UCLouvain

École polytechnique de Louvain

Machine learning analysis of cerebral microstructure to find biomarkers of vertigo

Author: **Rimez Dany**
Supervisors: **Macq Benoît, Dricot Laurence**
Readers: **Crèvecoeur Frédéric, Dessain Quentin, Grégoire Anais**
Academic year 2021–2022
Master [120] in Biomedical Engineering

Abstract

Machine learning and tractography are methods that have gained interest in the previous years to study brain microstructure. The purpose of this master thesis is to use those tools to find biomarkers for vertigo in hearing-impaired patients, while maintaining a maximum interpretability of the obtained results.

A workflow was created based on previously existing methods to produce bundles and analyse DTI measures in precise locations of these bundles. Those tract-wise analysis were also compared with regions-wise analysis of the microstructural metrics, structural lateralization of key bundles, and FSL's statistical analysis.

Acknowledgements

I would like to thank Benoit Macq and Laurence Dricot for giving me the opportunity to work on this great subject during my master's thesis.

I would also thank Quentin Dessain, Nicolas Delinte, Anaïs Grégoire and Naïma Deggouj for the time they took to answer my questions.

Contents

Abreviations

1 Introduction **1**

2 Theoretical background **2**
2.1 Magnetic resonance imaging 2
2.2 Tractography 5
2.2.1 Tractometry 9
2.3 A bit of neuroanatomy 13
2.3.1 Thalamo-cortical Network 16
2.4 Vertigo 17
2.4.1 Peripheric vestibular system 18
2.4.2 Symptoms and tests 19
2.4.3 Previous studies 21
2.5 Machine Learning 23
2.5.1 Preprocessing 25
2.5.2 Models 26
2.5.3 Evaluation 30

3 Methods **33**
3.1 Data preprocessing 34
3.2 Tractography 37
3.3 Tractometry 40
3.4 Analysis 45
3.4.1 Structural lateralization 45
3.4.2 Machine learning 46

4 Results **52**
4.1 Age influence 52
4.2 Elikopy statistical analysis 53
4.3 Machine learning analysis 56

4.3.1 White-matter parcellation 57
4.3.2 Tract-based analysis 61
4.4 Structural lateralization 66
4.5 Results summary and interpretation 68

5 Discussion and future perspectives **71**
5.1 Tractography 71
5.2 Tract segmentation and Tractometry 72
5.3 Machine Learning 73

6 Conclusion **77**

Bibliography **78**

Annexes **87**
6.1 White Matter Query Language 87
6.2 Cleaning for probabilistic tractogram 88
6.3 Supplementary images 92
6.3.1 Theoretical background 92

Abreviations

AD Axial diffusivity.

AF Arcuate Fasciculus.

AFQ Automated Fiber Quantification.

ANTs Advanced Normalization Tools.

ARC Arcuate Fasciculus.

ATR Anterior Thalamic Radiation.

BPPV Benign Paroxysmal Positional Vertigo.

BUAN BUndle ANalytics.

CC Corpus Callosum.

CSD Constrained Spherical Deconvolution.

CSF Cerebro Spinal Fluid.

CST Cortico-Spinal Tract.

dMRI Diffusion Magnetic Resonance Imaging.

DTI Diffusion Tensor Imaging.

FA Fractional anisotropy.

FBC Fiber to Bundle Coherence.

FDR False Discovery Rate.

FSL FMRIB Software Library.

IFOF Inferior Fronto-Occipital Fasciculus.

ILF Inferior Longitudinal Faciculus.

MD Mean diffusivity.

MdLF Middle Longitudinal Fasciculus.

MNI Montreal Neurological Institute-Hospital.

MRI Magnetic Resonance Imaging.

NODDI Neurite Orientation Dispersion and Density Imaging.

ODF Orientation Distribution Function.

OR Optic Radiation.

PCA Principal Component Analysis.

PIVC Parietal-Insular Vestibular Cortex.

RD Radial diffusivity.

RDF Random Forest.

RecoBundles RecoBundles.

RFBC Relative Fiber to Bundle Coherence.

ROI Region Of Interest.

SGL Sparse-Group Lasso.

SIFT Spherical-deconvolution Informed Filtering.

SLF Superior Longitudinal Faciculus.

SLR streamline-based registration.

SVM Support Vector Machine.

TBSS Tract-Based Spatial Statistics.

UF Uncinate Fasciculus.

VEMP Vestibular-evoked myogenic potentials.

VHIT Head Impulse test.

VNG Videonystagmograpy.

VOR Vestibulo-ocular Reflex.

WMQL White matter query language.

Chapter 1

Introduction

Vertigo is a common symptom that affects up to 20% of adults yearly, and lead to an interruption in daily activities 80% of the time [1]. This is a daily matter of concern for a few hearing-impaired patients included in an on-going clinical research. However, the vertigo symptoms were not yet investigated on this population.

The purpose of this master thesis is then to find biomarkers for vertigo in those patients. To reach this goal, machine-learning tools will be used to analyse the cerebral microstructure of the patients based on Diffusion Magnetic Resonance Imaging (dMRI). One of the main goals wll be to maintain a maximum interpretability of the results.

Tractography and tractometry, that have gained in interest in the recent years, will be used to produce the data to analyse, and a comparison with other common methods (based on brain parcellation, and Tract-Based Spatial Statistics (TBSS)) will be done.

This thesis will be divided in several parts. First, an overview of the theoretical background needed to understand the methods used. Second, the presentation of the means developed to obtain the results. Third, the presentation of those results, and finally their discussion.

Chapter 2

Theoretical background

2.1 Magnetic resonance imaging

Magnetic Resonance Imaging (MRI) is an imaging technique that is really an asset in medical applications as it allows various imaging modalities without implying any radiation. It relies on the magnetic properties of protons to create different types of images contrasts. As shown on figure 2.1, at rest protons have randomly oriented individual magnetic moments μ, which sum to 0 and result in a null magnetization M of the matter.

However, when they are placed in a magnetic field B_0, they align to this magnetic field and form two populations: parallel (low energy) and antiparallel (high energy) to the magnetic field. The parallel population will contain a few more protons than the other population, causing the magnetisation of the matter.

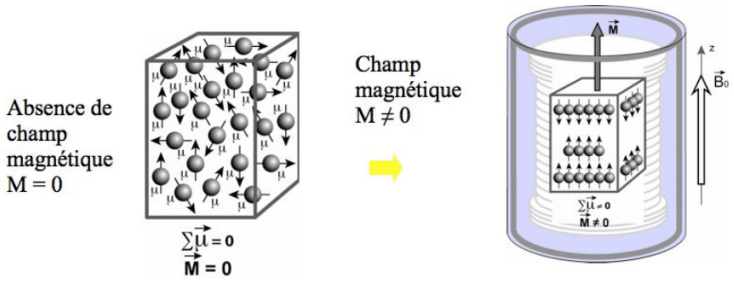

Figure 2.1: Effect of the magnetic field of protons' spin oriantation [2].

But protons are quantum particles, having an angular momentum (spin) that will interact with the magnetic field, resulting in a torque which causes the proton to precess around the magnetic field instead of perfectly aligning with it. This pre-

cession occurs at a particular frequency, the Larmor frequency, that is proportional to the magnetic field and the gyromagnetic ration (γ):

$$\omega_0 = \gamma B_0$$

When the protons are excited by a Radio-Frequence pulse, some will switch to a higher energy state, becoming in phase with one another and the mean magnetization switch from longitudinal (Mz on figure 2.2) to transverse (Mxy). Another precession is started around the transverse direction besides the precession around B_0. T1 and T2 relaxation times are unique for each tissue, tightly linked to the chemical nature of tissues. We can then measure the magnetization at different times to play with its relaxation profile and obtain different contrasts.

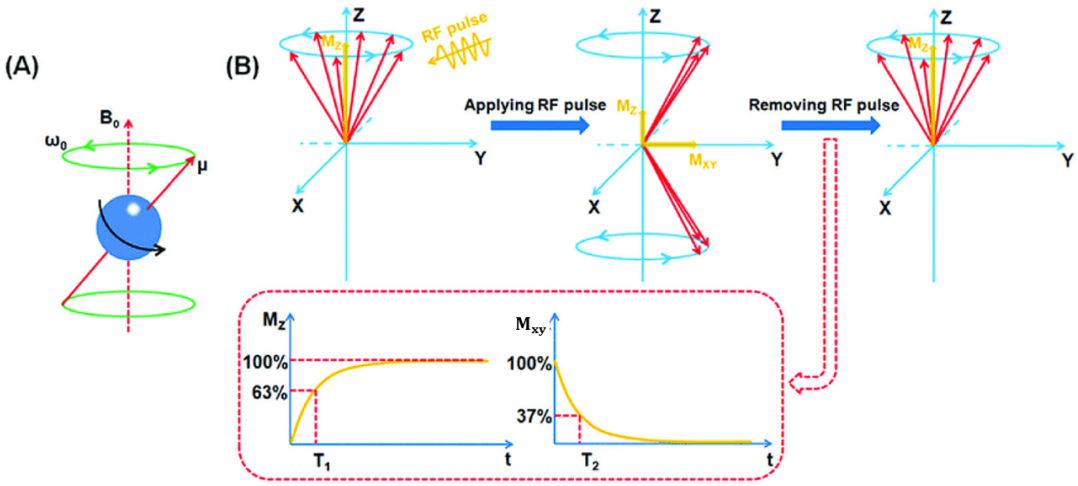

Figure 2.2: Schematic illustration of the mechanisms of MRI. (A) Protons precession, (B) after RF pulse excitation, and the graphical representation of T_1 relaxation and T_2 relaxation, adapted from [3].

There are a lot of different MRI sequences using unique sequences of excitations and measure, as well as gradient (secondary magnetic fields used to encode the voxels of the image) variations to obtain those contrasts.

