UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

Corso di laurea in Bioingegneria della Riabilitazione

PREDICTIVE MODELING OF POST-STROKE COGNITIVE IMPAIRMENT IN NUMERICAL AND FINANCIAL ABILITIES THROUGH BRAIN CONNECTIVITY ANALYSIS

Relatrice
Prof.ssa Bertoldo Alessandra

Laureando
Gatti Geremia

Correlatrice
Ing. Baron Giorgia

ANNO ACCADEMICO 2023 – 2024

5 Dicembre 2024

ABSTRACT

Stroke is one of the principal causes of death and disability in the world. Understanding how the human brain is altered by its occurrence is fundamental for the creation of effective rehabilitation programs to help patients in the recovery of the lost functions.
The aim of this thesis is to create a model for the prediction of numerical and financial deficits of post-stroke patients, using new promising connectivity-based approach and data reduction algorithms, thereby linking symptoms to brain networks. In this view, the structural MRI and rs-fMRI imaging data of 31 stroke patients were collected in order to extract their anatomical disconnections and functional connectivity that are processed through 2 different data reduction techniques (Principal Component Analysis and Uniform Manifold Approximation and Projection). The resulting scores and some confounding variables (age, schooling, lesion volume and parcel loads) are used as input in the Canonical Component Analysis, where they are correlated with the scores coming from specific tests ideated to infer numerical and financial abilities. The evaluation of the Canonical Correlation Analysis outputs through a hierarchical clustering highlights that a better performance is related with right stroke patients with small lesions, which show an important inter-hemispherical segregation, while inter-hemispheric integration seems to play a relevant role in patients with wider damage.
This work wants to be a helpful tool in the understanding of brain behaviour after stroke, highlighting the synergistic and additive nature of different types of network modalities, and their corresponding influence on behavioural performance after brain injury in order to create effectiveness rehabilitation programs or new prediction models for the evaluation of stroke patients abilities.

(IT) L’ictus è una delle principali cause di morte e disabilità nel mondo. Comprendere come il cervello umano viene alterato dalla sua comparsa è fondamentale per la creazione di programmi riabilitativi efficaci per aiutare i pazienti nel recupero delle funzionalità perdute.
Lo scopo di questa tesi è creare un modello per la previsione dei deficit numerici e finanziari dei pazienti post-ictus, utilizzando un nuovo e promettente approccio basato sulla connettività cerebrale e algoritmi di riduzione dei dati, al fine di collegare i sintomi alle reti cerebrali. In quest'ottica, sono stati raccolti i dati di imaging MRI strutturale e rs-fMRI di 31 pazienti con

ictus, al fine di estrarre le loro disconnessioni anatomiche e connettività funzionali che vengono elaborate attraverso 2 diverse tecniche di riduzione dei dati (Principal Component Analysis and Uniform Manifold Approximation and Projection). Le nuove dimensioni risultanti e alcune variabili additive (età, scolarizzazione, volume delle lesioni e carico dei pacchi) vengono utilizzate come input per la Canonical Component Analysis, dove sono correlate con i punteggi provenienti da test specifici ideati per dedurre abilità numeriche e finanziarie. La valutazione dei risultati derivanti dalla Canonical Component Analysis attraverso un clustering gerarchico evidenzia che una prestazione migliore è correlata ai pazienti con ictus destro e domesioni ridotte della lesione, i quali mostrano un'importante segregazione inter-emisferica, mentre l'integrazione inter-emisferica sembra giocare un ruolo rilevante nei pazienti con danno più ampio.

Questo lavoro vuole essere uno strumento utile nella comprensione del comportamento cerebrale dopo un ictus, evidenziando la natura sinergica e additiva di diverse reti neurali e la loro corrispondente influenza sulle prestazioni comportamentali dopo una lesione cerebrale, al fine di creare programmi di riabilitazione efficaci o nuovi modelli predittivi per la valutazione delle capacità dei pazienti con ictus.

INDEX

1 INTRODUCTION

1.1 The brain after stroke: a clinical overview

3 According to the Ministry of Health in Italy it represents the second cause of death (about 10% of total deaths) and the first cause of disability, only 25% of patients who survive manage to recover completely [1].

About 70% of strokes are caused by the occlusion of one of the major cerebral arteries, usually the middle cerebral artery, due to embolism or thrombosis, but also smaller vessels occlusion can lead to limited lesions, typically in the subcortical white matter or basal ganglia. In all these cases the events are called ischemic strokes.

Less common is the haemorrhagic stroke caused by the breaking of a vessel that can be intraparenchymal or it can also take place in the subarachnoid space. Even if it is less frequent than the ischemic one, its mortality rates reach almost 80% in low and middle-income countries [2].

Studying strokes, it is important to remember that they are events that change in time, in particular ischemic strokes are categorised into hyperacute (0-6 h), acute (6-24 h), subacute (24h to roughly 2 weeks) and chronic (more than 2 weeks) stages, after the onset of the illness [3], and the speed of progression of the death of neurons change for each person and depend on the location of the stroke in the brain, for example, cortical regions are more susceptible to ischemia than the caudal region.

After an insult a spontaneous recovery takes place in the brain to try to substitute the missing functions of the damaged area, altering its normal behaviour.

During the acute phase it can be interesting to study stroke patients because it gives the chance to analyse the normal activity of the brain before the reorganisation and to also detect

the effect of small lesions. Instead, usually after six months, in the chronic phase, the reorganisation of the brain is generally complete and this can lead to the opportunity to understand what function can't be substituted after the loss of some part of the brain. Moreover, chronic stroke patients are more stable and can undergo longer assessment sessions, so it is important to understand what patients include in a study depending on the research questions, methods and the availability of an adequate number of patients [4].

Assessing stroke with Magnetic Resonance Imaging (MRI) has a great advantage over other imaging techniques thanks to its non-invasive characteristic, good anatomical resolution and contrast.

The typical parameters needed to discriminate between different tissues are the spin-lattice relaxation time (T1) and the spin-spin relaxation time (T2). In addition to the standard techniques, MRI allows to compute more advanced imaging sequences like diffusion-weighted MRI (DWI), that is particularly appropriate to assess strokes [5].

For example, in acute phase the infarct is not visible for many hours after the onset, but it can be noticed using DWI, that scan the motion of water molecules in tissues, in which images a brighter (hyperintense) area means acute lesion and darker (hypointense) area indicates chronic one, so helping also to know the age of the lesion.

Fluid-attenuated inversion recovery (FLAIR) is another MRI technique, based on T2-weighted sequence, that can be useful for ischemic strokes identification and to recognize hypoperfusion area around the lesion in chronic stage. In fact, it is important to understand what happens near the damaged areas, where there could be edema that is the cause of a second type of injury on the tissue, due to inflammation, mechanical pressure, thrombin production or other. This could also alter the normal behaviour of the tissue increasing the functional deficit.

Focusing on chronic strokes, MRI techniques are preferred to computed tomography (CT) scan thanks to its non-invasive nature, the employment of radio frequency waves, the enhanced contrast and better signal-to-noise (SNR) ratio. FLAIR and T2 sequences are good to recognize changes in the white matter that have a relevant role on the identification of the symptom in patients [4].

All these techniques are valid solutions for the anatomical investigation of brain injury, but there are also solutions to infer the functional activation of the neurons' populations, not only as a reaction to a stimulus. Even in resting state, indeed, the human brain consumes a lot of energy, around 20% of the total amount of energy is used for communication between neurons

and to keep them and their supporting cells alive, whereas the increment in neuron metabolism due to task-related events is less than 5% [6]. So it seems reasonable to study spontaneous activity related to activation of patterns referred to as resting-state networks. These networks can be divided into several categories, they can be distinguished in 7 major patterns: visual, somatomotor, dorsal attention, ventral attention, limbic, frontoparietal and default, but they can be divided by other ways obtaining for example 17 different networks [7].

To assess them, resting state fMRI (rs-fMRI) is a functional imaging technique that measures the fluctuation of blood oxygenation level-dependent (BOLD) signals related to activation of populations of neurons in resting state brain. For its nature it is a low frequency signal (less than 0.1 Hz) because it manages to acquire indirectly the activity of brain regions, which has a dependency with the change of the concentrations of haemoglobin in the tissue. In rs-fMRI the acquisition is made in absence of a task so it can evaluate the functional connectivity related to activation of resting-state networks [8]. In fact, the disruption of parts of the brain can lead to the loss of specific functions and changes in the functional architecture of the brain due to damage such as stroke and its analysis can reveal important clues about brain functioning and its ability to recover [9].

Moreover, the increasing number of people suffering from stroke and the associated risk factors have led to the creation of new effective rehabilitation programs and to the need to know the processes that regulate the behaviour of the brain. In order to understand it, researchers cooperate with psychologists that are in charge to define the degree of impairment caused by the lesions using cognitive tests specifically created, which scores are then associated with information coming from neuroimaging techniques like structural and functional MRI. Regarding the aim of this thesis, for example, a group of stroke patients are examined through NADL Short and NADL-F Short tests, purposely ideated to investigate numerical and financial skills which will be discussed in the following chapter.

1.2 Post-Stroke Cognitive Impairments in Numerical and Financial Abilities

An important ability nowadays is to process numeric information. In fact, these skills are fundamental in everyday life, where each person is constantly required to know how to do calculations, understand fractions, proportions, remember codes, telephone numbers or addresses.

Arithmetic abilities are processed in our brain by activation of a distributed network that engages percentual, motor, spatial, and mnemonic functions and the most important region corresponds to the parietal lobes. Indeed, it has been demonstrated by neuroimaging methods that the intraparietal sulcus (IPS) is the core locus for numerical processing and it is activated just only on the left or right hemisphere or bilaterally depending on the task. However, as many other abilities, numerical skills can rely on the activation of a wide network that has interesting relations to many other areas, like language or memory. For example, solving a new numerical problem activates the IPS bilaterally, but when the same task is done a second time the angular gyrus of the left parietal lobes, related to memory, is involved [10].

In support of this, many researches have demonstrated that brain lesions can create deficits which can be very specific: some generic examples could be the selective impossibility to transcode from oral to writing number or vice versa, the difficulty to apply some specific simple operation but not others, like addition but not subtraction, or the impairment to understand particular signs or rules [11].

In fact, after a brain injury like stroke, patients can also experience some form of acalculia, an acquired disability that causes difficulties in the capability to understand numerical information or calculation. Since acalculia is not usually screened in patients, there is poor information about its nature and impact on the subjects [12] and it may also affect not only their numeric abilities but also financial ones.

As for the numerical skills, indeed, also the financial abilities are important in everyday life, since it directly affects the independence of people and so their quality of life. Usually this topic interests patients with Alzhaimer or mild cognitive impairment, that show difficulties managing financial conceptual knowledge [13], but since the dealing with money can be easily linked to the numerical skills some studies have been done on the topic.

In fact, some relations between financial and numerical abilities have been found in patients with different pathological conditions, showing that the two abilities are positively correlated,

but more detailed research shows that numerical abilities are more involved with basic financial tasks, rather than advanced ones that are more associated with abstract reasoning.
It can be concluded that even if the two abilities are related, there is not necessarily a dependency between them [14].

Due to this complexity, interesting hints can be derived from the analysis of brain lesioned patients to try to infer how numerical and financial abilities are processed and impaired by the lesion. On one hand, the location of the lesion can give some important information, in particular observing the side (left or right) and the site (parietal or non-parietal) of the damage, for instance parietal lesions tend to give more difficulties in oral and digital codes; oral codes would be more compromised with left parietal lesion that the right one; make use of alphanumeric codes involves the left hemisphere without dependency on parietal area due to its relation with language; the magnitude comparison ability is less impaired with non-parietal damage then parietal one, left subcortical lesions affect the knowledge of arithmetical facts; etc... [15]
On the other hand, connectivity analysis can provide fundamental insights into the altered mechanisms induced by the brain lesion, in fact for instance, analysis of anatomical and functional connectivities among different regions in the brain indicates that numerical cognition is supported by a widely distributed network involving the (intra)parietal and (pre)frontal cortices, as well as the hippocampus [16].

In the next section it will be provided a brief description of the most common methods used to analyse the link between lesion and behavioural and cognitive deficit.

1.3 Lesion to symptom mapping: voxel-based versus connectivity-based approaches

As mentioned in the previous section, the study of brain lesions is important for neuroscientists to understand the brain behaviour and connectivity as well as the recovery mechanisms after an injury. In fact, compromised brain areas can give important hints on the role and function they have, just looking at the cognitive loss that the patients experience. To infer the link that connects the behavioural change and cerebral damage the main tool used by researchers is lesion-symptom mapping (LSM) [17]. Initially the studies were made just comparing patients with analogous deficit overlapping their lesions and then also applying statistical measures on the single voxel of the brain imaging data. Indeed, traditionally all the analysis regarding the finding of a relation between lesion and symptom have been made just focusing on the damaged part of the grey matter. This method is called Voxel-based Lesion Symptom Mapping (VLSM) or "massively univariate approach" because a detailed topological map of the lesioned voxel is created. Based on this information, for each voxel of the brain, the patients can be divided in who have and who haven't lesions on that specific voxel and so it is possible to compare behavioural scores of the two groups with a statistical test [18].
In this way it can be related the presence or not of a damage in a specific region of the brain with its role in the patients behaviours.

However this method doesn't take into account that the lesion map is merged with the connectome and in fact many researches have found out relations between the interruption of white matter structure and cognitive and behavioural deficits, like language [19], visuo-spatial attention [20], motor function [21] and general cognition [22].
This leads to one of the limitations of the classic VLSM that is the fact that lesions with different degrees of damage or locations can affect the same anatomical structure [23].
Traditional VLSM studies also make the assumption that the strength of the relation between structural damage and behaviour is the same regardless of different behavioural domains measured, even if it is possible that the results of function at higher level like attention, active thinking or memory rather than sensory-motor function, are based on the contribution of different and distributed networks [24].
Moreover, there are also some statistical issues to overcome, like the necessity to introduce correction for the increase of false positives due to the many thousands of tests performed.

Another problem is the correlation between close voxels and their link to the nature of the lesions, that are tightly connected with the underlying vascular structure. Indeed, VLSM makes the assumption that the state of each voxel is independent of the damage of others, that it can't be true in the human brain. This can lead to grouping irrelevant voxels with relevant areas even if they are not related with the function of interest [17] [25].

In [17] authors pointed out the importance of considering the raise of a deficit as the result of a group of lesioned neurons or areas evaluated as a whole instead of associating a functional role to single voxels. Explaining this concept using their example, the difference of the two methods can be associated with the will to identify the presence of a city in a country measuring the numbers of cars every 100 yards, which is less precise than considering the cars in their context (in a city the cars are closer to each other).

Multivariate methods should be more able to determine functionally related areas, made up of various voxels that are important to define a neurological deficit, even if the selected voxels are in two distant parts of the brain and following this idea it can be considered not only that a specific function is processed by many areas, but also that their activation is governed by different combination rules.

In the same study the researchers demonstrate that a multivariate approach based on sparse canonical correlation analysis can overcome the results of VLSM regulated with different methods to achieve multiple comparison correction, even for different sample sizes (number of subjects varying from 20 to 131). Analogous results are reached when the model is used to study real scores of aphasia.

Supporting this idea, many other studies tried to create models to predict specific scores, integrating or comparing information about focal damage with connectivity, for example in [24] a ridge regression model was implemented to predict scores about attention, visual memory, verbal memory, language, motor, and visual domains and comparing VLSM and connectivity inputs. It was found that visual memory and verbal memory are better predicted by connectivity measures, for attention and language both have the same effect and motor and visual impairment are more explained by lesion topography.

In [26] the outcome of a connectivity-based LSM has demonstrated the connection between regions that are engaged in specific language functions, that are not found applying VLSM.

Also in [27] a complex multivariate model based on measures of graph theory of structural and functional connectivity and lesions information is implemented using a random forest analysis to evaluate four aphasia scores. The results, that are compared with VLSM show that the

traditional method and its limitations can be substituted by this new, more complete approach. In fact, the disconnections caused by the lesions can lead to dysfunctions that are related to regions placed far from the damage that can be seen on structural imaging and causing *diaschisis*, the dysfunction that belong to apparently intact cortices, induced by remote but connected neuronal populations [32]. For example, in patients with post-stroke aphasia it is not rare that the impairment mismatch the location of the classic clinical-topographic correlations [28] and this lack of matching can be explained by focusing on the connectivity data.
Anyway, it is important to remember that the computation of this kind of inferences on the disconnection it is often possible thanks to the creation of atlases that define the parcels of the grey matter and map the tracts of the connectome that link all the different parts of the brain, merging the imaging data of a large population of subjects [29] [30]. In this way researchers can work on reliable anatomical templates that then can be integrated with the information about the lesions, according to the same brain coordinate space as the atlases, to measure the impact of the related focal damages and disconnections [23].

Lesions like strokes can interrupt the structural anatomical pathway of white matter fibres that connect different region and this can also induce to complex processes of modification of the functional reorganisation [31], so now that it is clear the relevance of the disconnection inference over the voxel-based method, the next chapter will focus on the most important types of connectivity and how they are related to the lesions.

1.4 The connectivity of the brain: insights into structural and functional relationships

As previously mentioned, the brain organisation based on networks of connections distributed spatially in a variable complex way is not only visible from an anatomical point of view, but also reflects the functional organisation of the brain [32].
In order to understand how the brain works, it is possible to separate the study of connectivity into two main categories: structural and functional. Distinguishing the two connectivity is possible to assess the brain behaviour from two different points of view, understanding the affinity and differences among them and their informative contribution.

Structural connectivity (SC), or also called anatomical connectivity, is made up of the physical connections formed by synapses between various neuron pools. It defines bundles and tracts of nerve fibres that constitute the connectome and form the white matter of the brain. These connections are spatially distributed, linking also distant regions and are typically stable in a short time scale (seconds or minutes) [33].
There are many methods to evaluate SC: using some metrics from diffusion weighted imaging (DWI) like fractional anisotropy or apparent diffusion coefficient, using the microstructural profile covariance from histological studies, inspecting covariance among metrics derived from anatomy (for example grey matter volume or cortical thickness from T1 sequences of MRI) or considering the concentration of neuronal tracers in the axons by retrograde or anterograde neuronal tracing studies [32].
Typically DWI is used to infer in vivo white matter's architecture. This information is derived from the diffusion of water molecules in the brain tissues and it is used to detect the white matter tracts. An important limitation is that this technique can't differentiate between axons of different directions or roles (excitatory or inhibitory), despite some other invasive methods.
Anyway, many tools have been developed to estimate the connection between regions using DWI data, creating tractography that reconstruct structural networks and calculating measures that are used to quantify the anatomical connectivity, like the probability of connection, fibres' numerosity or fibres' length, even if there is still no consensus on measures of connectivity [34].

Functional connectivity (FC) is based on the time-dependent activation of different patterns of neurons in various areas of the brain and for this reason it is derived from the BOLD signal of

resting-state or task-based functional MRI or from the direct signal of neural activation coming from EEG data, electrocorticography or MEG records [32]. Starting from these time series, FC can be derived by inspecting the statistical dependence of the activation of different parts of the brain. This dependence is calculated using simple statistical measures such as correlation, variance, phase locking or spectral coherence and is highly time-dependent, so it varies based on the time scales adopted [33].

Pearson correlation among the signals originate from the brain regions is one of the most popular methods used and so the result is a symmetric matrix constituted by values of the statistical result of the measure adopted for every possible region combination and "1" all along the diagonal because each region is fully correlated with itself (Fig. 1).

One limitation in FC analysis is that the brain can be represented by different parcellation atlases that are used to define the regions' boundaries which differentiate in numerosity of regions, locations and shapes [34]. Moreover, also the methods used to define the parcellation can be based more on functional or anatomical information [35] and all these differences can hinder the comparison of results.

The study of functional and structural connectivity can provide complementary insights into brain organisation. Studying structural and functional network, it is possible to understand different organisation of the brain, for example looking at the SC it can be seen an assortative behaviour, so regions with similar properties are more prone to be connected, instead of the FC, more disassortative so with better affinity among regions with dissimilar attributes [36].

Another point of discordance between SC and FC concerns the detection of indirect connection between regions, since it can be found a strong functional relation between two regions that don't share direct anatomical connection [37].

However in the same paper it has been shown also that the existence of direct or indirect anatomical connection generally implies the presence of strong FC (Fig. 2) and in fact many researches in literature have found that the anatomy of the brain determines, at least partially, its function, for example demonstrating that the same functional areas are subserved by similar structural connectivity patterns, or finding strong functional connectivity in regions with anatomical connection and there are also other studies that manage to obtain SC from the corresponding FC [38], demonstrating that SC and FC can be two complementary sources of information, and confirming a strong association between functional and structural levels.

Not only spatial dependence, but also temporal scale is important to remember describing SC-FC relationships. In fact, differently from the SC that is stable in the short period of time, FC can be very different forming various functional patterns that can be explored during spontaneous neural activity. The relationship results stronger in particular with low frequency sampling periods, in the order of minutes. At higher frequency FC becomes more unstable, reflecting the rich underlying dynamic [39].
For this reason, typically the time of acquisition for rs-fMRI is about 5-7 minutes, even if it has been shown also that the time range can go up to 13 minutes to increase reliability [40].

As described in the previous chapters, many research prove that their informative contribution can be so important to overcome some of the traditional methods of analysis of the relation among brain lesions and symptoms, for example in [41] it has been found out that FC modification in the brain network caused by stroke can be better explained by the structural disconnection rather than just focal damage measures.
Discovering the relationships that link these types of connectivity is one of the challenges of neuroscience, in particular following a brain injury, it is not only possible to understand how these are related, but also how the brain is able to recover from lost or reduced functionality.

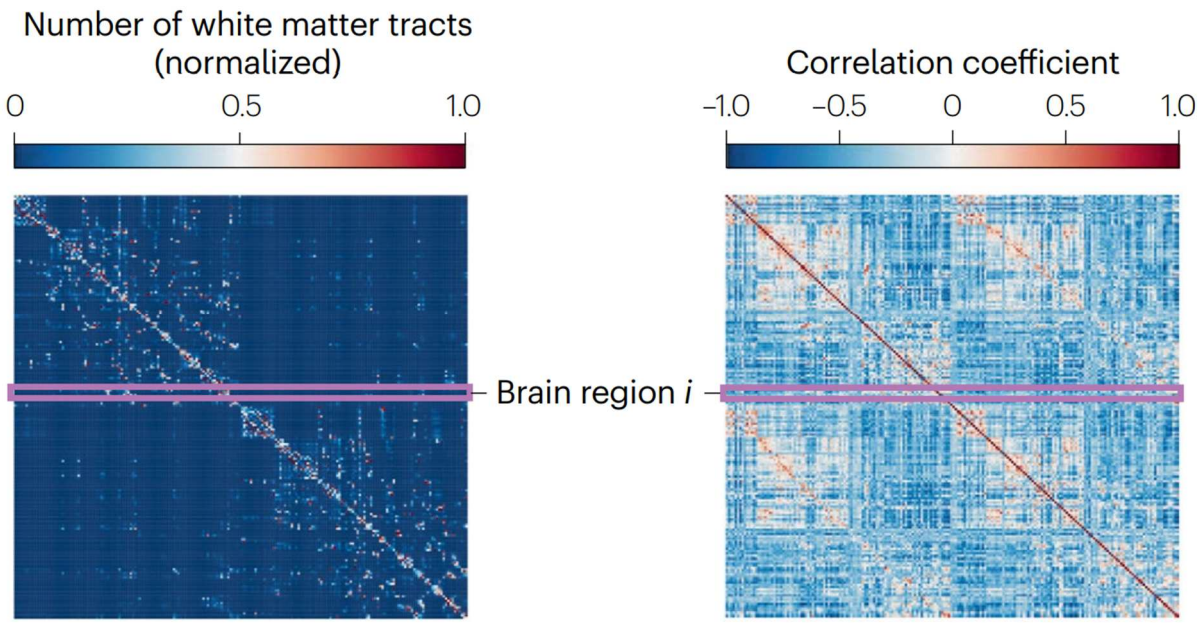

Fig. 1 – Two examples of SC (right) and FC (left) from [32], derived from DWI and fMRI. Rows and columns represent brain regions and cell's entries in the SC are the normalised number of white matter tracts, while entries in the FC's cells are values from a statistical correlation.

