University of Tartu

Faculty of Science and Technology

Institute of Technology

Heba Elshatoury

Disentangling the association between functional connectivity and genetics in Mild Cognitive Impairment via multivariate regression models

Master's thesis (30 ECTS)
Robotics and Computer Engineering

Supervisors:

Dr. Ilaria Boscolo Galazzo
Prof. Gloria Menegaz
Prof. Gholamreza Anbarjafari

Tartu 2021

Resümee

Kerge kognitiivse häire puhul funktsionaalse ühenduvuse ja geneetika vahelise seose lahti harutamine mitmemõõtmeliste regressioonimudelite abil

Selle uuringu eesmärk on analüüsida fenotüübi ja genotüübi vahelisi seoseid kerge kognitiivse häirega (MCI) ja kognitiivselt normaalsetel (CN) patsientidel, kasutades regressioonil põhinevat mudelit. Andmed koguti 177 katsealustelt, kelle andmed laeti alla Alzheimeri tõve neuropildistamise algatuse (ADNI) andmebaasist, neist 82 MCI ja 95 CN. Selles uurimistöös testiti mitmeid müra vähendamise kanaleid puhkeolekus funktsionaalse MRI andmete peal, et analüüsida funktsionaalset ühenduvust (FC) ja eraldada pildi atribuute. Kanalite jõudlust võrreldi ja valiti häiringregressioonil põhinev kanal. Pärast andmete eeltöötlemist, genereeriti kõigi katsealuste jaoks FC maatriksid. Rakendati atribuutide vähendamise tehnikat, et pakkida FC maatriksid 28 pildi atribuudiks. Lõpuks rakendati osaliste vähimruutude (PLS) mudelit piltide ja geneetiliste tunnuste peal, milles uuriti korrelatsioone. Mudeli olulisuse hindamiseks ($p < 0,05$) viidi läbi permutatsiooni test. Uuriti LASSO tulemustega PLS-mudelit ja esitati seosed piltide ja geneetiliste tunnuste vahel.

CERCS: T111 Pilditehnika; T121 Signaalitöötlus; B110 Bioinformaatika, meditsiiniinformaatika, biomatemaatika, biomeetrika

Märksõnad: arvutinägemine, funktsionaalne MRI, kerge kognitiivsete funktsioonide häire, funktsioonide vähendamine, regressioonanalüüs

Abstract

Disentangling the association between functional connectivity and genetics in Mild Cognitive Impairment via multivariate regression models

The aim of this study is to analyse links between phenotype and genotype for patients with Mild Cognitive Impariment (MCI) and Cognitively Normal (CN) using regression based model. Data was collected from 177 subjects downloaded from The Alzheimer's Disease Neuroimaging Initiative (ADNI) database out of which 82 MCI and 95 CN were considered. In this research work multiple noise reduction pipelines were tested on resting state functional MRI data to analyse functional connectivity (FC) and extract imaging features. The pipelines performance was compared and nuisance regression based pipeline was selected. After preprocessing of the data FC matrices were generated for all subjects. Feature reduction technique was implemented to summarize the FC matrices into 28 imaging features. Finally, Partial Least Squares (PLS) model was applied to imaging and genetic features in which correlations are examined. A permutation test was performed to evaluate the model significance ($p < 0.05$). PLS model with LASSO results were investigated and associations between the imaging and genetic features were reported.

CERCS: T111 Imaging, image processing; T121 Signal processing; B110 Bioinformatics, medical informatics, biomathematics, biometrics

Keywords: computer vision, functional magnetic resonance imaging, mild cognitive impairment, feature reduction, regression analysis

Acknowledgements

I would like to express my special gratitude to Dr. Ilaria Boscolo Galazzo for her absolute support, guidance and patience she has given me through out this research project. I would like to thank her for her time and dedication. Thank you for teaching me a lot of new skills and for being the mentor I look up to.

I would like to thank Professor Gloria Menegaz for her guidance and support and for giving me the opportunity to join her lab.

I would like to thank Professor Gholamreza Anbarjafari for his supervision and for his time and feedback through out this thesis project.

Finally, I would like to thank Professor Silvia Francesca Storti, Federica Cruciani, Francesco Zumerle, and Giorgio Dolci my colleagues at the Neuroimaging Lab in University of Verona for their assistance, advice and collaboration.

Content

List of Figures

List of Tables

Abbreviations and Acronyms

AD - Alzheimer's Disease

ADNI - Alzheimer's Disease Neuroimaging Initiative

BET - Brain extraction tool

BOLD - Blood oxygenation level dependent

CN - Cognitively Normal

CON - frontoparietal control network

CSF - Cerebrospinal fluid

DAN - dorsal attention network

DMN - default mode network

EMCI - Early Mild Cognitive Impairment

FA - Flip Angle

FC - Functional Connectivity

FOV - Field of view

fMRI - functional Magnetic Resonance Imaging

FSL - FMRIB Software Library

FWHM - Full width at half maximum

GLM - General Linear Model

GM - Gray Matter

GUI - Graphical user interface

GWAS - genome-wide association studies

Hz - Hertz

ICA - Independent Component Analysis

ICA-AROMA - ICA-based Automatic Removal Of Motion Artifacts

ICs - independent components

LASSO - least absolute shrinkage and selection operator

LH - left hemisphere

LIM - limbic network

LMCI - Late Mild Cognitive Impairment

MCI - Mild Cognitive Impairment

MNI - Montreal Neurological Institute

MRI - Magnetic Resonance Imaging

NIfTI - Neuroimaging Informatics Technology Initiative

NRP - Nuisance Regression pipeline

PLS - Partial Least Squares

PRS - Polygenic Risk Score

PVE - Partial Volume Estimate

RH - right hemisphere

ROI - Region of interest

rs-fMRI - resting-state functional Magnetic Resonance Imaging

RSNs - Resting State Networks

SMC - Significant Memory Concern

SMN - somatomotor network

SNPs - Single nucleotide polymorphisms

SVD - singular value decomposition

TE - Echo Time

TR - Repetition Time

VAN - ventral attention network

VIS - visual network

WM - White Matter

1 Introduction

1.1 Problem Overview

The human brain consists of a very complex and efficient network. Different regions in the brain are responsible for specific functions. The brain consists of structural and functional areas in which each has its role. Brain research is constantly ongoing with more discoveries made to understand this extremely complex organ. Neuroimaging and neuroscience studies have given the opportunity to utilize tools that can give us an insight into this system. As related to this research, neuroimaging techniques have given researchers and scientists the chance to visualize and analyse specific data to study brain anatomy, microstructure and functions and how diseases can affect it. In particular, blood oxygenation level dependent (BOLD) functional magnetic resonance imaging (fMRI) allows measuring functional activities that happen in the brain with the focus of specific regions or overall [1]. Functional connectivity (FC) is a type of analysis that can be extracted from fMRI, both during task or rest. One of the advantages of fMRI is that it is non-invasive method of evaluating neuronal activity in the brain either with the patient given a certain task or during rest referred to as task-based fMRI and resting state fMRI (rs-fMRI). The study of FC allowed to discovery of high temporal correlations in the timeseries measured in fMRI scan [2]. Biswal et al. [3] have researched FC in the motor cortex. Their study has concluded that measuring blood oxygenation levels during rest has showed functionally connected regions in the brain. With this discovery it was early on concluded that if timeseries of specific regions showed similar patterns in non-necessary anatomically connected parts of the brain then these regions are functionally coupled. Spontaneous, low frequency fluctuations (< 0.1 Hz) occurring in the BOLD signal at rest have proved the existence of spatially distinct brain areas sharing a similar BOLD activity, the so-called resting-state networks (RSNs). A graph representation of how these functionally connected networks are extracted is shown in figure 1.1. Heuvel and Pol [4] display the process of comparing timeseries in different nodes and calculating the correlation between them. FC could be represented using connectivity matrix which shows the level at which these networks and nodes are associated. In this study, the focus is on analysing rs-fMRI in order to investigate FC in the brain. Connectivity studies have been used widely to study how diseases like Alzheimer's Disease (AD) can affect functionally connected networks in the brain [1]. FC measures have been used to analyse changes in brain networks and compare between different groups of healthy controls, Mild Cognitive Impairment (MCI) and AD patients. These studies give promising results in understanding how neurodegenerative disorders develop in order to try to delay this process or inform patients from early signs.

Another factor that also affect the possibility of a person getting a disease is genetics. Studying the genetic variations between people can give us an insight into which genes give higher risks for developing neurodegenerative disorders. Particularly, genome-wide association studies (GWAS) make use of genetic variability data collected from large set of individuals to investigate the association of DNA with traits [5]. Specific genes have been identified by GWAS to be

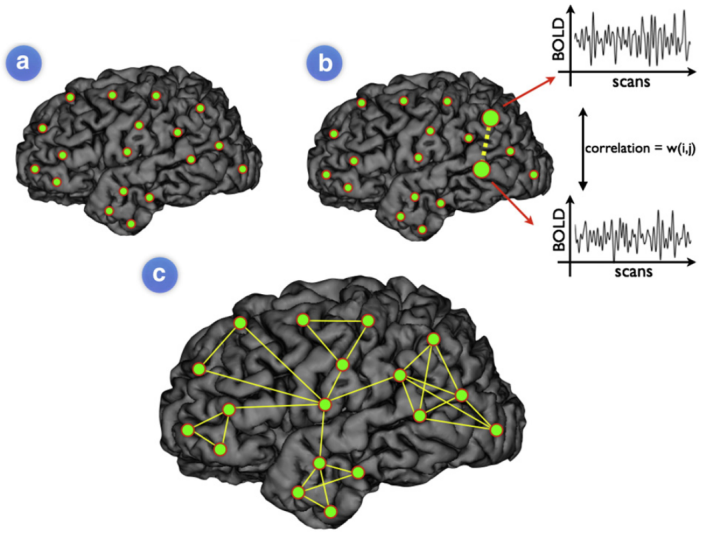


Figure 1.1: Steps of functional brain network modeling [4]

related to risk of an individual to develop AD [6]. Researchers have been combining genetics and imaging data in order to investigate the cause and mechanism of diseases like AD. One can assume that a person's DNA can have a role in the decline of cognitive ability. With the advancement of genetic sequencing and collection of large scale genetic material across populations it gives rise to possibility of examining this issue in more detail. Imaging genetics research is a discipline in which genetic variations and neuroimaging data is combined and surveyed [7]. This is an ongoing research field with constantly advancing tools; in particular one aspect that will help advance this research is collection of large data. Imaging genetics research work with multivariate sets of variables and with very high-dimensional data. Statistical methods are used widely to model the issue of co-linearity when it comes to imaging and genetic data; these regression approaches are suitable to handle the correlation patterns in such high dimensional data such as finding links between phenotype and genotype in imaging genetics [8].

1.2 Objectives and Roadmap

The goal of this research is to combine functional connectivity (FC) and genetic features for patients with Mild Cognitive Impairment (MCI) and Cognitively Normal (CN) to identify links between the phenotype and genotype. For this study data was collected from large public database named ADNI for deriving the fMRI data. To extract the imaging features, multiple noise reduction techniques were evaluated on resting-state functional MRI (rs-fMRI) in which the performance was tested. Preprocessing pipelines of rs-fMRI data are investigated to reduce noise and extract signals of interest. ICA-AROMA and nuisance regression pipeline (NRP) for motion and noise identification were tested on a subgroup from the data in order to generate the FC matrices and representative measures used as imaging features. Finally, a multivariate regression based model named Partial Least Squares (PLS) was used to analyse correlations between genetics and fMRI features.

2 Background

2.1 Functional magnetic resonance imaging (fMRI)

fMRI is used as a tool to map brain activation as well as FC between brain regions. fMRI scans are being captured across time by measuring the blood flow in the brain that translates to neural activity. BOLD fMRI uses blood oxygenation to measure this change [9]. Changes in the ratio of oxygenated hemoglobin to deoxygenated hemoglobin happen during neural activity which can be detected and classified as activation [9]. This signal is represented in activation maps that can be localised in different areas in the brain for further studies. The BOLD signal path as an effect of a stimulus is shown in figure 2.1 in which activation in the visual cortex was triggered after the experiment of showing a flickering checkerboard as illustrated by [10].

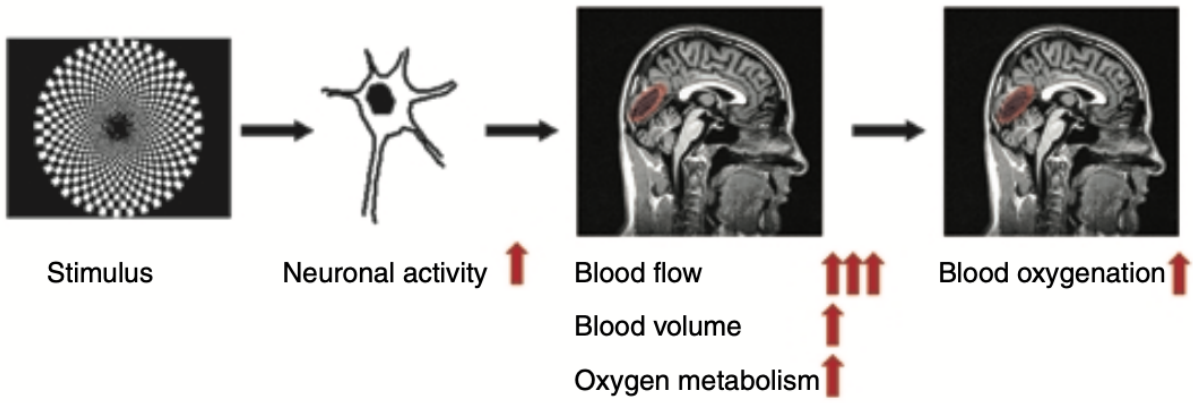


Figure 2.1: BOLD signal path [10]

Blood can flow in the brain indicating activity either for preforming a specific task or during rest [10]. The measured signal in the brain can be compared to the timeseries signal of a specific task to know which voxels were activated [9]. Many studies rely on fMRI data to study a healthy functioning brain and compare how a certain disease can affect functions of the brain.

BOLD fMRI data is acquired by getting a series of images over a period of time in order to see the change in neural activity. Each voxel in the brain has a signal that can be identified [9]. Brain activation can be analysed by looking at the timeseries of the signal and can be compared to the task that had been preformed or to another condition [1]. Example of a timeseries in a certain voxel can be shown in figure 2.2. Correlations between the BOLD fMRI signal and a certain stimulus could be examined to find out if there is any relation [10]. There are multiple techniques that study signal correlation with different brain regions either with a task or during rest which utilizes the measured BOLD signal. Statistical analysis methods is used to minimize

the noise and divide the signal of interest from the instrumental and physiological noise which are used as confounds [10].