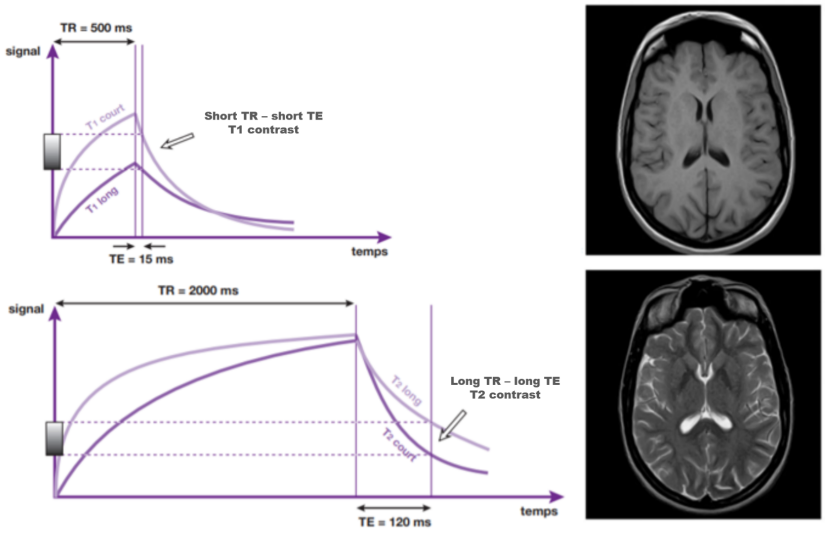

Figure 2.3: Examples of MRI sequences parameters (TR is the time between 2 RF pulse and TE the time between the RF pulse and the signal measurement) for T1 (above) and T2 (below) contrasts [4]

Diffusion MRI Diffusion MRI is just another MRI modality that measures the motion of water molecules in tissues. It is based on the use of two RF pulses: one that is followed by a gradient dephasing the protons, and a second one followed by a gradient that rephases the protons. If the protons are highly mobile, they will not be in the same position at the second gradient (measurement) than at the first one and the signal will be low (as gradients are used to encode voxels position), while it will be high in voxels in which water diffusion is limited.

From diffusion MRI, we can evaluate tissue's microstructure in many ways, making MRI a precious tool for non-invasive microstructure probing. Many mathematical models try to evaluate it using some assumption about the microstructure and its link with the dMRI signal.

Diffusion tensor imaging Diffusion Tensor Imaging (DTI) is one of those mathematical models. It is a model that fits a 2-dimensional tensor to the dMRI data in order to extract 3D information about the diffusion of water in cerebral tissues.

$$D = \begin{pmatrix} D_{xx} & D_{xy} & D_{xz} \\ D_{yx} & D_{yy} & D_{yz} \\ D_{zx} & D_{zy} & D_{zz} \end{pmatrix}$$

This tensor is symmetric ($D_{xy} = D_{yx}$) and we can get several metrics of interest from it, with λ_i the ith eigen value of the diffusion tensor:

- Axial diffusivity (AD) is the first eigen value (λ_1 on figure 2.4), the mobility of water along the preferred axis of diffusion (λ_1 on figure 2.4)
- Radial diffusivity (RD) is the mean of the two other eigen values, the mobility of water orthogonally to the preferred axis of diffusion ($\frac{\lambda_1+\lambda_2}{2}$)
- Mean diffusivity (MD) quantifies the mobility of water molecules, independently of their diffusion direction ($\frac{\lambda_1+\lambda_2+\lambda_3}{3}$)
- Fractional anisotropy (FA) quantifies how much water molecues diffuse along a common direction. it ranges from 0 (perfectly isotropic) to 1 (perfectly anisotropic) ($\sqrt{\frac{1}{2}\frac{(\lambda_1-\lambda_2)^2+(\lambda_2-\lambda_3)^2+(\lambda_3-\lambda_1)^2}{\lambda_1+\lambda_2+\lambda_3}}$)

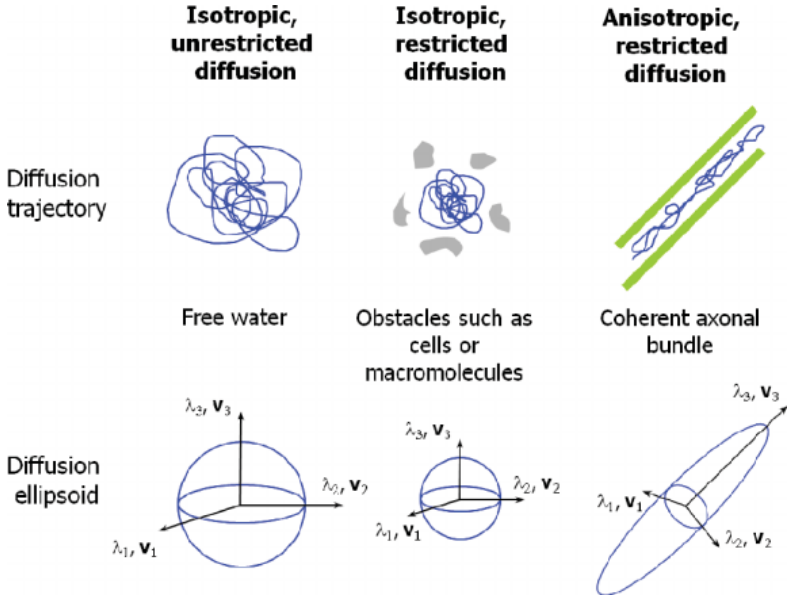

Figure 2.4: DTI metrics schematics [5].

2.2 Tractography

One way of studying dMRI data is using the information of water diffusion across brain structures to reconstruct the white matter fibres. Those reconstructed fibres would allow retrieving sub-voxel information, measuring information along tracts of interest, structural connectivity analysis, etc ...

There are many ways to achieve this reconstruction and obtain a so-called tractogram (see figure 2.5), using different approaches (local or global tracking) and mathematical algorithms based on microstructure metrics (DTI, Neurite Orientation Dispersion and Density Imaging (NODDI) or composite), or directly on dMRI images, complemented by graph-theory, mathematical or anatomical constraints to obtain the most relevant results.

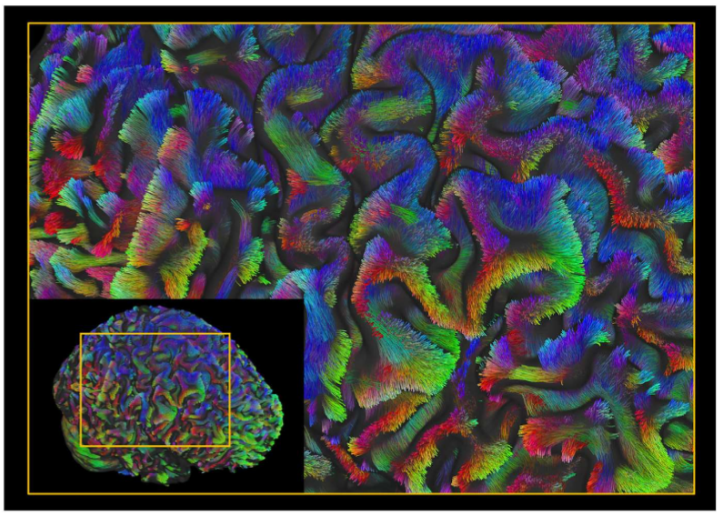

Figure 2.5: Example of a whole brain tractogram produced with high-quality data [6].

In this part will be presented the notions that will be used in the following sections when it comes to tractography.

Local tracking Local tracking implies propagating a streamline, starting from a seed point, along a direction given by the Orientation Distribution Function (ODF) and with a given step length. The ODF field can be interpreted as the probability, for each considered direction, that water diffuses in this specific direction in a given voxel (see figure 2.6).

At each iteration, we lengthen the streamline by both of its ends. But we also need to know when to stop a streamline growth as explained a bit later 2.2. The main drawback of this paradigm is the accumulation of trajectory errors along the streamline, we then need to find the right step size: a small step is often chosen for this reason, though we are limited by the resolution of the image and taking a step too small will only impair the results.

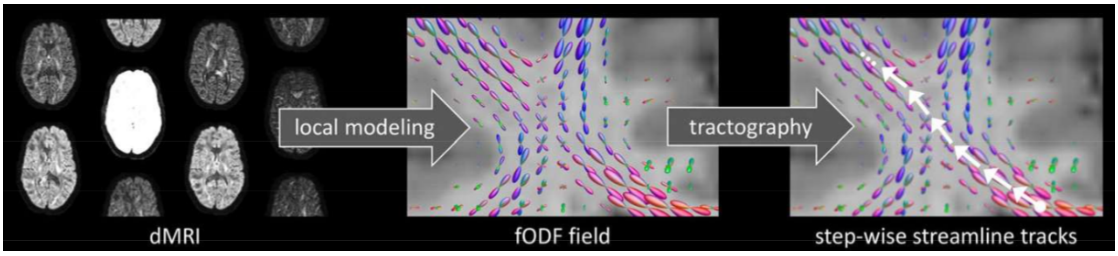

Figure 2.6: Local tracking algorithm: the algorithm follows the ODF the reconstruct a streamlines pathway [6].

From there, there will be two kinds of tracking algorithms: deterministic and probabilistic ones. With deterministic tracking, direction at each step is chosen as the most probable one (hence the name "deterministic" as a given seed will always give the same streamline), while probabilistic algorithms select at random the next direction according to the ODF: the most probable is more likely to be chosen, but it is unlikely (not to say impossible) to obtain the same streamline by using a given seed again and again. It is then usual to set a maximum for the angle change between two steps for probabilistic algorithms to have streamlines with reasonable curvature.

Constrained spherical deconvolution Constrained Spherical Deconvolution (CSD) is one of the mathematical models that can be used to compute the ODF. It uses a tissue response [1] to produce a sharp ODF based on tissue properties, by deconvolution of the dMRI images with this response. If the tissue response is accurately estimated, this leads to an improved ODF with less ambiguous directions where several tracts cross, allowing to produce not only simple short trajectories but also longer and complexer ones [6]. But a perfect estimation is difficult to find and often patient-specific, especially with low b-values and low resolution images.

