noori, Aziz M Mezlini, Bradley T Hyman, Alberto Serrano-Pozo, and Sudeshna Das. “Systematic review and meta-analysis of human transcriptomics reveals neuroinflammation, deficient energy metabolism, and proteostasis failure across neurodegeneration”. In: *Neurobiology of disease* 149 (2021), p. 105225.

[232] Alexander M Kleschevnikov, Jessica Yu, Jeesun Kim, Larisa V Lysenko, Zheng Zeng, Y Eugene Yu, and William C Mobley. “Evidence that increased Kcnj6 gene dose is necessary for deficits in behavior and dentate gyrus synaptic plasticity in the Ts65Dn mouse model of Down syndrome”. In: *Neurobiology of disease* 103 (2017), pp. 1–10.

[233] Ira T Lott and Elizabeth Head. “Dementia in Down syndrome: unique insights for Alzheimer disease research”. In: *Nature Reviews Neurology* 15.3 (2019), pp. 135–147.

[234] Heike Wulff and Boris S Zhorov. “K+ channel modulators for the treatment of neurological disorders and autoimmune diseases”. In: *Chemical reviews* 108.5 (2008), pp. 1744–1773.

[235] Abolfazl Heidari, Chanakan Tongsook, Reza Najafipour, Luciana Musante, Nasim Vasli, Masoud Garshasbi, Hao Hu, Kirti Mittal, Amy JM McNaughton, Kumudesh Sritharan, et al. “Mutations in the histamine N-methyltransferase gene, HNMT, are associated with nonsyndromic autosomal recessive intellectual disability”. In: *Human molecular genetics* 24.20 (2015), pp. 5697–5710.

[236] Diego Sepulveda-Falla, Alvaro Barrera-Ocampo, Christian Hagel, Anne Korwitz, Maria Fernanda Vinueza-Veloz, Kuikui Zhou, Martijn Schonewille, Haibo Zhou, Luis Velazquez-Perez, Roberto Rodriguez-Labrada, et al. “Familial Alzheimer’s disease–associated presenilin-1 alters cerebellar activity and calcium homeostasis”. In: *The Journal of clinical investigation* 124.4 (2014), pp. 1552–1567.

[237] Radka Václavıková, David J Hughes, and Pavel Souček. “Microsomal epoxide hydrolase 1 (EPHX1): Gene, structure, function, and role in human disease”. In: *Gene* 571.1 (2015), pp. 1–8.

[238] Hazal Haytural, Georgios Mermelekas, Ceren Emre, Saket Milind Nigam, Steven L Carroll, Bengt Winblad, Nenad Bogdanovic, Gaël Barthet, Ann-Charlotte Granholm, Lukas M Orre, et al. “The proteome of the dentate terminal zone of the perforant path indicates presynaptic impairment in Alzheimer disease”. In: *Molecular & Cellular Proteomics* 19.1 (2020), pp. 128–141.

[239] Vladimir Ilievski, Paulina K Zuchowska, Stefan J Green, Peter T Toth, Michael E Ragozzino, Khuong Le, Haider W Aljewari, Neil M O’Brien-Simpson, Eric C Reynolds, and Keiko Watanabe. “Chronic oral application of a periodontal pathogen results in brain inflammation, neurodegeneration and amyloid beta production in wild type mice”. In: *PloS one* 13.10 (2018), e0204941.

[240] Yi-Qian Sun, Rebecca C Richmond, Yue Chen, and Xiao-Mei Mai. "Mixed evidence for the relationship between periodontitis and Alzheimer's disease: A bidirectional Mendelian randomization study". In: *PloS one* 15.1 (2020), e0228206.

[241] Frances K Wiseman, Tamara Al-Janabi, John Hardy, Annette Karmiloff-Smith, Dean Nizetic, Victor LJ Tybulewicz, Elizabeth Fisher, and André Strydom. "A genetic cause of Alzheimer disease: mechanistic insights from Down syndrome". In: *Nature Reviews Neuroscience* 16.9 (2015), pp. 564–574.