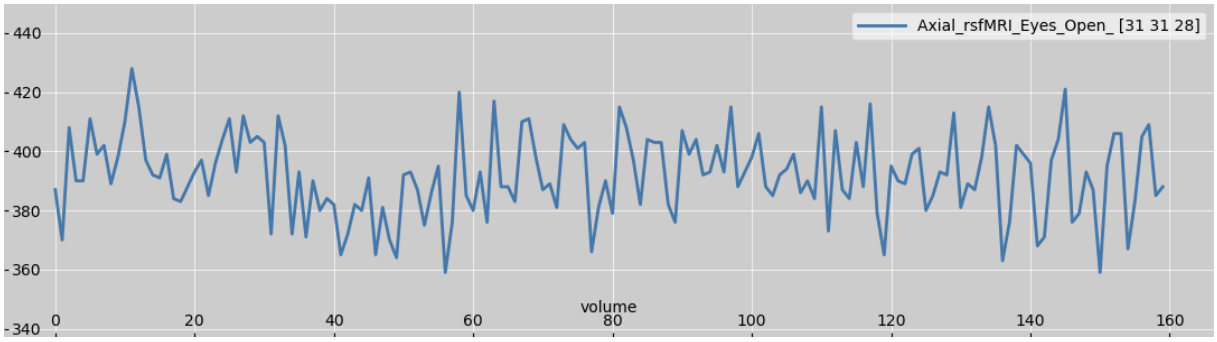


Figure 2.2: Timeseries in a certain voxel in fMRI data

2.2 Resting state fMRI (rs-fMRI) and functional connectivity (FC)

rs-fMRI is a type of fMRI that relies on studying parts of the brain that interact while not necessary being anatomically connected. This method is used in brain mapping in which a subject is told to relax and not to think of anything in particular while still being awake, in the meantime the scan is taking place [1]. Unlike task-based fMRI which relies on the patient doing a specific task that triggers neural activity; rs-fMRI can show us brain networks and connectivity that occur in the brain during rest [11]. Recently, many studies have been studying functional interactions between regions in the brain [4]. Temporal correlations detected between the timeseries in different parts of the brain is defined as FC which is extracted from rs-fMRI and are commonly named in the literature as low-frequency fluctuations in the BOLD signal [1].

After running multiple studies to report any kind of functionally connected brain networks it was discovered that there are some RSNs in the human brain that are functionally linked. An example of the most reported RSNs in multiple literature sources which were repeatedly discovered with different methods and databases are shown in figure 2.3. During rest studies have shown these RSNs characterized by high FC while being anatomically separated [4]. Studying RSNs have opened a lot of opportunities in understanding how these functionally connected networks could change or be affected with for example diseases or under other conditions [2]. Investigating brain networks in patients with MCI or AD during rest and comparing it to healthy controls have a potential for FC to serve as biomarker for such diseases [1].

2.3 BOLD rs-fMRI noise

While performing these studies looking for voxel-wise correlations across the brain, it is essential to minimise as much as possible noise that could affect the fMRI signal. There are many kinds of noise that could interfere and give false signals, some of these noise are related to the scanner and instruments used in this process [10]. Moreover, physiological effects are another type of noise that could corrupt the signal, for example heartbeat and respiration comes on top corresponding to approximately 1 Hz and 0.3 Hz respectively [10]. Since cardiac and respiratory signals are repetitive, one can detect such physiological signal by looking at the power

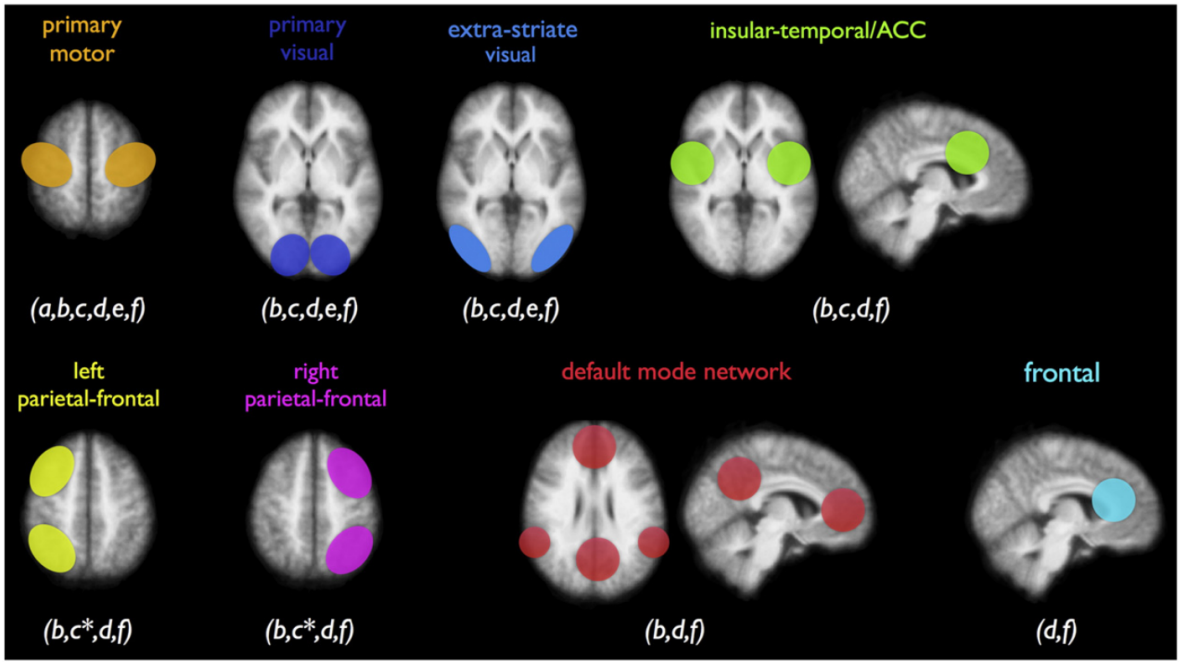


Figure 2.3: Resting-state networks reported by literature [4]

spectrum of the fMRI signal. The power spectrum is the Fourier transform of the timeseries for the fMRI data which illustrates the frequency distribution [10]. Filtering of rs-fMRI BOLD signal is crucial to get rid of noise that has high frequency fluctuations [12]. Another aspect that can create large perturbation of the fMRI signal is the motion of the subject. Motion can create noise frequency which could be larger than the BOLD signal of interest itself. It is critical to prevent it from happening in first place during a scanning session by making sure the person is comfortable but sometimes it's inevitable after long sessions and especially for patients with MCI or AD [1]. Such noise which result in rotation and changes in pixel location could be removed during image registration and there are other tools used to remove motion related artifacts [10], [11]. Head movement can affect the scan by changing the voxels in each image so that the voxels are not aligned properly anymore which makes the brain image become blurry, in return it's hard to process the data later on with such interference [13]. Multiple noise reduction techniques are becoming more popular and essential in neuroimaging research to include methods that will reduce such noise from rs-fMRI data in the preprocessing process [4].

In studying FC in rs-fMRI the most common noise that can greatly affect the signal is a participant head movement during the scanning session. It is very important to remove motion artifacts from the signal before running any analysis as it will corrupt the signal and can give false results. The two most common head motion noise reduction pipelines are nuisance regression based pipelines and independent component analysis (ICA)-based pipelines [14].

The nuisance regression based pipelines use a regression model in which 6 motion parameters which were identified during realignment process are used as regressors. The six motion parameters include three for displacement and three for rotations [1]. There exists many variations of this approach in which one can include as well the motion parameters derivatives and if needed additionally the square derivatives. Such that the regression model will be composed of 6, 12 or 24 regressors among other variants [13]. Adding to the motion parameters, non-brain tissues

are also regressed from the signal such as white matter (WM) and Cerebrospinal fluid (CSF).

The second main pipeline used is the ICA-based noise reduction pipelines. These kind of pipelines are data-driven in which they make use of ICA. This approach is applied on the fMRI data to divide it into independent components (ICs) composed of signal of interest and noise signals [1]. There are two ways to classify these identified components either as signal or noise, one way is manually labeling them or by using automated ICA classifier [1]. There are some available pipelines which already have classifiers implemented in which the components are classified automatically. One drawback in such models is that to apply this classification one needs to re-train it for each dataset used [14]. While other strategies require the classifications of such ICs manually. Two common ICA-based pipelines which automatically classifies the noise from the data are ICA-FIX short for (FMRIB's ICA-based X-noiseifier) [15] and ICA-AROMA short for (ICA-based Automatic Removal Of Motion Artifacts) [16]. One of the advantages of ICA-AROMA is that it doesn't need retraining on new data which makes it convenient and advantageous [1].

2.4 rs-fMRI processing and FC analysis

Following preprocessing and cleaning the data from noise, it is prepared for FC analysis. There are a couple of approaches for FC analysis. To observe and visualize FC, maps and matrices can be derived. Such FC maps can show us to what extent a specific region has similar activation to another part in the brain [4]. FC matrices are one method of visualizing FC after comparing timeseries for any correlation patterns. Generating the FC matrices rely on a graph-based connectivity modeling described as node-based connectivity analysis [1]. To generate the FC matrices the node-based approach is described in the following points and illustrated in figure 2.4 [1]:

1. Define region of interest (ROI) by grouping some voxels together
2. Extraction of mean signal for each ROI represented by the timeseries of the BOLD signal
3. Pairwise calculation of the correlation between timeseries
4. Build FC matrix using ROI-to-ROI information (each row/column represent a region and each element value describes strength of FC which are color-coded)

The ROI defined in step 1 in figure 2.4 are also commonly named nodes and parcels. The grouping of voxels and selection of nodes can be assigned based on atlas that exist in literature among other approaches [1]. The connectivity matrix extracted from the data is finally used for analysis or summary measures could be obtained from it [1]. Additionally, statistical methods could be applied to further analyse the matrices and report the results.

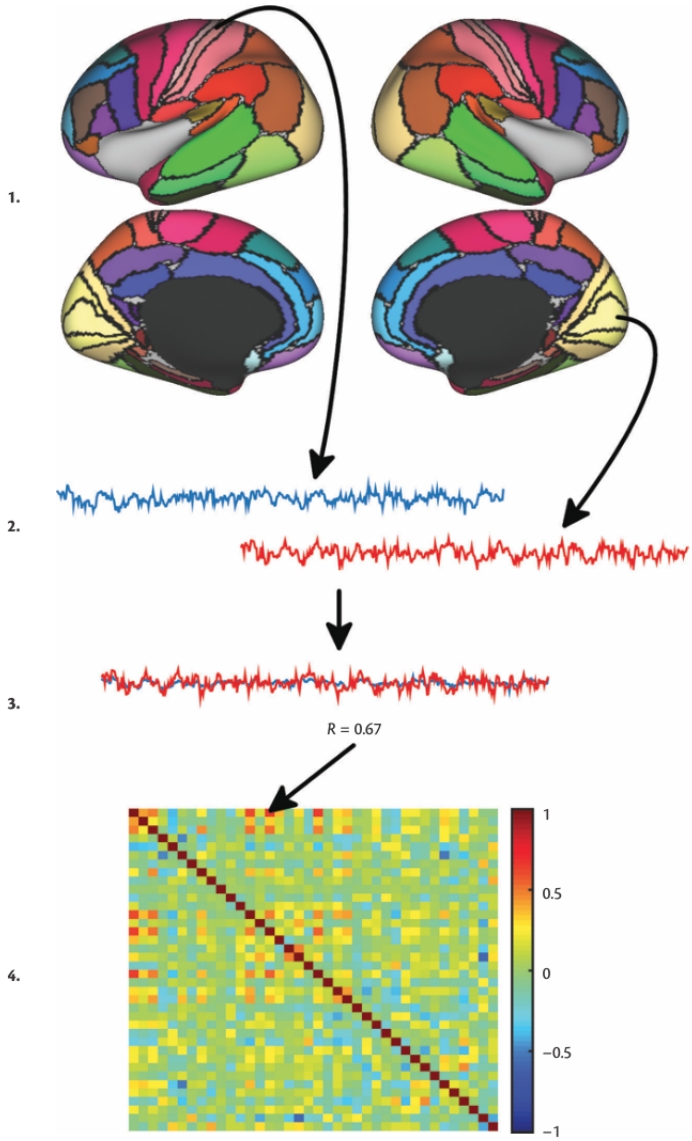


Figure 2.4: Steps of node-based connectivity analysis [1]

3 Literature Review

3.1 Data preprocessing and noise reduction pipelines

Studying FC in rs-fMRI has been heavily researched over the past years. fMRI allows the possibility to capture signals that translates to neural activity while a patient is doing a specific task or at rest. rs-fMRI measures signals in the brain during rest, the patient is told not to think of anything in particular. Nonetheless, there are still many perturbations and noises that are also captured and can interfere with the signal of interest. Raw data for rs-fMRI needs to be filtered out in order to get rid of as much noise as possible [1]. There is not a single pipeline that researchers use for noise reduction of BOLD fMRI data, rather many apply different pipelines that fit their data and needs. The choice of a preprocessing pipeline is heavily dependent on the data used. In this study Alzheimer's Disease Neuroimaging Initiative (ADNI) database is included, therefore it was particularly of interest to review the pipelines used in literature with the same data. Although not many studies used the ADNI phase 3 data, some will be presented in this chapter in which they worked with ADNI in general from multiple phases and one is surveyed that used another database. By comparing the pipelines mentioned in this chapter, an overview of the general approaches and preprocessing methodologies were examined and are reported in the next paragraphs.

Huang et al. (2018) [17] used rs-fMRI data from ADNI to study CSF biomarkers, structural MRI and functional MRI among other features and their ability to differentiate between groups of CN, MCI and AD. They included 130 subjects in their study. Preprocessing of their rs-fMRI data was done using FMRIB Software Library (FSL) in which they applied slice timing correction, motion correction, smoothing and high-pass temporal filtering as well as smoothing and elimination of non-brain tissues. Data was registered to MNI-space with 4-mm resolution in order to apply ICA analysis. 10 RSNs were extracted from the ICA components and FC values from default mode network (DMN) was calculated by extracting the signal correlations.

A study by Palmqvist et al. (2017) [18], has presented accumulation of amyloid-beta and levels of their existence in the brain of AD patients. Preclinical AD subjects from ADNI and BioFINDER are examined for further analysis. Accumulation of amyloid-beta and relationship to brain network changes is presented in this study. For image processing of structural MRI they used Freesurfer for segmentation. For the BioFINDER data multiple software like FSL, AFNI and ANTS were used for preprocessing of resting-state data. First 5 volumes were discarded and data was normalized to MNI space. Furthermore, skull stripping and segmentation of WM, gray matter (GM) and CSF were part of their pipeline. Slice timing correction and regressing of motion parameters were applied to their data for further noise reduction, and it was reported that no spatial smoothing was implemented. Finally the fMRI signal was filtered using a band-pass filter of 0.01-0.1 Hz to get rid of physiological noise.

Franzmeier et al. (2020) [19] published their study on rate of accumulation of tau protein in the brain and its relation to cognitive decline and AD. They use rs-fMRI and longitudinal tau-PET data from ADNI and BioFINDER database including AD subjects and healthy controls. FC was extracted from rs-fMRI data by including 400 regions of interest. FC is analysed in this research to study brain architecture and how changes in the connectivity can affect tau spreading in the brain. The preprocessing pipeline followed in this study and applied to ADNI data was motion correction; furthermore, the mean signal of WM and CSF and 6 motion parameters estimated during realignment procedure was regressed out from the signal. Detrending and band-pass filter of between 0.01 and 0.08 Hz and despiking was implemented. Fianlly to get rid of motion artifacts, scrubbing was used to eliminate spikes with frame-wise displacement higher than 0.5mm. If spikes was found the volumes were zero padded among with 1 volume before and 2 volumes after. Zero-padding of volumes instead of volume removal is used in order to keep the volumes consistent along subjects. Multiple non-linear registrations were applied to the rs-fMRI data to transform the data in MNI space or T1-weighted images. To compute the FC values, Schaefer fMRI atlas was used with 400 brain parcellations corresponding to 400 ROI and 7 RSNs.