There are again many ways to find estimate tissues (white and gray matter, Cerebro Spinal Fluid (CSF), ...) responses (see 2.7), but single-shell dMRI only allows single-tissue response estimation, we then try to find a good estimate for white matter. The most straightfull technique is to compute this response on voxels where the FA is above a certain threshold (but it is then needed to fine-tune this threshold), on voxels within the white-matter mask, or using iterative algorithms that start either from a response estimated with an aforementioned method (Tournier algorithm in MRtrix3 and Dipy or Robust Unbiased Model-BAsed Spherical Deconvolution in Dipy), or from a template response (a "*fat*" one

[1] only white matter for single-shell data, but multiple tissue responses for white and grey matter can be used with multi-shell data.

for the tax algorithm or a sharp one for the Tournier algorithm) [7, 8].

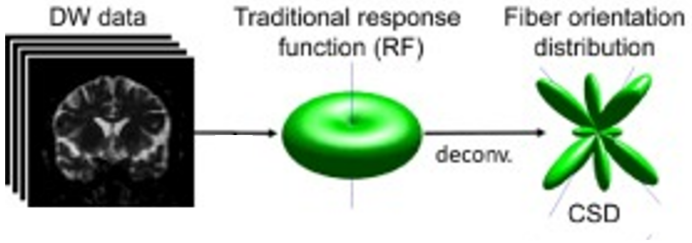

Figure 2.7: Response function and schematics, adapted from [9].

Seeding To generate all streamlines, many seeds are created in the image volume. As shown on figure 2.8, it is often done within the white-matter mask or at the white-matter/grey-matter interface. The latter has the advantage to increase the number of streamlines that gets in the gyri but will more likely produce streamlines that follow incoherent pathways or over-estimated lengths. Indeed, in that zone of the image, the FA is much lower than in deep white matter and chances to find a good seeding point are reduced.

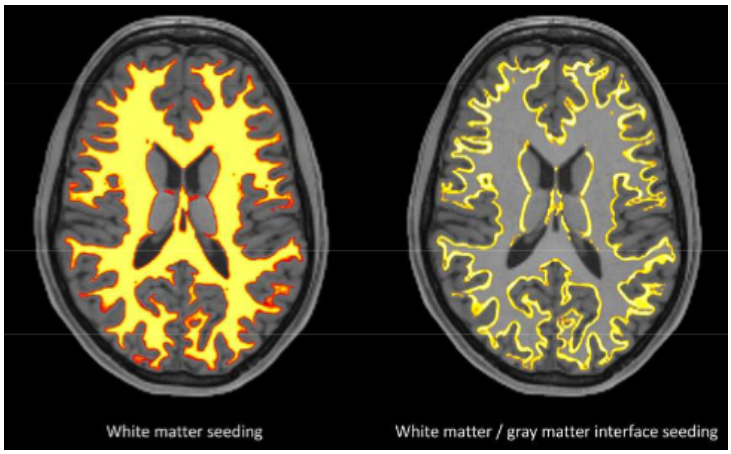

Figure 2.8: Modalities of seeding [6].

Stopping criterion Streamline growth can be stopped by several ways: if it exits the white-matter mask or if the streamline gets in a zone of the image where the FA is under a given threshold for instance. In the first case we would depend on the quality of the used mask as we should not terminate a streamline too soon, while it is always possible to crop it after tracking if it has gone too far in the gray matter. Another possibility is to use anatomical constraints : streamlines that get in CSF or do not connect two gray-matter volumes are either rejected or back-tracked (we

go back of a few iterations and choose new directions and find correct streamlines). It is also possible to get rid of tracts that are too short, knowing they are often spurious and not biologically coherent trajectories. Such anatomically-constrained stopping criterion for tractography has shown to reduce the number of incoherent streamlines in the final result and to improve post-tracking filtering.

That last method, Spherical-deconvolution Informed Filtering (SIFT), relies on data generated by CSD to accurately filter out streamlines where there are many of them, while keeping more where there are less, reducing the bias of tractography towards simple trajectories that are supported by the ODF and improving both fibre density and connection strength estimation. However, this implies to generate a lot of streamlines and keep only a small portion (often 1/10), though there are also seeding approaches that try to put the seeds where they are needed more to improve the ratio of kept streamlines after filtering [10].

Global tracking Like local tracking, global tracking starts by propagating streamlines from seeding points. But rather than lengthening them until a stopping criterion is reached, global tracking stops after a few steps, generating short streamlines that are reoriented and joined to represent the ODF the best. Those methods rely on probabilistic algorithms as Expectation-Maximisation, or graph analysis approaches. However, global tracking is less used and less accessible: nor Dipy or FSL do propose any of them and MRtrix3 has only one algorithm available for multi-shell datas.

2.2.1 Tractometry

Several approaches exist to treat tractography data, like connectomics (creating a graph from the structural connectivity information) or tractometry. Tractometry implies sampling microstructural information along tracts, and extracting features from there, either by taking the mean for a given bundle, or by taking the mean for each segments of a given bundle (see 2.9). Once extracted, statistical analysis can be run on those features to determine if there is a local change in any of the segments in any bundles.

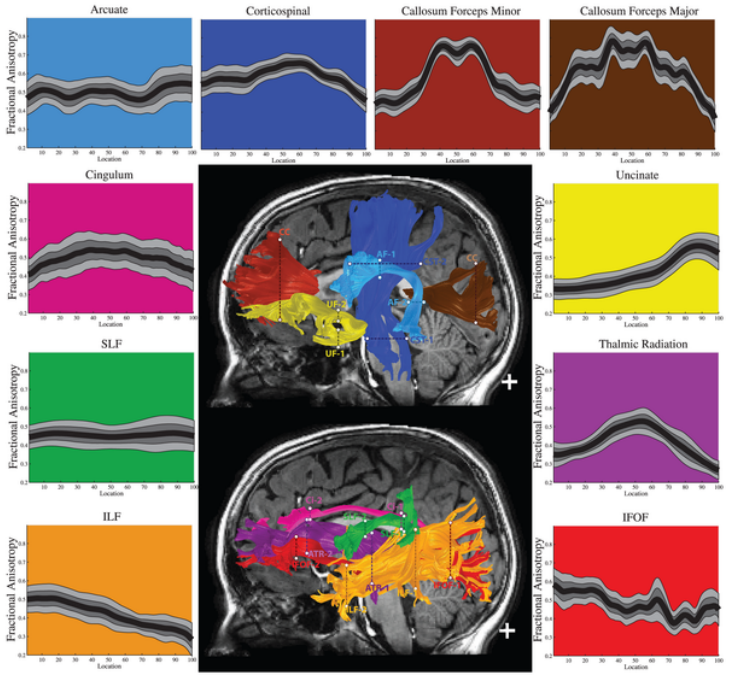

Figure 2.9: Tractometry analysis in a picture: bundles are reconstructed from the dMRI data and tract-profiles are computed for any metric of interest [11]

Tract segmentation The first thing that is needed to be able to evaluate metrics along a given bundle is obviously to identify all the streamlines that belong to that bundle. To do so, it is usual to use include and exclude regions: geometrical shapes like rectangles, disks or spheres that allow to select a streamline if it passes through the roi (include region) or to remove it (exclude region). Those rois can then be used together by using AND (all conditions must be fulfilled) and OR (any condition has to be fulfilled) relationship.

AFQ Automated Fiber Quantification (AFQ) was made to segment deterministic whole-brain tractogram into specific well-defined bundles like Cortico-Spinal Tract (CST) or Arcuate Fasciculus (ARC). The procedure follows several steps that are illustrated on figure 2.10.

First streamlines are attributed to a given bundle according to the Region Of Interest (ROI)s they cross or not. Those are defined in Montreal Neurological Institute-Hospital (MNI) space, streamlines then need to be registered to this space using deformable registration done with Dipy. Hence, all rois that are crossed can be listed for each streamline. Rois are mainly include regions with *AND* relationships, meaning that only streamlines that pass nearby (a distance threshold

can be defined) all ROIs associated with a given bundle are selected. The sole exception is the Superior Longitudinal Faciculus (SLF) that has 1 exclude region while all other bundles only have 2 include regions. Streamlines are also selected whether they cross the mid-line or not: it is only used for corpus callosum tracts as CST has a "do not cross the midline" feature, where "do not cross the midline" means the streamline goes from its start to its end without ever going from one hemisphere to the other.

An optional step after ROI selection is to use probability density maps to refine the selection of streamlines: only streamlines that have a mean probability (mean of the values of the density map that are encountered along the streamline points) above a chosen threshold are kept. In practice, this step is done before ROI selection to fasten the computations. Those density maps are defined a bit like brain atlases using several subjects, except that instead of determining anatomical ROIs, each voxel of a density map comprise the number of streamlines belonging to a given bundle that passes through this voxel. Such visitation count maps are then normalized so that they can be treated as probability, averaged and normalized across subjects to give the resulting probability density maps. Both AFQ and Natbrainlab provide density maps that are generated from deterministic tracts.

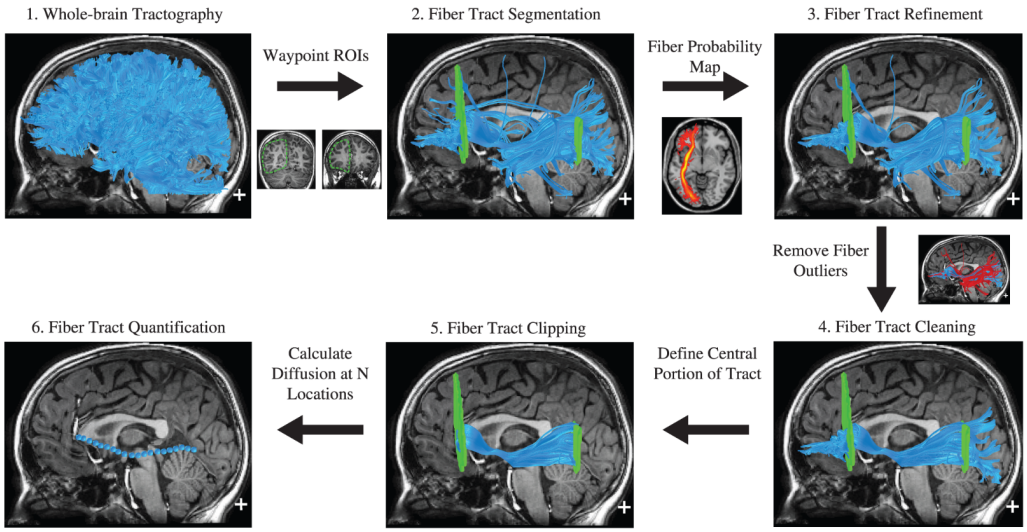

Figure 2.10: AFQ worflow [11].