[242] Matilde Cescon, Peiwen Chen, Silvia Castagnaro, Ilaria Gregorio, and Paolo Bonaldo. "Lack of collagen VI promotes neurodegeneration by impairing autophagy and inducing apoptosis during aging". In: *Aging (Albany NY)* 8.5 (2016), p. 1083.

[243] Michael Zech, Daniel D Lam, Ludmila Francescatto, Barbara Schormair, Aaro V Salminen, Angela Jochim, Thomas Wieland, Peter Lichtner, Annette Peters, Christian Gieger, et al. "Recessive mutations in the α3 (VI) collagen gene COL6A3 cause early-onset isolated dystonia". In: *The American Journal of Human Genetics* 96.6 (2015), pp. 883–893.

[244] Richard Sherva, Alden Gross, Shubhabrata Mukherjee, Ryan Koesterer, Philippe Amouyel, Celine Bellenguez, Carole Dufouil, David A Bennett, Lori Chibnik, Carlos Cruchaga, et al. "Genome-wide association study of rate of cognitive decline in Alzheimer's disease patients identifies novel genes and pathways". In: *Alzheimer's & Dementia* 16.8 (2020), pp. 1134–1145.

[245] Federica Cruciani, Antonino Aparo, Lorenza Brusini, Carlo Combi, Silvia F. Storti, Rosalba Giugno, Gloria Menegaz, and Ilraia Boscolo Galazzo. "Identifying the joint signature of brain atrophy and gene variant scores in the Alzheimer's Disease". In: *Journal of Biomedical Informatics* (In submission).

[246] Farouk S Nathoo, Linglong Kong, Hongtu Zhu, and Alzheimer's Disease Neuroimaging Initiative. "A review of statistical methods in imaging genetics". In: *Canadian Journal of Statistics* 47.1 (2019), pp. 108–131.

[247] Luigi Antelmi, Nicholas Ayache, Philippe Robert, and Marco Lorenzi. "Sparse Multi-Channel Variational Autoencoder for the Joint Analysis of Heterogeneous Data". In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. PMLR, 2019, pp. 302–311.

[248] Andres Diaz-Pinto, Nishant Ravikumar, Rahman Attar, Avan Suinesiaputra, Yitian Zhao, Eylem Levelt, Erica Dall'Armellina, Marco Lorenzi, Qingyu Chen, Tiarnan DL Keenan, et al. "Predicting myocardial infarction through retinal scans and minimal personal information". In: *Nature Machine Intelligence* 4.1 (2022), pp. 55–61.

[249] Clément Abi Nader, Nicholas Ayache, Giovanni B Frisoni, Philippe Robert, Marco Lorenzi, and Alzheimer's Disease Neuroimaging Initiative. "Simulating the out-

come of amyloid treatments in Alzheimer's disease from imaging and clinical data". In: *Brain communications* 3.2 (2021), fcab091.

[250] Brian B Avants, Nick Tustison, and Gang Song. "Advanced normalization tools (ANTS)". In: *Insight j* 2.365 (2009), pp. 1–35.

[251] Diederik P Kingma and Max Welling. "Auto-encoding variational bayes". In: *arXiv preprint arXiv:1312.6114* (2013).

[252] Scott M Lundberg and Su-In Lee. "A unified approach to interpreting model predictions". In: *Advances in neural information processing systems* 30 (2017).

[253] "SimulAD: a dynamical model for personalized simulation and disease staging in Alzheimer's disease". In: *Neurobiology of Aging* 113 (2022), pp. 73–83. ISSN: 0197-4580. DOI: https://doi.org/10.1016/j.neurobiolaging.2021.12.015. URL: https://www.sciencedirect.com/science/article/pii/S0197458021003699.

[254] Mine Öğretir, Siddharth Ramchandran, Dimitrios Papatheodorou, and Harri Lähdesmäki. "A Variational Autoencoder for Heterogeneous Temporal and Longitudinal Data". In: *arXiv preprint arXiv:2204.09369* (2022).