A research group with the goal of studying brain entropy maps based on rs-fMRI data from ADNI has presented their results and processing of data for subjects in MCI, AD and CN groups. Wang et al. (2017) [20] used imaging data from ADNI phase 2 to study permutation entropy in multiple cognitive groups rs-fMRI to examine fMRI signal characteristics that are altered. Data Processing Assistant for Resting-State fMRI (DPARSF) toolbox and the SPM8 package are used in the preprocessing of their rs-fMRI data. To start with the first 10 volumes were discarded from the fMRI images; moreover based on information from the last slice, slice-timing correction was applied. Motion correction to remove motion of subjects during scan was performed and data were registered to MNI space. For the last step, they got rid of signal drifts using a linear model. To calculate the Regional Homogeneity (ReHo) in the rs-fMRI brain maps after preprocessing, spatial smoothing of 8-mm FWHM smoothing kernel is applied. As suggested by previous studies regression of nuisance signals like WM, CSF, global signal and motion parameters was tested. Finally, segmentation of images was applied and an 8-mm FWHM Gaussian kernel is applied to smooth the GM images.

The aim of the study by Lee et al. (2016) [21] is to investigate FC in DMN in multiple MCI stages. To study the cognitive decline in MCI patients they used rs-fMRI from ADNI database. Preprocessing was done using SPM8 and Group ICA of fMRI Toolbox (GIFT) in which, FC is computed using ICA. Deletion of the first 2 volumes in the fMRI data and slice-timing correction is performed to the timeseries. Spatial reallignment and co-registration is applied between fMRI and T1 MRI images; furthermore the images were transformed to MNI space. The normalized fMRI images were smoothed using a 6-mm FWHM Gaussian kernel. For the structural T1 weighted images they have applied segmentation to divide WM, CSF and GM. The spatially normalised GM maps was smoothed out with a 8-mm FWHM Gaussian kernel.

Finally, Vos et al. (2018) [22] have investigated multiple measures of FC in rs-fMRI data. Such analysis was done with the aim of examining accuracy for classification between AD patients and healthy controls. The scans were captured at the Medical University of Graz as a part of the prospective registry on dementia (PRODEM). FSL is used for preprocessing of MRI data. Brain extraction and bias field correction were applied on the structural MRI, then non-linearly registered to standard MNI. The rs-fMRI data have been preprocessed with brain extraction, motion correction, temporal high-pass filter with 100s as cut-off and spatially smoothed. MCFLIRT is

used to remove head motion and the motion parameters were regressed from the signal. ICA-FIX is used as a tool to detect further noise components and remove them from the fMRI data. A 3-mm FWHM kernel is used for spatial smoothing which is recommended before running ICA. To derive the FC matrices, they have used temporal concatenation ICA in FSL MELODIC, in which the RSNs were obtained after registering the fMRI to standard MNI space. 10 RSNs and WM and CSF were used as confounds map to apply dual regression analysis using FSL in order to compute whole brain FC. The mentioned methods are parts of what this study included among other measures included in the classification process.

Table 3.1: Literature review on preprocessing summary

Authors	No. of Subjects	Pipeline	Dataset
Huang et al. [17]	130	ICA based	ADNI
Palmqvist et al. [18]	473/406	Nuisance regression based	ADNI-2/BioFINDER
Franzmeier et al. [19]	81/57	Nuisance regression based	ADNI-3/BioFINDER
Wang et al. [20]	124	Nuisance regression based	ADNI-2
Lee et al. [21]	130	ICA based	ADNI
Vos et al. [22]	250	ICA based	PRODEM

3.2 Regression analysis on genotype and phenotype features

The topic of imaging genetics has been studied recently with the availability of large data from DNA in populations as well as the advancement in neuroimaging and MRI technology. Research centers working in the field of genetics around the world collect DNA samples from the human population to understand genetics and research this topic more thoroughly. The possibility to investigate links between genotype and phenotype is very useful in understanding diseases which utilizes imaging genetics research [7]. Such studies can help researchers identifying risk factors and mechanisms that contribute to developing a certain disease [7]. For example one can assume the hypothesis that certain genes are associated with how the brain functions. With imaging genetics tools, such hypothesis can be investigated using statistical techniques and methods allowing to model the association between imaging and genetic features [8]. Since imaging genetics studies involve high-dimensional and multi-variable data, it's important to consider the model that will fit with such attributes. Some studies that used these regression analysis and statistical methods in linking phenotypic and genotypic features will be discussed in the next section.

Le Floch et al. (2012) [23] combined different methods of Partial Least Squares (PLS) regression and Canonical Correlation Analysis (CCA) with dimensionality reduction techniques to study the variations in imaging genetics data. The objective of this research is to examine the possibility of genetics features in their case Single Nucleotide Polymorphisms (SNPs) to describe the variability in the imaging features in this case it was fMRI. Since the genotypic and phenotypic features used in this research is of very high dimension, dimensionality reduction steps were applied. Principal component analysis (PCA) and filtering are the two methods used before applying the regression models on the dataset. The models compared in this research were trained on both a real and a simulated datasets. The significance of the model was estimated using a permutation test. Finally, it was reported that one of the methods used which

is filtering and sparse PLS (fsPLS) had resulted in the best performance compared to other strategies. The model has showed links between the genetic features corresponding to sets of SNPs and the fMRI features which were networks calculated during a reading task given to the subjects. Although, fsPLS was successful in reporting associations between multivariate data in regard to genotype and phenotype features; it is reported by the authors that results are hard to interpret and more understanding from neuroscience perspective is still needed for a more thorough analysis.

Meanwhile, Grellmann et al. (2015) [24] replicated a previous study in order to compare performance of multivariate models. The three models they tested are partial least squares correlation (PLSC), sparse CCA and Bayesian inter-battery factor analysis (Bayesian IBFA). The aim is to evaluate the ability of these techniques to find links between genotype and phenotype. In their study they included groups from healthy controls and patients with diagnosis of schizophrenia spectrum disorder, bipolar disorder or psychosis not specified. As genetic features SNPs were used and as fMRI features they have included the whole brain and not a specific region of interest. Results has shown that the fastest model was PLSC and with specific parameters the performance was better than that of sparse CCA. They have reported that PLSC is very applicable to work with multivariate data, although if many SNPs and all brain voxels are used one must apply dimensionality reduction techniques in order for the model to handle all variables.

Lorenzi et al. (2016) [25] used data from ADNI database in which 657 subjects divided in groups of healthy controls and patients with AD were included. The aim is to identify correlations between genotype and phenotype. In their study the genotype features was represented in SNPs and the phenotype features corresponded to structural volumes in the brain (whole brain, ventricles, hippocampi, enthorinal cortex and mid-temporal lobes). A PLS model was trained on the two groups together and then later validated on MCI subjects to verify whether the identified components could be able to classify MCI converters/non-converters. Finally, a permutation test is applied to evaluate the PLS model significance. The model has shown some correlations and anti-correlations between parts of the brain as well as identified which genetic features had the highest weights. The PLS model showed promising results and capability of handling large dimensional data.

Similar to the previous mentioned study, research by Lorenzi et al. (2018) [26] is described in which PLS was used to associate brain atrophy to the complete set of SNPs from AD patients. The aim of the study is to model the role of genes on brain atrophy in patients with AD. Imaging genetics methods were implemented to investigate the correlation between the sets of variables using multivariate statistical models. This research has uncovered a significant link between the TRIB3 gene and the pattern of GM loss in AD. The PLS components were calculated using singular value decomposition (SVD) of the covariance matrix. The PLS model was applied to the phenotype and genotype data collected from the ADNI database.

Looking at the most commonly used preprocessing pipelines in the literature using the ADNI database which were equally divided, it was decided to test the two approaches (ICA-based and nuisance regression-based) and evaluate the outcome. Moreover, PLS proved to be successful in modeling high dimensional multivariate data. The performance of the PLS in imaging genetics research perfectly fits the aim of this research.

4 Data

4.1 Database

4.1.1 The Alzheimer's Disease Neuroimaging Initiative (ADNI)

The database used in this research is from The Alzheimer's Disease Neuroimaging Initiative (ADNI). ADNI is a study that collects biomarkers for early detection of AD by collecting clinical, imaging and genetic data among other things from participating patients. The study aims to share this data with researchers around the world to contribute to advancements in research of early detection for AD. Many researchers and scientists have been using the data collected from ADNI to help study the progression of this disease and ways for early detection [27].

ADNI study has multiple phases along the years, ADNI 1, GO, 2 and 3 starting from 2004 to 2016. In this study subjects from the ADNI-3 phase are included. ADNI-3 phase started in 2016, some of the subjects in previous phases were enrolled again in this phase while also new ones were included. Among other things added in ADNI-3 is the addition of Systems Biology tools to identify AD genetics and AD biology to find the connection between them both [28]. More specifically one of the primary goals of ADNI-3 phase is the addition of functional imaging as a technique for clinical trials. The data offered by ADNI study for MRI include different characteristics with the aim to help researchers studying the relationship between structural and FC of the brain and how the anatomical and functional brain connectivity can be affected in an AD patient. The main sequences that are included in ADNI-3 are structural MRI, diffusion MRI (dMRI), rs-fMRI, ASL perfusion MRI, among others [29].

The subjects analysed in this project belong to different categories, more precisely: CN, significant memory concern (SMC), early MCI (EMCI), MCI and late MCI (LMCI). As mentioned on the ADNI website [30], CN are the healthy controls in which no signs of dementia or memory loss is reported. SMC is a new category added in order to study the gap between CN and MCI, the reported memory problems which the participants have shared are memory concerns which are measured by the Cognitive Change Index and the Clinical Dementia Rating. MCI subjects are divided into 3 stages, early and late levels of the MCI group is categorized by the Wechsler Memory Scale Logical Memory II. Although, daily activities are not interfered with and no signs of dementia is shown yet in MCI subjects; patients report memory concerns and are diagnosed by clinician [30].

4.1.2 Image acquisition

Images from ADNI-3 were obtained using a 3T Scanners across all sites using GE, Philips and Siemens scanners. More about the MRI protocol could be found here: `https://adni.loni.usc.edu/wp-content/uploads/2017/07/ADNI3-MRI-protocols.pdf`

Parameters used for acquiring rs-fMRI data are: TR (repetition time) = 3000 ms, TE (echo time) $\sim$ 30 ms, FA (flip angle) = 90°, 3.4 mm isotropic voxel size, FOV (field of view) = $220 \times 220 \times 163$ mm. fMRI volumes were diverse across subjects ranging between 160 to 200 volumes. Structural T1_weighted images were captured using 3D MPRAGE sequence with TR = 2300 ms, voxel size of $1 \times 1 \times 1$ mm and FOV = $208 \times 240 \times 256$ mm.

An example of the fMRI image for one subject in the LMCI group is shown in figure 4.1. The temporal resolution of the fMRI data is determined using the time needed to acquire a single volume, defined as TR [1]. Figure 4.1 shows the fMRI in 3 planes (sagittal, coronal and axial) and below is the timeseries of the voxel identified by the green crossing.

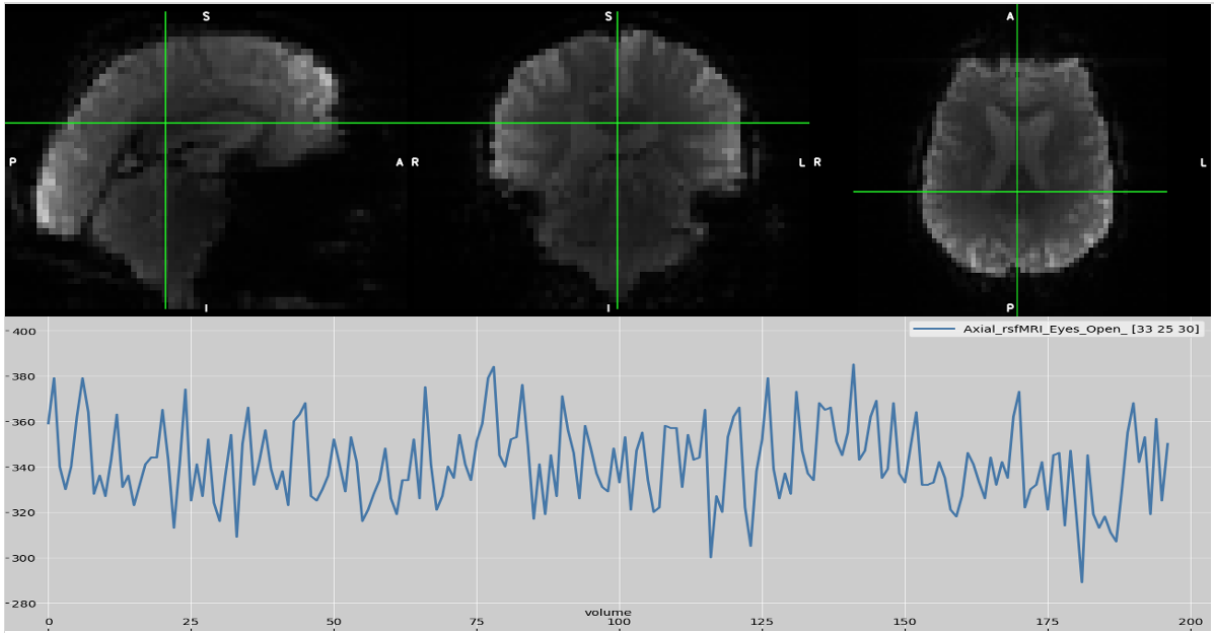


Figure 4.1: Example of fMRI image and the timeseries viewed using the *FSLeyes* tool in FSL

The structural MRI image for the same subject shown in previous section is shown in figure 4.2. Figure 4.2 shows the high resolution image of the brain extracted T1_weighted MRI scan. The structural MRI can inform on the different brain tissue and can highlight whether atrophy is present or not in some parts of the brain.

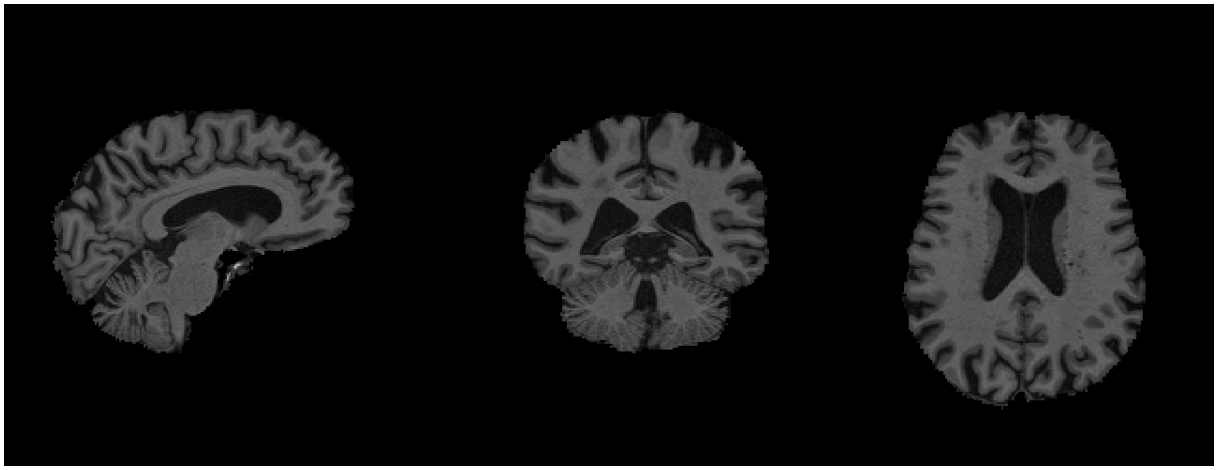


Figure 4.2: Example of brain extracted structural MRI image viewed using *FSLeyes* tool in FSL

4.2 Tools and Software

- **FMRIB Software Library (FSL)**

  The main software library used in this research is the FMRIB Software Library (FSL). FSL allows to preform brain image analysis on functional, structural and diffusion MRI data [31]. The tools in FSL allows to view the MRI images as well as the signal for the fMRI data. *FSLeyes* is one of the tools available in FSL that allows image viewing. There are many different field of views in which the data could be viewed. FSL has multiple tools for MRI data in which many analysis and operations could be applied to the data using this library. The tools available on FSL can be run both from the FSL GUI as well as from the command line [31]. FSL is a publicly available software tool for neuroimaging analysis developed by researchers in the Analysis Group at University of Oxford [32]. More information on tools and installation of the software could be find here: [33]

- **MATLAB**

  MATLAB a product by MathWorks is used through out this project when needed. MATLAB is a software used for programming and numeric computing. It's a matrix-based programming language used widely in industry and academia to do computational analysis and model development among many other features. More information about MATLAB could be found on their website: `https://www.mathworks.com/products/matlab.html`

- **Bash**

  In order to be able to run the preprocessing and generate the FC matrices and further imaging features on as many subjects as needed in a short time, a bash script was written. Bash is a scripting language; it's a command-line shell that allows running repeated tasks much faster, more efficiently and more robustly. Bash reference manual could be found in this link for more details: `https://www.gnu.org/software/bash/manual/bash.html`

  In order to run the analysis on all subjects 3 bash scripts were written. The first script incorporated all the steps to run basic preprocessing, the second one implemented the noise reduction pipeline and the last one was dedicated to the generation of the FC matrices and extraction of additional imaging features. These scripts allowed quick and easy implementation of the analysis on all the subjects. It took about 20 minutes to run the 3 scripts per subject.