Once streamlines are segmented into bundles, they are cleaned further. AFQ uses a distance measure to discard streamlines that follows significantly different trajectories from the other streamlines of the bundle. To do so, fibres tracts are

represented as a 3-dimensional gaussian distribution to detect outlier fibres using the Mahalanobis distance:

$$D_M(x) = \sqrt{(x-\mu)^T \Sigma^{-1} (x-\mu)}$$

where x is a 3D point, μ the mean point, and Σ the covariance matrix (which indicates the spread along each axis). Bundle is then splitted into a hundred segment, before using the Mahalanobis distance for each segment to evaluate the distance from other streamlines while taking the segment 3D shape into account by normalizing with the covariance matrix. A streamline is discarded if its mean Mahalanobis distance across segments is 5 standard deviations away from the mean distance of streamlines in the bundle, or if the streamline length is 4 standard deviations away from the mean length [2]. This procedure is followed five times to better filter the data while improving means and standard evaluations at each iteration as incorrect streamlines were discarded. In this process, streamlines are resampled to a hundred points to form the 100 segments over which Mahalanobis distance will be evaluated.

Finally, resulting bundles can[3] be filtered, given the ROIs of a cortical atlase (it uses the AAL atlas by default): a streamline is discarded if one of its endpoints is found farther than a distance threshold from those ROIs. Include and exclude regions can be added to the default ones, either defined in patient or MNI space.

After segmentation, AFQ samples microstructural metrics at each point of the resampled streamlines but does not offer any analysis of the sampled data, so that T-test or other statistical analysis is often done afterwards to detect differences between subjects.

BUAN and RecoBundles It follows the same principles as AFQ: segmentation in bundles, splitting into segments and statistical analysis. However, here the segmentation is performed differently. RecoBundles is the first step of the BUndle ANalytics (BUAN) profile workflow: using a subset of streamlines from each tractograms, the streamlines are registered onto a reference tractogram using streamline-based registration (SLR) (a streamline-wise distance metric between the two tractograms is minimized). Then centroids from the subject's whole tractogram are computed and matched with the centroids of the reference bundle to remove strongly inaccurate streamlines (far pruning), and the obtained streamlines are

[2] approximately 99% of the streamlines belonging to the Gaussian distribution are kept with 3 standard deviations, so that only the streamlines that do not belong to this distribution are removed

[3] this step is not mandatory

registered onto the reference bundle to achieve another stronger pruning (local pruning)[12]. A flowchart of this procedure can be found on figure2.11[12].

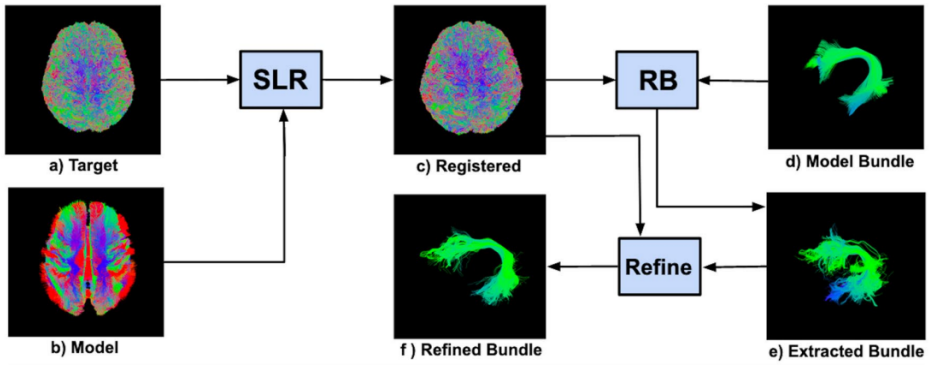

Figure 2.11: RecoBundles's bundle segmentation flowchart [13]. RB stands for RecoBundles.

The second step is to split the obtained bundles into segments. Here, instead of resampling the streamlines to 100 points to define the 100 segments, centroids that are defined along the bundle are used to assign streamlines' points to the right segment. After what the sampled microstructure values are averaged and a linear-mixed-model is used to find differences between patients [13].

2.3 A bit of neuroanatomy

This part presents the main anatomical denominations that will be used in the following of this thesis. Figures 2.12 and 2.13 shows the names of cortical areas.

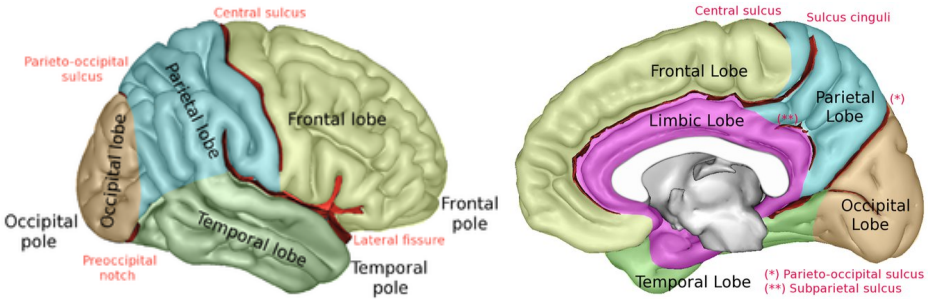

Figure 2.12: The lobes of the brain [14, 15].

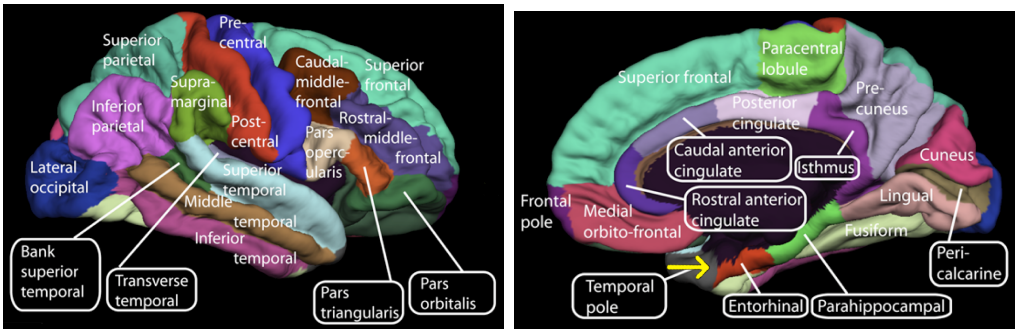

Figure 2.13: The Destrieux cortical atlas used in Freesurfer [16].

White matter bundles are often defined from such cortical parcellation to have anatomically relevant characterization of long-range connections. Many papers will use different names for those white matter bundle. Definitions presented here could then differ a bit from the literature, but are the ones considered for the rest of this work and are based on anatomical definitions in the White matter query language (WMQL) documentation and in the articles used as sources in the following paragraphs.

As shown on figure 2.14, the longitudinal fasciculus connects the parietal lobe with the temporal lobe (Middle Longitudinal Fasciculus (MdLF), Inferior Longitudinal Faciculus (ILF)) and the frontal lobe (SLF) as shown on figure 2.14. ILF may be involved a broad range of brain functions concerning the visual modality, including object, face and place processing, reading, lexical and semantic processing, emotion processing, and visual memory. They are involved in language, but also in other functions such as visually guided behaviours like reading and face recognition for the ILF [17], or motor behaviour and working memory for the SLF.

The SLF is often divided in three parts, though it can vary between articles [18]:

SLF I : connects to the superior parietal cortex with the premotor cortex, which suggests this branch is involved in motor behaviour regulation

SLF II : connects the prefrontal cortex to the inferior parietal cortex which controls spatial attention and visual and oculomotor functions. It is then thought to convey information about the visual space, and regulation of the attention in the visual field.

SLF III : connects the inferior parietal lobule with the inferior frontal lobe. It plays a role in language, working memory, orofacial and hand movements monitoring, and lesions of its posterior part near the inferior parietal lobule have been associated with gestural impairments

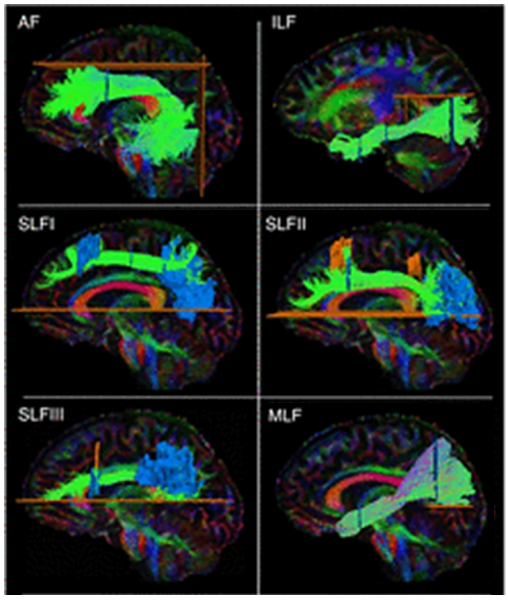

Figure 2.14: Bundles illustrations, adapted from [19].

Similarly, the Arcuate Fasciculus (AF) can be divided in three parts: an anterior segment that is similar to the SLFI, a posterior segment similar to MdLF, and a long segment that connects the prefrontal cortex to the superior temporal lobe. This last part has a crucial role in language by exchanging and modulating audio-spatial information.

Besides connections with the parietal lobes, the prefrontal cortex also connects with the occipital lobe through the Inferior Fronto-Occipital Fasciculus (IFOF), and with the temporal pole through the Uncinate Fasciculus (UF), shown on figure 2.15.