[255] Federica Cruciani, Ettore Cinquetti, Lorenza Brusini, Antonino Aparo, Carlo Combi, Ilaria Boscolo Galazzo, and Gloria Menegaz. "Exploring the potential of MCVAE for patients stratification and skewed data compensation across the AD continuum". In: (In preparation).

[256] Ke Wang, Jing Tian, Chu Zheng, Hong Yang, Jia Ren, Yanling Liu, Qinghua Han, and Yanbo Zhang. "Interpretable prediction of 3-year all-cause mortality in patients with heart failure caused by coronary heart disease based on machine learning and SHAP". In: *Computers in Biology and Medicine* 137 (2021), p. 104813.

[257] Raquel Rodrıguez-Pérez and Jürgen Bajorath. "Interpretation of machine learning models using shapley values: application to compound potency and multi-target activity predictions". In: *Journal of computer-aided molecular design* 34.10 (2020), pp. 1013–1026.

[258] Kjersti Aas, Martin Jullum, and Anders Løland. "Explaining individual predictions when features are dependent: More accurate approximations to Shapley values". In: *Artificial Intelligence* 298 (2021), p. 103502.

[259] Marthe S Veldhuis, Simone Ariëns, Rolf JF Ypma, Thomas Abeel, and Corina CG Benschop. "Explainable artificial intelligence in forensics: Realistic explanations for number of contributor predictions of DNA profiles". In: *Forensic Science International: Genetics* 56 (2022), p. 102632.

[260] Ronald Carl Petersen, PS Aisen, Laurel A Beckett, MC Donohue, AC Gamst, Danielle J Harvey, CR Jack, WJ Jagust, LM Shaw, AW Toga, et al. "Alzheimer's

disease neuroimaging initiative (ADNI): clinical characterization". In: *Neurology* 74.3 (2010), pp. 201–209.

[261] Naomi George, Edward Moseley, Rene Eber, Jennifer Siu, Mathew Samuel, Jonathan Yam, Kexin Huang, Leo Anthony Celi, and Charlotta Lindvall. "Deep learning to predict long-term mortality in patients requiring 7 days of mechanical ventilation". In: *PloS one* 16.6 (2021), e0253443.

[262] Sonu Subudhi, Ashish Verma, Ankit B Patel, C Corey Hardin, Melin J Khandekar, Hang Lee, Dustin McEvoy, Triantafyllos Stylianopoulos, Lance L Munn, Sayon Dutta, et al. "Comparing machine learning algorithms for predicting ICU admission and mortality in COVID-19". In: *NPJ digital medicine* 4.1 (2021), pp. 1–7.

[263] Qin-Yu Zhao, Huan Wang, Jing-Chao Luo, Ming-Hao Luo, Le-Ping Liu, Shen-Ji Yu, Kai Liu, Yi-Jie Zhang, Peng Sun, Guo-Wei Tu, et al. "Development and validation of a machine-learning model for prediction of extubation failure in intensive care units". In: *Frontiers in medicine* (2021), p. 654.

[264] Ahmed Salih, Ilaria Boscolo Galazzo, Federica Cruciani, Lorenza Brusini, and Petia Radeva. "Investigating Explainable Artificial Intelligence for MRI-based Classification of Dementia: a New Stability Criterion for Explainable Methods". In: *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2022, pp. 4003–4007.

[265] Serap Aydın, Çağdaş Güdücü, Fırat Kutluk, Adile Öniz, and Murat Özgören. "The impact of musical experience on neural sound encoding performance". In: *Neuroscience letters* 694 (2019), pp. 124–128.

[266] Alexandre Abraham, Fabian Pedregosa, Michael Eickenberg, Philippe Gervais, Andreas Mueller, Jean Kossaifi, Alexandre Gramfort, Bertrand Thirion, and Gaël Varoquaux. "Machine learning for neuroimaging with scikit-learn". In: *Frontiers in neuroinformatics* (2014), p. 14.

[267] Richard W Bohannon. "Motricity index scores are valid indicators of paretic upper extremity strength following stroke". In: *Journal of Physical Therapy Science* 11.2 (1999), pp. 59–61.