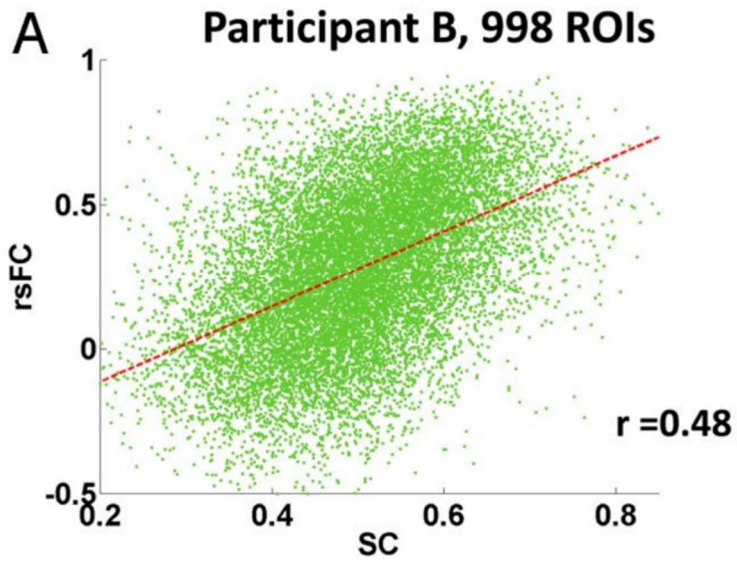

Fig. 2 - Scatter plot (single acquisition of 20 min) of resting state FC against SC at high resolution for participant B in [37], showing the correlation between the two connectivity.

1.5 Aim of the thesis

The aim of this study is to build a model based on T1-weighted structural MRI and rs-fMRI for the prediction of cognitive scores related to numerical and financial abilities, in first chronic stroke patients.

As it will be seen in the methods description, two different algorithms of data reduction are used: PCA that is a very common linear method and UMAP that is a relatively new method, non linear and less used. With this work it will be possible to compare the two algorithms and evaluate if UMAP can be a valid method in this topic of research.

Moreover, the results could show the importance of the connectivity based approach against the classic univariate voxel based one and help to shed light on the process that rules the loss of cognitive abilities that are still less often studied than motor impairment caused by stroke, in particular for financial abilities that are tightly related to the numerical domain, but less analysed and which researches are more focused on healthy subjects or neurodegenerative patients. These skills are fundamental nowadays and the loss of these abilities compromise the independence of the patients so the implementation of a model that could objectively predict financial and numerical skills could be an interesting tool to evaluate people's quality of life, guide treatments to enhance these specific capacities or to help in juridical scope.

2 MATERIALS

2.1 Patients

This thesis is made in cooperation with a project of the IRCCS San Camillo Hospital of Venice, where patients with different pathologies have been recruited. At the beginning 51 chronic stroke subjects were selected, with age between 18 and 85 years old, without other previous stroke events. Then 17 participants have been excluded because of the presence of bilateral lesions or in the cerebellum, due to the impossibility to evaluate the lesion or to assess the imaging data.
During the analysis, three of the remaining 34 patients have been removed because their lesions have invalidated the BOLD signal of too many regions (see also *Preliminary inspection* section of the results).

The main demographic information regarding the final selected cohort of 31 patients are reported in Tab. 1.

	Right	**Left**	**Total**
Numerosity	15	16	31
% Male	60	50	55
% Female	40	50	45
Mean Male Age	67.33 ± 9.50	65.38 ± 16.50	66.41 ± 16.50
Mean Female Age	70.5 ± 11.00	71.13 ± 15.00	70.86 ± 16.00

Tab.1 - Demographic patients' cohort information, divided in left and right lesioned subjects.

2.2 Cognitive tests

The analysis carried out in this thesis supports a project of the IRCCS San Camillo of Venice, in which new methodologies are studied for the rehabilitation of patients afflicted by different brain disorders. Part of the project regards the analysis of numerical and financial abilities of stroke patients and for this reason NADL Short and NADL-F Short tests are carried out and then evaluated to monitor patients' progress.

NADL Short

The Numerical Activities of Daily Living (NADL) has been created to inspect the degree of awareness of the numerical abilities of the patients.

Some specific deficits can impact in different ways patients' life, so NADL has been created with the purpose to be the basis for such an investigation.

The test is composed of two sections, the first part called "informal" is carried out to evaluate the numerical knowledge necessary in everyday life about various topics like Time, Measure, Transportation, Communication, Money and General Knowledge in which patients are asked to solve real world problems. The second part called "formal" is the one in which more theoretical and specific mathematical skills are considered. This section is composed of other four subtests: number comprehension, reading and writing arabic numerals, mental calculation and written calculation, where it is necessary to know the fundamental arithmetic rules and operation [11].

Because of the length of all the process (45 min), a shorter version was used in this study called NALD Short, reducing the formal part of the test to assess numerical abilities in about 15 minutes [42].

NADL-F Short

Similar to the previous test, NADL-Financial (NADL-F) investigates the ability to manage private finances according to self-interest in simulating real life situations.

It is easy to think that financial abilities impairment can lead to serious problems in persons' individual independence. Indeed, legally a person who can't manage his/her own finances is flanked by a financial guardian. Nevertheless the creation of a valuation method is very difficult because of the complexity of this topic. For instance some tests are based on the traditional neurophysiological principle, others want to highlight the importance of the evaluation of real-

life situations, others again focus on the subjects' decision-making abilities or are motivated by juridical purposes and so on.

The NADL-F test wants to be a clinical tool to assess a wide range of topics fundamental to the subject to be financially independent in his socio-cultural context, in particular simulating real life situations that can occur in a European environment.

Like the NADL test, also NADL-F requires a lot of time to be completed and for this reason has been used a shorter version that lasts about 15 minutes, but that retains the most important features of the original test. It will be called NADL-F Short [43].

It is composed of seven subsets: Counting currencies, Reading abilities, Item purchase, Percentage, Financial concepts, Bill payments, Financial judgements, with increasing difficulty [44]. In fact, the first four are then gathered into a single score called "Basic" and the remaining three represent the "Advanced" abilities.

To sum up, at the end of the cognitive assessment four scores are produced: Formal and Informal from the NADL Short test and Basic and Advanced from NADL-F Short. All of them are then normalised dividing by the maximum value that can be obtained from each of the four tests to have values in the range from zero to one.

	Mean - Right	**Std Dev - Right**	**Mean - Left**	**Std Dev - Left**
Basic	0.687	0.211	0.654	0.246
Advanced	0.620	0.278	0.500	0.197
Formal	0.753	0.148	0.685	0.230
Informal	0.707	0.111	0.707	0.144

Tab.2 - Mean and Standard Deviation of the four scores of the 31 patients considered in the analyses, grouped in left and right stroke lesioned subjects.

3 METHODS

3.1 Data acquisition

Structural and functional MRI data were collected before and after the rehabilitation at IRCCS San Camillo Hospital in Venice using a 3 T Achieva Philips scanner (Philips Medical Systems, Best, The Netherlands) with an 8-channel head coil. Participants' heads were immobilised accurately with head cushions.

The anatomical scan consisted of a 3-dimensional Magnetization Prepared T1 weighted (T1w) Rapid Acquisition Gradient Echo (MPRAGE) sequence acquired at 0.8 mm isotropic resolution (flip angle = 8°, repetition time (TR) = 9.8 ms, echo time (TE) = 4.5 ms, inversion time (TI) = 950 ms, field of view (FOV) = 250x250x200 mm^3, SENSE acceleration 2 and 2.6 along primary (Anterior-Posterior (AP)) and secondary (Right/ Left (RL)) phase encoding directions). Resting-state functional MRI scans with PA phase encoding direction were acquired using a single-shot Echoplanar Imaging (EPI) sequence (TR = 2.1 s, TE = 30 ms, flip angle = 90°, multiband factor = 3, SENSE factor = 1.2, and spatial resolution = 2.2x2.2x2.4 mm^3).

3.2 Structural analysis

3.2.1 Preprocessing and lesion tracing

Before processing MRI data, a preliminary step requires defining the lesion mask for each single patient.

The first segmentation of the lesions has been done exploiting LINDA (Lesion Identification with Neighborhood Data Analysis) [45], an automated brain lesion segmentation software that proceeds to the estimation of the lesion in T1w sequences using a hierarchical approach from low to high resolution, taking into account information of each voxel and the signal of neighbouring ones.

In fact, the training of the algorithm started from the low resolution images of 48 subjects in which some random forest (RF) models, taking into account the value of each voxel and its neighbours, compared them with the binarized lesion masks. After the training with a specific resolution, the model is applied to the same subjects to obtain some features that are passed to the next step where all the process is repeated with higher resolution of the images but including also the information about the features calculated on the lower resolution step. After all the training process is repeated at every resolution, the algorithm is ready to process a new subject, using the same hierarchical procedures of the training but with the already trained RF models. At the end, in the highest resolution step the features computed are converted in a discrete segmentation map (Fig. 4) [45].

One of the limitations of LINDA is that it manages to identify just lesions in the left hemisphere, so the right lesioned T1 images have been flipped before the segmentation.

In addition, since some consistent errors are made by LINDA during the segmentation process (for example some clusters of healthy tissue are identified as part of the lesion), small, noncontiguous clusters are removed, as suggested in [46]. Additionally, because one of the exclusion criteria for patients is the presence of lesions in the cerebellum, this region is subtracted to prevent the software from misclassifying voxels in that area.

The resulting lesion masks are then visually inspected by three researchers independently, and corrected manually using MRIcron software where needed, to ensure a better reliability of the segmentation process.

A second refinement step is then carried out, consisting in: the removal of small isolated clusters of voxels (< 1000 clusters), the filling of holes to ensure continuity of the lesion, smoothing of the lesion's edges, erosion of the boundaries and the application of the brain mask to avoid that the lesion exceed the cerebral margins.

The structural preprocessing pipeline then included bias field correction (N4BiasFieldCorrection [68]), skull-stripping [69] and nonlinear diffeomorphic registration [70] to the standard symmetric MNI space [71] through cost-function masking implemented in the Lesymap software. The Computed Anatomy Toolbox 12 (CAT 12) was used for brain tissue segmentation into white matter, grey matter and cerebrospinal fluid after removing the portion of lesioned tissue (stroke lesion correction option).

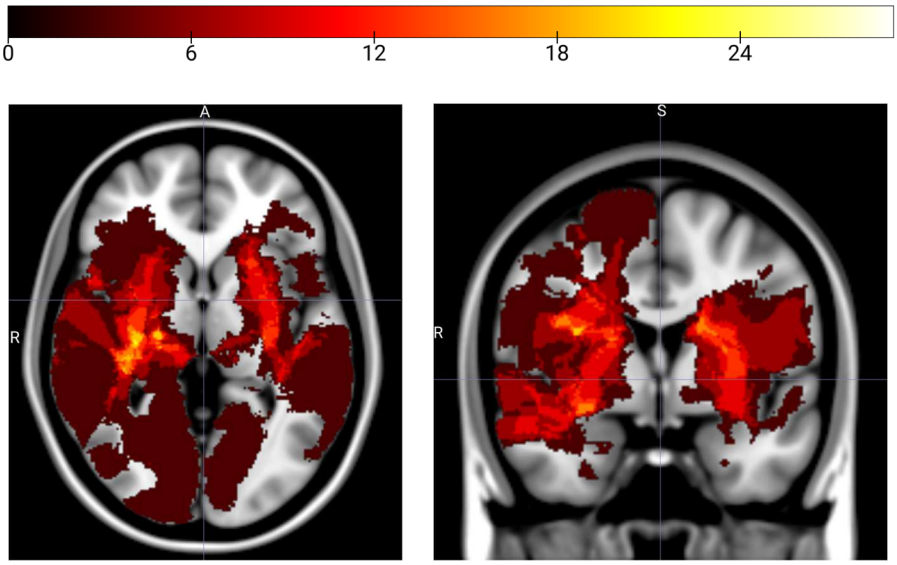

Fig. 3 - Axial and coronal sections of the lesions' frequency map of the cohort of patients. As indicated by the colorbar, brighter voxels are the most frequently affected (max 29.03%) and mostly on the right side where lesions are typically wider.

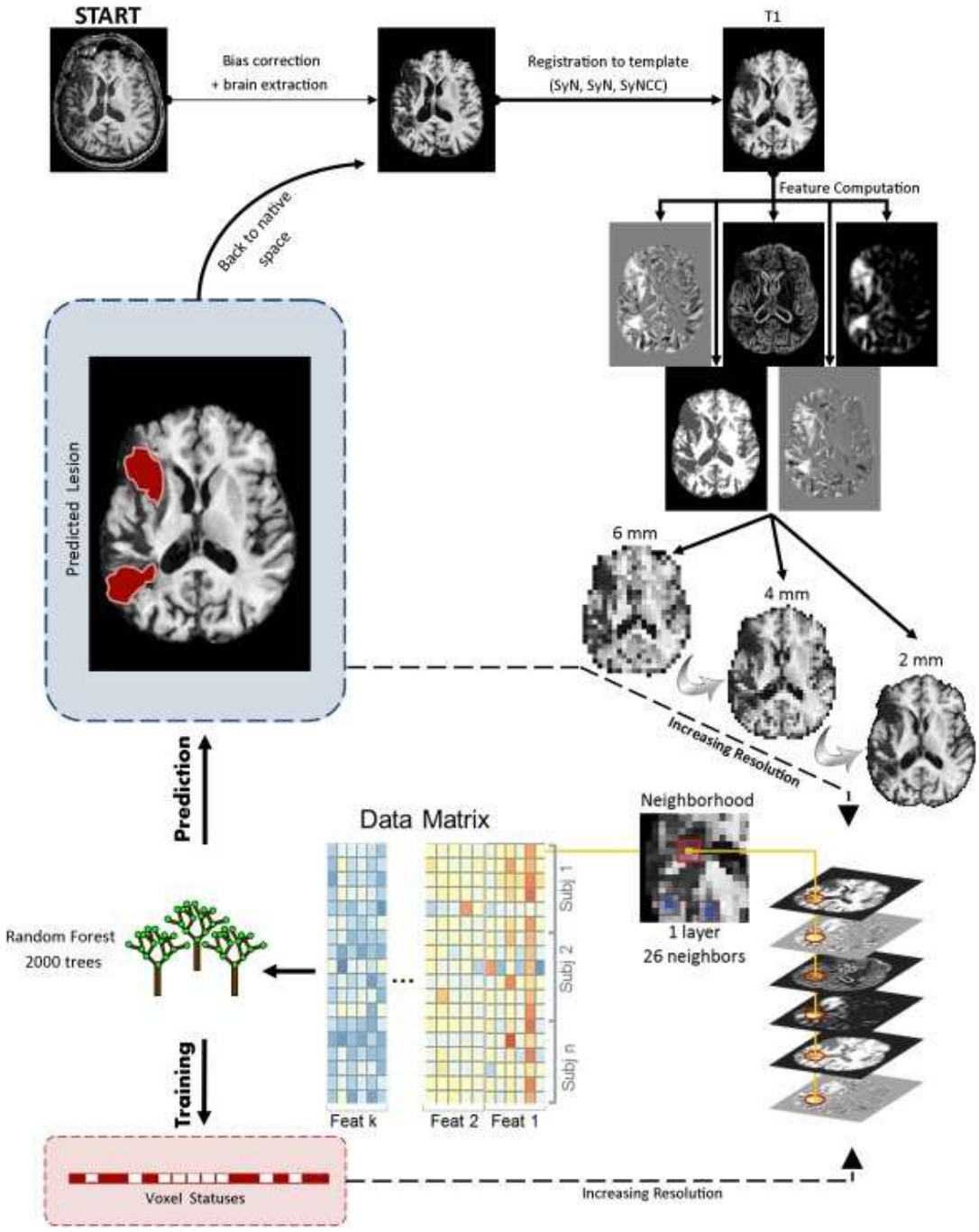

Fig. 4 - LINDA workflow. In the lower part is shown the multi-resolution Random Forest algorithm [45].

3.2.2 Analysis of the structural alteration: Structura Disconnectivity definition

To achieve information about the structural disconnection caused by the lesions, for each patient two matrices have been computed to inspect direct and indirect disconnection.
This was done using the Lesion Quantification Toolkit (LQT) [35], a MATLAB software package specifically designed to infer the impact of a lesion on grey and white matter.
The measures produced by the toolkit are based on a population-scale atlases of white matter tractography used to map the major tracts that constitute the connectome and so to estimate the disconnections. LQT uses the HCP-842 tractography atlas implemented in [49] with diffusion MRI data of 842 healthy subjects, managing to define 70 macroscale white matter tracts that are then also reconstructed in the MNI space.
Common tools used in neuroimaging research are the parcellations atlases that divide the grey matter in functionally or anatomically defined regions. Depending on the atlas used the numerosity and shape of the parcellation can be very different.
The toolkit needs the binary lesion segmentation as input, registered to the MNI coordinate space and the parcellation atlas desired, also registered to the MNI brain template and with the same image dimension of the lesion mask.
For this study, the lesion masks were registered to MNI152 space (dimensions 182x218x182; $1mm^3$ voxels) and the parcellation template chosen in the one developed by Schaefer and colleagues [47] using high quality rs-fMRI of 1489 healthy subjects, with 100 cortical parcels from 7 networks in Fig. 5, to which 12 subcortical regions from the third version of Automated Anatomical Labelling (AAL3) atlas [48] are then added. They are Thalamus, Caudate, Putamen, Pallidum, Hippocampus and Cerebellum for each of the two hemispheres.

Once that inputs are defined, LQT output different single-subject results:

- The “region-based damage”, which quantifies the percent of voxels in each grey matter parcel, overlapping the atlas with the segmented mask. Thanks to the fact that the chosen parcellation is based on the functional properties of the regions, the entity of the damage can be easily contextualised in the following analysis in terms of resting state networks. The output is a single vector of 112 values corresponding to the related patient, so then a matrix with all the subjects as rows can be easily created (Fig. 7).

- The “tract-based disconnection”, which is calculated by embedding the lesion in the tractography and measuring the number of streamlines that are intersected for each

tract. These numerosities are then converted in percentages of the total number of streamlines for each tract. This measure wants to be a better estimation of degree of damage of disconnection, than the classic "tract lesion load" or using probability concepts, which may underestimate the effect of the lesion as explained in [35]. Again the result is a vector of 70 percentage values, one for each tract.

- The "parcel-wise disconnection severities" measure offers another important tool that will be used in this thesis. The information obtained from this section is organised into matrices in which each cell contains the severity of disconnections between pairs of grey matter parcels. More specifically, a SC matrix is created based on the HCP-482 tractography and the chosen parcellation, in which just the number of streamlines that bilaterally terminate within both parcels are considered, selecting the specific command *end*. Then the lesion mask is embedded and filtering the relative streamlines, a "raw parcel-wise disconnection matrix" is created, where each entry is the number of disconnected streamlines among parcel pairs and it can also be converted into a percentage of the total streamlines connecting that parcels creating an analogous percent matrix, so that the matrix's cells correspond to a degree of direct disconnection severity. Differently from the "tract-based disconnection" that gives a hint on the disconnection estimating the white matter's tracts damaged, this measure focuses on the parcel disconnection, estimating the number of streamlines in each tract that are compromised (Fig. 6).

- In order to also consider the indirect connectivity between parcels the toolkit gives the opportunity to measure the "shortest structural path lengths" (SSPL) that is the minimum number of direct connections that are necessary to link a pair of grey matter parcels. This matrix has entry's values equal to "1" for regions with direct connection and values equal to the number of the required links to complete the indirect connection. To compute the matrix a binarized SC matrix is created, to define the presence or absence of connection among parcel pairs and also the percent spared connection matrix for each patient by subtracting the raw disconnection matrix previously created. Then a threshold has to be set to binarize it considering the minimum percentage of streamlines that have to be spared to consider a connection between two parcels already functionally viable. This threshold has been set to the default 50% value, which means

that at least half of the streamlines connecting a parcel pair have to be spared in order to consider it in the SSPL computation.

Then, a SSPL increase matrix is created by subtracting the atlas SSPL matrix in order to consider the presence of the lesion that leads to the interruption of some direct connections and so to the increase of the SSPL values. Finally, an indirect-only SSPL increase matrix is computed simply setting every direct connection to 0 [35] (Fig. 6).

In summary, using LQT, the information about the structural disconnectivity (SDC) were extracted from the patient's imaging data, in particular the percentage of parcel damage, the disconnection matrices and the indirect SSPL matrices were considered in the following analysis.

These last two matrices (direct disconnection and indirect SSPL) are vectorized in order to create a single matrix containing all the patients, without considering the diagonal and the lower triangular portion of the matrices thanks to their symmetric property. The resulting matrix has rows that correspond to patients and columns that coincide with all the regions and it is checked to ensure the absence of Nan and Inf values or identical rows.

The information of the percentage matrix of parcel damage is used as additional input to the canonical analysis that will be better explained further in this chapter.

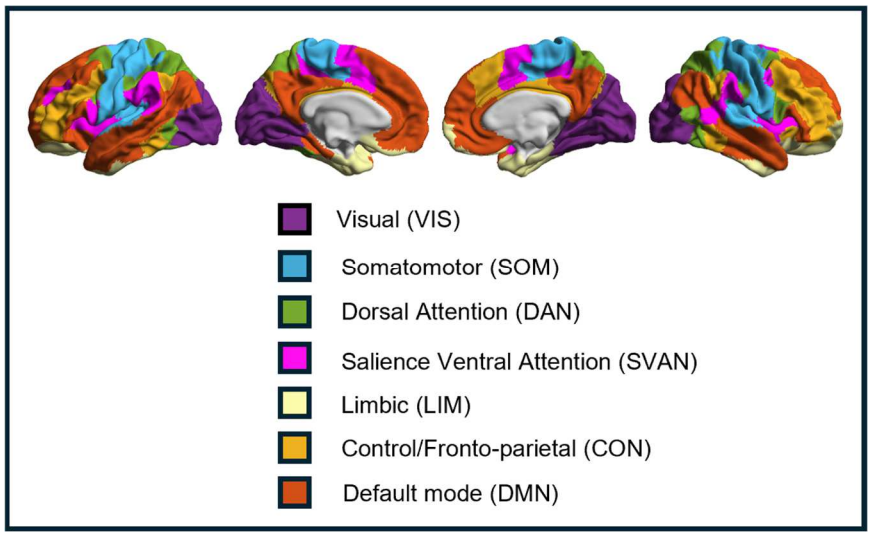

Fig. 5 - The 7 major networks from Schaefer atlas [4].