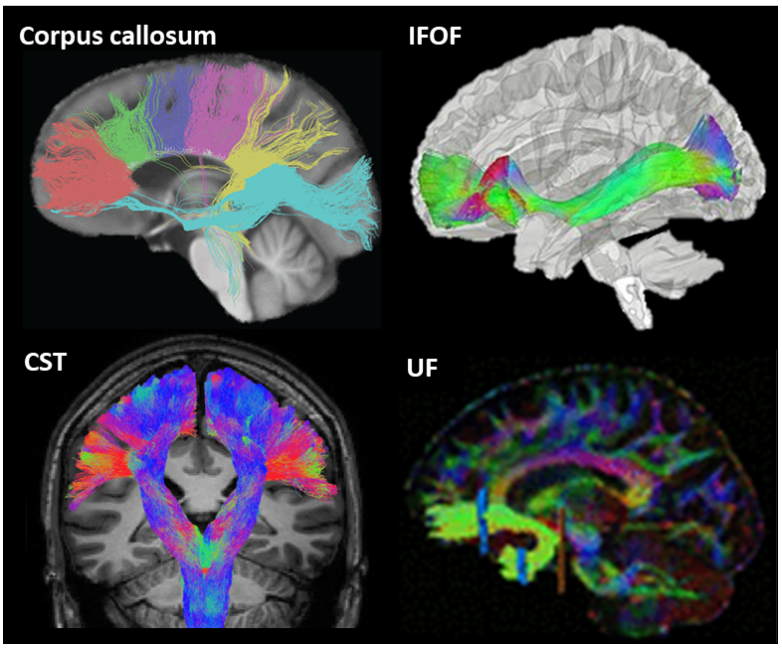

Figure 2.15: Bundles illustrations, adapted from [19] (UF), [20] (Corpus Callosum (CC)), [21] (CST), and [22] (IFOF).

The IFOF is thought to play a role in semantic processes and treatment of familiar stimuli [23], while the UF would be implied in working memory and treatment of competing memory representations [24], and is related to the hippocampal network [25].

Apart from aforementioned intra-hemispheric connections, there are also inter-hemispheric ones -called commissural fibres- that connect left and right hemispheres together. The corpus callosum is the main commissure, and is often subdivided in several parts, depending on the cortical regions that are connected.

Finally, the CST is the white matter bundle that vehiculates sensorimotor information between the sensorimotor cortex and spinal cord's motoneurons. It has then a crucial role in motor action planning and execution, with other white matter bundles which are closely related as the superior cerebellar peduncles and the thalamic projection with the sensorimotor cortex.

2.3.1 Thalamo-cortical Network

When it comes to structural connectivity, thalami are highly connected with a lot of cortical regions, allowing information transfer and its regulation between

cortical regions, as well as the integration of multisensory information. The vestibular network is not an exception: as it is shown on figure 2.16, the thalami are part of two pathways of the connections between the vestibular nuclei and the insula.

The posterior thalamocortical vestibular pathway is thought to provide neural correlate related to perceptual stability when both active and passive movements occur.

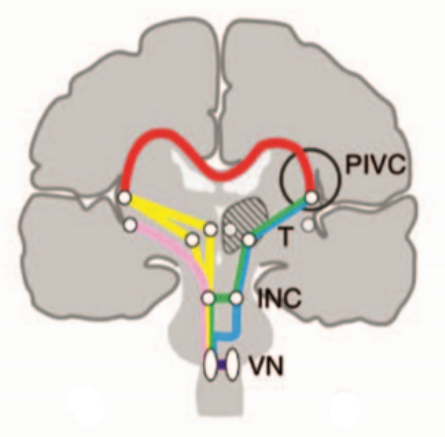

Figure 2.16: Left thalamo-cortical network. Ipsilateral projections are shown in yellow (passing through the thalamus) and pink (getting directly to the insular cortex). Contralateral projections are shown blue (crossing in the brainstem), green (crossing in the midbrain) and red (callosal tracts) [26].

Besides the connection between vestibular nuclei and the postero-lateral thalamus, the anterior thalamus is connected to head direction cells and the navigation networks. Unilateral lesions of the anterior thalamus were then found to cause dizziness and swaying unsteadiness, while brainstem and postero-lateral thalamus impairments was associated to rotational vertigo, nystagmus and posture imbalance [26].

2.4 Vertigo

The vestibular system is a crucial part of our neural system: it allows us to stand, walk, to perceive our position in space, stabilize our gaze, ... It is a highly multisensory network as it relies not only on vestibular inputs from the inner ear, but also on other sensory inputs such as visual, proprioceptive and auditive. Such a diversified multisensory integration requires both bilateral and hierarchical complex architecture of the network [27].

So, when this system happens to dysfunction, it can express itself by many symptoms and many things can be the cause of this deficit. This section will provide a glimpse of these possibilities, as well as the clinical test that can be performed to accurately diagnose the patients, and an overview of the vestibular system.

2.4.1 Peripheric vestibular system

The main peripheric vestibular organs are parts of the inner ear, as pictured on figure 2.17: the semicircular canals, the utricle and the saccule. Semicircular canals are responsive to rotations in 3 dimensions but not to linear displacements. Those linear stimuli are treated by the otolith organs: the utricule detects near-horizontal translations and the saccule the near-vertical ones [28].

These organs contain a liquid, the endolymph, that will flow during head rotation or translation. As shown on figure 2.17, this flow bends hair bundles of hair cells (present at the inner surface of saccule and utricle, and at the ampula at the ends of semicircular canals), inhibiting or exciting the cells. The signal received by each cell is transmitted by the vestibular nerve to the vestibular nuclei and the cerebellum to integrate the overall vestibular stimulation of the peripheral vestibular organs [28].

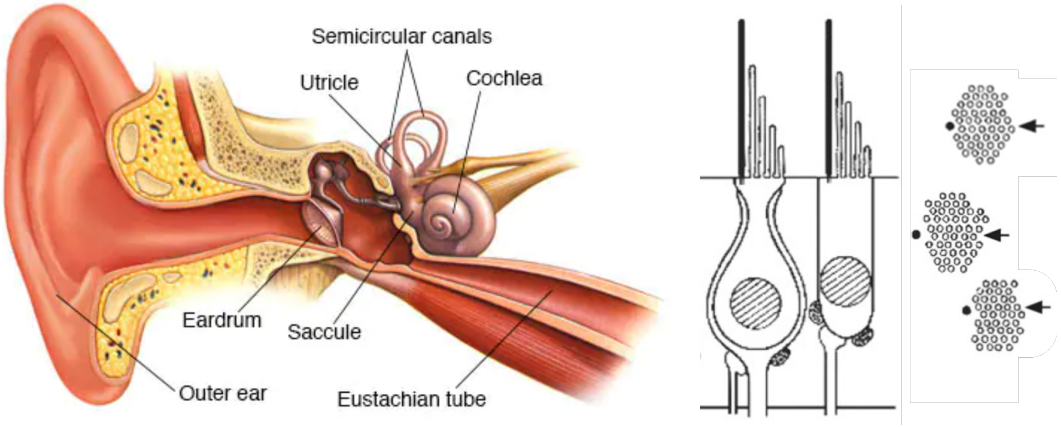

Figure 2.17: Left: peripheral vestibular system scheme [29]. Right: hair cells scheme, the arrow indicates the direction of an excitatory flow (adapted from [28]).

The right and left vestibular nuclei have connections with each other as already mentioned in section 2.3.1, allowing bilateral integration of the vestibular inputs. They also project back to vestibular organs where it can regulate the vestibular nerve excitation. Vestibular nuclei have projections with the cerebellum, to the

same regions that receive direct vestibular nerve inputs (nodulus and ventral uvula, lobules I to VII, the flocculus and ventral paraflocculus).

The rest of vestibular nuclei connections are either descending path involved in oculomotor, spinal and autonomic reflexes, or ascending path towards the thalami and cortex that were already pictured in 2.3.1.

2.4.2 Symptoms and tests

The symptoms of a dysfunction of the vestibular system can be numerous and varied:

- Nystagmus [4], it can be a sign of both central or peripheral deficit depending on the symptoms
- the false impression of moving
- the false impression that the external world is moving
- balance (postural) disorder, walk in a star-trajectory
- visual instability, etc ...

They can be accompanied by many other non-vestibular secondary symptoms as motion sickness, presyncope, mental confusion or depersonalization/derealisation [30]. Those symptoms can have many causes, peripheral or central.

In some cases, the symptoms can be reduced by some central compensation mechanism: the patient's neural system reorganizes itself to adapt to the lesions and circumvent the deficit. An example is Benign Paroxysmal Positional Vertigo (BPPV): it has a peripheral origin, but there occurs central compensation in some patients [31]. Apart from BPPV, diseases that are often encountered in the articles studying the vestibular network comprise stroke, tumours, vestibular migraine (a migraine subtype that is characterized by its vestibular symptoms), etc ...

Central causes to vestibular symptoms could be any lesion of the sensory inputs implied in the vestibular function (auditive, visual, propriceptive, speed or pressure perception, ...), of the vestibulo-ocular or vestibulospinal pathways themselves, or even of the cerebellum.

[4] inability to fix the gaze: when fixing a target, the patient's eye slowly drift away from the target before coming back to it with a corrective saccade

Lastly, with age comes an increased propency to develop vestibular deficits because of the diminution of sensory information integration: the number of long connections is known to reduce with age while shorter connections will form to compensate for the loss of long connections.

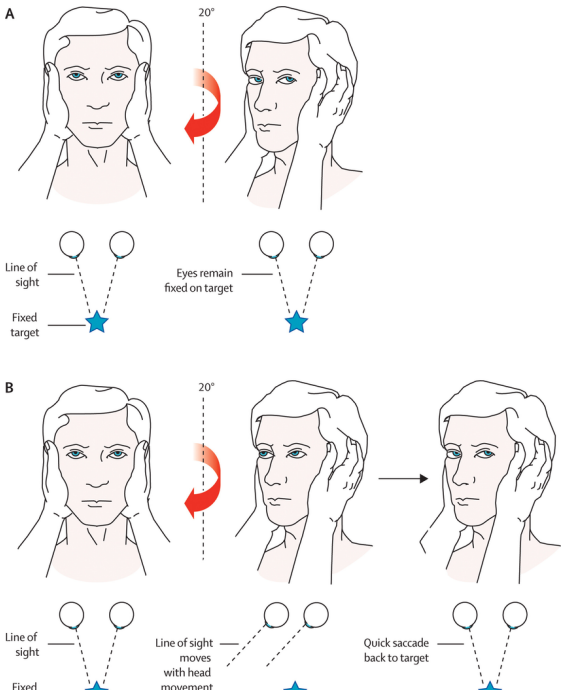

Figure 2.18: Head Impulse test (VHIT) procedure [32].