[268] Silvia Campagnini, Piergiuseppe Liuzzi, Andrea Mannini, Benedetta Basagni, Claudio Macchi, Maria Chiara Carrozza, and Francesca Cecchi. "Cross-validation of predictive models for functional recovery after post-stroke rehabilitation". In: *Journal of NeuroEngineering and Rehabilitation* 19.1 (2022), pp. 1–11.

[269] Ceren Tozlu, Dylan Edwards, Aaron Boes, Douglas Labar, K Zoe Tsagaris, Joshua Silverstein, Heather Pepper Lane, Mert R Sabuncu, Charles Liu, and Amy Kuceyeski. "Machine learning methods predict individual upper-limb motor impairment

following therapy in chronic stroke". In: *Neurorehabilitation and neural repair* 34.5 (2020), pp. 428–439.

[270] Scott M Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. "From local explanations to global understanding with explainable AI for trees". In: *Nature machine intelligence* 2.1 (2020), pp. 56–67.

[271] Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. "Explainable ai: A review of machine learning interpretability methods". In: *Entropy* 23.1 (2020), p. 18.

[272] Bo Norrving and Brett Kissela. "The global burden of stroke and need for a continuum of care". In: *Neurology* 80.3 Supplement 2 (2013), S5–S12.

[273] Thomas Truelsen, B Piechowski-Jóźwiak, Ruth Bonita, Colin Mathers, Julien Bogousslavsky, and Gudrun Boysen. "Stroke incidence and prevalence in Europe: a review of available data". In: *European journal of neurology* 13.6 (2006), pp. 581–598.

[274] Stefano Paolucci, Maria Grazia Grasso, Gabriella Antonucci, Maura Bragoni, Elio Troisi, Daniela Morelli, Paola Coiro, Domenico De Angelis, and Francesco Rizzi. "Mobility status after inpatient stroke rehabilitation: 1-year follow-up and prognostic factors". In: *Archives of Physical Medicine and Rehabilitation* 82.1 (2001), pp. 2–8.

[275] Samar M Hatem, Geoffroy Saussez, Margaux Della Faille, Vincent Prist, Xue Zhang, Delphine Dispa, and Yannick Bleyenheuft. "Rehabilitation of motor function after stroke: a multiple systematic review focused on techniques to stimulate upper extremity recovery". In: *Frontiers in human neuroscience* 10 (2016), p. 442.

[276] Melissa Yeo, Hong Kuan Kok, Numan Kutaiba, Julian Maingard, Vincent Thijs, Bahman Tahayori, Jeremy Russell, Ashu Jhamb, Ronil V Chandra, Mark Brooks, et al. "Artificial intelligence in clinical decision support and outcome prediction–applications in stroke". In: *Journal of medical imaging and radiation oncology* 65.5 (2021), pp. 518–528.

[277] Anna K Bonkhoff and Christian Grefkes. "Precision medicine in stroke: towards personalized outcome predictions using artificial intelligence". In: *Brain* 145.2 (2022), pp. 457–475.

[278] Silvia Campagnini, Chiara Arienti, Michele Patrini, Piergiuseppe Liuzzi, Andrea Mannini, and Maria Chiara Carrozza. "Machine learning methods for functional recovery prediction and prognosis in post-stroke rehabilitation: a systematic review". In: *Journal of NeuroEngineering and Rehabilitation* 19.1 (2022), pp. 1–22.

[279] Takeshi Imura, Yuji Iwamoto, Tetsuji Inagawa, Naoki Imada, Ryo Tanaka, Haruki Toda, Yu Inoue, Hayato Araki, and Osamu Araki. “Decision tree algorithm identifies stroke patients likely discharge home after rehabilitation using functional and environmental predictors”. In: *Journal of Stroke and Cerebrovascular Diseases* 30.4 (2021), p. 105636.

[280] DS Ouellette, C Timple, SE Kaplan, SS Rosenberg, and ER Rosario. “Predicting discharge destination with admission outcome scores in stroke patients”. In: *NeuroRehabilitation* 37.2 (2015), pp. 173–179.