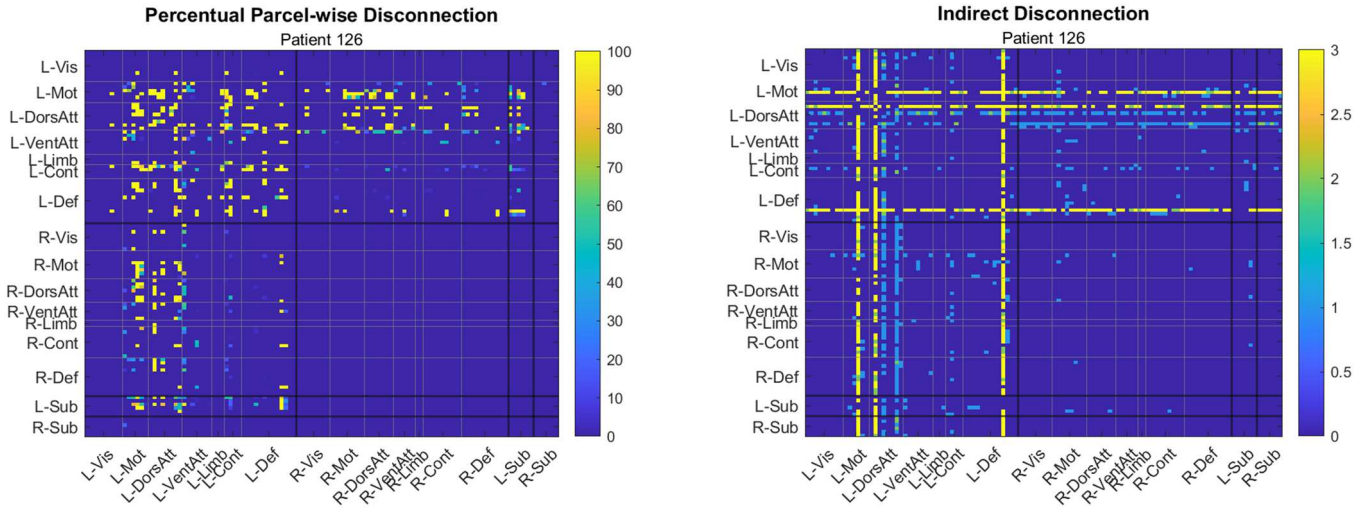

Fig. 6 - Percentual parcel-wise direct disconnection (left image) and indirect disconnection (right image) of the patient 126. The black lines indicate the division between left (0-50), right (51-100) and subcortical (101-112) parcels.

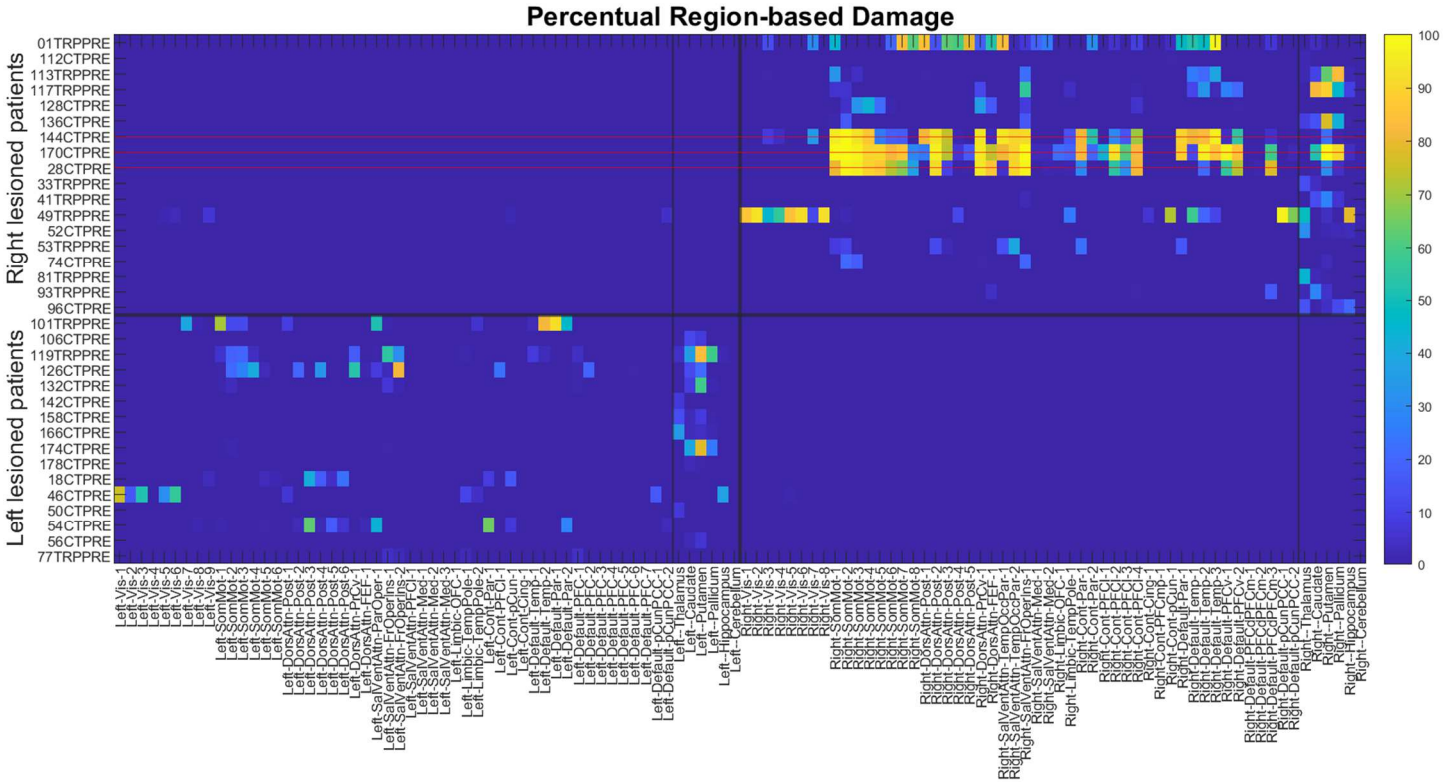

Fig. 7 - Percentual region-based damage matrix for 34 patients. The lesions of three patients have been evaluated too wide (red lines), compromising the BOLD time series, so they will be removed from the analysis (see also *Preliminary inspections* in the results section).

3.3 Functional analysis

3.3.1 Preprocessing

Concerning functional preprocessing, fMRI volumes were first corrected for slice timing disparities using the method described by Smith et al. [51]. Realignment to the median volume was then performed with FSL's MCFLIRT tool [52], followed by correction for magnetic field distortions using FSL's TOPUP [53]. To further reduce confounding influences, nuisance regression was applied using the CONN toolbox [54]. This included the removal of the following regressors: five principal components extracted from white matter and cerebrospinal fluid [55] after linear registration to the EPI space, subject motion parameters (three translation and three rotation parameters along with their first-order derivatives), and a variable number of additional noise components corresponding to outlier scans identified with ART tool included in the toolbox. In these steps, the lesion's mask was excluded during time series extraction to ensure that only BOLD signal coming from healthy voxels was analysed. High-pass filtering was subsequently applied through CONN, post-regression, to avoid frequency mismatches in the nuisance regression process [56].

Finally, low-pass filtering (cut-off = 0.1 Hz) was applied to the time series to focus on slow-frequency fluctuations while minimising the influence of residual physiological noise, head motion, and other artefacts. Volume censoring was performed to discard volumes affected by significant head motion (framewise displacement greater than 0.4 mm). After performing nonlinear normalisation to EPI space, the time series were projected onto the cortical surface using Schaefer's atlas [47]. Time series were extracted for each of the 112 brain regions in the parcellation by averaging the preprocessed fMRI BOLD signals within each node at each time point, excluding lesioned and non-gray matter voxels, as well as those affected by BOLD signal dropout.

3.3.2 Functional Connectivity definition

As described in the introduction, there are many methods to define the FC of a patient. In this work a similar approach to the seed-based method is used, but in a different way. As it has been explained in [6], a seed-based FC is created defining the lesion as seed, and correlating it with all the other regions of the brain.

A limitation of this method can be seen considering that lesions can affect both white and grey matter, leading to a mixed signal from tissues with different properties and characteristics. This blending of signals can complicate interpretation, as the method does not distinguish between the different types of tissue involved. Additionally, it is important to differentiate between lesion types, as they can consist of necrotic tissue (as in ischemic strokes), blood (as in haemorrhagic strokes) or edema. Each of these introduces distinct properties to the acquired signal, further complicating the accuracy and clarity of the results.

For these reasons, as it has been made in [58][59], the FCs are made by computing the correlation between every parcel pair and excluding the lesion from the analysis, so after removing the noisy volumes, the BOLD signals of all the parcels are used to create the FC matrix of each patient, with MATLAB *corr* function. Using the default setting of the function, Pearson correlation is obtained among every pair of parcels, as shown in Fig. 8.

As for SDC matrices, a matrix with the vectorised FCs of all the patients is created, considering only the upper triangular portion and removing the diagonal.

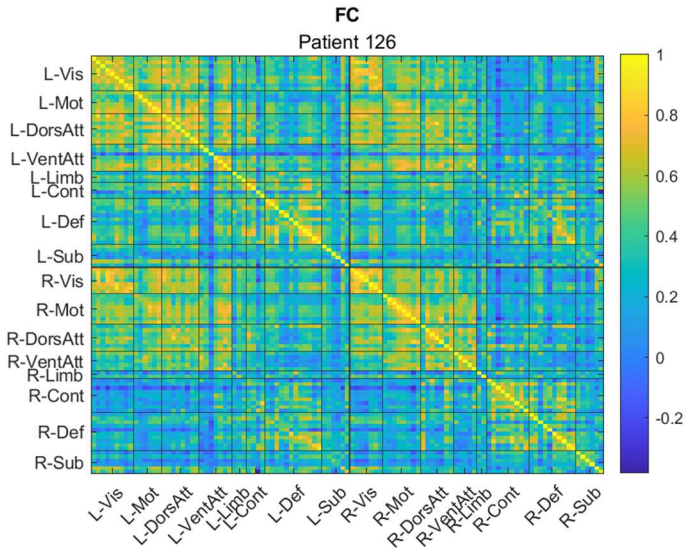

Fig. 8 – FC matrix of Pearson correlation between 112 brain regions. The major diagonal is made up of all "1" representing the correlation among a region and itself. Two secondary diagonals are created by the strong relation between homotopic regions that share the same function in the two hemispheres (e.g. left motor area and right motor area).

3.4 Data analysis

In this section it will be presented the method used to create the prediction model. In broad terms, the procedure is composed of a part of data reduction, using two different methods: Principal Component Analysis (PCA) and Uniform Manifold Approximation and Projection (UMAP), that will be then explained. This part is fundamental because from the analysis of the images many outcomes can be provided, as it has been seen in the LQT paragraph and each of these outcomes is composed of hundreds of elements. When all the needed information is collected, a unique matrix made of 31 rows, corresponding to patients and more than 18.000 columns of imaging data is made, so it is necessary a data reduction process, but in order to use the two algorithms some parameters have to be setted like the number of PCs and UMAP's neighbours and output dimensions.

To find the best values for these parameters and reduce the computational cost of the model, a simple classification algorithm is used to try to categorise the patients with good performance on each cognitive score from the bad ones. The accuracy of the classification is evaluated for every parameter, so that the parameters that lead to bad accuracy can be then removed. Then the Canonical Correlation Analysis (CCA), implemented with embedded permutations tests from the CCA/PLS Toolkit [61], is used to find the relationships between imaging data and cognitive score from NADL Short and NADL-F Short tests, reducing again the computational costs and selecting the best models. In the next section the two data reduction techniques used in this thesis will be presented in more detail.

All the following analysis has been done with MATLAB software (R2023b Update 5).

3.4.1 Data dimensionality reduction techniques

Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is one of the most common methods used for dimensionality reduction and features extraction, where data points are projected in a new set of dimensions called principal components (PCs) maximising their variance.

Letting x_n be the starting dataset of *n* variables or observations, it can be projected into a scalar y_n thank to the vector *u* as

$$y_n = u^T x_n \, .$$

The variance of the new data can be calculated as

$$\frac{1}{N}\sum_{n=1}^{N} \left\{u^T x_n - u^T \underline{x}\right\} = u^T \cdot S \cdot u$$

that can be maximised considering to set the constraint $u^T u = 1$ whose solution can be achieved with the Lagrange multiplier method. It can be found out that the variance of the new projected data is maximised when *u* is the eigenvector that maximise the related eigenvalue in the relation:

$$S \cdot u = \lambda \, u \, .$$

The starting data x_n can be now generally represented as

$$X = U * L^T + \underline{X}$$

where *U* are the eigenvectors or components or scores, *L* are the loading (the coefficients of the linear combination that define the components) and $\underline{X}$ is the mean values matrix.

Principal components are uncorrelated (orthogonal) to each other and capture the variance in the data, with the first component accounting for the maximum possible variance. Each subsequent component explains the largest portion of the remaining variance, with each one being orthogonal to all previous components.

Uniform Manifold Approximation and Projection (UMAP)

UMAP (Uniform Manifold Approximation and Projection) [50] is a novel manifold learning algorithm for dimension reduction. It competes with t-SNE algorithm, overcoming it for visualisation quality, the preservation of the global structure of data, reduced time consumption performance and the possibility to manage data with wide dimensionality. Thanks to its great scalability, it is usually applied for bioinformatic, material science and general machine learning fields, but as explained in [66] it can be applied also in the study of brain connectivity with the aim to find intrinsic low dimensional and non-linear surfaces (manifolds) where the data lies, preserving important geometric relationships among data points.

While UMAP is grounded in a rigorous mathematical and theoretical framework, for the purposes of this thesis, we will focus on a more general and practical explanation.

UMAP is based on the assumption that the data lie on a locally connected manifold where they are uniformly distributed and that the manifold's topological structure should be preserved. Even if it is based on topological data analysis and simplicial complexes, UMAP is considered an algorithm of the same class of k-neighbour graph learning type because it can be described in two phases: in the first one a weighted k-neighbour graph is computed and then a low dimensional layout of the created graph is calculated [50][63].

Phase I: Graph Construction

Initially the weighted graph has to be defined and it is done basing it on the topological analysis and the concept of “simplices”, which are simple ways to build a k dimensional object depending on a selected number k of vertices (see Fig. 9). Each data point given to the algorithm becomes a vertex of the simplices built on them. The ensemble of all the simplices creates a “simplicial complex” that can be used as a graph made up of data points as the vertices of the complex. One first problem to face is the choice of the right number of vertex/data points needed to create the single simplices. In fact, varying the number of vertices, the complexity of the simplex changes, with the risk of the creation of high dimensional simplices and graphs. This problem wouldn’t exist if the data were uniformly distributed in the manifold, because a fixed distance could be set from each data point and would be considered just the neighbour points not beyond this threshold distance, limiting the number of vertices and edges of the simplices. Since the density varies, to overcome this issue the definition of distance is changed and it is defined based on the *k*-nearest neighbour settled by the user. In this way a local distance function is achieved for each data point, maintaining the concept of uniformity of the manifold.

Moreover the *k* value assumes an important meaning, easy to interpret: small *k* means focusing on the finer details of the structure of the data, whereas a large *k* means capturing the global structure of the dataset, but losing the details.

In more mathematical terms, assuming $X = \{x_1, \ldots x_n\}$ as the input dataset, a dissimilarity matrix is created, with the user-selected distance function *d*, which is then transformed into a binary adjacency matrix **A** by the *k*-nearest neighbour algorithm, choosing the appropriate value of *k*. **A** defines a directional graph $\bar{G} = (V, E, w)$ where V are the vertices or data sample, E are the edges connecting each vertex and *w* are the weights associated to the edges. In fact, now it is also possible to associate a weight on the edge of the graph, based on how far the data points are.

Again, in a more mathematical form, the weights are given by:

$$w(x_i, x_j) = exp\left(\frac{-max(0, d(x_i, x_j) - \rho_i)}{\sigma_i}\right)$$

where x_j refers to the *j*-th nearest neighbour, with $j = \{1..k\}$, of the data point x_i, $d(x_i, x_j)$ is the dissimilarity value with distance function *d* and ρ_i and σ_i are normalisation values related to the specific x_i calculated as

$$\rho_i = min\{d(x_i, x_j) \mid 1 \leq j \leq k, d(x_i, x_j) > 0\}$$

$$\sum_{j=1}^{k} exp\left(\frac{-max(0, d(x_i, x_j) - \rho_i)}{\sigma_i}\right) = log_2(k)$$

The original dissimilarity values among the neighbours are now normalised by exponential curves in the range [0,1].

But since the distance function *d* changes locally for each data point, two different weights are calculated, indeed it can be thought that two vertices are now connected by two edges with different direction and weights. To solve this problem, is computed a symmetrized version of the weighted adjacency matrix **A** as follow:

$$B = A + A^T - A \circ A^T$$

where $\circ$ represent the pointwise product.

If $\mathbf{A}_{ij}$ represents the probability of existence of the edge directed from x_i to x_j, $\mathbf{B}_{ij}$ will represent the probability of existence of at least one of the double directional edges built from x_i to x_j or from x_j to x_i. In this way a new symmetric graph G can be considered, with non-directional weighted edges defined by the adjacency matrix **B** and this will be the basis of the UMAP method [50][63][66] (Fig. 10) .

Phase II: Graph Layout

Now it has to be find a good low dimensional representation of the dataset. To do it, UMAP creates a set of attractive and repulsive forces, based on the edges' weights, that are used to locate the data in the new space with the user-selected number of final dimensions.

To optimise the representation of the dataset, the algorithm considers the value of cross entropy, calculated considering the edges of the graph G and the ones of an equivalent weighted graph H that belongs to the low dimensional space to reach. In a more mathematical statement, it can be written $w_h(e)$ as the weight associated to the edge e of a high dimensional representation and $w_l(e)$ as the weight associated to the edge e in a low dimensional case. The cross entropy can be measured as:

$$\sum^{e} = w_h(e)\, log\left(\frac{w_h(e)}{w_l(e)}\right) + (1 - w_h(e))\, log\left(\frac{1 - w_h(e)}{1 - w_l(e)}\right)$$

Where the first addend can be seen as an attractive force on the data points that will be minimised when $w_h(e)$ is high (so when the distance among the points is small) and the second addend is a repulsive force that will be stronger as $w_h(e)$ is high (so when $w_h(e)$ is small and the distances among data point is high, it will be minimised).

Optimising the total cross entropy calculated on all the edges of the lower and higher dimensional graph, the optimal representation of the dataset is reached [50][63][66].

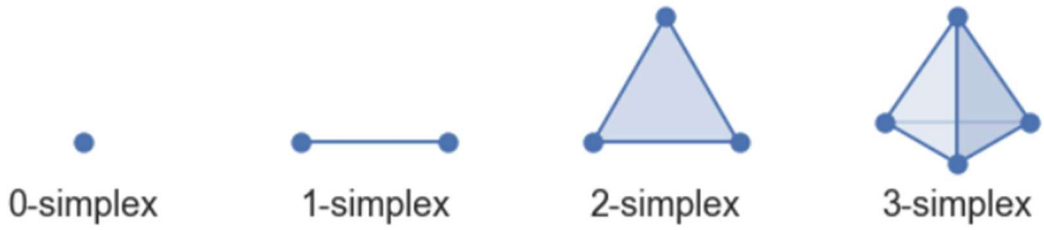

Fig. 9 - Examples of low dimensional simplices [63]. More are the vertices considered and more they become complex.

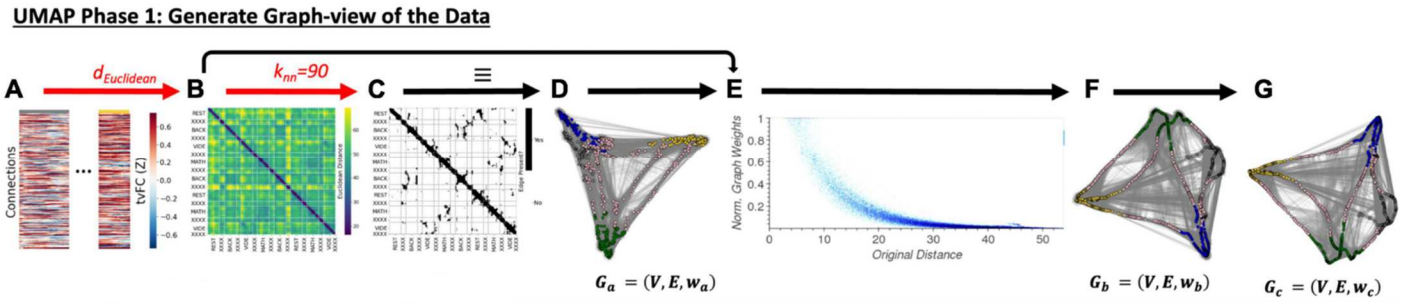

Fig. 10 - Example image from [66] that represents how UMAP reduces the data of time-varying FC (tvFC) (A). The dissimilarity matrix (B) is created with Euclidean distance and it is binarized in the adjacency matrix with number of neighbours k = 90 (C) in order to create the directional graph in (D). The exponential transformation of the weights is performed (E) and from the resulting graph (F) is calculated the related unidirectional graph (G) that is the input of the optimization phase.

3.4.2 K-Nearest Neighbours classifier

A first evaluation of the effectiveness of the PCA and UMAP algorithms is made by inserting their output scores in a nearest neighbours (KNN) classification model and checking the efficiency of the methods. In fact their output can vary based on the choice of some parameters. While the parameter to choose for the PCA methods is just the number of PCs, for UMAP there are many input options to consider. First of all, UMAP doesn't have restrictions on the number of output dimensions, starting from 2, for the visualisation purpose of the algorithm. Then, as it has been explained in the previous section, UMAP required the setting of a *k* number of neighbours that has an notable impact on the final output because large value of *k* means getting the global structure of the dataset, while for small *k* UMAP will focus on the detailed part of the dataset. Other important parameters for UMAP are the *min_dist* value that controls how tightly the algorithm is allowed to group the points together. In this study, this parameter is setted to 0.1, following the work of other researchers [57] and confirmed by other evaluations made during the continuation of the analysis. Other parameters have been kept on their default values.
The purpose of the KNN implementation is to get a better understanding of the algorithms and the impact of their parameters, limiting their range of values and reducing the computational costs of the final models of the next canonical analysis.