- **Python**

  For the regression analysis model in this research Python is used to apply the PLS model on the imaging and genetic features. Many libraries from Python are used in order to upload the data, run and fit the model as well as view the graphs and figures among other things. Example of the libraries used are *pandas*, *NumPy*, *Matplotlib*, *scikit-learn* and *SciPy*. More information about Python and its libraries are found on here: `https://www.python.org/`

5 Methodology

In this chapter the methodology of this research is presented. Figure 5.1 shows a summary of the steps taken during this research. First basic preprocessing was applied to the data, then two noise reduction pipelines were tested on an initial small subgroup of subjects. The NRP pipeline was chosen to continue with in which bash script was written to apply the pipeline and extract the imaging features quickly on as many subjects as needed. Finally, statistical analysis was used to model the correlation between imaging and genetics features using PLS.

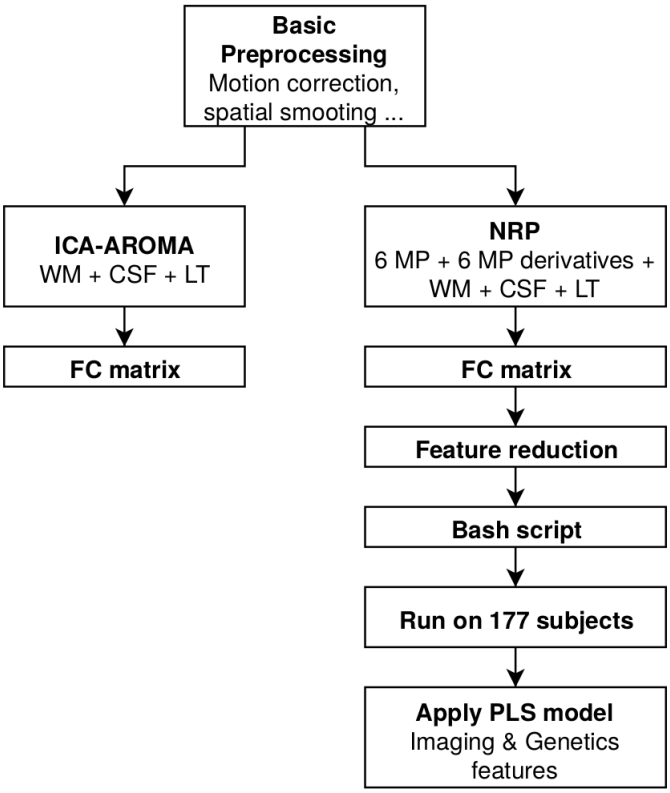


Figure 5.1: Overview of methodology steps

5.1 Data Pre-processing

In this section preprocessing of the rs-fMRI data is described. The noise reduction pipelines tested NRP and ICA-AROMA will be explained in subsections. Preprocessing of the rs-fMRI data is very important to remove noise and artifacts and in order to prepare the data for processing and computing the FC [11], [1].

5.1.1 Basic preprocessing

Basic preprocessing was applied on the data before applying the movement related noise reduction techniques. Preprocessing procedure was implemented on the data using *FEAT* analysis tool on FSL. More details on this tool: `https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FEAT/UserGuide`

The preprocessing steps are:

- **Removal of first 5 volumes:** Discarding the first volumes to make sure the signal has been stable and steady-state was reached during the scan.
- **Motion realignment (MCFLIRT):** This step corrects head motion of the subject. MCFLIRT [34] uses linear registration to remove motion artefacts. The algorithm for motion correction registers all the fMRI images according to the timeseries of the middle volume as a reference image.
- **4D mean intensity normalization:** In order to intensity normalize the volumes in the data a single scaling factor was used for all volumes in order to get rid of intensity differences across volumes.
- **Spatial smoothing with a 6-mm FWHM kernel:** The spatial smoothing is a preprocessing step that aims to reduce noise while keeping activation signal in the brain. The goal is to enhance signal to noise ratio [1]. Gaussian function is applied with selected kernel size to specify the full width at half maximum (FWHM) in which it defines how much the signal is smoothed.
- **Non-linear registration:** using BBR cost function with 12 degree-of-freedom (DOF), co-registration to the structural MRI data and spatial normalization to standard space were applied. The fMRI image was registered to the T1_weighted structural image of the same subject. Additionally the structural image was normalized to the standard MNI152 space. Finally the fMRI data was registered to the standard MNI152 space. Registration betwen spaces (functional, structural and standard) is an important step to match the difference in image and voxel sizes along the spaces. This step allows to calculate the transformation between different spaces [1].
- **Slice-timing correction:** This step is important because it fixes the delays and shifts that could have accumulated during the rapid measuring process of the slices. Since the assumption holds that all the slices are acquired half-way through the TR (3 seconds in this case) it's important to take into consideration this temporal delay and apply this correction to the timeseries of each voxel [1]. For ICA-AROMA no slice timing correction was applied; this was made following the suggestion of the pipeline mentioned in [16]. While for NRP "Interleaved" slice-timing correction was used since it was mentioned in the ADNI protocol that this is how the slices were acquired. Interleave slice measuring is when one slice is captured then the next one skipped then the missing ones are obtained again on an even/odd pattern.

In the preprocessing part using *FEAT* analysis tool on FSL, no highpass temporal filtering was applied to the data as suggested by [16].

Some of the anatomical data for certain subjects had problem with the brain extraction. For those structural T1_weighted images, parts of the skull and eyes were still not removed from

the brain extracted anatomical images. This issue has affected the registration of the data in different spaces. Therefore, in this case brain extraction tool (BET) [35] on FSL was applied to the structural images with the whole head to remove non-brain tissue from the images. BET was used with fractional intensity threshold of 0.2 and choosing parameters to remove eye and neck residuals for these subjects. After applying BET on those subjects with BET problem, FEAT analysis was carried out in which previous mentioned preprocessing steps are executed.

After applying these basic preprocessing steps on the raw rs-fMRI the output fMRI image after applying *FEAT* analysis was defined to as "filtered fMRI". The filtered fMRI will be the data used from now on to apply further preprocessing needed which will be discussed in the coming sections.

5.1.2 General Linear Model (GLM)

GLM is used widely in neuroimaging field for modelling and testing statistical hypothesis. In particular, it is used extensively due to its ability in handling timeseries and removing noise related components. The GLM model allows the possibility of modelling signals extracted from imaging experiments like fMRI being a timeseries measurement. The multiple regression model, GLM, models extracted signals in terms of explanatory variables known as regressors. These regressors correspond to signal patterns that could be found in the measured data and removed if it's classified as noise or confounds. The set of regressors are collected in a design matrix. Each regressor is scaled by a scaling parameter which is fitted using the GLM. Difference between the data and the fitted model is defined as residual error or residuals. In order for the GLM model to find the best fit, the residuals needs to be minimised. [36]

$$Y = X\beta + \varepsilon \tag{5.1}$$

The linear regression equation is shown in equation 5.1. In regard to GLM applied with a single regressor in case of neuroimaging data; Y represents the data, X represents the regressor, β is the scaling parameter and ε is the residual error.

$$Y = X_1\beta_1 + X_2\beta_2 + X_3\beta_3 + \varepsilon \tag{5.2}$$

In case of multiple regressors the multiple regression model will fit the model by adding the scaled regressors as seen in equation 5.2. An illustration of how GLM model is applied on fMRI timeseries is shown in figure 5.2 as described by [36].

In regard to equation 5.1 and 5.2, X corresponds to the matrix that have all the regressors and is called the design matrix. In the design matrix each column represents one regressor or covariate. β is a vector with all the scaling parameters in respect to the regressor. The GLM model is useful to remove noise related signals. For example, signal confounds such as age is one covariate that is usually regressed from the signal to avoid bias. A very important aspect when building the design matrix with all the regressors is demeaning. Demeaning is the process of calculating the mean for each regressor then subtracting it from each element, this will result in the mean of the regressor to be 0. All regressors in the design matrix need to be demeaned before the GLM model is applied in order to be able to interpret the results. [36]

In the following noise reduction pipelines that are described in the next subsection, linear regression of a set of nuisance variables is preformed using GLM. Either ICs classified as noise or

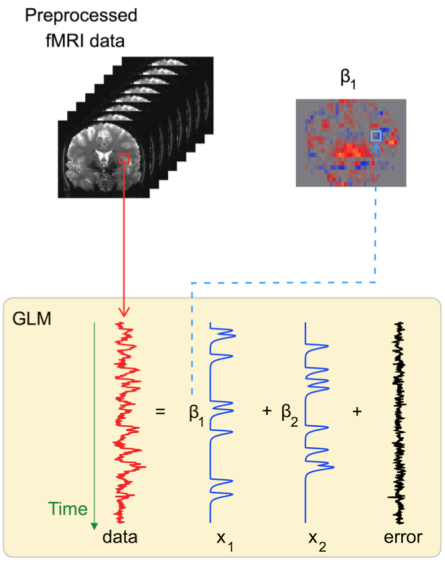


Figure 5.2: Example of GLM on timeseries data from fMRI [36]

tissue-based component are regressed using GLM where the residuals represent the new filtered fMRI data.

5.1.3 Nuisance Regression Pipeline (NRP)

NRP pipeline for movement related noise removal was applied to the data. The pipeline was followed according to [19] in which similar methodology was used on ADNI phase-3 data as well.

First step in this pipeline was to extract the signal of WM and CSF from the filtered data after the basic preprocessing. The Partial Volume Estimate (PVE) of WM and CSF tissues were transformed into functional space using the transformation matrix from *FEAT* analysis. This allowed registration of T1_weighted PVE images to the fMRI space in which the BOLD fMRI image of the subject is used as reference. Secondly, erosion was applied to the WM and CSF images. Erosion is the process of cutting off one or more voxel from all edges; this step is important to remove boundary voxels that could include GM [1]. After erosion, binarisation was applied on WM and CSF tissues using a conservative threshold (0.8). This step resulted in a WM and CSF masks which are used later to extract their timeseries. The three stages explained above for WM is displayed in figure 5.3. Using the WM and CSF masks the corresponding timeseries were extracted from the filtered fMRI using the command-line based utility *fslmeants* on FSL.

Additionally, 6 motion parameters estimated during motion realignment procedure *(MCFLIRT)* in the basic preprocessing step were included along with their derivatives calculated using the *gradient* function in MATLAB. Finally, a linear trend column was added as further regressor, and this was composed of an equally ascending numbers from -1 to 1 and its purpose is to remove non-motion related noise.

Therefore, the regressors included in the design matrix in NRP were WM, CSF, linear trend, 6 motion parameters and their derivatives. All columns in the design matrix were demeaned before applying GLM. *fsl_glm* command on FSL was used to regress out all the nuisance components from the filtered fMRI data, and the residuals resulting from this analysis were taken

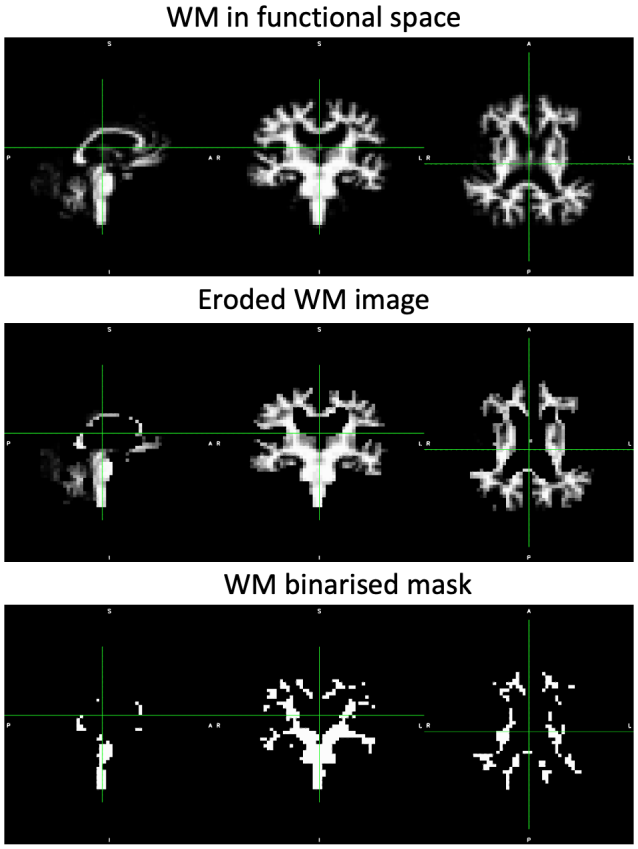


Figure 5.3: 3 stages of white matter preprocessing

as new cleaned data for all the subsequent operations. A band-pass filter with cut-off 0.01-0.08 Hz was applied to this new data. Finally, as suggested by [19] scrubbing was applied to remove any additional motion artifacts, in case of finding frames that exceed 0.5 mm frame-wise displacement. If this was found zero-padding of the high-motion volume as well as one preceding and two succeeding was applied to keep the number of volumes consistent.

5.1.4 ICA-AROMA pipeline

ICA-AROMA was applied to the data according to the pipeline suggested by [16]. ICA-AROMA package could be found here: `https://github.com/maartenmennes/ICA-AROMA`
Basic preprocessing as discussed in previous section and according to the suggestions by the authors was applied on the data. After basic preprocessing ICA-AROMA was applied using the output of *FEAT* analysis. ICA-AROMA is an ICA-based method for automatically classifying noise and signal, this strategy does not need to be retrained on new data. ICA-AROMA extracts ICs that were classified particularly as motion related noise or other kind of noise and utilizes GLM to regress these components from the data. The output of ICA-AROMA is the denoised functional data in which the classified noise was removed and further preprocessing could be applied on it as needed. After ICA-AROMA, WM and CSF signals were extracted from the output denoised functional data. WM and CSF masks were generated after applying erosion and binarisation as described in the NRP section. WM, CSF and linear trend signals are regressed from the data using *fsl_glm* command on FSL after demeaning the WM and CSF signals. Finally, the residual outcome was used as the cleaned data and a high-pass filter was applied using high-pass cut-off of 0.01 Hz.