[281] Vivek P Gupta, Andrew LA Garton, Jonathan A Sisti, Brandon R Christophe, Aaron S Lord, Ariane K Lewis, Hans-Peter Frey, Jan Claassen, and E Sander Connolly Jr. “Prognosticating functional outcome after intracerebral hemorrhage: the ICHOP score”. In: *World neurosurgery* 101 (2017), pp. 577–583.

[282] JoonNyung Heo, Jihoon G Yoon, Hyungjong Park, Young Dae Kim, Hyo Suk Nam, and Ji Hoe Heo. “Machine learning–based model for prediction of outcomes in acute stroke”. In: *Stroke* 50.5 (2019), pp. 1263–1265.

[283] Jeoung Kun Kim, Yoo Jin Choo, and Min Cheol Chang. “Prediction of motor function in stroke patients using machine learning algorithm: Development of practical models”. In: *Journal of Stroke and Cerebrovascular Diseases* 30.8 (2021), p. 105856.

[284] Wan-Yin Lin, Chun-Hsien Chen, Yi-Ju Tseng, Yu-Ting Tsai, Ching-Yu Chang, Hsin-Yao Wang, and Chih-Kuang Chen. “Predicting post-stroke activities of daily living through a machine learning-based approach on initiating rehabilitation”. In: *International journal of medical informatics* 111 (2018), pp. 159–164.

[285] Jeanette Plantin, Marion Verneau, Alison K Godbolt, Gaia Valentina Pennati, Evaldas Laurencikas, Birgitta Johansson, Lena Krumlinde-Sundholm, Jean-Claude Baron, Jörgen Borg, and Påvel G Lindberg. “Recovery and prediction of bimanual hand use after stroke”. In: *Neurology* 97.7 (2021), e706–e719.

[286] B Bobath. “Adult hemiplegia: evaluation and treatment London”. In: *William Heinnemann* (1978).

[287] Cathy M Stinear, Winston D Byblow, Suzanne J Ackerley, Marie-Claire Smith, Victor M Borges, and P Alan Barber. “PREP2: A biomarker-based algorithm for predicting upper limb function after stroke”. In: *Annals of clinical and translational neurology* 4.11 (2017), pp. 811–820.

[288] Rinske HM Nijland, Erwin EH van Wegen, Barbara C Harmeling-van der Wel, and Gert Kwakkel. “Presence of finger extension and shoulder abduction within 72 hours after stroke predicts functional recovery: early prediction of functional outcome after stroke: the EPOS cohort study”. In: *stroke* 41.4 (2010), pp. 745–750.

[289] Gert Kwakkel, Boudewijn J Kollen, Jeroen van der Grond, and Arie JH Prevo. "Probability of regaining dexterity in the flaccid upper limb: impact of severity of paresis and time since onset in acute stroke". In: *Stroke* 34.9 (2003), pp. 2181–2186.

[290] Cathy M Stinear, P Alan Barber, Matthew Petoe, Samir Anwar, and Winston D Byblow. "The PREP algorithm predicts potential for upper limb recovery after stroke". In: *Brain* 135.8 (2012), pp. 2527–2535.

[291] Marie-Claire Smith, Suzanne J Ackerley, P Alan Barber, Winston D Byblow, and Cathy M Stinear. "PREP2 algorithm predictions are correct at 2 years poststroke for most patients". In: *Neurorehabilitation and neural repair* 33.8 (2019), pp. 635–642.

[292] Sarah B Zandvliet, Gert Kwakkel, Rinske HM Nijland, Erwin EH van Wegen, and Carel GM Meskers. "Is recovery of somatosensory impairment conditional for upper-limb motor recovery early after stroke?" In: *Neurorehabilitation and neural repair* 34.5 (2020), pp. 403–416.

[293] David J Gladstone, Cynthia J Danells, and Sandra E Black. "The Fugl-Meyer assessment of motor recovery after stroke: a critical review of its measurement properties". In: *Neurorehabilitation and neural repair* 16.3 (2002), pp. 232–240.

[294] David A Belsley, Edwin Kuh, and Roy E Welsch. *Regression diagnostics: Identifying influential data and sources of collinearity*. John Wiley & Sons, 2005.