To proceed to the classification, the cognitive scores are binarized choosing the median of each of the four scores as threshold: the KNN classifier has to identify if a subject has good or bad performance for each score, starting from the input PCA or UMAP information scores.
The matrix made concatenating the SDC and FC is z-scored and given to PCA and UMAP algorithms to compute the data reduction, using the *pca* and *run_umap* MATLAB functions.
The classification is made on a 4-fold partition of the full dataset (*cvpartition* function is used, with *stratify* option setted as true), employing one fold as the test set and the remaining ones as training set.
The inspection of the PCs is made on all the possible components (30 PCs, since the subjects are 31) and for consistency the same number of output scores is chosen also for UMAP. The number of neighbours inspected span from the minimum possible (3 neighbours) till 31 (the total number of patients) and all the analyses were iterated 30 times (with the seed of the random folding that changes through the iterations using *threefry* method). The KNN model is

processed by *fitcknn* function, with the training folds as inputs and number of neighbours setted as 4. This value was chosen as the square root of the numerosity of the training set, as explained in [73]. The balanced accuracy of the classification results with PCA and UMAP is calculated as:

$$A = 0.5 \cdot \left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP} \right)$$

where TR are the true positive values, TN the true negative values, FP the false positive values and FN the false negative values.
Since the UMAP results are repeated 30 times, to assess the variability of the results depending on the random seed initialization, the accuracy values for the four cognitive scores are obtained averaging on the iterations.
For a more complete inspection, the overall numerical and financial accuracies have been obtained averaging the Formal and Informal scores of the NADL Short test and Basic and Advanced scores from NADL-F Short test, for both PCA and UMAP methods.

When the cognitive scores are binarized a frequent problem is that the majority of them become either 0 or 1, meaning that almost every patient has a bad or good performance on the test made for a specific score calculation and this can affect the efficiency of the classification. The balanced formula used and the stratification option of the *cvpartition* function are chosen to manage this problem.

3.4.3 Canonical Correlation Analysis (CCA)

The Canonical Correlation Analysis (CCA) [60] is a widely used method to compare cognitive behaviour and informative variables derived from the brain.

Considering **X** and **Y** two multivariate modalities, having observation per row and normalised features/variables per column, each modality can be represented as a linear combination of its own variable in two latent variables, also called "canonical variables" **U** and **V**, so that **U = X** $\mathbf{w_x}$ and **V = Y** $\mathbf{w_y}$.

CCA aims to find the pair of canonical weights or coefficients $\mathbf{w_x}$ and $\mathbf{w_y}$ that maximises the correlation between **U** and **V** as:

$$max_{w_x, w_y}\, corr\left(X\, w_x\, , Y\, w_y\right)$$

This method is commonly used in literature thanks to its ability to find linear association between multiple measures, in particular in the field of neuroimaging [62].

It can be easily noticed that the value of each canonical coefficient gives some information about the importance of a specific variable in **X** or **Y**, indeed it can be thought that the canonical variables **U** and **V** quantify the association across the two modalities, also referred to as associative effect. Solving the CCA iteratively, once the two weights $\mathbf{w_x}$ and $\mathbf{w_y}$ are obtained (representing the first associative effect), another pair of weights that represent a second layer of associative effect can be calculated, removing the information about the first effect through a process called deflation [64]. The new weights will represent the new associations among the variables.

In the present thesis, it is implemented CCA through the CCA/PLS Toolkit [61], which incorporates multiple multivariate latent variable models and it allows to perform a permutations test on the input data, which is also used to inspect various layers of associative effects.

By randomly shuffling $j = \{1..J\}$ times the row of **X** or **Y**, a new set of canonical correlations r_j between canonical variable and new statistics λ_j are computed. The p-value is calculated as:

$$p = \frac{1}{J} \sum_{j=1}^{J} \quad I\,[\lambda_1 \geq \lambda_j]\,,$$

where I is the Kronecker function, λ_1 is the statistic of the original unpermuted data and the null hypothesis Hp^0 of the permutation test is that the two populations come from the same distribution and so it should be rejected.

Regarding this thesis, it can be considered **X** as the multivariate matrix constituted by the output scores coming from the data reduction step, that contain the SDC and FC information, concatenated with some additional variables which can confound the prediction that are the patients' age, years of schooling and lesion volume and the information coming from the region-based damage. Instead, in Y there are the cognitive scores, so Formal and Informal scores taken together to inspect the numerical abilities from NADL Short test or Basic and Advanced scores from the NADL-F Short test, to evaluate the financial skills.

The region-based damage is a vector output from LQT and quantifies the damage of each of the 112 parcels of the brain of the specific patient analysed, so a cumulative matrix for all the patients has to be done. The resulting matrix has 31 rows and 112 columns, so a dimensionality reduction is needed again in order to try to keep the number of variables in input in the CCA less than the number of observations. To do this the first idea was to summarise the parcel region-based damage in a network-based damage matrix made of 8 columns averaging the parcel loads of each of the 7 cortical networks, considering the 6 subcortical regions as a whole network and removing the controlesional networks of each patient. In this way a good reduction of the original matrix was performed, but this was not enough to be considered as good input in the CCA. In fact, the canonical analysis is sensible to the multicollinearity problem: the canonical weights $\mathbf{w_x}$ and $\mathbf{w_y}$ become unstable when the input variables are correlated [64]. Since the network-based matrix showed high Pearson correlation values (Fig. 11), it is chosen to apply the PCA on the original region-based matrix, because the algorithm ensures the PCs to be uncorrelated. At the end 8 PCs are taken, managing to explain the 84% of the variance, creating a final new [31x8] matrix that is chosen as input for the CCA.

Since the ranges of parameters (PCs for PCA and output dimensions and neighbours for UMAP) still are too wide after the KNN process, the CCA model has been implemented in two phases: a first descriptive step, where the statistical power of CCA for all the possible parameters is inspected and a second predictive step where it is checked if the canonical variables associated to the selected parameters chosen during the first step are able to generalise the models in a cross-validation framework.
More in detail, in the descriptive phase the canonical variables are computed using all the 31 patients, setting the simple *cca* as chosen model of analysis implemented by the CCA/PLS Toolkit, with *permutation* framework option and 100 as number of permutations. No

normalisation is required because the Toolkit applies a z-score by default on the **X** and **Y** entries.

During this descriptive phase also the associative effects are inspected: setting the alpha threshold of the associative effect as 1 all the possible layers of associative effect are calculated, in this specific case there are two layers. So, for every parameter, the CCA is applied two times, calculating two pairs of canonical weights and two associated p-values, one for the first layer of associative effect and the other one for the second layer calculated after the deflation. The second layer always has higher p-values than the first one, so just the p-values related to the first associative effect have been considered in the following analysis. Using all the patients and inspecting all the parameters' ranges, the descriptive models become a sort of "best-case models" with which compare the results of the corresponding models coming from the predictive step. In this second step, instead, without changing the setting of the Toolkit, the generalizability of the models is assessed with a Leave One Out (LOO) method: one subject is removed from the dataset and the 30 remaining patients become the training set of the model. Once the canonical weights are computed on the training set, they are applied on the test set made of the removed subject calculating the corresponding canonical variables, then the process is repeated removing each time a different patient. At the end the two canonical variables of the descriptive step are compared with the corresponding two canonical variables coming from the predictive step: ideally it would be expected that they align along the identity line [65].

In this case the mean and standard deviation values of the training set is saved before the Toolkit application, in order to normalise the test set with the training statistical descriptive indices.

Different models have been inspected during the descriptive analysis, changing the dimensionality reduction method (PCA or UMAP), the score assessment (NADL or NADL-F) and inspecting the effect of the additional variables in the CCA (age, schooling, lesions' volume and parcel damage) removing them one at a time and using the parameters' ranges selected in the KNN analysis for the the descriptive step and the optimal ranges selected in the descriptive step for the sequent predictive step.

The parameters selection in the descriptive phase is done controlling the p-values coming from the permutation test. In fact the CCA is computed for each parameter and so the permutation test, too. The p-value of a certain permutation test describes how strong the null hypothesis is rejected and so a p-value close to 0 from a permutation test indicates that the original dataset significantly differs from the permuted ones, suggesting that under those specific conditions,

the model is more likely to effectively capture and utilise the input information. This means that when it becomes too high (alpha threshold setted at 0.1) the parameter associated with the computation of that specific p-value can be excluded.

With the same procedure can be also compared the p-values from different models, in particular removing one at a time the additional variables, it can be easily seen which one of them acts as a confounding variable raising the p-values.

Regarding the UMAP-based models, to reduce the random effects of the algorithm introduced by the seed initialization and considering the computational effort, the analysis has been repeated 5 times.

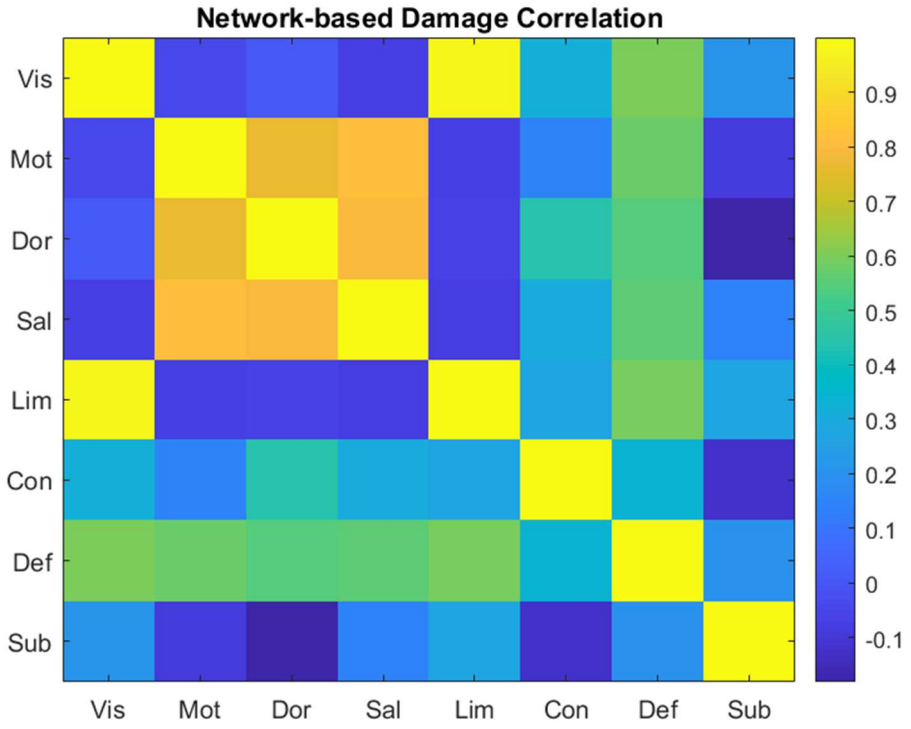

Fig. 11 - Pearson correlation on the columns of the first [31x8] damage matrix calculated averaging the parcels of each network (the subcortical regions are averaged together, called “Sub”). Some of the networks look highly correlated leading to multicollinearity problems.

3.4.4 Hierarchical clustering

Once that the optimal models have been selected, an interpretation of their explanatory power can be done through a clustering procedure: the canonical variables U and V of both the models (i.e., 4 input features in total) were used as input for a hierarchical clustering made by the *linkage* MATLAB function, with *average* and *cosine* as method and distance metric, respectively. In this procedure, the canonical variables of the third iteration have been chosen for the UMAP model because they showed the highest correlation with respect to the other 4 iterations.

The clusterization has been done changing the number of clusters between 2 and 10 clusters (*cluster* function with *MaxClust* approach) and the best clustering solution has been selected evaluating them with three different methods offered by the *evalcluster* function: Silhouette (to maximise), Calinski-Harabasz (to maximise) and Davies-Bouldin (to minimise) criteria. When the best number of clusters is chosen, CCA model interpretation can be carried out by observing the average pattern of features of the subjects within each cluster.

4 RESULTS

4.1 Preliminary inspections

Right and left lesioned subjects' performances

At the beginning of the analysis some speculations have been made in order to organise and evaluate the feasibility of the works. One important issue to solve was to define the presence of relation between the side of the lesion (right or left) with the cognitive impairment studied to decide if the patients' cohort has to be divided in two groups during the analysis.

A preliminary observation of the scores' distribution has been made by simply their histograms inspection that reveals that they don't have a normal shape (Fig. 12).

This could be caused by the intrinsic nature of the impairments studied, but it is also a common behaviour that recur in many studies where the number of patients is limited.

Following the work of another similar research [67], it has been decided to compute the Kruskal-Wallis Test, a non parametric test used to evaluate the distribution of two populations basing on their medians, instead of the ANOVA test that requires the normal distribution of the variables. The test is computed by *kruskalwallis* MATLAB function for every single score, previously z-scored, defining the right and left groups and testing the null hypothesis of similar distribution of the two groups.

Setting the threshold for p-value to 5%, the results show no significant difference between the left and right lesioned patients in three scores except for the Informal one (Formal: 0.488; Informal: 0.0236; Basic: 0.606; Advanced: 0.150), but after the correction for multiple comparisons made with Bonferroni-Holm's method all the scores accepted the null hypothesis with these corrected p-values: 0.975 (Basic), 0.451(Advanced), 0.975 (Formal), 0.095 (Informal). Considering these results and that the number of patients is very poor, the analyses that follow are computed joining left and right patients in a unique group.

Lesions evaluation and further patients exclusion

Another important evaluation has been made, that consists in the inspection of the number of lesioned voxels for each region. As shown in Fig. 7 and from the frequency map of all the lesions in Fig. 3, the right lesioned patients look more compromised than the left ones.

This raises the problem that if the majority of a parcel is overlapped to the lesion the signal that comes from the fMRI of that parcel has to be considered altered by the lesion too. For this reason, three right lesioned patients too widely damaged are excluded from every analysis while three regions (Left Default Parietal Network 1, Right Visual Network 8, Right Default Temporal Network 3) with a high percentage (all more than 92%) of lesioned voxels have been removed before the vectorisation of the FCs.

K-means inspection of the variability of UMAP scores

UMAP has been chosen for this research thanks to its non-linear nature and strong scalability, but it is also a new algorithm and some preliminary attempts have been necessary to understand how it works. The initialization of the algorithm can be random, but in practice a spectral layout algorithm is implemented to stabilise it, increasing the convergence to the results [7]. However, it has been noticed that its output can be very different among trials and it is important to set its parameters properly to obtain good results.

Since UMAP is comparable to a k-nearest neighbour algorithm, the main parameter is the number of k neighbours to consider. It is fundamentals because it changes the balance between local and global structure: if *k* is too small UMAP will concentrate on the local properties, potentially missing the big picture, while large *k* will push UMAP to focus on the relation between a huge number of data points, losing the fine details of the structure [63].

In order to understand how their variability can affect the UMAP's output, some preliminary tests are performed: the matrices of structural direct and indirect disconnections were z-scored and given as input of the algorithm, while iteratively changing the number of neighbours in the range between 3 (minimum value for *k*) and the number of patients analysed.

The resultant first two output scores are then clusterize with a *k*-means model (*kmeans* function, with a range of [2-10] possible number of clusters, whose optimum is chosen maximising the silhouette value). The k-means inspection is repeated multiple times observing how the outcomes can change on various trials, as it can be seen in Fig. 13.

From the same figure it is possible to understand that a good clustering is reached for the lower values of *k*, in particular 4 neighbours are considered to be a good compromise from the visual inspection of the results, confirmed by the choice made in [17], where about 11% of the input data is used as value for *k*. Given the demonstrated dependence of the embedding output depending on the number of neighbours, a further assessment is necessary to better understand the optimum number of neighbours, so the KNN classifier was lately chosen as a more robust method of analysis.

Another important parameter in UMAP is *min_dist* that is setted to 0.1, following the work of other researchers [17] and confirmed by other evaluations made during the continuation of the analysis (see *K-Nearest Neighbors results*).

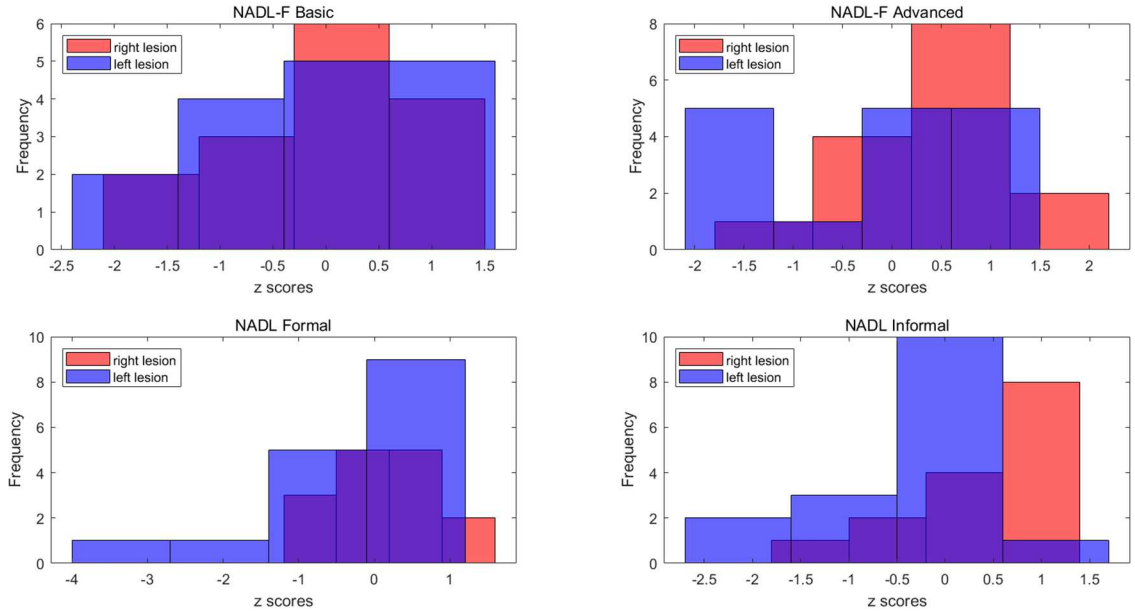

Fig. 12 - Histogram showing the distributions of the values of the 4 scores, after z-scoring.

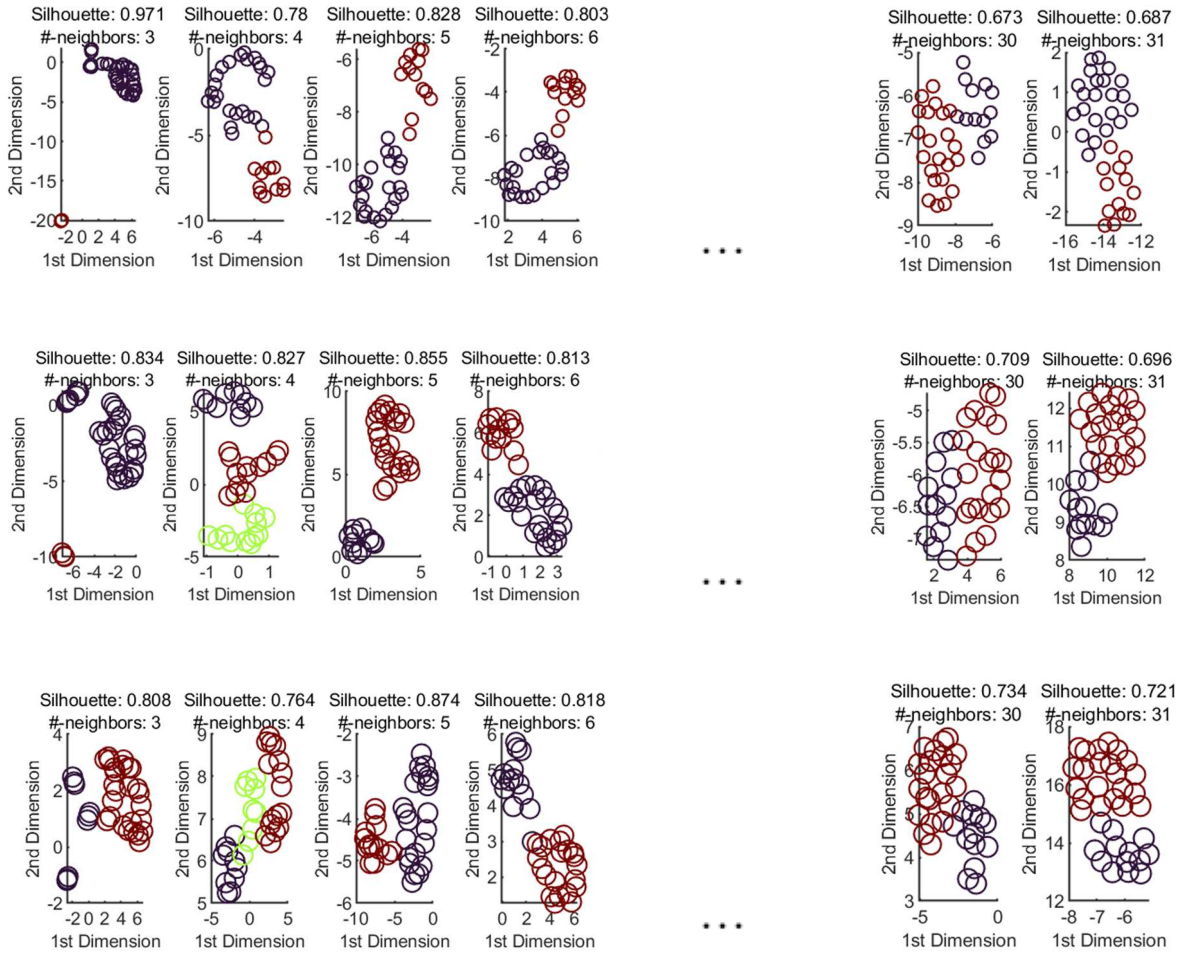

Fig. 13 - The two scores of UMAP's output are plotted during the "preliminary inspection", called respectively 1st Dimension and 2nd Dimension in the figures, with the colours that indicate the cluster in which every subject is located by the k-means model. In each row of the figure there are some examples of all the *k* neighbours checked, in three different trials: incrementing *k* the distribution of the data points is more sparse. The optimum number of clusters is chosen in the range [2-10] maximising their Silhouette values.

4.2 K-Nearest Neighbour results

The KNN analysis has been done to try to reduce the parameters' ranges of the two data reduction algorithms: PCA and UMAP.
The method is used to inspect the accuracy related to each parameter and this is done for each one of the four scores. The overall numerical and financial abilities are evaluated averaging the accuracy of the respective scores (so averaging the Formal and Informal scores' accuracies in a unique variable called NADL and averaging the Basic and Advanced scores' accuracies in a unique variable called NADL-F), in order to find a first selection of the optimal parameter ranges reducing the computational cost of the following analyses.