5.1.5 Including global signal as regressor

For both pipelines, the inclusion of a global signal represented by a mean GM signal was tested and added in the design matrix as a further regressor. GM signal was extracted from the filtered fMRI after erosion and binarisation as described in the NRP section but for the GM mask a threshold of 0.7 was used. All the previous mentioned steps were repeated with demeaned GM signal in the design matrix. Including the regression of global signal is a controversial step approach as it's reported that it may corrupt the signal and introduce a lot of negative correlations [37], [1]. This is due to the fact that the global signal may carry a lot of the neural signal and removing it can affect FC measures [13], [1]. Nevertheless, including the GM signal as a covariate in the design matrix may remove some of the motion artifacts as well as respiratory related signals that are present in the global signal [37].

5.2 Data Processing

5.2.1 Functional Connectivity (FC) matrix

The FC matrix for each subject was generated using the Schaefer functional atlas [38] with 100 parcels and 7 RSNs. As the Schaefer atlas is defined in the MNI space, the filtered denoised data resulting from the two pipelines were spatially normalized to the 2-mm MNI space (non-linear registration) before extracting the mean timeseries for each region. The connectivity matrix for each subject was calculated using Pearson's correlation coefficient [39]. For generating the FC matrices the function *corrcoef* in MATLAB was used. FC matrices are generated with the methodology described in figure 2.4.

The 7 RSNs included in this atlas are visual network (VIS), somatomotor network (SMN), dorsal attention network (DAN), ventral attention network (VAN), limbic network (LIM), frontoparietal control network (CON), default mode network (DMN). Each network has its own number of parcels according to the left and right brain hemisphere as defined by [38].

Figure 5.4 give an explanation on how the FC matrix is organized, the networks division and order and how it is interpreted.

5.2.2 Feature Reduction

The FC matrix is a 100×100 symmetric matrix. Half of the matrix could be used as imaging features which will be composed of 5000 values for each subject. Considering the high number of fMRI features given the number of subjects and the fact that single FC connections are very noisy, summary measures were derived from the full FC matrices. Performing some kinds of feature reduction is a very important step in order to lower the number of features while retaining as much information as possible from the original FC matrix.

In this study, two possible approaches were evaluated to aggregate the information inbuilt in the FC matrices, which will be briefly described below.

28 features

The aim of finding 28 features that summarizes the FC matrix is to retain as much information as possible from the original FC matrix and reduce its dimension. In order to summarize the FC

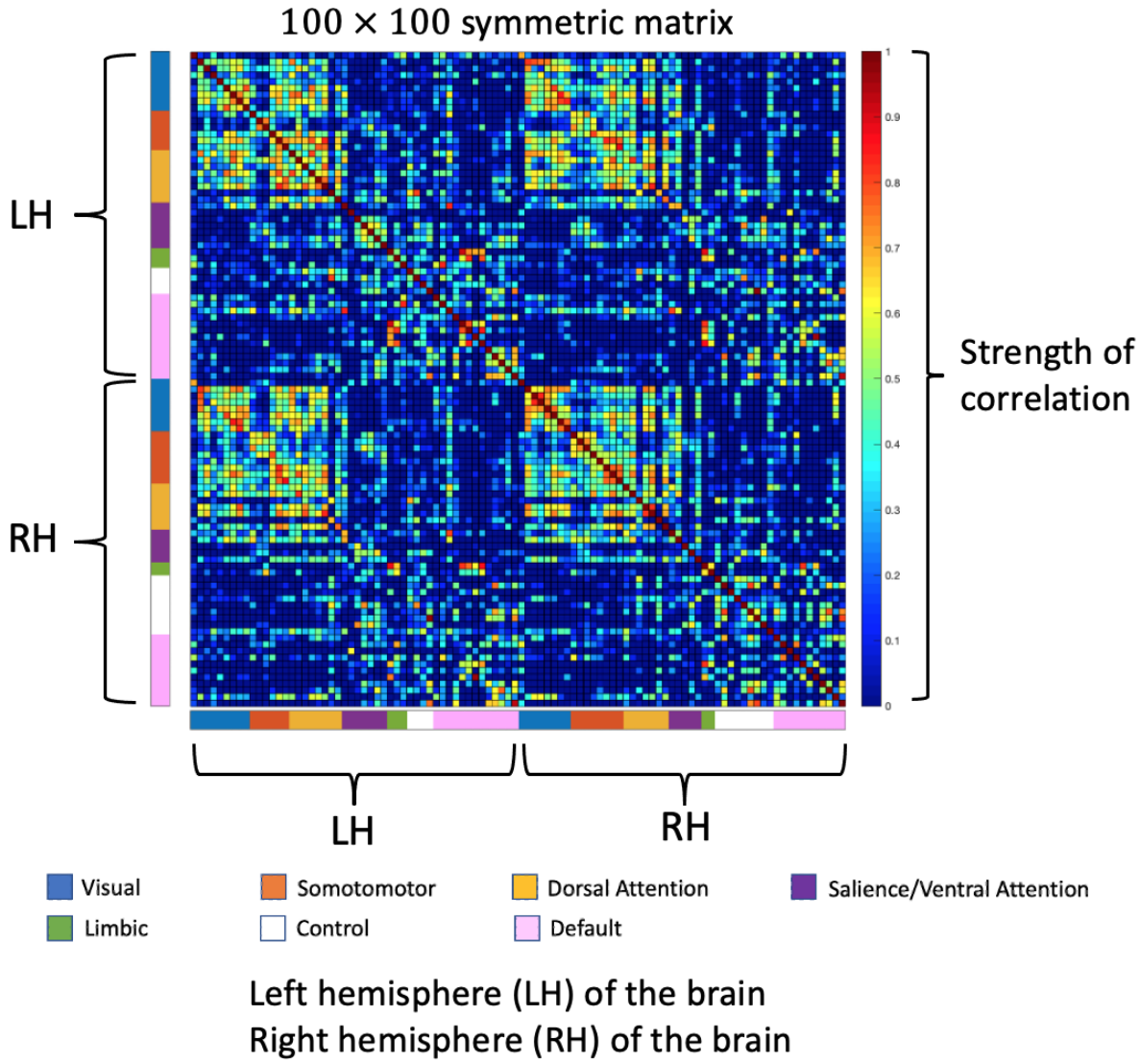


Figure 5.4: FC matrix illustration

matrix and generate a smaller 7×7 for each subject the following methodology was used. Mean values of within and between network connectivity was calculated using *mean* function in MATLAB. This process was done by taking into consideration the 7 RSNs in the FC matrices viewed in figure 5.4. Summary measures representing the mean connectivity value inside a given network (within-network FC/ intra-hemispheric connectivity) and across edges connecting regions belonging to different networks (between-network FC/ inter-hemispheric connectivity) were derived from the full matrices. Within-network FC was calculated as the mean value of all the region-to-region connectivity 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., DMN-SMN or DMN-VIS). The diagonal of the 7×7 matrix represents the global mean value for each main network (VIS, SMN, ...). Those 7 values represent the intra-hemispheric network connectivity which represent the global mean of the networks separately within the left and right hemisphere of the brain. Meanwhile, the rest of the 7×7 matrix represents the inter-hemispheric connectivity. This approach for the summary of FC matrix was adopted by [40] and [41] in which this study follows the same idea to generate the smaller matrix. Since the produced 7×7 matrix is symmetric, the diagonal and half of the

matrix was used as imaging features composed of 28 values. Figure 5.5 show both original FC matrix and summary matrix for one subject.

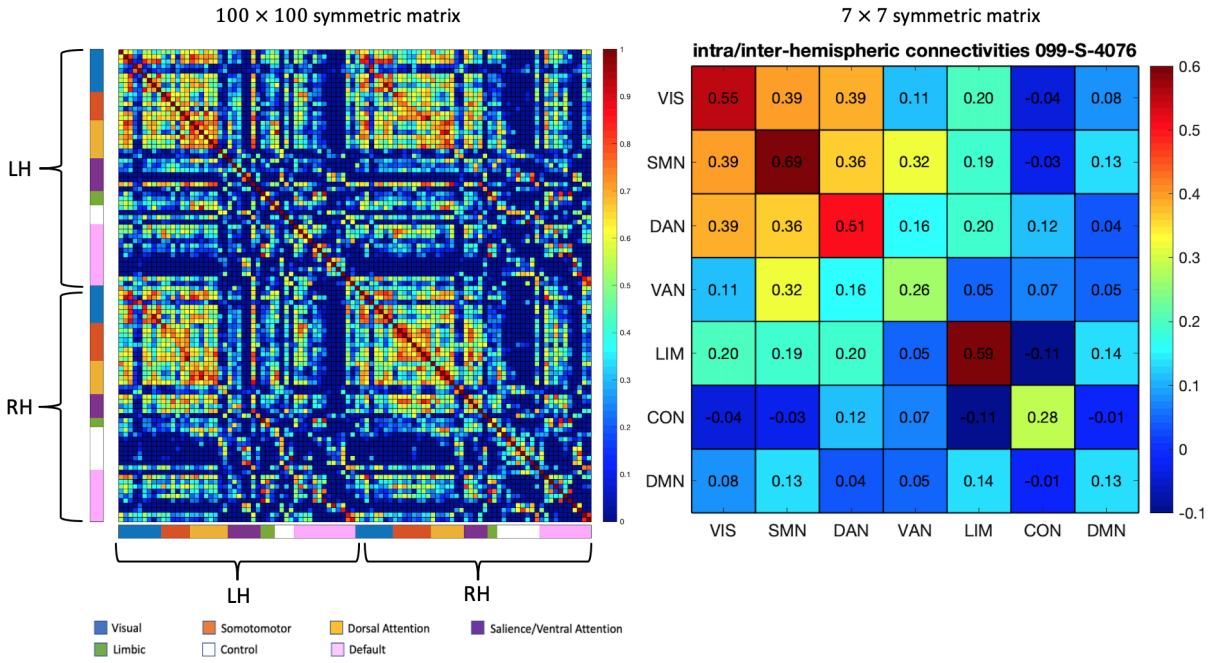


Figure 5.5: Summary of FC matrix for feature reduction (28 features)

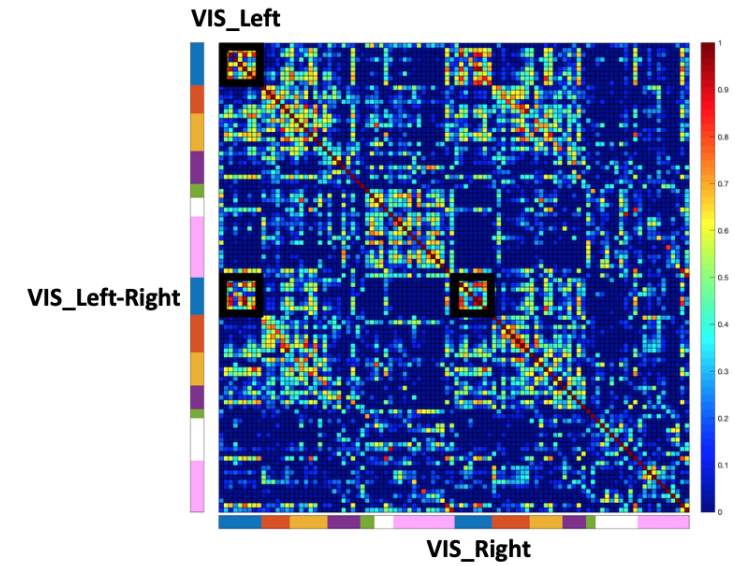
21 features

Another way that is proposed to summarize the FC matrix for smaller set of features is described in figure 5.6. The idea here is to calculate the global mean for each possible connection in each network separately. There are 7 RSNs in the FC matrix in which the left and right hemisphere is presented as shown in figure 5.4. This allows the possibility to calculate the mean for each possible connection being left, right and left-right per network which makes up 21 values as illustrated in figure 5.6 which the visual network possible connectivity is marked in black on the FC matrix and in the table the global mean of the three connections are highlighted accordingly. The mean was calculated using the *mean* function in MATLAB.

5.2.3 Genetic features

Information along the human genome that correspond to genetic variability are collected in polygenic risk score (PRS). PRS are used to estimate a person receptibility of certain diseases [42]. Recently studies have shown that PRS that is calculated on the basis of identifying AD is able to predict diagnosis and could be related to imaging biomarkers related to the disease [6]. Calculating PRS is based on combining variables in an individual's genome which represent a risk for disease development. These variables could be identified using single-nucleotide polymorphism (SNPs) that are present in a person's DNA. SNPs are genetic variation in human genome that identifies the variability among people [5].

For this study, two PRSs were used PRS1 and PRS2 as genetic features. For each subject PRS1 and PRS2 were identified according to the calculated score as reported by [6]. Altmann et al. (2020) [6] have used data from ADNI to compute the PRSs for each individual subject.



	VIS	SMN	DAN	VAN	LIM	CON	DMN
Left	0.50896	0.58446	0.48501	0.34862	0.27621	0.53613	0.36845
Right	0.53934	0.49533	0.43397	0.47434	0.45884	0.25841	0.19128
Left-Right	0.44339	0.48498	0.38329	0.1717	0.26167	0.15695	0.093728

Figure 5.6: Extraction of 21 features

As reported by them, SNPs were selected by passing a P-value threshold of $1e-5$ and 0.5 for PRS1 and PRS2 respectively. For PRS1 55 SNPs have passed the threshold and for PRS2 101,450 SNPs. This computation was done using the software PRSice v2.1.9. PRS is calculated according to equation 5.3 and as reported by PRSice software [43]:

$$PRS_j = \sum_i \frac{S_i \times G_{ij}}{M_j} \tag{5.3}$$

where S is the effect size which is identified from GWAS statistic summary, G is the number of effective allele observed in the sample and M is the total number of SNPs included in the sample in which i is the SNPs count and j is the individual sample count [43].

5.3 Partial Least Squares (PLS)

PLS is a regression based model that works with multivariate data. PLS can be used to to model the relationship between two sets of variables [44]. PLS was first introduced by statistician Herman Wold around the year 1975 [45]. PLS is also commonly described as Projection to Latent Structures as another naming for the method. The aim of PLS regression is to maximize the co-variance between two sets of variables. It achieves this by projecting the data into a latent space in which maximum co-variance is applied to the projected data.

In this research, PLS was used to model the variations between imaging features and genetic features. PLS model was applied to evaluate associations between the phenotype and genotype of the subjects included in this study.

Let X be a matrix of dimension $K \times N$ and Y a matrix of dimension $K \times M$ in which K is the number of observations in this case the number of subjects and x_n $(n = 1, ..., N)$, y_m $(m = 1, ..., M)$ are distinct sets of features.

$$w_x, w_y = \underset{w_x, w_y}{\operatorname{argmax}}\, cov(Xw_x, Yw_y) \tag{5.4}$$

$$cov(Xw_x, Yw_y) = \frac{w_x^T S_{XY} w_y}{\sqrt{w_x^T w_x}\sqrt{w_y^T w_y}} \tag{5.5}$$

X and Y are two matrices representing imaging features and genetic features respectively, w_x weighted component of X, w_y weighted component of Y and S_{XY} cross-covariance matrix. The calculation of the weights matrices and the covariance is illustrated in equations 5.4 and 5.5.

One alternative to implement the PLS model is using the SVD and cross-product matrix method. The SVD approach is if M is matrix that is expressed in $M = X^T Y = U \Lambda V^T$ in which U and V are orthogonal and their columns vectors are the eigen-components that show the variability in the X and Y sets of features; Λ is a diagonal matrix composed of the eigenvalues that give the amount of this variability represented by each component [46], [25]. A representation of the PLS model adopted in this study, including the proper dimension of the different input variables (imaging/genetics) is reported in figure 5.7.