[295] Leo Breiman. "Random forests". In: *Machine learning* 45.1 (2001), pp. 5–32.

[296] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *Random forests, The elements of statistical learning*. 2009.

[297] Leo Breiman. "Manual on setting up, using, and understanding random forests v3. 1". In: *Statistics Department University of California Berkeley, CA, USA* 1.58 (2002), pp. 3–42.

[298] Evangelia Christodoulou, Jie Ma, Gary S Collins, Ewout W Steyerberg, Jan Y Verbakel, and Ben Van Calster. "A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models". In: *Journal of clinical epidemiology* 110 (2019), pp. 12–22.

[299] Yaar Harari, Megan K O'Brien, Richard L Lieber, and Arun Jayaraman. "Inpatient stroke rehabilitation: prediction of clinical outcomes using a machine-learning approach". In: *Journal of neuroengineering and rehabilitation* 17.1 (2020), pp. 1–10.

[300] Carlos Fernandez-Lozano, Pablo Hervella, Virginia Mato-Abad, Manuel Rodrıguez-Yáñez, Sonia Suárez-Garaboa, Iria López-Dequidt, Ana Estany-Gestal, Tomás Sobrino, Francisco Campos, José Castillo, et al. "Random forest-based prediction of stroke outcome". In: *Scientific reports* 11.1 (2021), pp. 1–12.

[301] Carlo Dindorf, Wolfgang Teufl, Bertram Taetz, Gabriele Bleser, and Michael Fröhlich. “Interpretability of input representations for gait classification in patients after total hip arthroplasty”. In: *Sensors* 20.16 (2020), p. 4385.

[302] Angela Lombardi, Domenico Diacono, Nicola Amoroso, Alfonso Monaco, João Manuel RS Tavares, Roberto Bellotti, and Sabina Tangaro. “Explainable deep learning for personalized age prediction with brain morphology”. In: *Frontiers in neuroscience* (2021), p. 578.

[303] Ahmed Salih, Ilaria Boscolo Galazzo, Zahra Raisi-Estabragh, Steffen E Petersen, Polyxeni Gkontra, Karim Lekadir, Gloria Menegaz, and Petia Radeva. “A new scheme for the assessment of the robustness of explainable methods applied to brain age estimation”. In: *2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE. 2021, pp. 492–497.

[304] Xinlei Mi, Baiming Zou, Fei Zou, and Jianhua Hu. “Permutation-based identification of important biomarkers for complex diseases via machine learning models”. In: *Nature communications* 12.1 (2021), pp. 1–12.

[305] Elvio Amparore, Alan Perotti, and Paolo Bajardi. “To trust or not to trust an explanation: using LEAF to evaluate local linear XAI methods”. In: *PeerJ Computer Science* 7 (2021), e479.

[306] Miguel Monteiro, Ana Catarina Fonseca, Ana Teresa Freitas, Teresa Pinho e Melo, Alexandre P Francisco, Jose M Ferro, and Arlindo L Oliveira. “Using machine learning to improve the prediction of functional outcome in ischemic stroke patients”. In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 15.6 (2018), pp. 1953–1959.

[307] Ching-Heng Lin, Kai-Cheng Hsu, Kory R Johnson, Yang C Fann, Chon-Haw Tsai, YU Sun, Li-Ming Lien, Wei-Lun Chang, Po-Lin Chen, Cheng-Li Lin, et al. “Evaluation of machine learning methods to stroke outcome prediction using a nationwide disease registry”. In: *Computer methods and programs in biomedicine* 190 (2020), p. 105381.

[308] Jungsoo Lee, Heegoo Kim, Jinuk Kim, Won Hyuk Chang, and Yun-Hee Kim. “Multimodal Imaging Biomarker-Based Model Using Stratification Strategies for Predicting Upper Extremity Motor Recovery in Severe Stroke Patients”. In: *Neurorehabilitation and Neural Repair* 36.3 (2022), pp. 217–226.