Concerning the PCA method the results are shown in Fig. 14: the maximum accuracy is reached with 8 PCs for NADL and 10 PCs for NADL-F. After these values the classifier seems to slowly decrease the performance till about 25 PCs, while looking at the single scores the slopes are very different from each other. For example just the NADL Formal score present high performance since the very beginning (the first PC show barely the same accuracy of the highest one calculated with 8 PCs), just the NADL-F Basic score has an evident single peak with very high accuracy and then proceeds with a flat trend, while all the other seems to have a local minimum peak around 25 PCs and then gain more accuracy in the latest PCs.
It is also hard to see similarity among scores that originate from the same test (Advanced and Basic from NADL Short or Formal and Informal from the NADL-F Short), probably due to the intrinsic differences of the assessments of the two types of numerical and financial abilities, as explained in the "Cognitive Tests" chapter.

Considering the exploratory aim of the KNN evaluation, which algorithm won't be actually included in the prediction model, wide ranges have been considered as valid, that are from 5 to 25 PCs, for the PCA-based models.

Regarding the UMAP data reduction inspection, there are two parameters that have to be considered: the number of output scores (also called dimensions) and the number of neighbours. The first ones have been evaluated in the range [2-30] for consistency with the number of PCs considered in the PCA model and also the second ones have been widely inspected, from 3 to 31, in order to inspect all the possible cases.

Looking at Fig. 15 can be seen that NADL and NADL-F are quite different in accuracy: NADL-F has roughly half of the accuracy of NADL. The number of neighbours lead to variable accuracy around the first half of the range, while increasing its values the efficiency of the KNN is more stable. Instead, the number of output scores (or dimensions in the figures) always show the same trend: really flat in all the range, except considering 2 or 3 scores where there are the highest accuracy values, probably due to the fact that UMAP has been created with the purpose of visualisation of big datasets. Only the NADL Informal case shows the higher maximum value with 8 dimensions, although the trend is really flat along the dimensions (the peaks of accuracy always reach similar values), while it is more influenced by the number of neighbours.

Due to the high temporal and computational cost of this inspection no other values has been checked, except the *min_dist* parameter in the UMAP setting where a value of 0.01 has been setted, but the related results show very similar accuracy values and trends for every scores, confirming the effectiveness of the choice of the default value (0.1).

At the end of all the evaluation, the UMAP's range of the number of the output scores is reduced to the first 2 and 3 dimensions, while the range of neighbours values has been taken from 6 to 30, due to the more accentuated variability of the accuracy results.

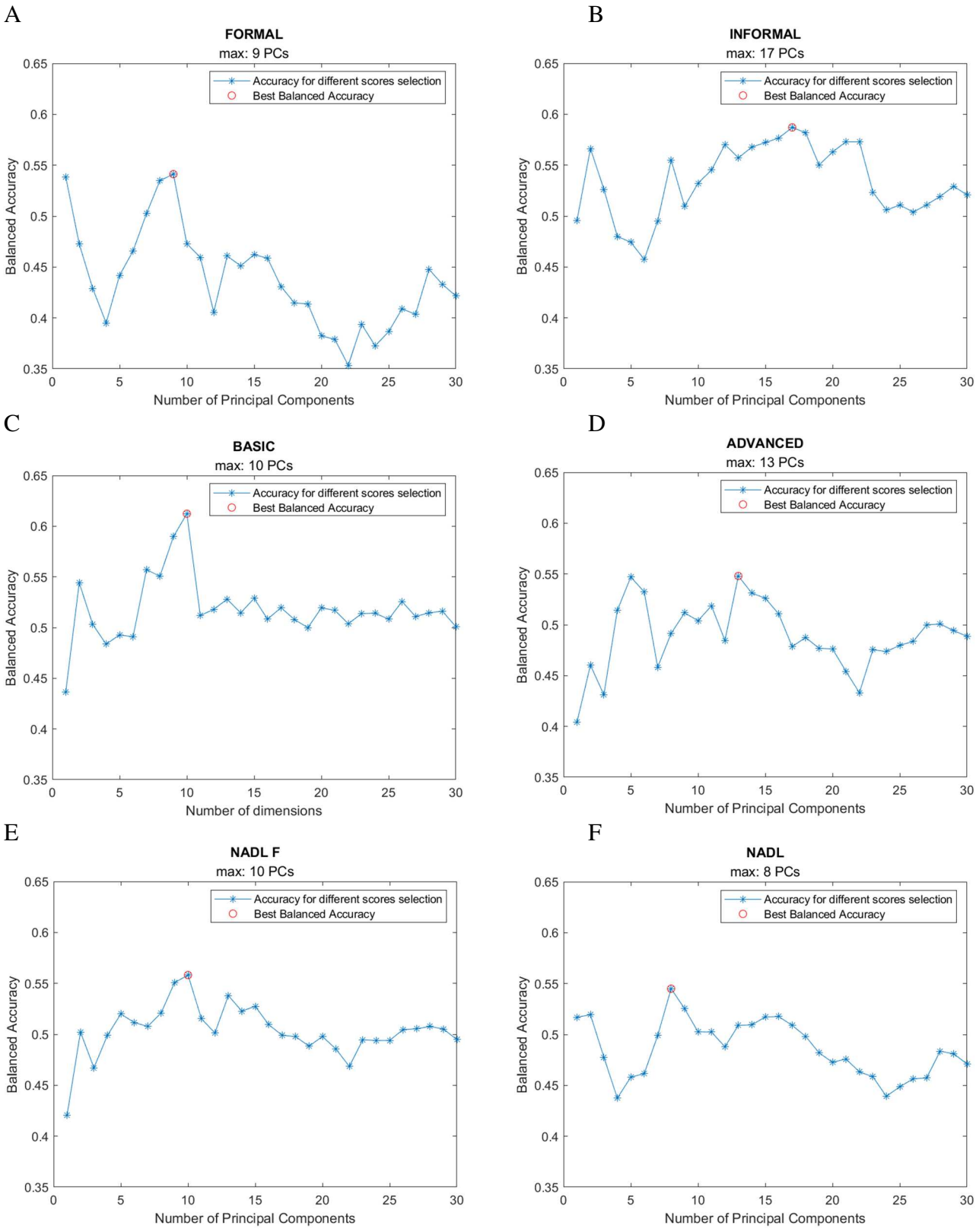

Fig. 14 - Balanced accuracy values from KNN classifier with PCA, calculated for the four separated scores: Formal (A), Informal (B), Base (C), Advanced (D). The overall financial abilities accuracy is calculated averaging Advanced and Basic scores (E) and the numerical one averaging on the Formal and Informal scores (F).

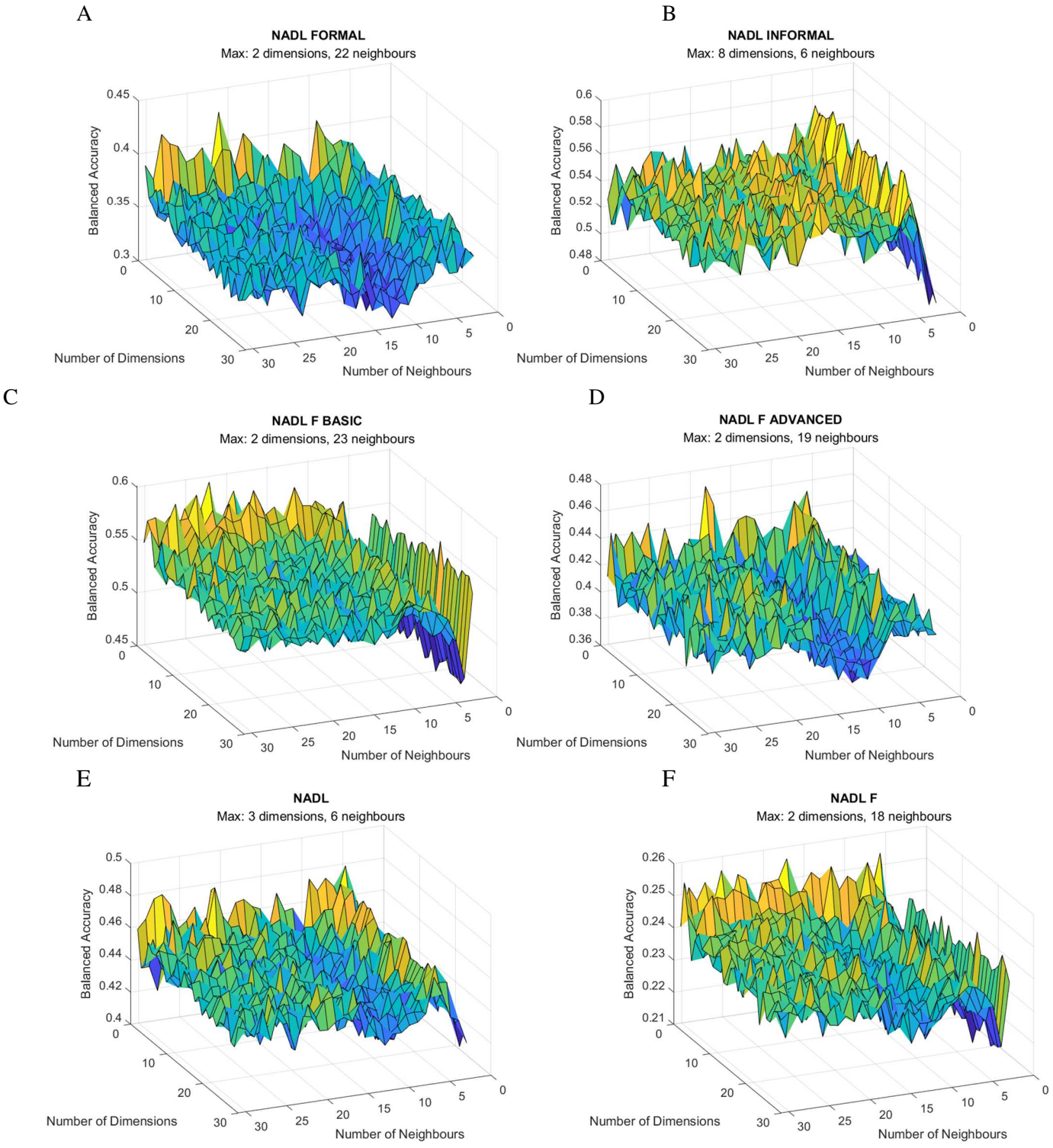

Fig. 15 - Balanced accuracy values from KNN classifier with UMAP, calculated for the four separated scores: Formal (A), Informal (B), Basic (C), Advanced (D). All the assessment has been repeated 30 times and then averaged on the iteration to have reliable results, for all the four single scores. Then the overall numerical and financial scores have been obtained averaging the Formal and Informal scores of the NADL Short test (E) and Basic and Advanced scores from NADL-F Short test (F).

4.3 Canonical Correlation Analysis results

In the first phase, a descriptive analysis of the full dataset is computed. From the inspection of the p-values calculated iteratively for each parameter, it is possible to select the models that best describe the associative effects between input and output variables.

Two separate CCAs were conducted: the numeric abilities are inspected using as input the Formal and Informal scores (again this analysis will be called NADL), while the financial abilities are evaluated using Basic and Advanced scores as input (the results will be referred as NADL-F).

For each of the two abilities, the impact of each additional variable (age, schooling, lesion volumes and parcel damage) has been evaluated removing them one at a time from the CCA input. The results showed that only the parcel damage has a significant effect on the analysis, so for each ability, a model with all the additional variables (called "All") and a model without the parcel damage information (indicated as "-Loads") were computed, in the ranges of parameter identified in the previous KNN classification analysis. The Fig. 16 shows two summary schemes of all the models that were assessed in this descriptive step.

From the inspection of the p-values related to the associative effects in the descriptive phase, it has been noticed that they are usually quite high, as a result, given the limited sample size, the 0.1 threshold was selected for almost every inspection.

Starting from the PCA-based models, the ones without the parcel damage information have clearly better performance, showing lower p-values for more PCs than the models with all the variables (see Fig. 17). For better clarity, Tab. 3 reports the selected models (i.e., the corresponding number of PCs) that remain under the 0.1 threshold.

			Ranges:
PCA	NADL	All	5, 6, 12, 14, 16
		-Loads	5-11, 14-18, 21
	NADL-F	All	11, 16
		-Loads	5-13, 17, 18

Tab. 3 - All PCA-based models are evaluated in the predictive step with the ranges of PCs listed in the last column.

Regarding the UMAP-based models, the averaged p-value on the 5 iterations of all the NADL models ("All" and "-Loads") and NADL-F "All" models exceeded the 0.1 threshold for all the number of neighbours and both dimensions, so they are excluded from the predictive step analysis.

Just the NADL-F "-Loads" model performs well, but all the neighbours remain almost always under the threshold both for 2 and 3 dimensions (Fig. 18).

In order to select the best parameters, it was considered the frequency of time that the p-value for a specific number of neighbours remains under the 0.05 threshold in all the 5 iterations. When the threshold is not exceeded in all the iterations, then the corresponding number of neighbours is selected (Fig. 19). Even in this case, no differences could be observed for the "2D" and "3D" models because they have a similar number of recurrences, so both were analysed in the predictive step.

Again, the selected neighbours for the two models are listed in Tab. 4.

				Ranges:
UMAP	NADL-F	-Loads	2D	12, 19, 21, 29
			3D	12, 13, 27, 29

Tab. 4 - Just two UMAP-based models are evaluated in the predictive step with the ranges of neighbours listed in the last column. Both ranges are very similar.

Now that the computational costs have been reduced in the descriptive step, the 6 models selected were evaluated in the predictive phase.

Fig. 20 shows the correlation between the canonical variables (U and V) calculated in the test set extracted from the predictive step. Ideally their correlation should reach the values of correlation calculated among the same canonical variables of the descriptive step. The threshold of 0.5 is settled, meaning that the models should be good enough to have at least roughly half of the predictive power of the models of the descriptive step, which correlations are close to the value of 1. This threshold is chosen after observing similar results in paper [27].

Regarding the prediction of the numerical abilities, just the NADL "-Loads" model based on PCA data reduction has good performance, while the NADL "All" never reaches the threshold.

Concerning the financial abilities just the two UMAP-based models get closer to the threshold

value, but if the mean values on the 5 iteration are considered just the “2D” overcome the 0.5 value. Moreover the standard deviation of the “2D” model is always less than the “3D” one, so the former case was chosen as the best model for the prediction of the financial skills.

Finally, the parameters corresponding to the maximum value of the two selected models were chosen: 6 PCs for the PCA-based prediction model for numerical skills and 12 neighbours and 2 dimensions for the UMAP-based prediction model for financial abilities.

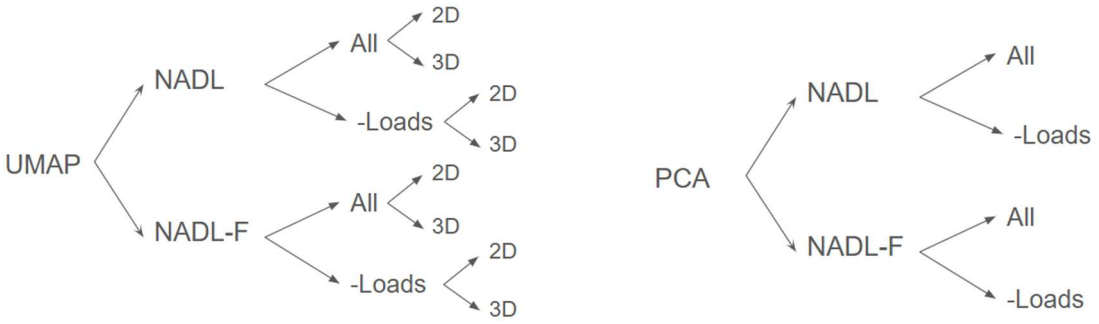

Fig. 16 - Schemes of the models analysed in the descriptive CCA, for UMAP and PCA methods of data reduction. NADL and NADL-F are models that want to assess numerical and financial skills, respectively. “All” and “-Loads” refer to the type of input in the CCA: with all the additional variables or removing the parcel damage information. In the UMAP-based models are evaluated the number of neighbours in the range [6-30], with 2 or 3 number of output scores (“2D” or “3D”), while in the PCA-based model are evaluated the number of PCs in the range [5-25].

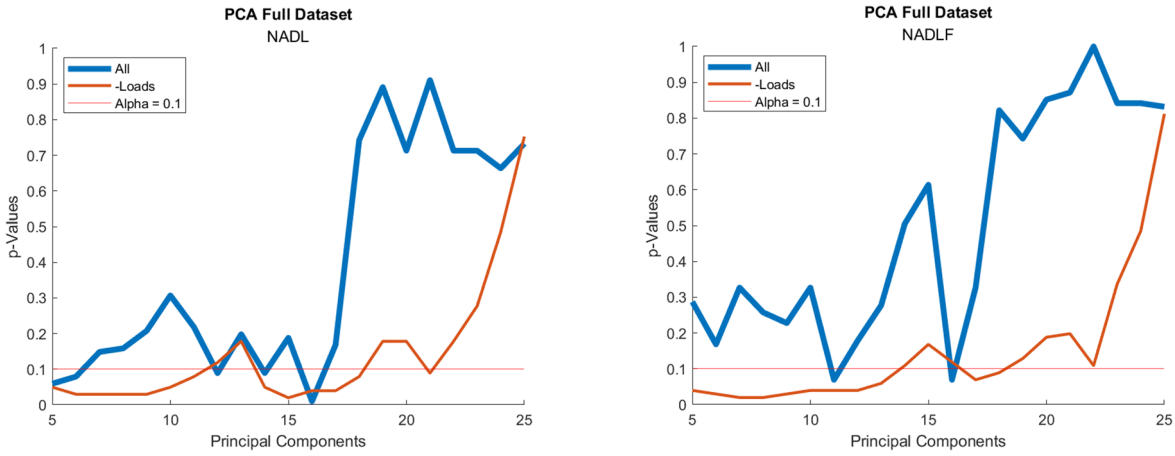

Fig. 17 - Plot of the p-values of each principal component considered in the descriptive CCA based on PCA data reduction, for NADL (left) and NADL-F (right). The blue line indicates the model with all the confounders variables, while the red line represents the p-values of the model without the parcel damage data.

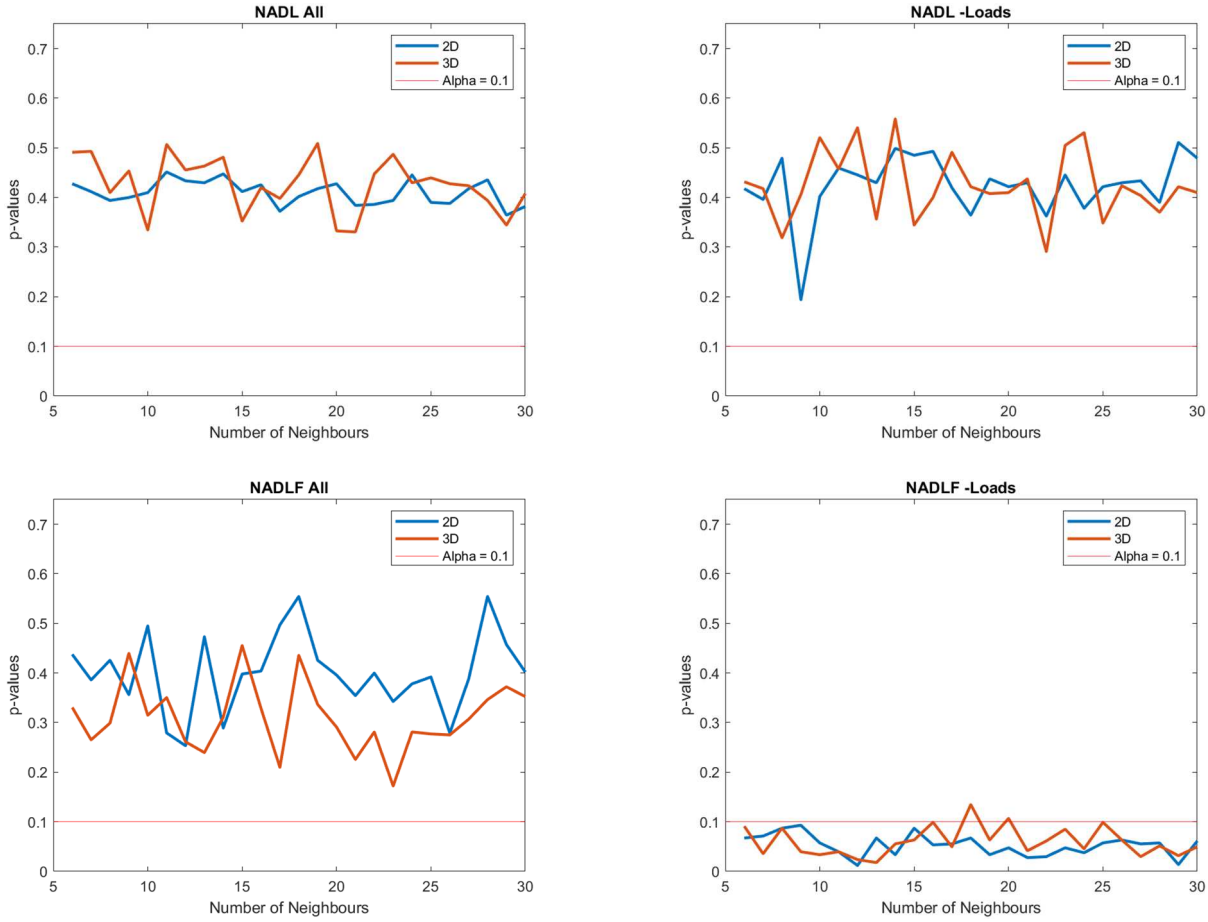

Fig. 18 - Plot of the p-values of each neighbour considered in the descriptive CCA based on UMAP data reduction. The number of UMAP's output scores (blue lines for 2 dimensions and red lines for 3 dimensions) wasn't significant, while the removal of the parcel damage information was fundamental in the NADL-F "-Loads" models.

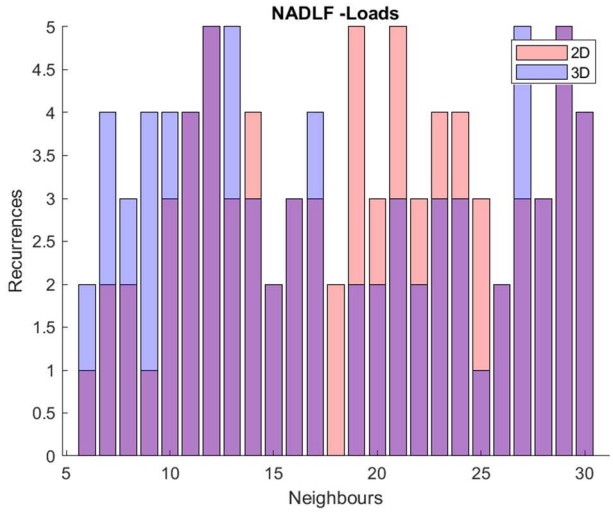

Fig. 19 - Frequency of recurrences of the p-values remaining under the 0.05 threshold on the 5 iterations, for the NADL-F "-Loads" models. No significant differences on the "2D" model (red) and "3D" model (blue) can be seen, so both models were analysed in the predictive step, for the number of neighbours that always remain under the threshold.