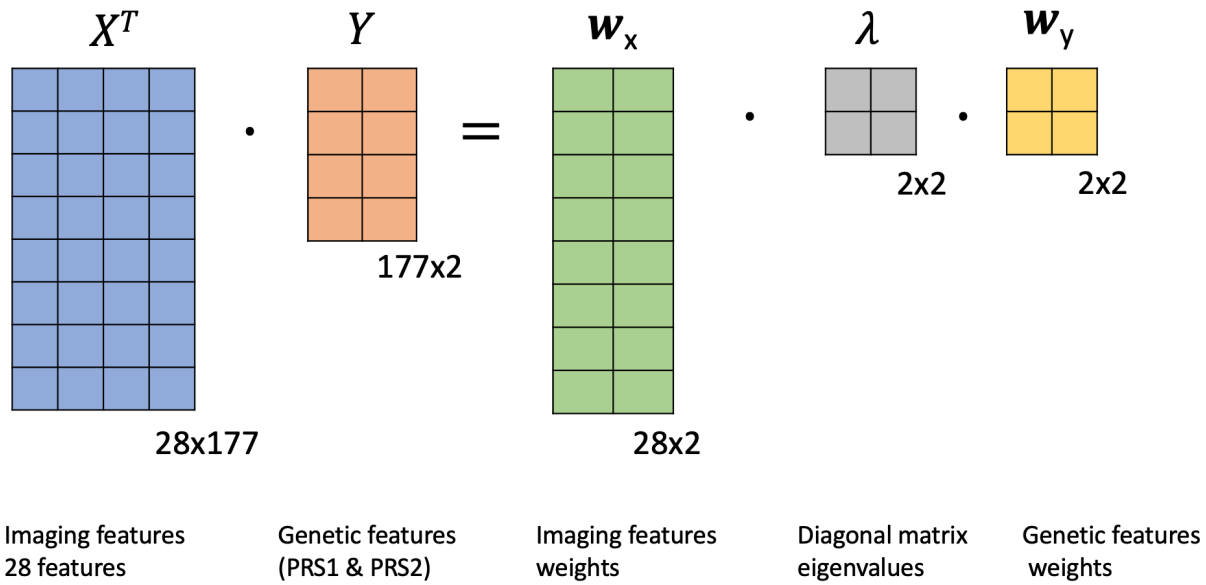


Figure 5.7: PLS model on imaging and genetics features

LASSO Regression

LASSO short for least absolute shrinkage and selection operator is a regularization technique used in linear regression models. LASSO reduces the number of features by estimating sparse coefficients. By shrinking some of the coefficients in the variables LASSO performs variable selection which make the model more interpretable. Penalty term in LASSO will control the number of coefficients set to zero. As the penalty increases, more coefficients become equal to

zero in which feature selection is taking place. [47]

Figure 5.8 shows a block diagram of the steps which were carried out to prepare the data before applying the PLS model. First, loading of the fMRI imaging features composed of 28 features and the genetic features made up of PRS1 and PRS2 for each subject. Along with the features the confounding data which included age and sex were loaded according to the corresponding information for each subject in this study. Second step is standardization of the data, the mean was subtracted from the data and then divided by the standard deviation. For the third step, deconfounding was applied to the data to remove affects from age and sex to remove any bias. Additionally, 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 [6]. The PLS model was applied and PLS component's weights as well as the projection of the data in the latent space was presented. The eigen-values were extracted in order to assess the variability in each component. Finally, in order to verify the significance of the model a permutation test was run, by permuting the Y matrix representing the genetic features and running the model for 1000 times and testing how many times the eigen-values obtained were higher than the original eigen-values. Additionally, the PLS model was applied with LASSO optimizer in which the same procedure was done and model significance ($p < 0.05$) was assessed with 1000 repetitions. Multiple parameters for the penalty term were tested, ranging between $0.11 - 0.3$.

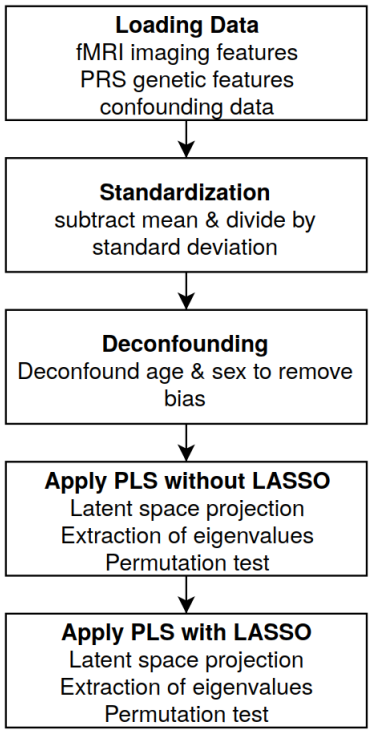


Figure 5.8: PLS methodology steps

6 Results

6.1 Noise reduction pipelines for single subjects

In this section the results for applying noise reduction pipelines on 5 subjects are presented. Two main noise reduction pipelines are experimented ICA-AROMA and NRP. Preprocessing was applied on 5 subjects in order to assess the performance and select the one to proceed with. The same 5 subjects were processed for both NRP and ICA-AROMA in order to evaluate the performance and view the FC matrices which were produced. The 5 single subjects that were tested belonged to different cognitive groups; 3 were CN, 1 SMC and 1 MCI. FC matrices and histogram of FC values were generated in order to visualize the results and compare the output FC matrices in regard to each pipeline.

6.1.1 ICA-AROMA pipeline

ICA-AROMA pipeline for removal of head movement noise in fMRI data was applied on 5 subjects in which FC matrices and histogram of FC values were generated. The FC matrices for 3 subjects are displayed in figure 6.1 in which ICA-AROMA was applied as preprocessing pipeline. The 3 subjects shown are from 3 different cognitive groups (CN, SMC and MCI).

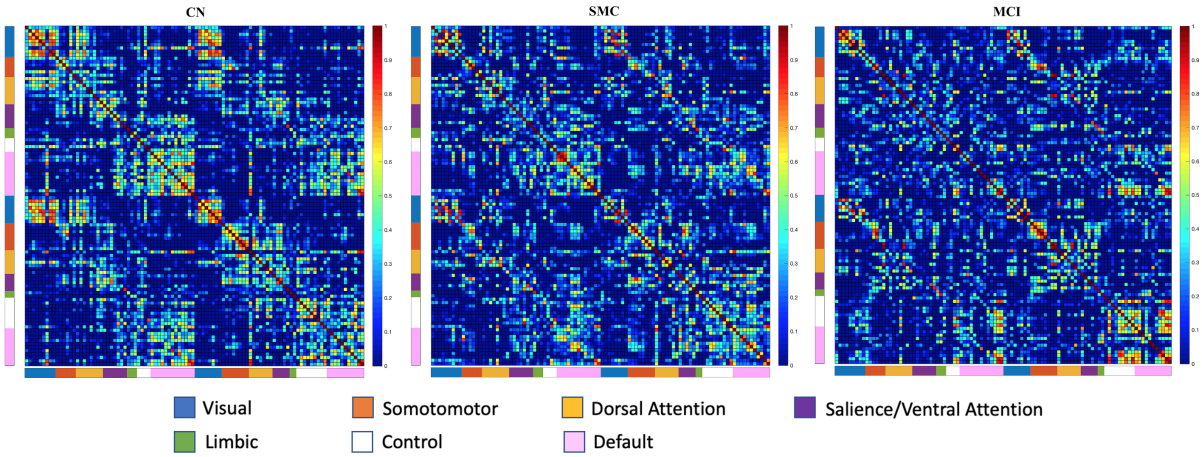


Figure 6.1: FC matrices for 3 subjects after ICA-AROMA pipeline

Applying ICA-AROMA on the ADNI data in this case did not show very promising results. ICA-AROMA FC matrices have a lot of negative correlations and the FC values seems to be centered around zero.

6.1.2 Nuisance Regression Pipeline (NRP)

NRP for head motion related noise removal was applied on 5 subjects and FC matrices were generated to assess the performance. Histogram of FC values were also presented in order to look at the distribution of the values. Figure 6.2 show the FC matrices for the 3 subjects after applying NRP on the filtered fMRI data.

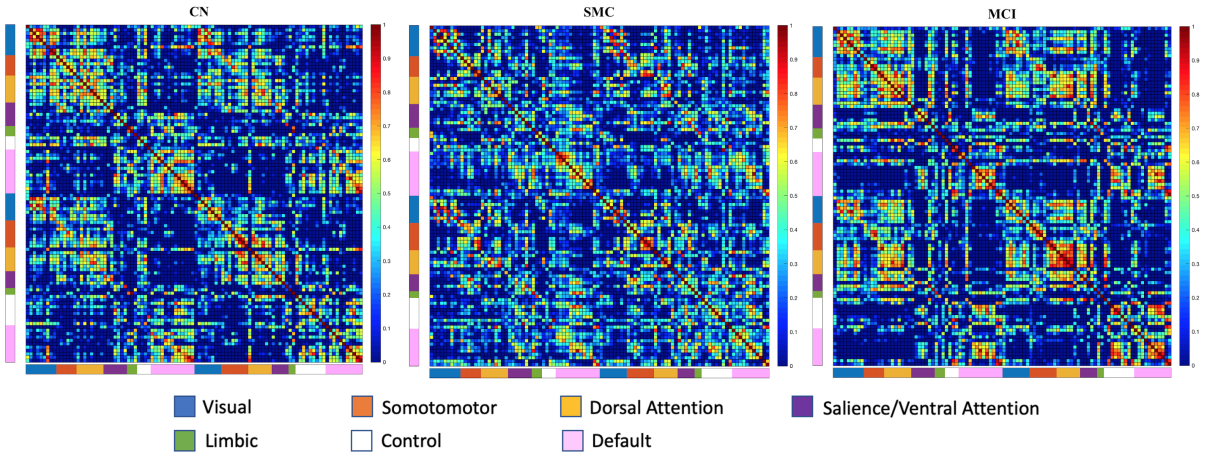


Figure 6.2: FC matrices for 3 subjects after NRP pipeline

The results shown by the FC matrices after applying NRP on the data have shown very promising outcome. The FC matrix patterns are clearly visible. The 5 subjects had also exhibited consistent results after applying this pipeline.

The FC matrices in figure 6.1 and 6.2 show the results of the noise reduction pipelines both for applying ICA-AROMA and NRP respectively. The ICA-AROMA FC values had a lot of negative correlations and patterns were not so visible compared to NRP. Meanwhile, NRP has shown better FC matrices and connectivity patterns. Consequently, it has been decided to move forward with NRP for the rest of the downloaded subjects.

Including global signal as regressor

After testing the two approaches with and without GM for NRP and ICA-AROMA, the results didn't show a big difference in the FC matrices. The FC matrices generated while including GM in the design matrix had more negative correlations in the FC values. Since there wasn't a big difference in the FC matrices and including GM as a regressor showed more negative correlations this approach was discarded and not moved forward with.

6.2 Feature reduction for imaging data

In this section the results for the feature reduction is reported. There are two ways in which feature reduction was computed in order to summarize the FC matrix into smaller set of features. Two sets of features have been extracted after applying NRP on the data; one set composed of 28 values for each subject, and the second set composed of 21 values for each subject.

6.2.1 28 imaging features

28 imaging features were extracted from the FC matrices after applying NRP for the subjects. A summary 7×7 symmetric matrix was calculated in which half of the values in this matrix composed of 28 values are used later as imaging features from the fMRI data. The 28 features represent the mean values for within and between network connectivity.

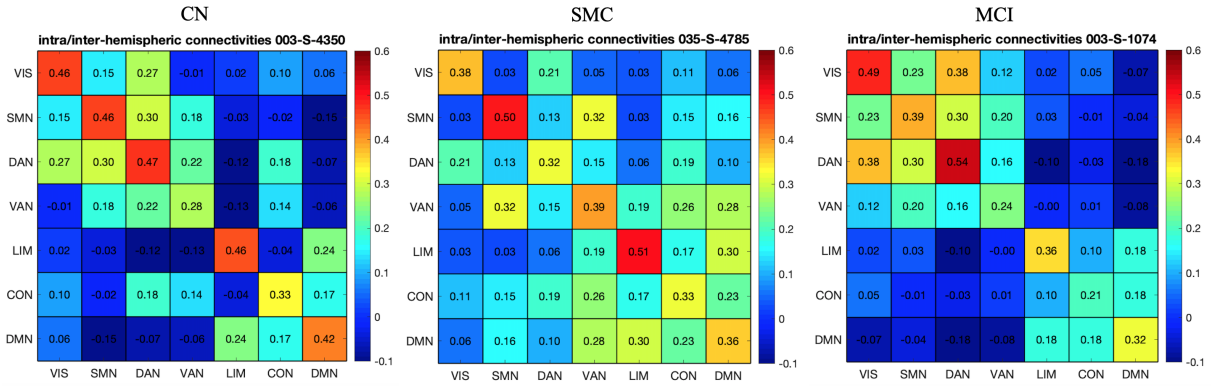


Figure 6.3: Summary matrices with 28 features for 3 subjects

The results of the feature reduction for the 3 subjects shown in figure 6.2 can be seen in figure 6.3. Figure 6.3 shows the 7×7 symmetric matrices for 3 subjects in three cognitive groups (CN, SMC and MCI), half of the values in this matrix was used as 28 imaging features.

6.2.2 21 imaging features

As another imaging features represented from the FC matrices 21 values were selected. For each subject in this study after applying NRP, 21 features are calculated as described in the methodology section and in figure 5.6. Since 7 RSNs are used in this study; global average of left, right and left-right are extracted for each network making up 21 values.

The vector for each subject composed of 21 values is shown in figure 6.4. Figure 6.4 show the extracted global average values results for the MCI subject which FC matrix is shown in figure 6.2.

	VIS	SMN	DAN	VAN	LIM	CON	DMN
Left	0.5492	0.5396	0.6162	0.2894	0.3341	0.3510	0.3363
Right	0.5004	0.3805	0.5841	0.4232	0.6217	0.2287	0.4125
Left_Right	0.4515	0.3350	0.4753	0.1395	0.2868	0.1470	0.2767

$M^{3\times7}$

0.5492	0.5004	0.4515	0.5396	...	0.1470	0.3363	0.4125	0.2767

$V^{21\times1}$

Figure 6.4: Vector of 21 features from the original values

This set of features are not considered further. Although it was tested and extracted during the analysis of subjects and could have been considered for the statistical modeling, later in the study the 21 features were discarded. The 21 features were not used further because only part of the FC matrix is represented using these 21 values. On the other hand, the 28 features represent the intra/inter-hemispheric connectivity so it has been decided to move forward with this set of features.

6.3 NRP for all considered subjects

NRP was applied on 177 subjects in total. Basic preprocessing and NRP were applied in order to remove any noise in the data. For each subject FC matrices were extracted and feature reduction has been implemented to extract the 28 imaging features. The imaging features were put into a vector and saved into each subject's folder.

The previous process mentioned was applied on 177 subjects using the bash script. For each subject it took about 20 minutes to run the preprocessing, processing and extract everything needed (FC matrix, summary matrix, feature vectors).

The subjects included in this study were divided into multiple cognitive groups. The arrangement of these groups is shown in table 6.1:

Table 6.1: Subjects grouping

Group	No. of Subjects
CN	52
SMC	43
EMCI	52
MCI	4
LMCI	26
Total	**177**

Mean matrices

The mean FC matrices for controls and MCI subjects is shown in figure 6.5. Control subjects including CN and SMC and MCI including EMCI, MCI and LMCI as shown in table 6.2. The mean matrices were generated by calculating the mean after combining each subject's individual FC matrix using the *mean* function on MATLAB.