[309] Marialuisa Gandolfi, Ilaria Boscolo Galazzo, Rudy Gasparin Pavan, Federica Cruciani, Alessandro Picelli, Silvia Francesca Storti, Nicola Smania, and Gloria Menegaz. “eXplainable AI allows predicting upper limb rehabilitation outcomes in subacute stroke patients”. In: *IEEE Journal of Biomedical and Health Informatics* (2022).

[310] Colin Studholme, Derek LG Hill, and David J Hawkes. “An overlap invariant entropy measure of 3D medical image alignment”. In: *Pattern recognition* 32.1 (1999), pp. 71–86.

[311] Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. *Captum: A unified and generic model interpretability library for PyTorch*. 2020. arXiv: 2009.07896 [cs.LG].

[312] Hao Zhang, Jiayi Chen, Haotian Xue, and Quanshi Zhang. “Towards a unified evaluation of explanation methods without ground truth”. In: *arXiv preprint arXiv:1911.09017* (2019).

[313] Federica Cruciani, Lorenza Brusini, Mauro Zucchelli, G Retuci Pinheiro, Francesco Setti, I Boscolo Galazzo, Rachid Deriche, Leticia Rittner, Massimiliano Calabrese, and Gloria Menegaz. “Explainable Deep Learning for decrypting disease signatures in Multiple Sclerosis”. In: *Explainable Deep Learning AI Methods and Challenges*. Elsevier, 202.

[314] Federica Cruciani et al. “An interpretability framework for MCVAE: preliminary results of the application to Imaging Genetics in Alzheimer’s disease”. In: *VIII Congress of the National Group of Bioengineering (GNB) 2023*. GNB. In submission.

[315] Paul S Aisen, Ronald C Petersen, Michael C Donohue, Anthony Gamst, Rema Raman, Ronald G Thomas, Sarah Walter, John Q Trojanowski, Leslie M Shaw, Laurel A Beckett, et al. “Clinical Core of the Alzheimer’s Disease Neuroimaging Initiative: progress and plans”. In: *Alzheimer’s & Dementia* 6.3 (2010), pp. 239–246.

[316] E Ozarslan, C Koay, TM Shepherd, SJ Blackb, and PJ Basser. “Simple harmonic oscillator based reconstruction and estimation for three-dimensional q-space MRI”. In: (2009).

[317] Saad Jbabdi and Heidi Johansen-Berg. “Tractography: where do we go from here?” In: *Brain connectivity* 1.3 (2011), pp. 169–183.

[318] Denis Le Bihan and Steven J Warach. *Diffusion and perfusion magnetic resonance imaging: applications to functional MRI*. 1995.

[319] Jörg Kärger and Wilfried Heink. “The propagator representation of molecular transport in microporous crystallites”. In: *Journal of Magnetic Resonance (1969)* 51.1 (1983), pp. 1–7.

[320] Edward O Stejskal and John E Tanner. “Spin diffusion measurements: spin echoes in the presence of a time-dependent field gradient”. In: *The journal of chemical physics* 42.1 (1965), pp. 288–292.

[321] John E Tanner and Edward O Stejskal. “Restricted self-diffusion of protons in colloidal systems by the pulsed-gradient, spin-echo method”. In: *The Journal of Chemical Physics* 49.4 (1968), pp. 1768–1777.

[322] Silvia De Santis, Andrea Gabrielli, Marco Palombo, Bruno Maraviglia, and Silvia Capuani. “Non-Gaussian diffusion imaging: a brief practical review”. In: *Magnetic resonance imaging* 29.10 (2011), pp. 1410–1416.

[323] Sylvain L Merlet and Rachid Deriche. “Continuous diffusion signal, EAP and ODF estimation via compressive sensing in diffusion MRI”. In: *Medical image analysis* 17.5 (2013), pp. 556–572.

[324] Mauro Zucchelli, Rutger Henri Jacques Fick, Rachid Deriche, and Gloria Menegaz. “Ensemble average propagator estimation of axon diameter in diffusion MRI: implications and limitations”. In: *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2016, pp. 465–468.

[325] Mauro Zucchelli, Samuel Deslauriers-Gauthier, and Rachid Deriche. “A closed-form solution of rotation invariant spherical harmonic features in diffusion mri”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2019, pp. 77–89.