Head impulse test This test quantifies the Vestibulo-ocular Reflex (VOR) gain (eye velocity in s^{-1} divided by head velocity in s^{-1}) of the patient to evaluate its capacity to maintain gaze on a target when his head his suddenly moved.

This test then assesses the healthiness of semicircular canals that measures the head rotations, but also of the parts of the nervous system that can be implied in the VOR: vestibular nuclei, superior colliculus, visual system and oculomotor system, cerebellum, etc ... [28]

Videonystagmography Videonystagmograpy (VNG) uses caloric stimulation to stimulate the eye and records its movements using infrared cameras to measure the responsiveness of the vestibular system and how symmetric the responses are between left and right [31, 33]. Caloric irrigation through the auditory canal selectively stimulates the inner ear by creating a flow within the semicircular canals.

Vestibular-evoked myogenic potentials Vestibular-evoked myogenic potentials (VEMP) is a vestibular function test performed by measuring evoked potentials after repetitive stimulation of the ear. There are 2 types of VEMP testing: cervical (cVEMP) and ocular (oVEMP). Even if they are mostly used to diagnose peripheral disorders (affection of the saccule and utricle), they can be linked to the central network: cVEMP with the ipsilateral vestibulo-spinal tract relaying information from the vestibular nuclei, and oVEMP with the pathways linked to the VOR (mostly the MdLF). Lesions involving the vestibular nuclei can present abnormalities of both cVEMPs and oVEMPs, delayed responses of VEMPs are characteristic of multiple sclerosis and reduced VEMP responses can be observed in patients with vestibular migraine [34, 35].

2.4.3 Previous studies

Many papers have studied the vestibular network over the years, and some have conclusions that seem contradictory regarding functional or structural connectivity (the way brain regions are associated in time to perform a task, or by their structural connections) and lateralization of the vestibular system. As examples, we can cite the structural (mainly at the brainstem level) and functional lateralization of the vestibular system using PET imaging and vestibular caloric stimulation [36], the functional lateralization of only some vestibular functions using fMRI [27], and structural lateralization of some regions of the vestibular system to the left and others to the right [37].

But it is to be kept in mind that some of these studies only investigated functional connectivity, while others opted for structural connectivity. Furthermore, functional connectivity studies sometimes look for task-related activity, and the unique nature of this task will therefore limit the ability to generalize to other studies without putting the results into the right context .

The posterior insula appears to be multimodal and is implicated in processing a variety of stimuli. It does not merely receive information from the vestibular organ, but also integrates stimuli from the visual and proprioceptive systems [38].

However, most of the articles agree to say the insula and surrounding frontal and parieto-temporal regions as well as the thalami are playing a crucial role in the vestibular function, which makes those regions as well as the association tracts that connect them a major point of interest when it comes to studying the vestibular network.

Structural connectivity Previous studies have shown that regions implied in the vestibular system are organized in modules (groups that are highly connected together but less with other modules), and they are more modules in the left hemisphere than in the right (images can be found in the annexes, figures 6.5 and 6.6), suggesting a higher multisensorial integration in the right hemisphere [39, 37]. They also showed significant lateralization of the regions that are parts of those modules.

The structural and functional separation of the two thalamic nuclei complexes allowed a lateralization of the right and left hemispheric functions to develop. Furthermore, it made possible the simultaneous performance of sensorimotor and cognitive tasks, which require different spatial reference systems in opposite hemispheres, for example, egocentric manipulation of objects (handedness) and allocentric orientation of the self in the environment by the multisensory vestibular system.

Another reason that could explain the contradictory results concerning lateralization is that some articles test if the regions are lateralized at all before inferring the side of the lateralization from the data, while in other articles they directly test the hypothesis of either right or left lateralization, or even compute an overall lateralization index by summing the lateralization of all intra-hemispheric connections.

Functional connectivity The same kind of results were obtained with functional connectivity, the vestibular system is divided in a symmetric bilateral and highly connected network, and a second asymmetric one (opposite to the hand) and less connected:

- asymmetric and poorly connected (correlated with fewer than 12 RSN): inferior frontal gyrus, superior and medial temporal gyrus (hearing, motion, spatial localization), insula (spatial orientation, spatial reasoning, coordinate transformation and further vestibular coordinate input processing functions) [27].
- symmetric (opposite to the hand) and highly connected: anterior insula (correlated with temporal and frontal lobes, linked to attentional salience and adaptive cognitive control of cognition and behaviour), sensorimotor areas, superior temporal gyrus (hearing, touch, pain, visceral sensation and tactile attention), temporo-parietal regions (limbic, visual, auditory and somatosensory percept in self-reference and localization of self-other relationships, as well as reflective thinking, introception and theory of mind) [27].

Other papers investigated changes of the functional connectivity in patients with vestibular system impairements:

- decreased ipsilateral functional connectivity between the thalamus and the left cingulate, bilateral insular and right supplementary motor cortex in patients with vestibular migraine [40].
- decreased controlateral functional connectivity between the left thalamus right precuneus and middle frontal gyrus, and between the right thalamus and the superior parietal gyrus in patients with vestibular migraine [40].

Those findings are in a manner consistent with the multisensory nature of the vestibular system and the fact that the thalami is a center of connection relays and that the vestibular network comprises insular, cingulate, frontal and parietal regions.

Attention Networks A last paper has shown that the parieto-insular vestibular cortex is inhibited when attention networks in the parieto-occipital cortex are activated by visual stimuli during attentive tracking. This shows vestibular sensations can be reduced through attention networks.

2.5 Machine Learning

Apart from statistical analysis as seen so far in the tractometry section, other use machine learning algorithms to find features that are related to a patient condition. Most of the information given in this section are inspired from [41, 42, 43]. The main problem for all models is the tendency to overfit[5] when there are much more features than subjects, but there are many ways to circumvent this drawback. This comes either by a selection of the features to reduce information redundancy, dimensionality reduction techniques, or by the inner structure of the models itself.

Cross-validation Cross-validation is a precious tool to limit overfitting. Data are first divided into two sets: one on which model is trained and one on which it is tested. This way, we can estimate its performances on data it had never seen before, evaluating the *generalization error*. However, a better estimation of this generalization error can be obtained by averaging the errors on different couples of train and test sets, this is called cross-validation. As models often

[5] *overfitting* means we are giving too much importance to the specificity of each subject, fitting the noise of the data. The model is then highly efficient on data it was trained on, but useless on new data. This often happens when we have too many parameters compared to the amount of data at our disposal.

have hyperparameters that must be optimized, the train set is itself splitted into train and test set to first select the best hyperparameters before evaluating the performances on other data. This methodology is depicted on figure 2.19. There are many types of cross-validation, of which stratified ones that keep the same ratio between classes in both train and test sets, repeated that repeat the cross-validation by ensuring sets in each run are independent, Shuffled that draw train and test sets at random after shuffling, K-fold that ensure no overlap between all sets, etc.

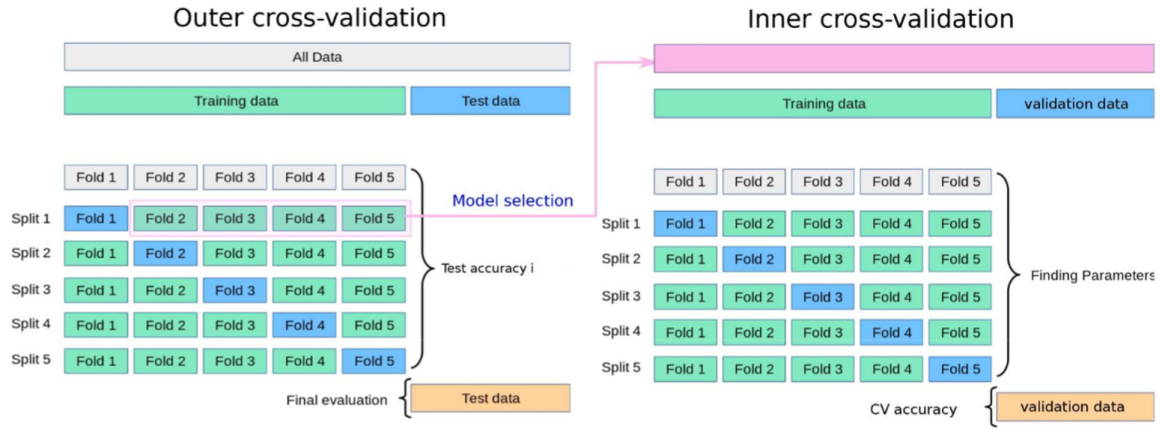

Figure 2.19: Nested cross-validation scheme. Adapted from [44].

Ensemble models This family of models uses other models that are trained on subsets of data, drawn with replacement (bootstrap) or not. By averaging the results of all trained models, one can have a more accurate and robust evaluation of model performances, and increase the ability to detect overfitting. This is the Bootstrap Aggregation (Bagging for short). Other Ensemble models use weighted averaging of the models prediction to "trust" more the ones that are more often accurate, like AdaBoost that uses exponential weights to tune this ponderation until the most difficult subjects are correctly classified. Though, as this weighted average implies another set of hyperparameters (the weights), it is necessary to tune them by using a cross-training loop within the nested cross-validation process, which can be highly computationally intensive.

Class Imbalance When dealing with data where a class is underrepresented compared to others, this class imbalance can bias the predictions of the model by giving more importance to the overrepresented class. To tackle with that issue, one can use weight subjects depending on the inverse size of the class they belong to. This is useful for linear models but is often inefficient with some non-linear models like Random Forest that are not sensitive to weights attributed to features or

samples. In this case we can use resampling (oversampling of the underrepresented classes and undersampling of the overrepresented classes), which has virtually the same impact as weighting except it is as efficient on Random Forest and linear models.

2.5.1 Preprocessing

Preprocessing in machine learning mainly consists in preparing the data so that the models could be the most efficient. This means it will often be model-specific and application-specific: depending on the data at our disposal and the models that are to be used, different preprocessing steps could lead to optimal results. Ideally, each preprocessing steps should be included in the cross-validation process.