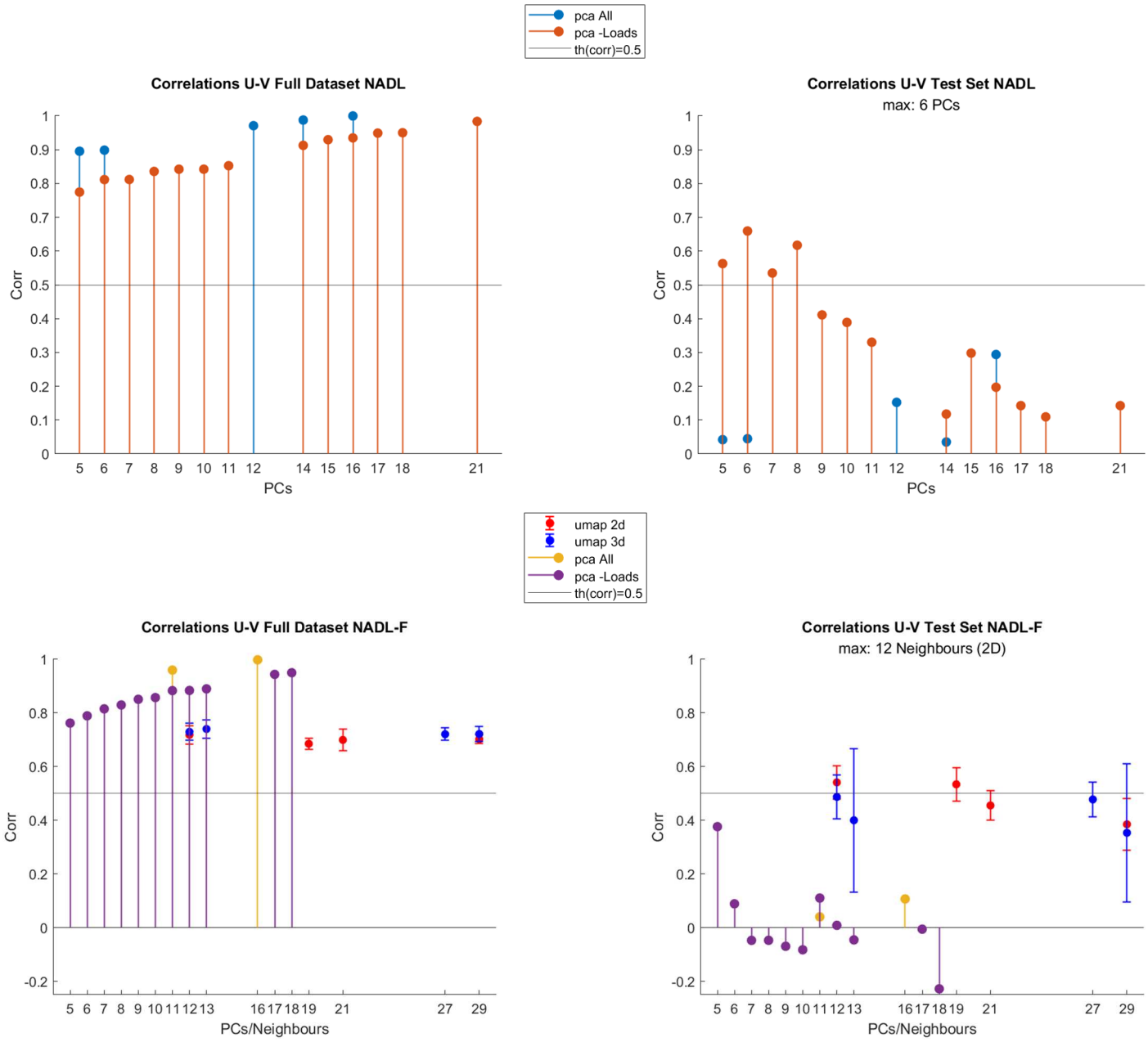

Fig. 20 - Figures of the correlation between canonical variables of each model selected for the predictive CCA. The points of the UMAP models represent the average values on the 5 iterations and the tails show the respective standard deviations. On the left there are the correlations of the variables calculated in the descriptive step (using the full dataset) and on the right there are the correlations among the variables calculated with the LOO method (using the test set). Just the PCA-based NADL "-Loads" and UMAP-based "-Loads" (2D and 3D) pass the 0.5 threshold.

4.4 Models' interpretation by hierarchical clustering

Thanks to the clusterization procedure it is possible to to give an interpretation of the models. First of all the optimal number of clusters has to be chosen: two over three of the methods used showed 3 as the best hierarchical clustering solution (Silhouette and Calinski-Harabasz), while the Davies-Bouldin method suggests 9 as a good choice, but considering that in this latter case there is a local minimum near the value 3 of number of clusters, 3 was chosen as the optimal choice (Fig. 21).

Moreover, the selection of fewer clusters can be helpful to simplify the interpretation of the models. A more accurate investigation can be done on further improvement of the models' effectiveness.

Fig. 22 represents the canonical variables of the two models, in which each subject included in the three clusters is marked with different colours. In addition, in the same figure the right lesioned patients are separated from the left lesioned ones, so that can be noticed that the cluster 1 has almost only right subjects (85.7%), while the other two clusters are more heterogenoeus in terms of lesion location (63.6% and 61.5% of left lesioned subjects in cluster 2 and cluster 3).

In order to investigate the characteristics of the 3 groups of patients, the average of the input variables used in the CCA has been computed: Fig. 23 show the mean values calculated over the subjects included in each cluster of all the variables used in the CCA for the two models (i.e. the 6 PCs for the NADL PCA-based prediction model, the 2 dimensions for NADL-F UMAP-based prediction model, the 3 confounding variables common to both models and the 4 cognitive scores).

All the variables were z-scored before the calculation of the mean values.

Starting from the inspection of the images from 24 to 30, some consideration on the 3 groups of patients can be done.

Regarding the PCA-based NADL "-Loads" model:

- The first group (cluster 1, green) has the best numerical performance (Fig. 23 D), in particular in the Formal score. In fact, also looking at the weights calculated by the CCA (Fig. 24 B), the Formal score seems to have more explanatory power than the Informal one.

This positive behaviour can be also mathematically confirmed looking at the (almost all) negative values of the canonical weights that have to be multiplied by the PCA scores, which mean values are alway negative again, in order to get positive values of the canonical variables **U** and **V** visible in Fig. 22 (the green patients are always in the positive first quadrant of the correlation plane).

In this group the schooling value of the subjects is really high (Fig. 23 C), more than in the other 2 groups and the lesions are the smallest in size, confirmed by the lesion frequency map in Fig. 28.

- The second group (cluster 2, blue) shows opposite characteristics of the first one since it includes the patients with the worst numerical abilities. Again the NADL Formal score seems to be more involved in the correlation than the Informal one (Fig. 23 D), confirmed by the related CCA weight (Fig. 24 B).

  Looking at Fig. 23 (C) these bad performances can be associated with the volume of the regions, in fact the second group has the highest mean lesion volume, even if this seems not to have a strong impact on the prediction because the related canonical weight is quite low (Fig. 24 A). The same consideration can be done for the age: here the patients have the highest age, but in this case the associated canonical weight (the one with the highest absolute value in the confounding variables group) probably leads to a stronger association between the bad performance and the age, rather than the lesion volume.
- The third group (cluster 3, red) seems to have mixed features: the numerical skills of the subjects are just a little above the average values (Fig. 23 D), probably due to the fact that they are the youngest group but with the lowest number of years of schooling (Fig. 23 C). The value of the fourth PC in Fig. 23 (A), the one with the highest canonical weight, is around the average value, while the other two groups seem to be characterised by patients that show high absolute values of that PC. Instead, the second PC looks more important in the identification of the characteristics that define this cluster.

Regarding the UMAP-based NADL-F "-Loads" model, similar considerations to the previous model can be done, but it has to be noticed that the canonical weights (Fig. 24) are very different from the ones of the PCA-based model: in this case the confounding variables have more predictive power, with an opposite trend to the previous model. In fact, considering just the three confounding variables, in the linear based model the age has a stronger impact than volume, while in the non-linear one the volume weight overcomes the one of the age. Moreover

it is easy to note that in the PCA-based model the confounding variables have less considerable weights respect to the fourth PC, while in the UMAP based one they have the strongest impact, with values similar or more high than the two UMAP scores. So, in this case, the UMAP-based model seems to focus more on the additional variable rather than the connectivity information, with particular attention on the volume of the lesion.

Repeating the inspection of the characteristics that distinguish the 3 groups, it can be observed that:

- The first group of patients (cluster 1, green) has good performance in the financial test, (Fig. 23 D) in particular in the Advanced NADL-F score. The confounding variables are the same as the PCA-based model and analogous observations can be done also in this case. So, in this group there are the subjects that show positive values on the two UMAP output scores (Fig. 23 B), with high schooling and small lesions (Fig. 23 C).
- The second group (cluster 2, blue) has bad performance in the Basic NADL-F score, while the Advanced one has values that are similar to the average (Fig. 23 D). The subjects are characterised by wider lesions and high age values (Fig. 23 C), with UMAP scores really close to the mean values (Fig. 23 B).
- The third group (cluster 3, red) has good performance in the Basic NADL-F test, similar to the one of the first group, but low scores in the Advanced test (Fig. 23 D). The patients have generally negative UMAP scores (Fig. 23 B) and are characterised by very poor schooling values and younger age (Fig. 23 C).

Looking at the connectivity related to the 3 clusters, it can be noticed that the mean direct SDC matrix (Fig. 25) and indirect SDC (Fig. 26) of the first group show clear high values of damage in the right hemisphere (indeed 85.7% of patients are right lesioned in this group, see fig. 28 A) and inter-hemispheric disconnections. This leads to the intra-hemispheric segregation that is visible in the mean FC (Fig. 27). The most disconnected networks are the Somatomotor, Dorsal Attention and Ventral Attention.

In the second group the mean direct SDC matrix shows disconnection in particular among regions of the same hemisphere that cause a high amount of indirect disconnection visible in Fig. 26, involving many networks. The mean FC matrix (Fig. 27) has instead very low activity, coherent with the poor performance of the numerical and financial tests.

The last group has direct SDC that doesn't show evident spatial patterns, while comparing the indirect SDC with the one of the second group can be noticed that there a complementary configuration emerges where less disconnections characterise the Left Sensorimotor, Left

Dorsal Attention and Right Visual Networks (Fig. 26), that are instead highly disconnected in the second group. Moreover the mean FC has higher values coming from the inter-hemispheric networks, especially highlighting links between sensory and cognitive areas (Fig. 27).

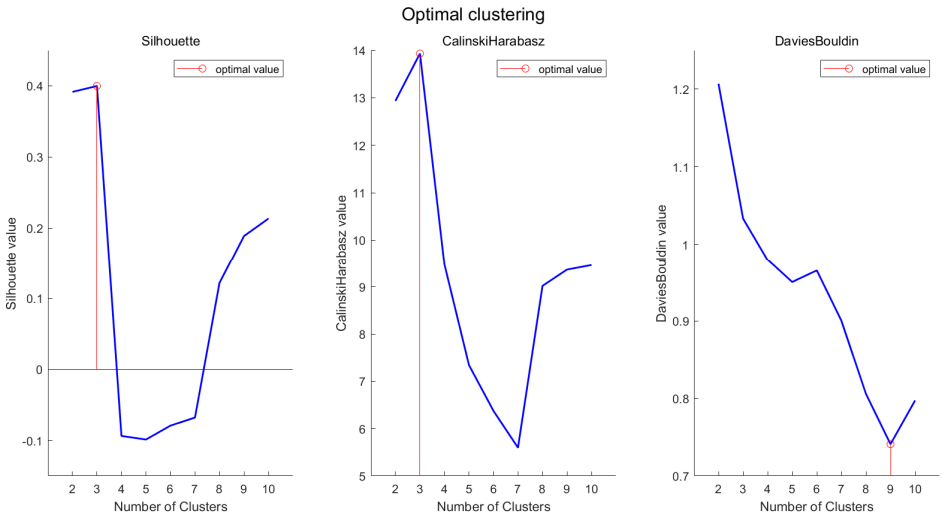

Fig. 21 - Plot of the values coming from the three methods used to find the best clusterization: two over three methods show that 3 is the optimal number of clusters. Also in the Davies-Bouldin method that has the best value with 9 clusters, seems to show a local minimum near 3.

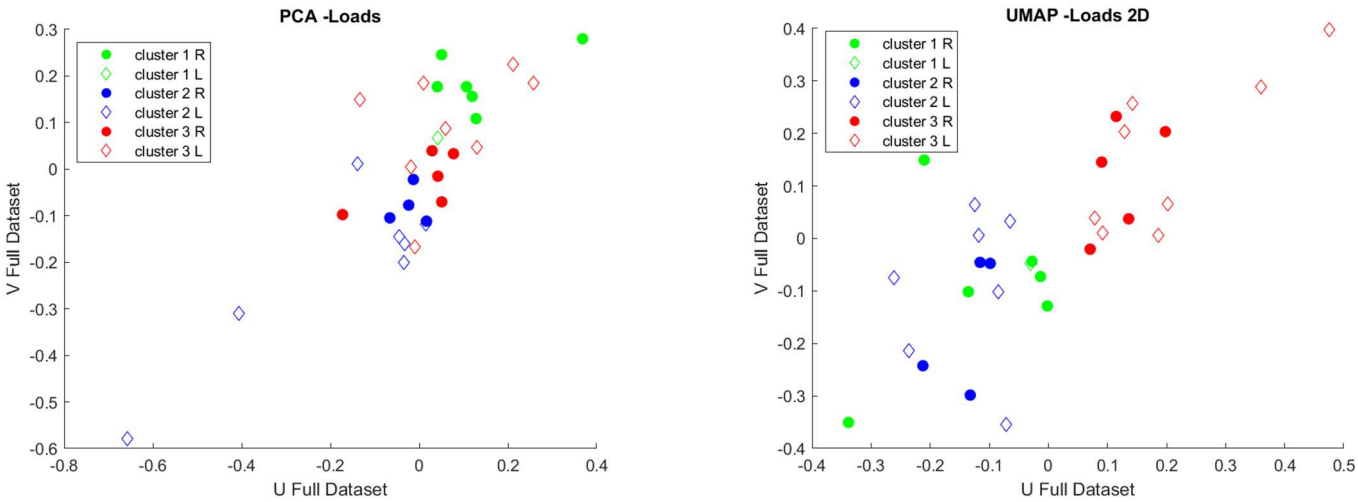

Fig. 22 - Figure of the canonical variables for the two models. The colours identify the patients of each of the 3 clusters, while the points indicate the right lesioned patients and the diamonds the left lesioned ones.

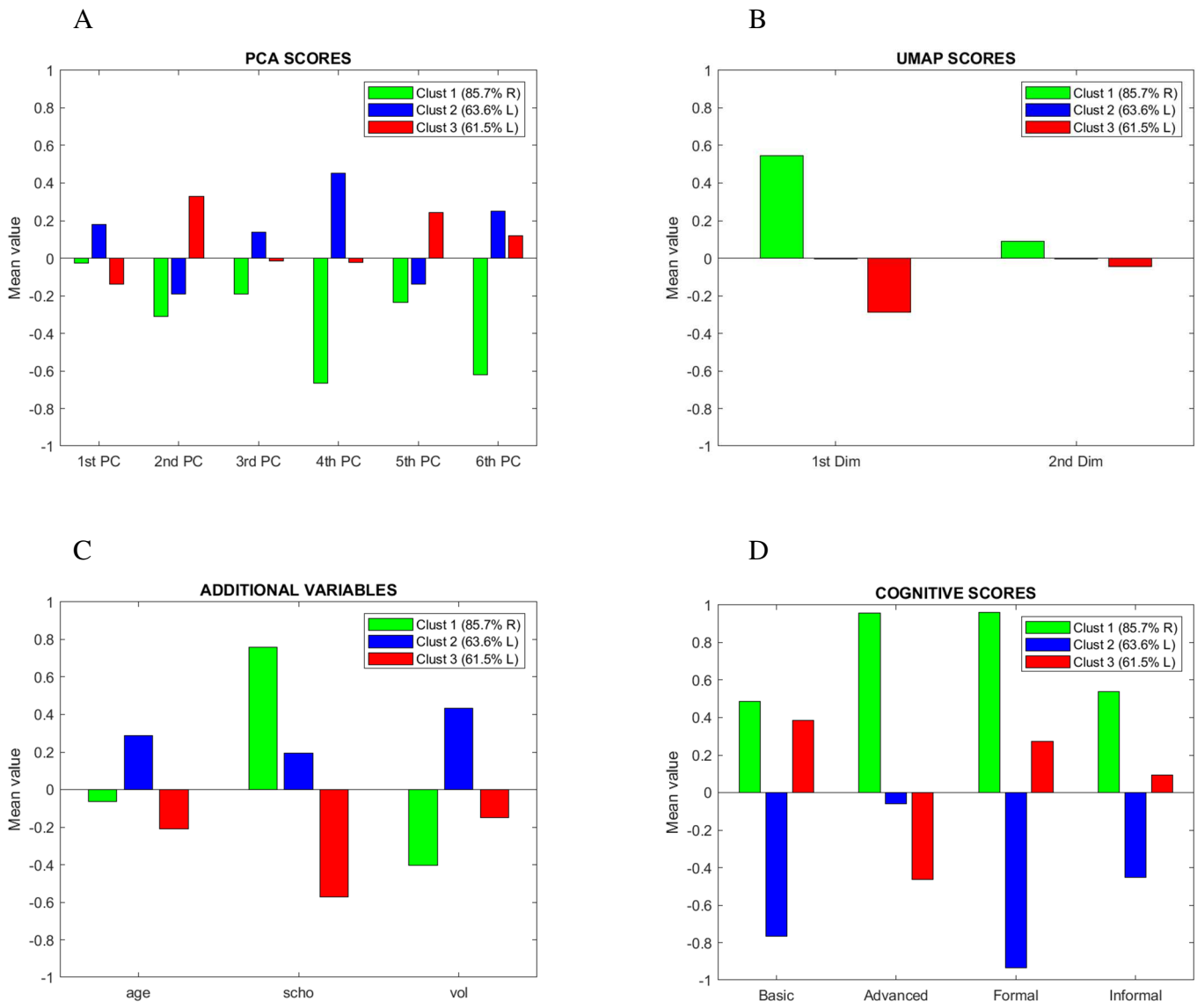

Fig. 23 - Mean values of the variables used in CCA, calculated for each cluster. Figures A, B, C represent the X input (6 PCs for the NADL "-Loads" model in A, 2 dimensions for the NADL-F "-Loads" model in B and the three confounding variables that are used in both models in C) and figure D shows the four cognitive score (Basic and Advanced for NADL-F "-Loads" model and Formal and Informal for NADL "-Loads" model). The legend also shows the percentage of right lesioned patients in the first cluster and the left lesioned patients in the latter two.

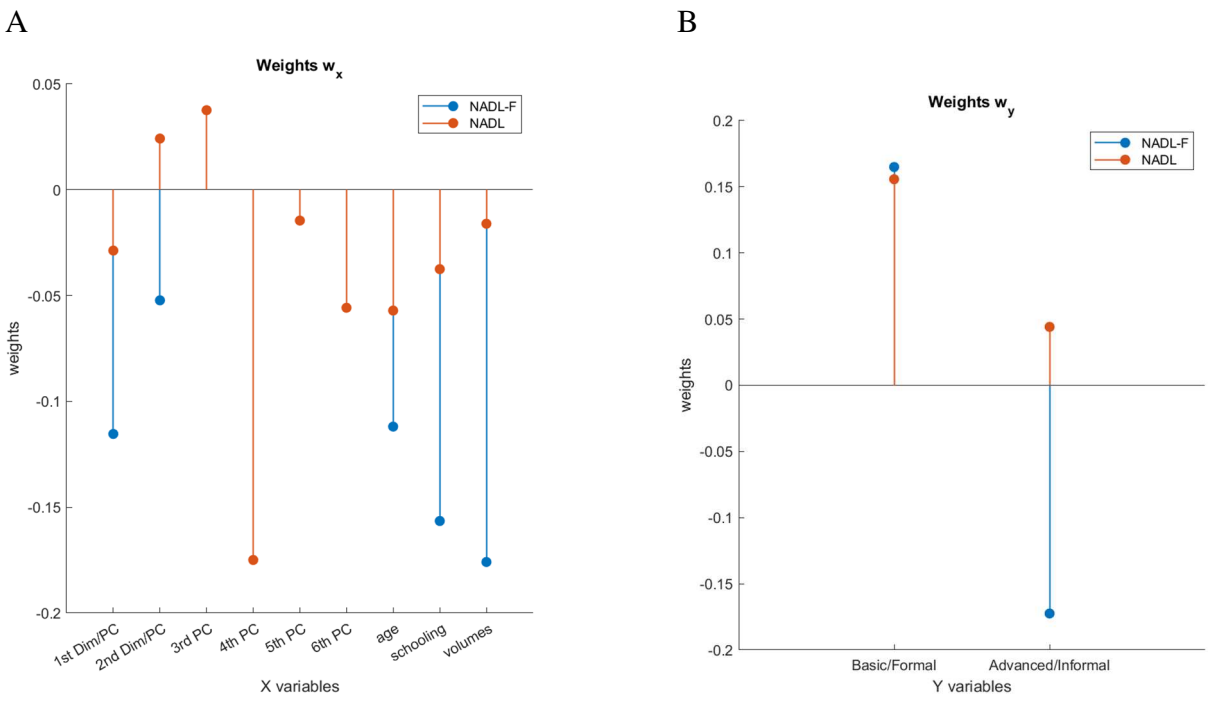

Fig. 24 - Figure of the canonical weights $\mathbf{w_x}$ (A) and $\mathbf{w_y}$ (B) for the two selected models: NADL PCA-based "-Loads" with 6 PCs and the third iteration of NADL-F UMAP-based "-Loads" with 12 neighbours.