Table 6.2: CN and MCI counts

Groups	No. of Subjects
CN/SMC	95
EMCI/MCI/LMCI	82
Total	**177**

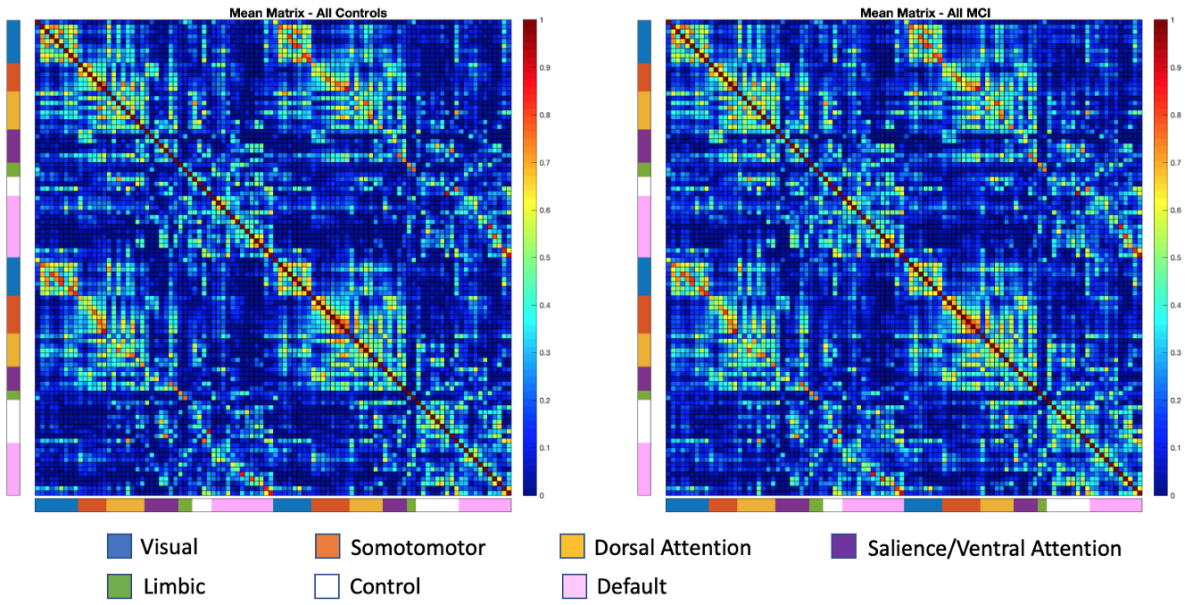


Figure 6.5: Mean FC matrices for CN and MCI

6.4 PLS model on imaging and genetics data

In this section PLS model results will be presented. PLS was applied on imaging and genetics features extracted from each subject. The aim is to find statistical correlations between phenotype and genotype. The phenotype was represented in the 28 features extracted from the FC matrices after preprocessing of the data and the genotype features were two PRSs (PRS1 & PRS2) which are a summary of the subjects SNPs. As described in the methodology section PLS was applied without LASSO first then with LASSO, the results for each test will be presented below. The PLS weights for both models were plotted and the projection of the data in the latent space were previewed. By extracting the eigen-values it's reported how much variability each component present. A permutation test was done to assess the significance of the PLS model.

PLS without LASSO

Figure 6.6 show the PLS component's weights for both the imaging and genetics features after applying the PLS model without LASSO. There are two components which are the two latent factors. Looking at the eigen-values it can be concluded that component 1 represents approximately 55% of the total variability in the data. While component 2 show about 45% of the variability. After running the permutation test as explained in the methodology section with 1000 repetitions and assessing the extracted eigen-values, the resulted *p-value* for this model is 0.845.

The imaging features shown in the PLS weights figures (figure 6.6 and 6.7) are displayed in multiple colors according to each column displayed in figure 5.5. Darker colors represent the intra-hemispheric connectivity (the diagonal of the 7×7 matrix) while the lighter shades show the inter-hemispheric connectivity.

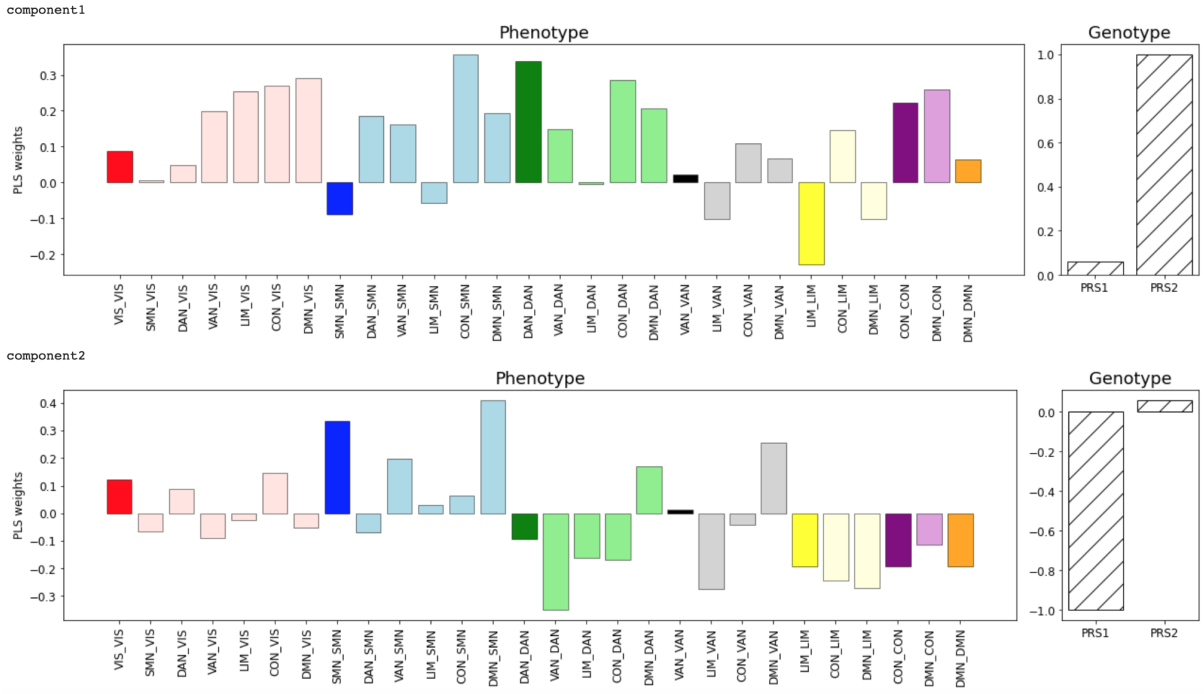


Figure 6.6: PLS component's weights for imaging and genetic features without LASSO

PLS with LASSO

For applying PLS with LASSO, penalty = 0.15 showed the optimal results in regard to number of features retained and the assessment of the permutation test.

Results for applying PLS model with LASSO with penalty = 0.15 is presented in PLS component's weights figure 6.7. From the eigen-values, component 1 show 54% of the variation in the data and component 2 show 46%. The projection of the data in the latent space is presented in figure 6.8 showing each component; in which figure 6.8a and 6.8b are the same figures but one show all the groups separated and the other show the combined groups into two MCI and CN as mentioned in table 6.2 respectively. A permutation test was applied to assess the model significance and resulted in *p-value* = 0.044.

In the first component in figure 6.7, while all the coefficients had the same sign, differences could be appreciated in terms of weights across the FC features. The imaging feature of between-network connections CON_DAN and CON_VAN had the highest representation. 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 (LIM_LIM feature) which reached the lowest value. It can also be noted that, within network connectivity (shown in darker colors) have generally lower weights compared to the between network connectivity features.

The second component in figure 6.7 most of the FC features had zero weights (or close to zero, as for LIM_VIS). There exists an opposite trend across networks, differently than shown before. 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, features related to the DAN and VAN networks, which reached high coefficient values in the first component,

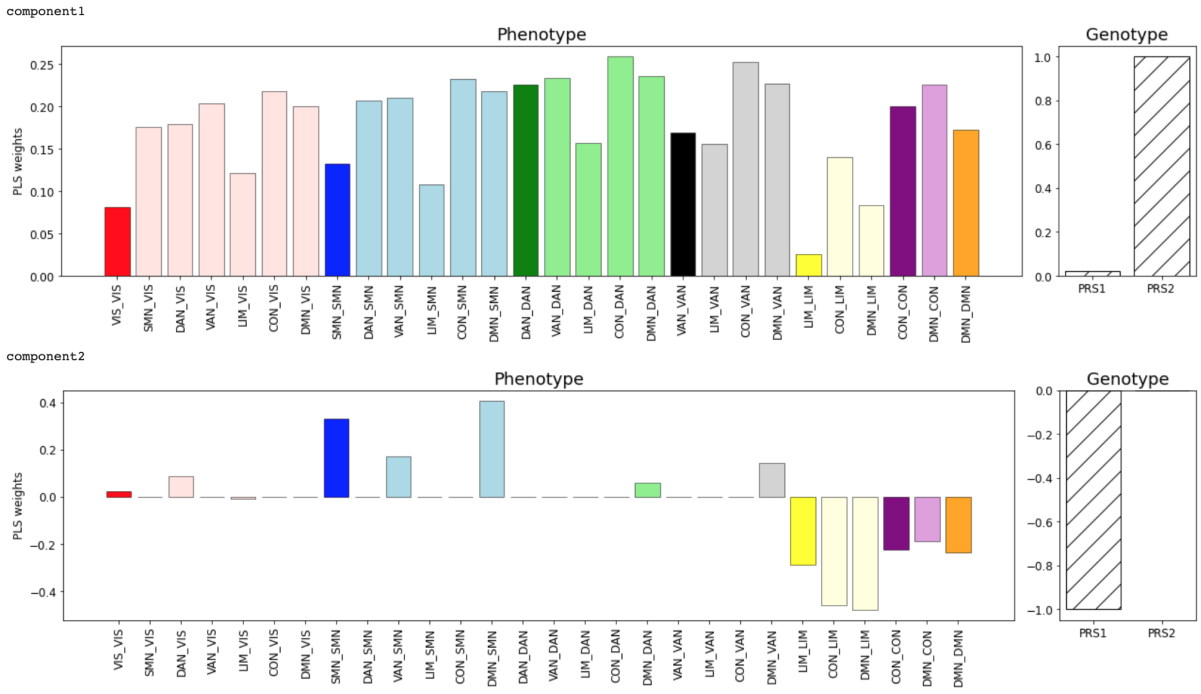
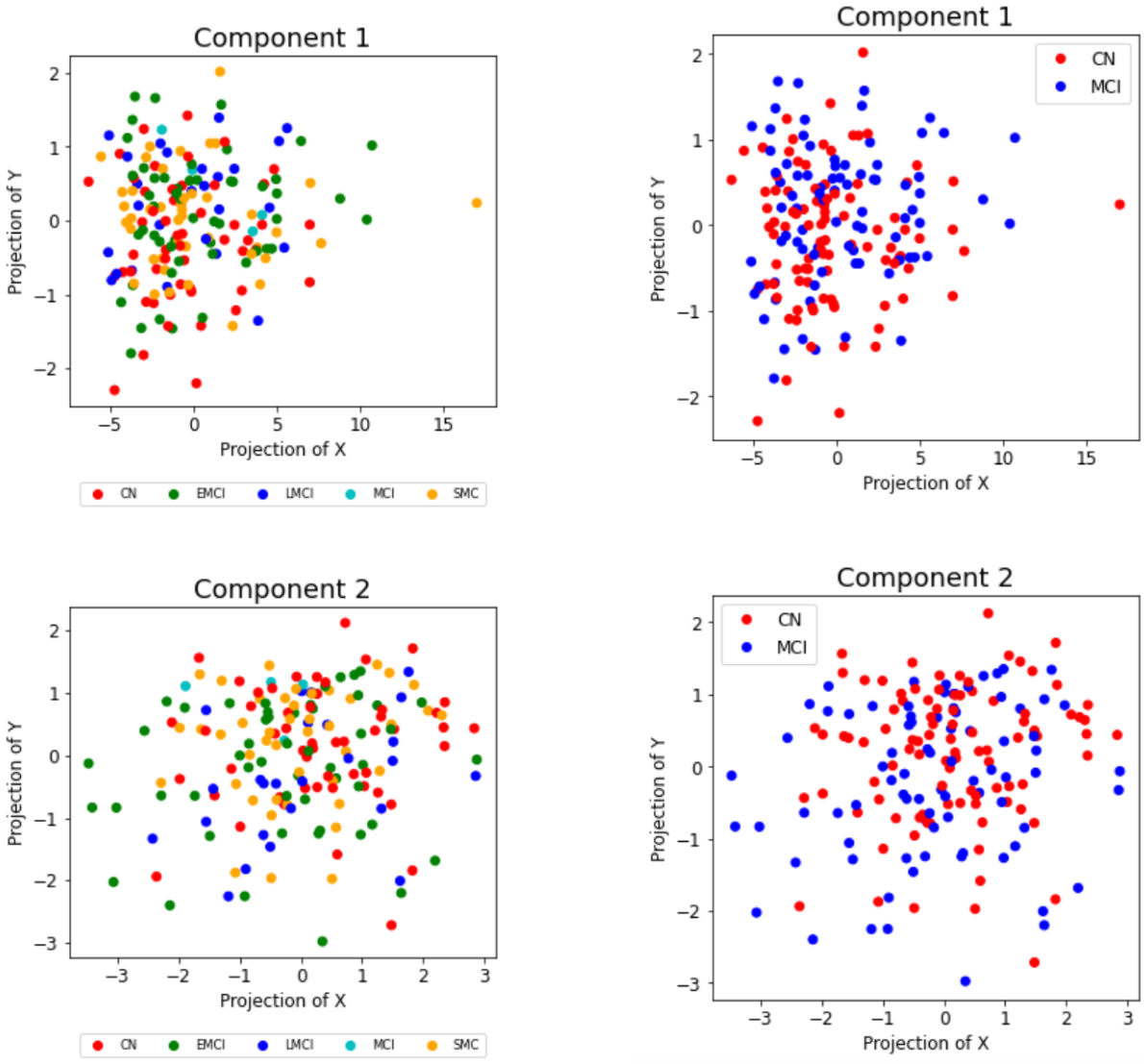


Figure 6.7: PLS component's weights for imaging and genetic features with LASSO

appeared to have a negligible contribution in this second one. Finally, the DMN seems to be the network with most surviving features either from within or between network connectivity with 6 features represented in the second component include DMN connection (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 imaging features involving LIM, CON and DMN. In the second component, SMN_SMN, VAN_SMN and DMN_SMN show anticorrelation with PRS1.



(a) All groups

(b) CN and MCI

Figure 6.8: Latent space projection of the data after PLS with LASSO

7 Analysis and Discussion

In this study multiple noise reduction pipelines were tested, namely ICA-AROMA and NRP both in which addition of GM was investigated. This test was done on 5 single subjects first to evaluate the performance and select the pipeline most suitable to use for all other subjects. Basic preprocessing and head motion noise reduction pipelines were applied to the subset and FC matrices were generated. The FC matrices generated from ICA-AROMA shown in figure 6.1 had a lot of negative correlations and the FC patterns are not clearly visible compared to NRP. NRP pipeline was simple to implement and showed consistency with the FC matrices along subjects with less negative correlations and clearly visible FC patterns, so it was selected to move forward with. Regarding the global signal represented in the GM, the matrices extracted were not very different from the pipelines without GM in the design matrix; although more negative correlation were present. Therefore, this approach was discarded since it is a controversial step and did not prove useful in this research.