Imputing This is simply the fact of "*filling the holes*" in the data. When data is missing for some features of some subjects, several ways may allow replacing this missing piece of information by a synthetic value created from values of subjects for whom it is not missing. All methods have different behaviours, drawbacks and advantages, but similar (and good) performances when there are a lot of data and not too many features. However, here we are in the opposite case. A constant imputing should then be preferred: missing values are replaced by a constant, classically 0.

Filtering The first approach to filtering is to use statistical analysis to discard subjects that are considered as outliers: their features are so different from other subjects' that they are likely to impair the results. This can also be done with the features by removing those that always have the same value in all subjects for example.

Once done, features can be selected to use only those that contain more information. This can be done either by:

- ranking them by mutual information, that considers even non-linear dependencies and reflects the amount of information that is brought by a feature when knowing the value of others, so that it tries to find the best compromise between information content and minimizing redundancy.

- ranking them according to the p-value of a F-test (an ANOVA test for classification). Such ranking is then based on correlation measures between the features (microstructural measures here) and the target (do patients have vertigo or do not?), which is more appropriate for linear models. However, as this does not include minimization of redundancy, it is usual to remove all but

one features that are too much correlated together (as they are considered to bring nearly the same information)

PCA Another approach to reduce the number of features is dimensionality reduction. Here instead of simply selecting the features, we first transform them so that this selection can be made by loosing the least information as possible. There are many dimensionality reduction methods, but Principal Component Analysis (PCA) is the most simple and common one. It uses correlation measures to find the axis in the feature space along which the correlation is maximized, while ensuring no correlation between those directions. With a less mathematical language, this means each PCA axis is linear combination of the features, each axis being ranked according to the difference it finds between subjects: first axis has maximal correlation and last ones only have what's left of information. We can then select the first components as they contain more information than the last ones. As it is based on correlation analysis, PCA will perform well with linear models, by removing parts of the data that are similar for all subjects and providing features that contains more information related to their differences.

Scaling Features often have different scales, units, etc, ... which means that using them as they can bias the results because a feature that has a value between 0 and 10 could be considered as more significant than another between 0 and 1 for instance, just because of the scale. The basic approach is the same as the one used with a T-test: standardization. This implies transforming the features towards a standard normal distribution (mean of 0 and standard deviation of 1) by subtracting their mean and dividing by their standard deviation. This "centring" is also the first step of the PCA algorithm, but is not always the best choice for all models and data.

Other scaling methods exist to match the scales of all features and reduce (limited) loss of information produced by centring: the mean and scale of features can also carry precious information, and though centring allows efficiently comparing the shape of features distribution, it does not compare their mean and standard deviations

2.5.2 Models

A lot of different models exist but they can be gathered in two main categories: linear and non-linear ones. All models that will be presented here have the advantage to provide feature importance estimation via their inner structure.

Support Vector Machine [45] Classical linear classification approaches imply finding a linear regression that fit a line for each class, minimizing the squared distances from the line in each class while keeping those lines parallel, and separating classes according to the midline. On the other hand, Support Vector Machine (SVM) uses only a few subjects (support vectors) for each feature so that it can find the best linear regression to separate two classes (where classic linear regression only gives one solution that is not necessarily optimal).

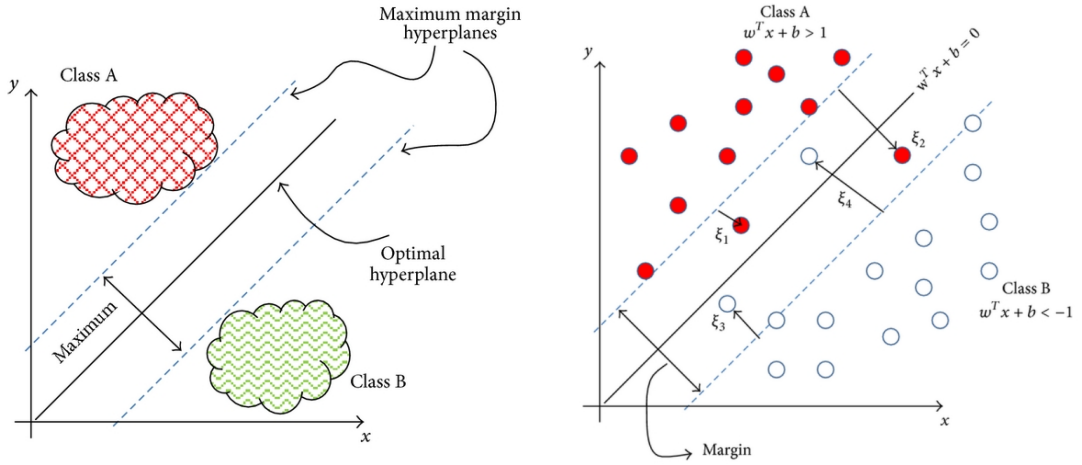

Figure 2.20: SVM schematic representation [45].

This is done by finding the linear regression that gives the largest marge between the two classes, while minimizing the number of misclassified subjects (ξ on figure 2.20) as shown in equation 2.5.2. Weights associated to each feature can be interpreted as the importance it has on the prediction. Non-linear variants of the SVM allows to use the same computations on transformed features (either a polynomial or exponential transform for instance), but this lower interpretation power. As it relies on a few data to make predictions, SVM is known to be resistant to overfitting, but also sensitive to unbalance between the number of subjects in each class (see 2.5.2).

$$\frac{1}{2}\sum_{j=1}^{p}\beta_j^2 + C\sum_{i=1}^{n}\xi_i$$

$$\text{subject to} \quad y_i(\sum_{j=1}^{p}\beta_j\phi_j(x_i) + b) \geq 1 - \xi_i$$

Where β_j are the weights associated with the p features, ξ_i the distance within the margin of misclassified data. In the case of overlapping classes, support vectors are then misclassified data.

Random Forest [46] Random Forest (RDF) is no more than another type of ensemble model that uses bagging on both samples and features (but features are not drawn with replacement). It fits decision trees on such constructed data subsets. Decision trees that use an impurity criterion (Gini or mutual information) to determine the most significant feature within the subset, split it according to the feature's values and rehearse with all newly-created subsets. It then works with logical operations (*AND*), and a feature importance can be estimated via the gain in purity when classing the patients according to this feature. This is *scikit-learn*'s implementation that takes the mean of such feature importance across the trees of the forest.

Logistic Regression [43] Logistic regression is actually a family of models. They are constrained linear regression that uses the log-loss instead of the mean-squared error to estimate the probability of belonging to each class. The constraints used during regression have the purpose to limit overfitting by limiting the l_2 *norm* of the weights (Ridge model), their l_1 *norm* (Lasso model) or both (ElasticNet model). The error criteria that are minimized by the models are then as follows, with β the weights of the regression and X_i and y_i the features and class of the i^{th} subject, respectively:

$$\begin{aligned}
\text{Lasso}: \quad & \sum_{i=1}^{n}\sum_{h=1}^{k} \delta_h \, log\left(\sigma(\beta, X_i)_h\right) + \boxed{\lambda_1 \sum_{j=1}^{p} |\beta_j|} \\
\text{Ridge}: \quad & \sum_{i=1}^{n}\sum_{h=1}^{k} \delta_h \, log\left(\sigma(\beta, X_i)_h\right) + \boxed{\lambda_2 \sum_{j=1}^{p} \beta_j^2} \\
\text{ElasticNet}: \quad & \sum_{i=1}^{n}\sum_{h=1}^{k} \delta_h \, log\left(\sigma(\beta, X_i)_h\right) + \boxed{\lambda_1 \sum_{j=1}^{p} |\beta_j|} + \boxed{\lambda_2 \sum_{j=1}^{p} \beta_j^2} \qquad (1) \\
& \underbrace{\text{binary logistic regression error criterion}} \quad \underbrace{l_1 \text{ penalization}} \quad \underbrace{l_2 \text{ penalization}}
\end{aligned}$$

Where δ_h is 1 when patient i belongs to class h and 0 otherwise, σ_h is the sigmoid function that translates the linear regression on the features into pseudo-probabilities of membership to the class h:

$$\sigma(\beta, x)_h = \frac{1}{1 + e^{-z}} \ , \qquad z = \sum_{s=1}^{p} \beta_s x$$

ElasticNet-like models are well known in the statistic field as alternatives of statistical analysis models (like ANOVA or T-test) that are efficient when dealing with a lot of features. The use of logistic error criterion is furthermore known to reduce the impact of outliers and of the shape of features' underlying distribution (if they are normally distributed or not) by using the logarithmic loss.

Sparse Group Lasso [47, 48] Sparse-Group Lasso (SGL) is another type of logistic regression. Its particularity is to apply penalties specific to groups in which features are clustered in. The Group Lasso model will then apply the Ridge penalty in a group-wise manner, implying ridge penalization within groups and Lasso penalization between them (the second is implicitly imposed by the grouped structure of data when the first is applied). On the other hand, SGL applies both the Lasso and group-wise Ridge penalties. With these improvements, equations of the Ridge and ElasticNet model become:

$$\text{Group Lasso}: \quad \sum_{i=1}^{n}\sum_{h=1}^{k} \delta_h \; log\left(\sigma(\beta, X_i)_h\right) \quad + \quad \boxed{\lambda_2 \sum_{j=1}^{p}\sum_{k=1}^{K} w_k \delta_k \beta_j^2}$$

$$\text{Sparse Group Lasso}: \quad \sum_{i=1}^{n}\sum_{h=1}^{k} \delta_h \; log\left(\sigma(\beta, X_i)_h\right) \quad + \quad \boxed{\lambda_2 \sum_{j=1}^{p}\sum_{k=1}^{K} w_k \delta_k \beta_j^2} \quad + \quad \boxed{\lambda_1 \sum_{j=1}^{p} |\beta_j|}$$

binary logistic regression error criterion

group-wise l_2 penalization

l_1 penalization

with δ_k equal to 1 within the group k and 0 outside, and ω_k the weight of the group k (by default the square root of the group length). Lasso will evaluate each variable individually, while Group Lasso will select groups as a whole and SGL is able to select variables on both individual and group levels.