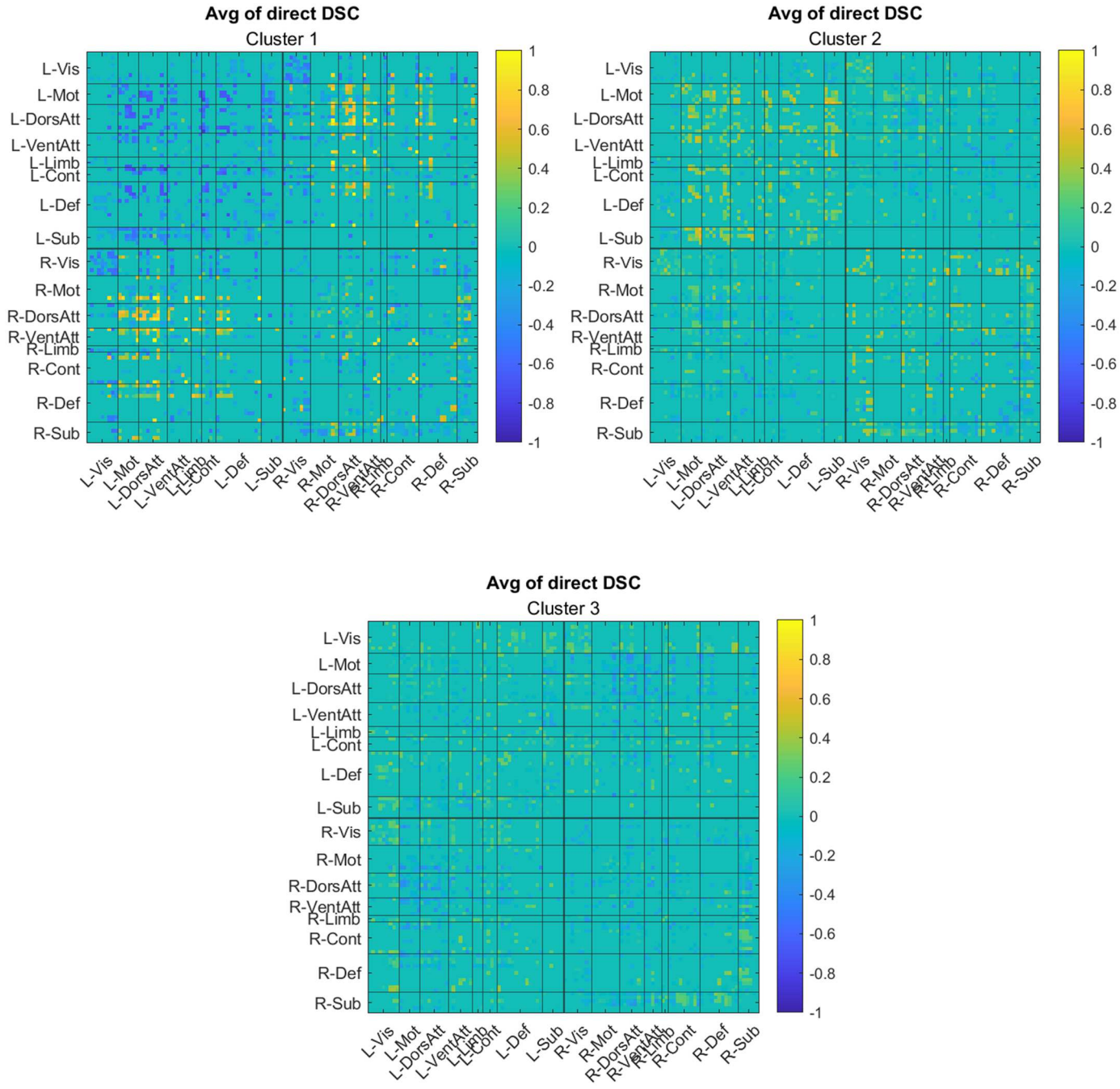

Fig. 25 - Average of the z-scored direct SDC matrices of the patients of each cluster.

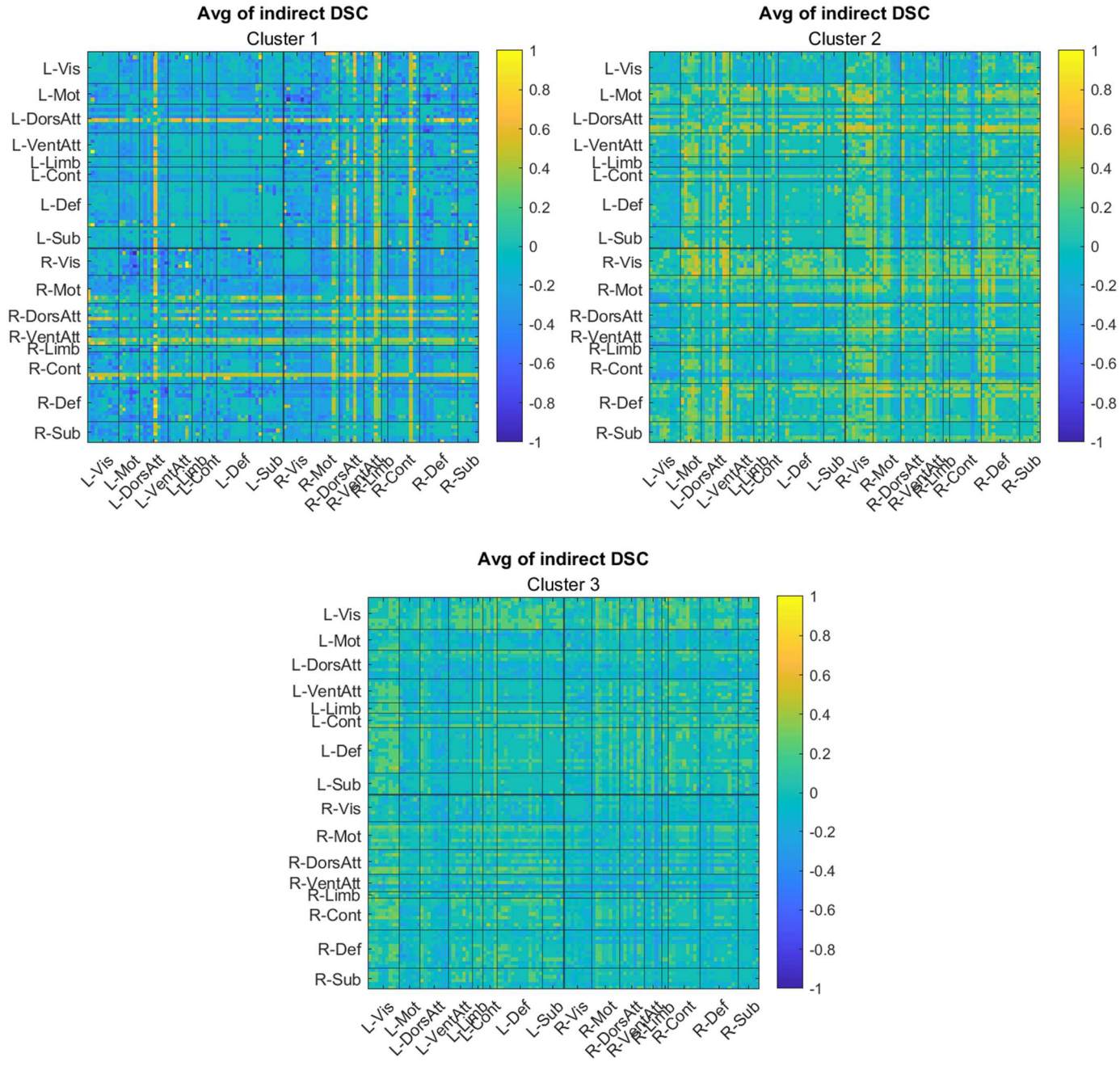

Fig. 26 - Average of the z-scored indirect SDC matrices of the patients of each cluster.

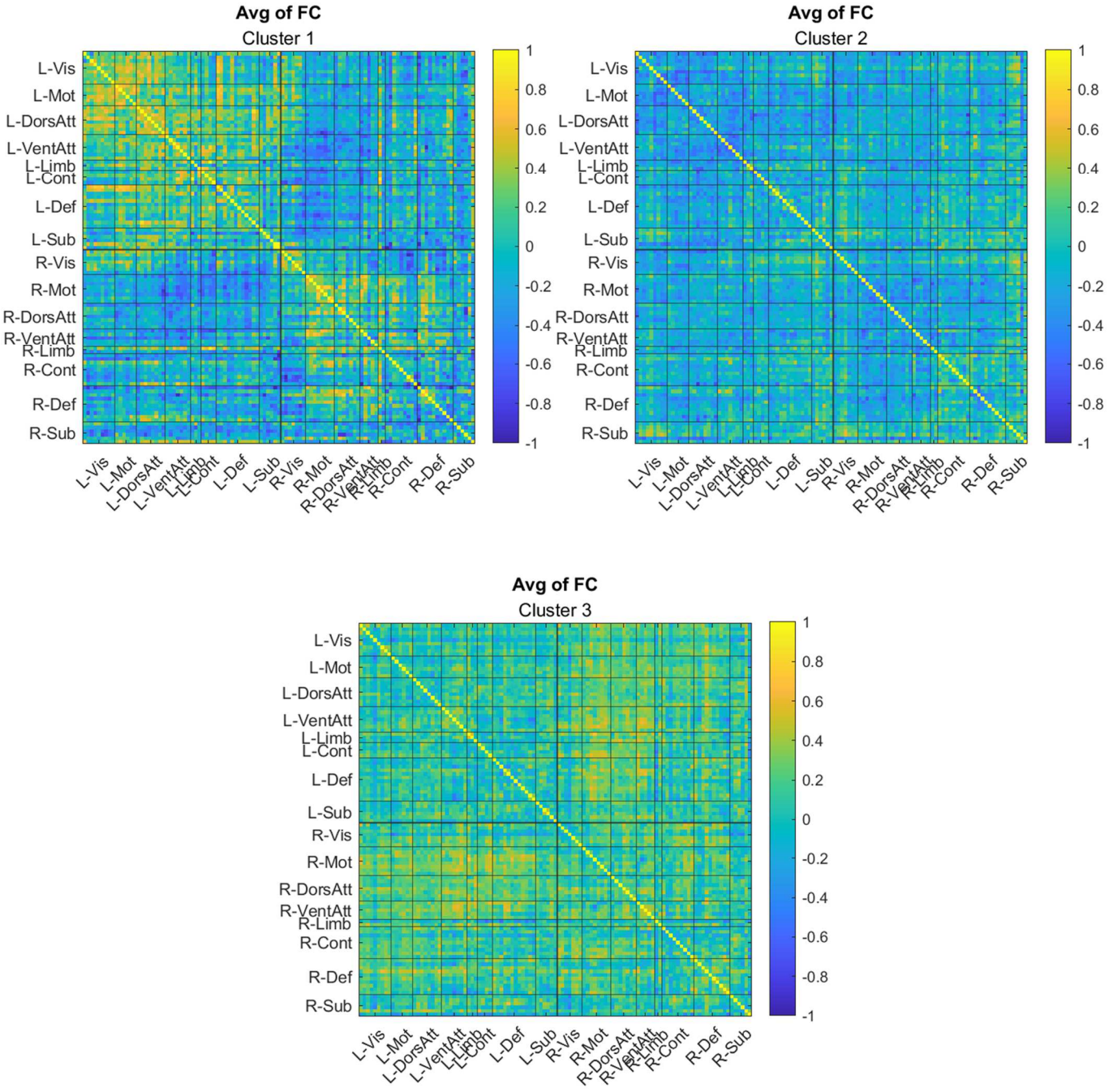

Fig. 27 - Average of the z-scored FCs of the patients of each cluster.

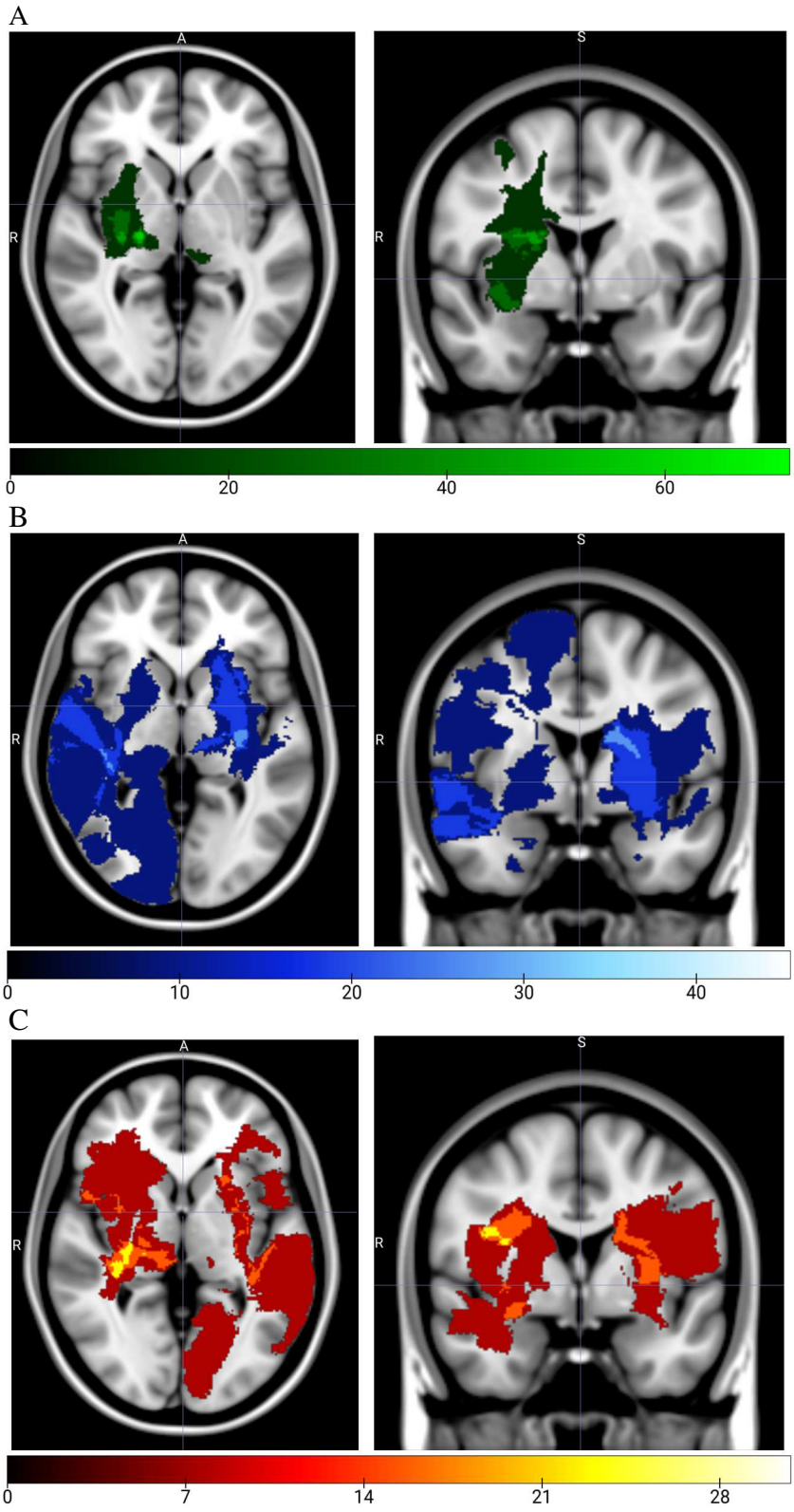

Fig. 28 - Frequency maps of the patients' lesions divided in the 3 groups (cluster 1 in A, cluster 2 in B and cluster 3 in C). The colorbar indicates the frequency of lesioned voxel, with maximum values of 71.43% in cluster 1, 45.45% in cluster 2 and 30.77% in cluster 3.

5 DISCUSSION

The purpose of this work was to create a model for the prediction of numerical and financial abilities, embedding and comparing the effectiveness of the linear PCA and the non-linear UMAP data reduction algorithms, so as to highlight features helpful in the understanding of the brain behaviour dealing with numerical and financial tasks. Considering the results of the work that has been done, many interesting aspects can be noticed. First of all, while the numerical abilities can be predicted using a model based on the linear PCA data reduction, the financial skills can be assessed using a model based on UMAP data reduction, that is a non-linear method. This means that probably non-linear approaches are able to identify peculiar characteristics in the connectivity that linear methods can't find.

To better understand the impact of the two data reduction methods a more detailed inspection could be done to interpret the PCs by checking their loads and the UMAP dimensions observing the graph created by the algorithm. In fact, interesting conclusions could be deduced by its structure like has been done in [74], where it has been found out that UMAP distributes the disconnectome profiles of the patients based on the lesion location and the commonly disconnected white matter tracts.

Considering the confounding variables, both models show that the information about the parcel damage can hinder the good results of the analysis, but since the two abilities are made up of two very different scores (Basic and Advanced for financial skills and Formal and Informal for numerical skills evaluate abilities that differ in complexity) more detailed inspection should be made on each of the four scores. Regarding the other variables (age, schooling and lesion volumes) the canonical weights show very different behaviours in the two models, indeed they have a notable impact in the UMAP-based model for the prediction of the financial abilities, while they are less considered in the PCA-based model for the assessment of numerical skills.

The optimal number of neighbours in the UMAP-based model is 12 that is quite high considering the number of patients, but confirmed from the works in other research topics where taking less than 10 neighbours has been seen to produce too much focus on the structural details of the dataset [72] and near to the algorithm's default value of 15.

The strong scalability of UMAP is confirmed considering that just 2 dimensions are needed to create the related models against the 6 PCs used in the selected PCA-based model for the numerical skills prediction.

From the inspection of the p-values related to the associative effects in the descriptive CCA step, it has been noticed that they are usually quite high, so the 0.1 threshold is selected for almost every inspection. It can be noticed also that the p-value of the final selected UMAP-based model for financial skills prediction has lower p-value (0.0119) than the PCA-based model for numerical skills (0.0297) indicating the superiority in statistical effectiveness of the non-linear method also for few observations. Probably adding more patients in the analyses will increase significance allowing the inspection of parameters with more statistical relevance.

From the inspection of the clustering analysis results it seems clear that, despite the statistical inference made in *Preliminary inspection*, there is a link between the side of the lesion and the patients' performance in the tests. In fact the first cluster, the one that shows the best performance, is composed of almost only right lesioned subjects, characterised by smaller lesions, high schooling values and greater functional activity on the left hemisphere. In fact, intra-hemispheric segregation that is visible in the mean FC is a common event in stroke patients [24] [77], in this case in particular in the left hemisphere, that compensate for the lost functionalities in the right part. Some studies also demonstrated that the preservation of the functionality in the left hemisphere is essential for numerical abilities, correlating the deficit in numerical processing with the presence of lesions in the left insular cortex [75] and left angular gyrus [10] [75]. Moreover as it has been told in the introduction, solving numerical tasks involves many other areas of the brain, linked with parietal and frontal networks, or related to long-term or working memory [10], so smaller lesions not affecting frontal regions may have a lower impact on numerical deficits. Similar consideration can be done also for the financial abilities, since these patients have good performance also in the NADL-F Short test and from literature can be noticed that financial skills have been associated again with regions located in the angular gyri [76].
On the other hand, the second cluster is composed of patients with the worst performance, larger lesions and older subjects and it is confirmed by the FC matrix where a very low activation is visible at a whole-brain level. Looking at the frequency map (Fig. 28 B), the lesions seem to be very close to the Corpus Callosum, essential in the connection among the two hemispheres [75]. This group also has a wider amount of direct disconnection within the left hemisphere, that lead to indirect disconnections, in particular in the Sensorimotor and Dorsal Attention Networks, the latter one associated with the IPS, core locus of numerical processing [10]. The low degree of indirect disconnections within sensorimotor, Dorsal Attention and Ventral Attention seems to be fundamental in the third group, whose FC seems

to be focused on inter-hemispheric integration. This phenomenon probably enhances the performance in numerical skills and Basic financial abilities with respect to cluster 2 (as can be suggested in [78] for motor impairment and in [75] where it is noticed that connection between left and right hemispheres is crucial in calculation tasks), although in this group are gathered patients that have very bad Advanced NADL-F scores.

6 LIMITATIONS

Despite such preliminary interesting findings, several limitations need to be pointed out. The implementation of the models still is not concluded: an accurate evaluation of the cognitive score effectively predicted is needed through a validation step by increasing the sample size.
In fact, the most important limitation regards the number of patients, which is too limited to create a good, reliable model, leading to overfitting issues. Moreover, the aim of this thesis is to create a prediction model for numerical and financial skills that, as explained in the introduction, rely on widely spread networks over the cortex. As a result, cognitive deficits caused by the stroke can be very different from patient to patient. Creating a model with such a heterogeneous sample can lead to difficulties that are increased by the small number of available patients. An L1 or L2 regularisation of the models can be done in order to damp these problems, but the inclusion of new patients is essential for the enhancement of the results.
Indeed, despite their promising results, non-linear methods have to be well understood in order to be properly applied. A lot of time has to be spent in the inspection of the best parameters in order to understand their impact on the analysis. Moreover, additional computational cost and time is needed to have stable results, iterating various times to avoid fluctuation on the outputs. All these issues can hinder the reproducibility of the analyses, so it is important to have a clear understanding of the chosen algorithm.

7 CONCLUSIONS

From these first analyses it can be asserted that nonlinear dimensionality reduction methods can represent a good solution to reach a further assessment in the investigation of brain functional behaviour, especially when complex cognitive scores like the financial one, have to be inspected, showing superior results to the common and well consolidated PCA. Moreover, the great scalability property of algorithms like UMAP can be fundamental to reduce the dimensionality of the dataset (for instance the selected UMAP-based model use just 2 dimensions instead of the PCA-based one that need 6 PCs), even if its non-linear nature lead to interpretation difficulties of the outputs. On the other hand, linear methods like PCA are simpler to use and their outcomes are more stable and easier to understand.

Considering the results of the two models, it seems clear that right stroke patients have more probability to preserve numerical and financial abilities which appear to be tightly connected. Further investigation could be interesting in order to better distinguish the differences among patients of the second and third clusters, since the assessment done for these two groups of patients seems to highlight important differences in the SDCs and FCs, with an interesting involvement of indirect interactions within and between Left Sensorimotor, Left Dorsal Attention Networks and Right Visual Network and inter-hemispheric functional and structural connectivity that also seems to play an important role.

8 REFERENCES

[1] Paolo Bellisario, Daniela Galeone, last update 26/01/2024, last consultation 31/11/2024, https://www.salute.gov.it/portale/alleanzaCardioCerebrovascolari/dettaglioSchedeAlleanzaCardioCerebrovascolari.jsp?lingua=italiano&id=28&area=Alleanza%20italiana%20per%20le%20malattie%20cardio-cerebrovascolari&menu=malattie#:~:text=L'ictus%20%C3%A8%20pi%C3%B9%20frequente,nelle%20donne%205%2C9%25).

[2] Iadecola C, Buckwalter MS, Anrather J. *Immune responses to stroke: mechanisms, modulation, and therapeutic potential*. J Clin Invest. 2020 Jun 1;130(6):2777-2788. doi: 10.1172/JCI135530.

[3] Kloska, S.P., Wintermark, M., Engelhorn, T. et al. *Acute stroke magnetic resonance imaging: current status and future perspective*. Neuroradiology 52, 189–201 (2010). https://doi.org/10.1007/s00234-009-0637-1.

[4] Pustina, D., & Mirman, D. (Eds.) (2022). *Lesion-to-Symptom Mapping: Principles and Tool*s. Neuromethods; Vol. 180. Springer. https://doi.org/10.1007/978-1-0716-2225-4

[5] K.M.A. Welch, Yue Cao, Vijaya Nagesh, *Magnetic resonance assessment of acute and chronic stroke, Progress in Cardiovascular Diseases*, Volume 43, Issue 2, 2000, Pages 113-134, ISSN 0033-0620, https://doi.org/10.1053/pcad.2000.9029.