In order to summarize the FC matrices and preform feature reduction, two summary methods were used which obtained imaging features composed of 21 and 28 features described in the methodology section about feature reduction and showed in figures 5.5 and 5.6. The 21 features derived were discarded and the 28 features were used further as imaging features since more information from the original FC matrices composed of within and between network connectivity are presented in the 28 features summary values.

The data was prepared to apply the PLS model on the imaging and genetic features. First PLS without LASSO was tested, the results were not promising and the permutation test failed to proof significance ($p < 0.5$). Secondly, PLS model with LASSO optimizer was applied in which variable selection took place. Analysis of the PLS with LASSO model weights showed associations between specific imaging features and one of the PRSs. In particular, all FC features were correlated with PRS2 in the first component, 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, CSF-tau levels and with progressive atrophy in AD [6]. The 28 FC features were differently represented in the two components and the PRSs had a differential association as seen in figure 6.7, suggesting these PRSs for AD might serve as potential unique patterns with association to FC features which require further investigation. Features including connectivity involving DAN, VAN, CON and DMN were those featuring 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. Connectivity with DMN were among the most surviving features in the second component. The RSN DMN has been largely investigated in the literature, and both within and between network changes have been reported between CN, MCI and AD patients [48], [49].

By looking at the latent space projection in figure 6.8, it can be seen that the groups are not clustered. There are a couple of reasons that could explain why the data is not clustered after applying the PLS model and finding maximum covariance between the genetic and imaging features. One reason could be that the groups of MCI and CN are very similar in terms of FC, which is indeed the case by looking at the FC matrices and the mean matrices shown in figure 6.5 it can be seen that there are very small difference between the imaging features represented in the mean matrices between the two groups. Taking into consideration also the organization of the groups in table 6.1 the majority of MCI subjects belong to the EMCI category which is a preliminary MCI stage and might have contributed to a lot of the overlap. Finally, another reason that might have contributed to this issue is the number of subjects included in this study. This study included 177 subjects but with more data the model could drastically improve and more separation between the groups could be achieved if tested.

This research attempted to exploit a multivariate PLS model for linking the PRS for AD with brain FC measures in MCI subjects. Within/between-network FC features are used as imaging features and associations with PRSs were reported. Evidence in favor of the suitability of PRS scores for explaining in a selective way the genetic role of such changes was found. Such research suggests the important role in investigating further the affect of AD on the RSNs while incorporating genetic information.

8 Conclusion and Future Work

The aim of this thesis is to investigate associations between neuroimaging phenotype and genetic features for CN and MCI. The phenotypic features were presented in terms of FC derived from rs-fMRI scans. The data used in this research was collected from ADNI database in which both structural and functional MRI for each subject was processed. In order to generate the imaging features multiple pipelines were tested to evaluate the preprocessing and noise reduction outcome. Two major pipelines were selected for testing ICA-AROMA and NRP with a couple of variations in both. As preliminary test assessment of single subjects first took place to test performance and choose the best pipeline. Raw rs-fMRI data was preprocessed with basic preprocessing procedure then ICA-AROMA or NRP were applied to remove further noise artifacts in particular head motion related noise. FC matrix for each subject was generated to map FC in the brain. The generated FC matrices have been compared in which NRP was chosen to continue running the analysis on the rest of the data. To summarize the FC matrix, feature reduction methods are implemented in which two approaches were tested one summarized the matrix into 28 features and the second to 21 features. The 28 features composed of the within and between network connectivity were selected as imaging features. In order to run on as many subjects as needed, a bash script was written to run the basic preprocessing, NRP and generate the FC matrices and summary features for all subjects included in this study. Overall analysis has been run on 177 subjects, divided into 5 cognitive groups (CN, SMC, EMCI, MCI and LMCI). As genetic features two PRSs scores proposed as summary of genetic variations were used for each subject (PRS1 & PRS2). Finally, multivariate regression method was used to investigate links between the imaging and genetic features. PLS model was applied to maximize the covariance between the two sets of data and tested with LASSO regularization retaining the most relevant features. PLS weights and projection of the data into latent space were observed. By surveying the outcome of the PLS model with LASSO specific features were among the most represented. Particularly the features representing the intra hemispheric connectivity were generally less presented compared to the between networks features. Analysis of the PLS component's weights showed associations between specific imaging features and one of the PRSs. Among all the networks features that include DMN connection have survived the most. A permutation test was done and PLS with LASSO model proved to be significant $p < 0.05$.

Despite these promising preliminary results, the small sample size represents the main limitation of this study. Including more subjects will improve the model and give additional insights into links between FC and genetic features. In addition, the lack of AD subjects causes another constraint. Taking into consideration AD patients will allow investigating associations between genetic and imaging features across all stages of neurodegenerative diseases. Moreover, testing the PLS model with variations of imaging features could be done. The 28 features selected in this study were one possibility of summarizing the FC matrix, ideally the FC matrix could be used but as it will be composed of 5000 features to avoid over-fitting a lot more subjects need to be included. Nevertheless, other summary approaches for feature reduction can be tested.

References

[1] Janine Bijsterbosch, Stephen M Smith, and Christian F Beckmann. *An introduction to resting state fMRI functional connectivity*. Oxford University Press, 2017.

[2] David M Cole, Stephen M Smith, and Christian F Beckmann. Advances and pitfalls in the analysis and interpretation of resting-state fmri data. *Frontiers in systems neuroscience*, 4:8, 2010.

[3] Bharat Biswal, F Zerrin Yetkin, Victor M Haughton, and James S Hyde. Functional connectivity in the motor cortex of resting human brain using echo-planar mri. *Magnetic resonance in medicine*, 34(4):537–541, 1995.

[4] Martijn P Van Den Heuvel and Hilleke E Hulshoff Pol. Exploring the brain network: a review on resting-state fmri functional connectivity. *European neuropsychopharmacology*, 20(8):519–534, 2010.

[5] National Institutes of Health Genome-Wide Association Studies Fact Sheet. `https://www.genome.gov/about-genomics/fact-sheets/Genome-Wide-Association-Studies-Fact-Sheet`. Accessed: 01-05-2021.

[6] 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. *Brain communications*, 2(1):fcz047, 2020.

[7] Kristin L Bigos and Daniel R Weinberger. Imaging genetics—days of future past. *Neuroimage*, 53(3):804–809, 2010.

[8] Li Shen and Paul M Thompson. Brain imaging genomics: integrated analysis and machine learning. *Proceedings of the IEEE*, 108(1):125–162, 2019.

[9] Paul M Matthews and Peter Jezzard. Functional magnetic resonance imaging. *Journal of Neurology, Neurosurgery & Psychiatry*, 75(1):6–12, 2004.

[10] Kâmil Uludag, David J Dubowitz, and Richard B Buxton. Basic principles of functional mri. *Clinical MRI. Elsevier, San Diego*, pages 249–287, 2005.

[11] Stephen M Smith, Diego Vidaurre, Christian F Beckmann, Matthew F Glasser, Mark Jenkinson, Karla L Miller, Thomas E Nichols, Emma C Robinson, Gholamreza Salimi-Khorshidi, Mark W Woolrich, et al. Functional connectomics from resting-state fmri. *Trends in cognitive sciences*, 17(12):666–682, 2013.

[12] Mark J Lowe, Mario Dzemidzic, Joseph T Lurito, Vincent P Mathews, and Micheal D Phillips. Correlations in low-frequency bold fluctuations reflect cortico-cortical connections. *Neuroimage*, 12(5):582–587, 2000.

[13] César Caballero-Gaudes and Richard C Reynolds. Methods for cleaning the bold fmri signal. *Neuroimage*, 154:128–149, 2017.

[14] Raimon HR Pruim, Maarten Mennes, Jan K Buitelaar, and Christian F Beckmann. Evaluation of ica-aroma and alternative strategies for motion artifact removal in resting state fmri. *Neuroimage*, 112:278–287, 2015.

[15] Gholamreza Salimi-Khorshidi, Gwenaëlle Douaud, Christian F Beckmann, Matthew F Glasser, Ludovica Griffanti, and Stephen M Smith. Automatic denoising of functional mri data: combining independent component analysis and hierarchical fusion of classifiers. *Neuroimage*, 90:449–468, 2014.

[16] Raimon HR Pruim, Maarten Mennes, Daan van Rooij, Alberto Llera, Jan K Buitelaar, and Christian F Beckmann. Ica-aroma: A robust ica-based strategy for removing motion artifacts from fmri data. *Neuroimage*, 112:267–277, 2015.

[17] Chun-Chao Huang, Wei-Ming Huang, Chia-Hung Chen, Zong-Yi Jhou, Ching-Po Lin, Alzheimer's Disease Neuroimaging Initiative, et al. The combination of functional and structural mri is a potential screening tool in alzheimer's disease. *Frontiers in aging neuroscience*, 10:251, 2018.

[18] Sebastian Palmqvist, Michael Schöll, Olof Strandberg, Niklas Mattsson, Erik Stomrud, Henrik Zetterberg, Kaj Blennow, Susan Landau, William Jagust, and Oskar Hansson. Earliest accumulation of β-amyloid occurs within the default-mode network and concurrently affects brain connectivity. *Nature communications*, 8(1):1–13, 2017.

[19] 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. *Nature communications*, 11(1):1–17, 2020.

[20] Bin Wang, Yan Niu, Liwen Miao, Rui Cao, Pengfei Yan, Hao Guo, Dandan Li, Yuxiang Guo, Tianyi Yan, Jinglong Wu, et al. Decreased complexity in alzheimer's disease: resting-state fmri evidence of brain entropy mapping. *Frontiers in aging neuroscience*, 9:378, 2017.

[21] Eek-Sung Lee, Kwangsun Yoo, Young-Beom Lee, Jinyong Chung, Ji-Eun Lim, Bora Yoon, and Yong Jeong. Default mode network functional connectivity in early and late mild cognitive impairment. *Alzheimer Disease & Associated Disorders*, 30(4):289–296, 2016.

[22] Frank de Vos, Marisa Koini, Tijn M Schouten, Stephan Seiler, Jeroen van der Grond, Anita Lechner, Reinhold Schmidt, Mark de Rooij, and Serge ARB Rombouts. A comprehensive analysis of resting state fmri measures to classify individual patients with alzheimer's disease. *Neuroimage*, 167:62–72, 2018.

[23] 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. *Neuroimage*, 63(1):11–24, 2012.

[24] 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. *Neuroimage*, 107:289–310, 2015.

[25] Marco Lorenzi, Boris Gutman, Derrek P Hibar, Andre Altmann, Neda Jahanshad, Paul M Thompson, and Sebastien Ourselin. Partial least squares modelling for imaging-genetics in alzheimer's disease: Plausibility and generalization. In *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, pages 838–841. IEEE, 2016.

[26] Marco Lorenzi, Andre Altmann, Boris Gutman, Selina Wray, Charles Arber, Derrek P Hibar, Neda Jahanshad, Jonathan M Schott, Daniel C Alexander, Paul M Thompson, et al. Susceptibility of brain atrophy to trib3 in alzheimer's disease, evidence from functional prioritization in imaging genetics. *Proceedings of the National Academy of Sciences*, 115(12):3162–3167, 2018.

[27] ADNI. About ADNI. `http://adni.loni.usc.edu/about/`. Accessed: 15-01-2021.

[28] ADNI. ADNI3: What's New. `http://adni.loni.usc.edu/adni-3/`. Accessed: 20-02-2021.

[29] 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. *Alzheimer's & Dementia*, 13(5):561–571, 2017.

[30] ADNI. Study Design. Background & Rationale. `http://adni.loni.usc.edu/study-design/`. Accessed: 04-04-2021.

[31] Mark Jenkinson, Christian F Beckmann, Timothy EJ Behrens, Mark W Woolrich, and Stephen M Smith. Fsl. *Neuroimage*, 62(2):782–790, 2012.

[32] Analysis Research. Wellcome Centre For Integrative Neuroimaging. Medical Sciences Division. University of Oxford. `https://www.win.ox.ac.uk/research/analysis-research/analysis-research`.

[33] FMRIB Software Library. FSL. `https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/`. Accessed: 07-01-2021.

[34] Mark Jenkinson, Peter Bannister, Michael Brady, and Stephen Smith. Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage*, 17(2):825–841, 2002.

[35] Stephen M Smith. Fast robust automated brain extraction. *Human brain mapping*, 17(3):143–155, 2002.

[36] Mark Jenkinson and Michael Chappell. *Introduction to neuroimaging analysis*. Oxford University Press, 2018.

[37] Kevin M Aquino, Ben D Fulcher, Linden Parkes, Kristina Sabaroedin, and Alex Fornito. Identifying and removing widespread signal deflections from fmri data: Rethinking the global signal regression problem. *Neuroimage*, 212:116614, 2020.

[38] 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. *Cerebral cortex*, 28(9):3095–3114, 2018.

[39] Ronald Aylmer Fisher. Statistical methods for research workers. In *Breakthroughs in statistics*, pages 66–70. Springer, 1992.

[40] Bianca De Blasi, Lorenzo Caciagli, Silvia Francesca Storti, Marian Galovic, Matthias Koepp, Gloria Menegaz, Anna Barnes, and Ilaria Boscolo Galazzo. Noise removal in resting-state and task fmri: functional connectivity and activation maps. *Journal of Neural Engineering*, 17(4):046040, 2020.

[41] Jian Zhai and Ke Li. Predicting brain age based on spatial and temporal features of human brain functional networks. *Frontiers in human neuroscience*, 13:62, 2019.

[42] Shing Wan Choi, Timothy Shin-Heng Mak, and Paul F O'Reilly. Tutorial: a guide to performing polygenic risk score analyses. *Nature Protocols*, 15(9):2759–2772, 2020.

[43] Shing Wan Choi and Paul F O'Reilly. Prsice-2: Polygenic risk score software for biobank-scale data. *Gigascience*, 8(7):giz082, 2019.

[44] Svante Wold, Michael Sjöström, and Lennart Eriksson. Pls-regression: a basic tool of chemometrics. *Chemometrics and intelligent laboratory systems*, 58(2):109–130, 2001.

[45] Herman Wold. Soft modeling: the basic design and some extensions. *Systems under indirect observation*, 2:343, 1982.

[46] Kim-Anh Lê Cao, Debra Rossouw, Christele Robert-Granié, and Philippe Besse. A sparse pls for variable selection when integrating omics data. *Statistical applications in genetics and molecular biology*, 7(1), 2008.

[47] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[48] Michael D Greicius, Gaurav Srivastava, Allan L Reiss, and Vinod Menon. Default-mode network activity distinguishes alzheimer's disease from healthy aging: evidence from functional mri. *Proceedings of the National Academy of Sciences*, 101(13):4637–4642, 2004.

[49] 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. *Proceedings of the National Academy of Sciences*, 104(47):18760–18765, 2007.

Non-exclusive licence to reproduce thesis and make thesis public

I, Heba Elshatoury

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to:

   1.1 reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright, and

   1.2 make available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work from 15/06/2022 until the expiry of the term of copyright,

   "Disentangling the association between functional connectivity and genetics in Mild Cognitive Impairment via multivariate regression models"

   supervised by Dr. Ilaria Boscolo Galazzo, Prof. Gloria Menegaz and Prof. Gholamreza Anbarjafari

2. I am aware of the fact that the author retains the rights specified in p. 1.

3. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Heba Elshatoury
18.05.2021