[326] Emmanuel Caruyer and Ragini Verma. “On facilitating the use of HARDI in population studies by creating rotation-invariant markers”. In: *Medical image analysis* 20.1 (2015), pp. 87–96.

[327] Risi Kondor. “A novel set of rotationally and translationally invariant features for images based on the non-commutative bispectrum”. In: *arXiv preprint cs/0701127* (2007).

[328] Ramakrishna Kakarala and Dansheng Mao. “A theory of phase-sensitive rotation invariance with spherical harmonic and moment-based representations”. In: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE. 2010, pp. 105–112.

[329] Herbert HH Homeier and E Otto Steinborn. “Some properties of the coupling coefficients of real spherical harmonics and their relation to Gaunt coefficients”. In: *Journal of Molecular Structure: THEOCHEM* 368 (1996), pp. 31–37.

[330] L Shapley. “Quota solutions of n-person games”. In: *Edited by Emil Artin and Marston Morse* (1953), p. 343.

[331] Stefano Nembrini, Inke R König, and Marvin N Wright. “The revival of the Gini importance?” In: *Bioinformatics* 34.21 (2018), pp. 3711–3718.

Sommario

L'adozione di modelli di Intelligenza Artificiale (IA) in medicina e nelle neuroscienze ha il potenziale per svolgere un ruolo significativo non solo nel portare progressi scientifici, ma anche nel contribuire al processo decisionale clinico. Tuttavia, esistono molti dubbi e preoccupazioni a riguardo, dovute agli eventuali bias che l'IA potrebbe avere e che potrebbero avere conseguenze di vasta portata, soprattutto in un campo critico come quello della biomedicina.
É difficile ottenere modelli IA utilizzabili perché non solo é fondamentale imparare dai dati, estrarre la conoscenza e garantire la capacitá di generalizzazione, ma é anche necessario districare i fattori esplicativi sottostanti per comprendere a fondo le variabili che portano alle decisioni finali. Da qui la richiesta di approcci che aprano la "scatola nera" dell'IA per aumentare la fiducia e l'affidabilitá delle capacitá decisionali degli algoritmi di IA. Tali approcci sono comunemente definiti Explainable IA (XAI) e stanno iniziando ad essere applicati in campo medico, anche se non ancora pienamente sfruttati.
Con questa tesi intendiamo contribuire a rendere possibile l'uso dell'IA in medicina e nelle neuroscienze compiendo due passi fondamentali: (i) pervadere praticamente i modelli di IA con XAI (ii) proporre metodi di validatione per i modelli XAI.
Il primo passo é stato raggiunto da un lato concentrandosi sulla tassonomia XAI e proponendo alcune linee guida specifiche per le applicazioni di IA e XAI nel settore delle neuroscienze. Dall'altro lato, abbiamo affrontato questioni concrete proponendo soluzioni XAI per decodificare le modulazioni cerebrali nella neurodegenerazione basandoci sui cambiamenti morfologici, microstrutturali e funzionali che si verificano in diversi stadi della malattia e sulle loro connessioni con il substrato genotipico.
Il secondo passo é stato raggiunto definendo innanzitutto quattro attributi relativi alla validazione delle XAI, ovvero stabilitá, coerenza, comprensibilitá e plausibilitá. Ciascun attributo si riferisce a un aspetto diverso della XAI, che va dalla valutazione del-

la stabilitá delle spiegazioni tra diversi metodi XAI, o tra input altamente collineari, all'allineamento delle spiegazioni ottenute con lo stato dell'arte della letteratura. Abbiamo quindi proposto diverse tecniche di validazione che mirano a soddisfare praticamente tali requisiti.
Con questa tesi, abbiamo contribuito all'avanzamento della ricerca sulla XAI, con lo scopo di aumentare la consapevolezza e l'uso critico dei metodi di IA, aprendo la strada ad applicazioni reali che consentano lo sviluppo di una medicina e di un trattamento personalizzati, adottando un approccio oggettivo e guidato dai dati all'assistenza sanitaria.