[6] Smitha K, Akhil Raja K, Arun K, et al. *Resting state fMRI: A review on methods in resting state connectivity analysis and resting state networks*. The Neuroradiology Journal. 2017;30(4):305-317. doi:10.1177/1971400917697342.

[7] Yeo BT, Krienen FM, Sepulcre J, Sabuncu MR, Lashkari D, Hollinshead M, Roffman JL, Smoller JW, Zöllei L, Polimeni JR, Fischl B, Liu H, Buckner RL. *The organization of the human cerebral cortex estimated by intrinsic functional connectivity*. J Neurophysiol. 2011 Sep;106(3):1125-65. doi: 10.1152/jn.00338.2011. Epub 2011 Jun 8.

[8] Smadar Ovadia-Caro, MSc, Daniel S. Margulies, PhD, and Arno Villringer, MD,The Value of Resting-State *Functional Magnetic Resonance Imaging in Stroke*,Stroke, Volume 45, Number 9,10 July 2014, https://doi.org/10.1161/STROKEAHA.114.003689.

[9] Desowska, Adela and Turner, Duncan L.. *Dynamics of brain connectivity after stroke,* Reviews in the Neurosciences, vol. 30, no. 6, 2019, pp. 605-623. https://doi.org/10.1515/revneuro-2018-0082

[10] Brian Butterworth, Vincent Walsh, *Neural basis of mathematical cognition*, Current Biology, Volume 21, Issue 16, 2011, Pages R618-R621, ISSN 0960-9822, https://doi.org/10.1016/j.cub.2011.07.005.

[11] Semenza C, Meneghello F, Arcara G, Burgio F, Gnoato F, Facchini S, Benavides-Varela S, Clementi M, Butterworth B. *A new clinical tool for assessing numerical abilities in neurological diseases: numerical activities of daily living*. Front Aging Neurosci. 2014 Jun 20;6:112. doi: 10.3389/fnagi.2014.00112.

[12] Benn, Y., Jayes, M., Casassus, M., Williams, M., Jenkinson, C., McGowan, E., & Conroy, P. . *A qualitative study into the experience of living with acalculia after stroke and other forms of acquired brain injury*. Neuropsychological Rehabilitation, 33(9), 1512–1536, 2022. https://doi.org/10.1080/09602011.2022.2108065

[13] Silvia Benavides-Varela, Francesca Burgio, Luca Weis, Micaela Mitolo, Katie Palmer, Roberta Toffano, Giorgio Arcara, Antonino Vallesi, Dante Mantini, Francesca Meneghello, Carlo Semenza, *The role of limbic structures in financial abilities of mild cognitive impairment patients*, NeuroImage: Clinical, Volume 26, 2020, 102222, ISSN 2213-1582, https://doi.org/10.1016/j.nicl.2020.102222.

[14] Burgio, F., Danesin, L., Wennberg, A. et al. *Financial and numerical abilities: patterns of dissociation in neurological and psychiatric diseases*. Neurol Sci 45, 4779–4787 (2024). https://doi.org/10.1007/s10072-024-07610-9.

[15] Mayer E, Reicherts M, Deloche G, et al. *Number processing after stroke: Anatomoclinical correlations in oral and written codes*. Journal of the International Neuropsychological Society. 2003;9(6):899-912. doi:10.1017/S1355617703960103.

[16] Moeller Korbinian , Willmes Klaus , Klein Elise, *A review on functional and structural brain connectivity in numerical cognition, Frontiers in Human Neuroscience*, Volume 9, 2015, doi:10.3389/fnhum.2015.00227.

[17] Dorian Pustina, Brian Avants, Olufunsho K. Faseyitan, John D. Medaglia, H. Branch Coslett, *Improved accuracy of lesion to symptom mapping with multivariate sparse canonical correlations*, Neuropsychologia, Volume 115, 2018, Pages 154-166, ISSN 0028-3932, https://doi.org/10.1016/j.neuropsychologia.2017.08.027.

[18] Bates, E., Wilson, S., Saygin, A. et al. *Voxel-based lesion–symptom mapping*. Nat Neurosci 6, 448–450 (2003). https://doi.org/10.1038/nn1050.

[19] Griffis, J.C., Nenert, R., Allendorfer, J.B., Szaflarski, J.P., 2017a. *Damage to white matter bottlenecks contributes to language impairments after left hemispheric stroke*. NeuroImage Clin. 14, 552–565. https://doi.org/10.1016/j.nicl.2017.02.019.

[20] Malherbe C, Umarova RM, Zavaglia M, Kaller CP, Beume L, Thomalla G, Weiller C, Hilgetag CC. *Neural correlates of visuospatial bias in patients with left hemisphere stroke: a causal functional contribution analysis based on game theory*. Neuropsychologia. 2018 Jul 1;115:142-153. Doi: 10.1016/j.neuropsychologia.2017.10.013. Epub 2017 Oct 12.

[21] Findlater SE, Hawe RL, Mazerolle EL, et al. *Comparing CST Lesion Metrics as Biomarkers for Recovery of Motor and Proprioceptive Impairments After Stroke*. Neurorehabilitation and Neural Repair. 2019;33(10):848-861. doi:10.1177/1545968319868714.

[22] Kuceyeski A, Navi BB, Kamel H, Raj A, Relkin N, Toglia J, Iadecola C, O'Dell M. *Structural connectome disruption at baseline predicts 6-months post-stroke outcome*. Hum Brain Mapp. 2016 Jul;37(7):2587-601. doi: 10.1002/hbm.23198. Epub 2016 Mar 26.

[23] Joseph C. Griffis, Nicholas V. Metcalf, Maurizio Corbetta, Gordon L. Shulman, *Lesion Quantification Toolkit: A MATLAB software tool for estimating grey matter damage and white matter disconnections in patients with focal brain lesions*, NeuroImage: Clinical, Volume 30, 2021, 102639, ISSN 2213-1582, https://doi.org/10.1016/j.nicl.2021.102639.

[24] Siegel JS, Ramsey LE, Snyder AZ, Metcalf NV, Chacko RV, Weinberger K, Baldassarre A, Hacker CD, Shulman GL, Corbetta M. *Disruptions of network connectivity predict impairment in multiple behavioral domains after stroke*. Proc Natl Acad Sci U S A. 2016 Jul 26;113(30):E4367-76. doi: 10.1073/pnas.1521083113. Epub 2016 Jul 11.

[25] Yee-Haur Mah, Masud Husain, Geraint Rees, Parashkev Nachev, *Human brain lesion-deficit inference remapped, Brain*, Volume 137, Issue 9, September 2014, Pages 2522–2531, https://doi.org/10.1093/brain/awu164.

[26] Yourganov G, Fridriksson J, Rorden C, Gleichgerrcht E, Bonilha L., *Multivariate Connectome-Based Symptom Mapping in Post-Stroke Patients: Networks Supporting Language and Speech*. J Neurosci. 2016 Jun 22;36(25):6668-79. doi: 10.1523/JNEUROSCI.4396-15.2016.

[27] Pustina, D., Coslett, H.B., Ungar, L., Faseyitan, O.K., Medaglia, J.D., Avants, B. and Schwartz, M.F. , *Enhanced estimations of post-stroke aphasia severity using stacked multimodal predictions*. Hum. Brain Mapp., 2017, 38: 5603-5615. https://doi.org/10.1002/hbm.23752.

[28] Alexandre Croquelois, Julien Bogousslavsky; *Stroke Aphasia: 1,500 Consecutive Cases*. Cerebrovasc Dis 1 March 2011; 31 (4): 392–399. https://doi.org/10.1159/000323217.

[29] Rojkova K, Volle E, Urbanski M, Humbert F, Dell'Acqua F, Thiebaut de Schotten M. *Atlasing the frontal lobe connections and their variability due to age and education: a spherical deconvolution tractography study*. Brain Struct Funct. 2016 Apr;221(3):1751-66. doi: 10.1007/s00429-015-1001-3. Epub 2015 Feb 15.

[30] B. T. Thomas Yeo, Fenna M. Krienen, Jorge Sepulcre, Mert R. Sabuncu, Danial Lashkari, Marisa Hollinshead, Joshua L. Roffman, Jordan W. Smoller, Lilla Zöllei, Jonathan R. Polimeni, Bruce Fischl, Hesheng Liu, and Randy L. Buckner, *The organization of the human cerebral cortex estimated by intrinsic functional connectivity*, Journal of Neurophysiology 2011 106:3, 1125-1165, https://doi.org/10.1152/jn.00338.2011.

[31] Umi Nabilah Ismail, Noorazrul Yahya, Hanani Abdul Manan, *Investigating functional connectivity related to stroke recovery: A systematic review*, Brain Research, Volume 1840, 2024, 149023, ISSN 0006-8993, https://doi.org/10.1016/j.brainres.2024.149023.

[32] Fotiadis, P., Parkes, L., Davis, K.A. et al. *Structure–function coupling in macroscale human brain networks*. Nat. Rev. Neurosci. 25, 688–704 (2024). https://doi.org/10.1038/s41583-024-00846-6

[33] E. W. Lang, A. M. Tomé, I. R. Keck, J. M. Górriz-Sáez, and C. G. Puntonet, *Brain connectivity analysis: A short survey*, Comput. Intell. Neurosci., vol. 2012, 2012, doi: 10.1155/2012/412512.

[34] Batista-García-Ramó K, Fernández-Verdecia CI. *What We Know About the Brain Structure-Function Relationship*. Behav Sci (Basel). 2018 Apr 18;8(4):39. doi: 10.3390/bs8040039.

[35] Joseph C. Griffis, Nicholas V. Metcalf, Maurizio Corbetta, Gordon L. Shulman, *Lesion Quantification Toolkit: A MATLAB software tool for estimating grey matter damage and white matter disconnections in patients with focal brain lesions*, NeuroImage: Clinical, Volume 30, 2021, 102639, https://doi.org/10.1016/j.nicl.2021.102639.

[36] Laura E. Suárez, Ross D. Markello, Richard F. Betzel, Bratislav Misic, *Linking Structure and Function in Macroscale Brain Networks*, Trends in Cognitive Sciences, Volume 24, Issue 4, 2020, Pages 302-315, ISSN 1364-6613, https://doi.org/10.1016/j.tics.2020.01.008.

[37] C. J. Honey, O. Sporns, L. Cammoun, X. Gigandet, J. P. Thiran, R. Meuli, and P. Hagmann. *Predicting human resting-state functional connectivity from structural connectivity*, Proceedings of the National Academy of Sciences, Vol. 106, No. 6, February 10, 2009. Doi: https://doi.org/10.1073/pnas.0811168106.

[38] Lu Zhang, Li Wang, Dajiang Zhu, *Predicting brain structural network using functional connectivity, Medical Image Analysis*, Volume 79, 2022, 102463, ISSN 1361-8415, https://doi.org/10.1016/j.media.2022.102463.

[39] Honey CJ, Thivierge JP, Sporns O. *Can structure predict function in the human brain?* Neuroimage. 2010 Sep;52(3):766-76. doi: 10.1016/j.neuroimage.2010.01.071. Epub 2010 Jan 29.

[40] Rasmus M. Birn, Erin K. Molloy, Rémi Patriat, Taurean Parker, Timothy B. Meier, Gregory R. Kirk, Veena A. Nair, M. Elizabeth Meyerand, Vivek Prabhakaran, *The effect of scan length on the reliability of resting-state fMRI connectivity estimates*, NeuroImage, Volume 83, 2013, Pages 550-558, ISSN 1053-8119, https://doi.org/10.1016/j.neuroimage.2013.05.099.

[41] Griffis JC, Metcalf NV, Corbetta M, Shulman GL. *Structural Disconnections Explain Brain Network Dysfunction after Stroke*. Cell Rep. 2019 Sep 3;28(10):2527-2540.e9. doi: 10.1016/j.celrep.2019.07.100.

[42] Burgio, F., Danesin, L., Benavides-Varela, S. et al. *Numerical activities of daily living: a short version*. Neurol Sci 43, 967–978 (2022). https://doi.org/10.1007/s10072-021-05391-z.

[43] Toffano, R., Burgio, F., Palmer, K. et al. *Numerical Activities of Daily Living – Financial: a short version*. Neurol Sci 42, 4183–4191 (2021). https://doi.org/10.1007/s10072-021-05047-y.

[44] Arcara, G., Burgio, F., Benavides-Varela, S., Toffano, R., Gindri, P., Tonini, E., … Semenza, C. (2017). *Numerical Activities of Daily Living – Financial (NADL-F): A tool for the assessment of financial capacities*. Neuropsychological Rehabilitation, 29(7), 1062–1084. https://doi.org/10.1080/09602011.2017.1359188.

[45] Pustina, D., Coslett, H.B., Turkeltaub, P.E., Tustison, N., Schwartz, M.F. and Avants, B. (2016), *Automated segmentation of chronic stroke lesions using LINDA: Lesion identification with neighborhood data analysis*. Hum. Brain Mapp., 37: 1405-1421. https://doi.org/10.1002/hbm.23110.

[46] Thye M, Szaflarski JP, Mirman D. *Shared lesion correlates of semantic and letter fluency in post-stroke aphasia*. J Neuropsychol. 2021 Mar;15(1):143-150. doi: 10.1111/jnp.12211. Epub 2020 May 15.

[47] Alexander Schaefer, Ru Kong, Evan M Gordon, Timothy O Laumann, Xi-Nian Zuo, Avram J Holmes, Simon B Eickhoff, B T Thomas Yeo, *Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI*, Cerebral Cortex, Volume 28, Issue 9, September 2018, Pages 3095–3114, https://doi.org/10.1093/cercor/bhx179.

[48] Edmund T. Rolls, Chu-Chung Huang, Ching-Po Lin, Jianfeng Feng, Marc Joliot, *Automated anatomical labelling atlas 3*, NeuroImage, Volume 206, 2020, 116189, ISSN 1053-8119, https://doi.org/10.1016/j.neuroimage.2019.116189.

[49] Yeh FC, Panesar S, Fernandes D, Meola A, Yoshino M, Fernandez-Miranda JC, Vettel JM, Verstynen T. *Population-averaged atlas of the macroscale human structural connectome and its network topology*. Neuroimage. 2018 Sep;178:57-68. doi: 10.1016/j.neuroimage.2018.05.027. Epub 2018 May 24.

[50] Leland McInnes, John Healy, James Melville. *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*, 2020, 1802.03426, arXiv, stat.ML, https://arxiv.org/abs/1802.03426.

[51] Smith SM, Jenkinson M, Woolrich MW, Beckmann CF, Behrens TEJ, Johansen-Berg H, et al. *Advances in functional and structural MR image analysis and implementation as FSL.* Neuroimage 2004;23. https://doi.org/10.1016/j.neuroimage.2004.07.051.

[52] Mark Jenkinson, Peter Bannister, Michael Brady, Stephen Smith, *Improved Optimization for the Robust and Accurate Linear Registration and Motion Correction of Brain Images*, NeuroImage, Volume 17, Issue 2, 2002, Pages 825-841, ISSN 1053-8119, https://doi.org/10.1006/nimg.2002.1132.

[53] Jesper L.R. Andersson, Stefan Skare, John Ashburner, *How to correct susceptibility distortions in spin-echo echo-planar images: application to diffusion tensor imaging*, NeuroImage, Volume 20, Issue 2, 2003, Pages 870-888, ISSN 1053-8119, https://doi.org/10.1016/S1053-8119(03)00336-7.

[54] Susan Whitfield-Gabrieli and Alfonso Nieto-Castanon, *Conn: A Functional Connectivity Toolbox for Correlated and Anticorrelated Brain Networks, Brain Connectivity*, Vol. 2, No. 3,2012, doi: https://doi.org/10.1089/brain.2012.0073.

[55] Xiaoqian J. Chai, Alfonso Nieto Castañón, Dost Öngür, Susan Whitfield-Gabrieli, *Anticorrelations in resting state networks without global signal regression*, NeuroImage, Volume 59, Issue 2, 2012, Pages 1420-1428, ISSN 1053-8119, https://doi.org/10.1016/j.neuroimage.2011.08.048.

[56] Michael N. Hallquist, Kai Hwang, Beatriz Luna, *The nuisance of nuisance regression: Spectral misspecification in a common approach to resting-state fMRI preprocessing reintroduces noise and obscures functional connectivit*y, NeuroImage, Volume 82, 2013, Pages 208-225, ISSN 1053-8119, https://doi.org/10.1016/j.neuroimage.2013.05.116.

[57] Bing Liu, Xiaohan Tian, Yingjie Peng et al. *Deciphering complex brain spatiotemporal dynamics shaping diverse human behavior*, 13 October 2023, PREPRINT (Version 1) available at Research Square, https://doi.org/10.21203/rs.3.rs-3344208/v1.

[58] van Meer MP, van der Marel K, Wang K, Otte WM, El Bouazati S, Roeling TA, Viergever MA, Berkelbach van der Sprenkel JW, Dijkhuizen RM. *Recovery of sensorimotor function after experimental stroke correlates with restoration of resting-state interhemispheric functional connectivity*. J Neurosci. 2010 Mar 17;30(11):3964-72. doi: 10.1523/JNEUROSCI.5709-09.2010.

[59] Siegel JS, Snyder AZ, Ramsey L, Shulman GL, Corbetta M. *The effects of dynamic lag on functional connectivity and behavior after stroke*. Journal of Cerebral Blood Flow & Metabolism. 2016;36(12):2162-2176. doi:10.1177/0271678X15614846.

[60] Hotelling, H. (1936). *Relations Between Two Sets of Variates*. Biometrika, 28(3/4), 321–377. https://doi.org/10.2307/2333955.

[61] University College London, last consultation 31/11/2024, https://mlnl.github.io/cca_pls_toolkit/.

[62] Anderson M. Winkler, Olivier Renaud, Stephen M. Smith, Thomas E. Nichols, *Permutation inference for canonical correlation analysis, NeuroImage*, Volume 220, 2020, 117065, ISSN 1053-8119, https://doi.org/10.1016/j.neuroimage.2020.117065.

[63] Leland McInnes, last consultation: 31/11/2024, https://umap-learn.readthedocs.io/en/latest/how_umap_works.html.

[64] Agoston Mihalik, James Chapman, Rick A. Adams, Nils R. Winter, Fabio S. Ferreira, John Shawe-Taylor, Janaina Mourão-Miranda, *Canonical Correlation Analysis and Partial Least Squares for Identifying Brain–Behavior Associations: A Tutorial and a Comparative Study*, Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, Volume 7, Issue 11, 2022, Pages 1055-1067, ISSN 2451-9022, https://doi.org/10.1016/j.bpsc.2022.07.012.

[65] Alessandra Griffa, Enrico Amico, Raphaël Liégeois, Dimitri Van De Ville, Maria Giulia Preti, *Brain structure-function coupling provides signatures for task decoding and individual fingerprinting*, NeuroImage, Volume 250, 2022, 118970, ISSN 1053-8119, https://doi.org/10.1016/j.neuroimage.2022.118970.

[66] Gonzalez-Castillo J, Fernandez IS, Lam KC, Handwerker DA, Pereira F, Bandettini PA. *Manifold learning for fMRI time-varying functional connectivity*. Front Hum Neurosci. 2023 Jul 11;17:1134012. doi: 10.3389/fnhum.2023.1134012.

[67] Jimenez-Marin, A., De Bruyn, N., Gooijers, J. et al. *Multimodal and multidomain lesion network mapping enhances prediction of sensorimotor behavior in stroke patients*. Sci Rep 12, 22400 (2022). https://doi.org/10.1038/s41598-022-26945-x.

[68] Tustison NJ, Avants BB, Cook PA, Zheng Y, Egan A, Yushkevich PA, Gee JC. *N4ITK: improved N3 bias correction*. IEEE Trans Med Imaging. 2010 Jun;29(6):1310-20. doi: 10.1109/TMI.2010.2046908. Epub 2010 Apr 8.

[69] Tustison NJ, Avants BB. *Explicit B-spline regularization in diffeomorphic image registration*. Front Neuroinform. 2013 Dec 23;7:39. doi: 10.3389/fninf.2013.00039.

[70] Avants BB, Tustison NJ, Song G, Cook PA, Klein A, Gee JC. *A reproducible evaluation of ANTs similarity metric performance in brain image registration*. Neuroimage. 2011 Feb 1;54(3):2033-44. doi: 10.1016/j.neuroimage.2010.09.025. Epub 2010 Sep 17.

[71] Vladimir Fonov, Alan C. Evans, Kelly Botteron, C. Robert Almli, Robert C. McKinstry, D. Louis Collins, Unbiased average age-appropriate atlases for pediatric studies, NeuroImage, Volume 54, Issue 1, 2011, Pages 313-327, ISSN 1053-8119, https://doi.org/10.1016/j.neuroimage.2010.07.033.

[72] Diaz-Papkovich, A., Anderson-Trocmé, L. & Gravel, S. *A review of UMAP in population genetics*. J Hum Genet 66, 85–91 (2021). https://doi.org/10.1038/s10038-020-00851-4.

[73] Amit Pandey, Achin Jain, *Comparative Analysis of KNN Algorithm using Various Normalization Techniques*, International Journal of Computer Network and Information Security(IJCNIS), Vol.9, No.11, pp.36-42, 2017. DOI:10.5815/ijcnis.2017.11.04.

[74] Lia Talozzi, Stephanie J Forkel, Valentina Pacella, Victor Nozais, Etienne Allart, Céline Piscicelli, Dominic Pérennou, Daniel Tranel, Aaron Boes, Maurizio Corbetta, Parashkev Nachev, Michel Thiebaut de Schotten, *Latent disconnectome prediction of long-term cognitive-behavioural symptoms in stroke*, Brain, Volume 146, Issue 5, May 2023, Pages 1963–1978, https://doi.org/10.1093/brain/awad013.

[75] Margaret Jane Moore, Jason B. Mattingley, Nele Demeyere, *Multivariate and network lesion mapping reveals distinct architectures of domain-specific post-stroke cognitive impairments*, Neuropsychologia, Volume 204, 2024, https://doi.org/10.1016/j.neuropsychologia.2024.109007.

[76] Griffith HR, Stewart CC, Stoeckel LE, Okonkwo OC, den Hollander JA, Martin RC, Belue K, Copeland JN, Harrell LE, Brockington JC, Clark DG, Marson DC. *Magnetic resonance imaging volume of the angular gyri predicts financial skill deficits in people with amnestic mild cognitive impairment*. J Am Geriatr Soc. 2010 Feb;58(2):265-74. doi: 10.1111/j.1532-5415.2009.02679.x. Epub 2010 Jan 26.

[77] Julian Klingbeil, Max Wawrzyniak, Anika Stockert, Dorothee Saur. *Resting-state functional connectivity: An emerging method for the study of language networks in post-stroke aphasia*, Brain and Cognition, Volume 131, 2019, Pages 22-33, https://doi.org/10.1016/j.bandc.2017.08.005.

[78] Brancaccio, A.; Tabarelli, D.; Belardinelli, P. *A New Framework to Interpret Individual Inter-Hemispheric Compensatory Communication after Stroke*. J. Pers. Med. 2022, 12, 59. https://doi.org/10.3390/jpm12010059.