Model comparison Figure 2.21 shows the fraction of predicted positives (False Positives + True Positives, the subjects predicted with vertigo for example) in function of the mean predicted probability of appartenance to the positive class. This for instance shows the tendency of Random Forest and SVM towards false positive when the predicted probability is high, while Logistic Regression have the opposite behaviour. This is a reason why Random Forest and SVM are more sensitive to class imbalance than logistic regression, as an over-representation of the positive class will bias them towards high-probability prediction and high False-Positive rate.

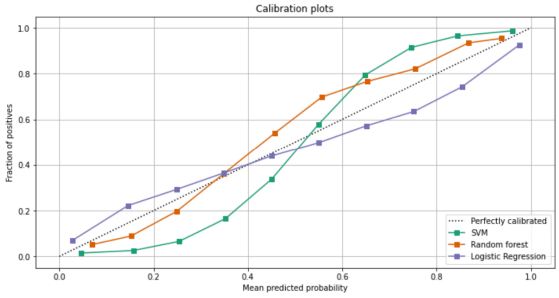

Figure 2.21: Calibration curve of models, adapted from [49].

Figure 2.22 shows the results of *Richie et al* on several datasets. They used Logistic Regression model without or with PCA (PCR Lasso and SGL) to predict age (regression) or (classification). The main conclusions were that [47]:

- SGL was either more accurate, more robust or both.
- Logistic regression was more efficient when dealing with the mean of segments instead of the mean of the whole bundle.

We can also see those performances depend on the studied data, and it is legitimate to think SGL performances would strongly rely on the underlying sparsity of the data (if biomarkers are located in a specific bundle or are more distributed).

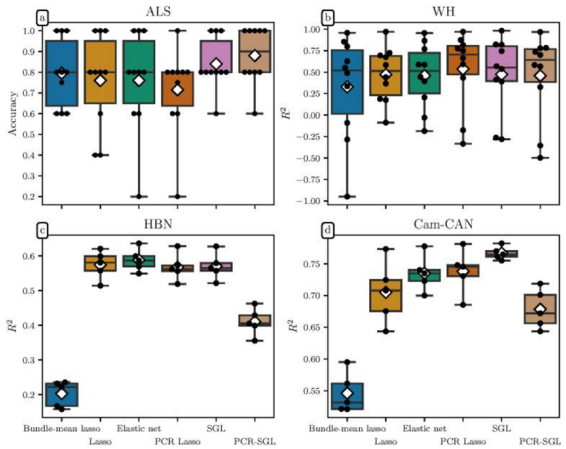

Figure 2.22: Comparison of logistic regression models on different datasets, with classification (scored by accuracy) or regression tasks (scored by coefficient of determination, R^2) [47].

2.5.3 Evaluation

Depending on the models used there are many strategies to both evaluate model performances and feature importance (how much a given feature impact the prediction of a given model).

Model performance Here, we ask "How well does my model perform?". This is a legitimate question as if we want to find relevant pieces of information in our data, we can only trust models that really found patterns related to our research question (does a patient have vertigo?) in it, in other words the models that are able to correctly predict the conditions of a majority of patient.

There are many metrics to evaluate a classification model such as accuracy (proportion of correctly classified subjects), precision/recall (mean of the proportion of of in the predictions/instances of a given class) or log-loss (that gives an idea of how much the model is right/wrong when it gives a good/bad prediction). All of which can be adapted to account for class imbalance. These are measured on the test set during nested cross-validation to avoid overfitting, and we seek for models that have the better scores while having the least variance across folds (a model that would be efficient on all folds can be expected to generalize better than one efficient on some and awful on others).

Another way to assess a model's performances is to use random permutations of subjects' labels: for each cross-validation fold, the model is fitted on the train set and evaluated on the test set.

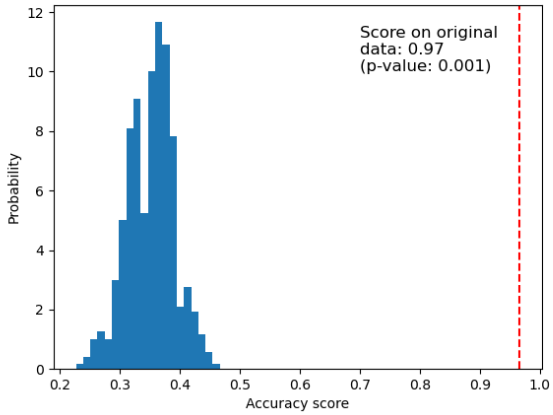

Figure 2.23: Histogram of the scores of an example model for all permutations [50].

This is done several times by randomly permuting the classes of subjects. Performances with the real data are then compared to those obtained with the random permutations to compute a p-value that estimates the probability the model's predictions on real data are due by chance: a small p-value indicates the results are reliable.

Feature importance As for model performances, features impact on models' predictions can be evaluated by many ways. Some models have a structure that allows a direct interpretation of a feature impact on their predictions: linear models have coefficients interpretable as correlation links while RDF have so-called "gini-feature-importance" that uses the gini impurity criterion averaged on the

forest's trees features to give an insight on the non-linear links between features and patient's class.

However, this impurity importance is biased by a simple mean across trees as all features are not present in each of them [51], but cannot be corrected by taking into account the number of time a feature is selected, knowing this already conveys information about feature importance as features subsets for all trees are only conserved if they respect certain constraints (a minimum number of splits must be possible for example).

Permutations can again help to compensate this problem and to have a measure that is: similar for all models (and then comparable), independent of the training task (and then less prone to overfitting and statistical biases), applicable for all models (even those that do not inherently support feature importance estimation) and that takes model performances into account. One way is to train the model on a train set and assess its performances on a test set, while replacing at each iteration the values of a unique feature by random values that follows a distribution inferred from the training data, which is done n times for each feature. Feature importance is then considered as the mean difference between the performance metric on the original data and on permuted data. Many other approaches exist based on the same principles, training and testing on different sets, using different noise model or feature importance evaluation.

Chapter 3

Methods

This part will introduce the methods used to process and analyse data. Analysis were done based on deeply and bilaterally deaf patients that had either vertigo symptoms, vertigo symptoms with the ability to compensate, or no vestibular deficit. Most of them were right-handed, and they were all included in a study that received the approval of an ethics committee. Their dMRI and T1 images were used. dMRI were acquired with a Siemens 3T. dMRI images were acquired with 64 directions, a b-value of 800 s/mm^2, an in-plane resolution of 1.95mm and a slice thickness of 2mm.

In the following pages, there will be a few moments where registration is mentioned. If not presented otherwise, it will consist in a rigid affine registration (6 degrees of freedom: 3 translations and 3 rotations) computed by FMRIB Software Library (FSL) with the cross-correlation metric, and applied by Advanced Normalization Tools (ANTs) with the GenericLabel interpolation. This method was inspired by *Yousef et al.*'s work, that identified FSL's affine registration as slightly more efficient than ANTs's , and by *Shaerers*'s work underlying the efficiency of ANTs's GenericLabel interpolation for label images [52, 53].

The codes used as well as some supplementary figures can be found on Github: `https://github.com/Dany546/TFE_vertigo.git`. All images of the following sections were produced using Freeview application [54] or Fsleyes's comandline interface [55] (for anatomical views), TrackVis [56] (for tractograms), or python scripts.

Some patients were not taken into account as their dMRI images were not whole.

3.1 Data preprocessing

On one side, data were preprocessed with Elikopy [57]. Basic preprocessing was performed: brain extraction [58], denoising using Marchenko-Pastur PCA [59], motion correction [60, 61, 62] and bias field correction [63]. It was chosen not to apply a gibbs correction as no gibbs artefact seemed to appear in most of the patients.

On the other side, white-matter segmentation and parcellation were computed from the T1 images using Freesurfer. The parcellation was done based on the Destrieux cortical atlas, assigning white-matter voxels to the closest atlas region to form white-matter regions corresponding to the cortical atlas. The segmentation includes a peak around for grey-matter, one for white-matter and a last one for CSF. As it incorrectly identified parts of the skull as white-matter, it was corrected by hand.

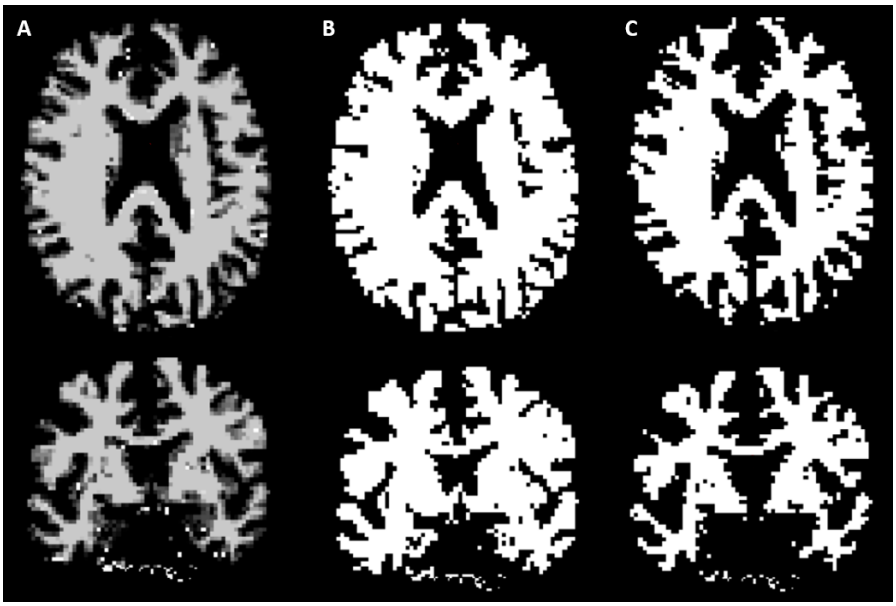

Figure 3.1: Freesurfer white-matter mask. A: Freesurfer white-matter segmentation, B: binarized segmentation with a large binarization interval, C: binarized segmentation with a small binarization interval. All images were registered to the dMRI space with Dipy. Binarization was achieved before registration (with nearest-neighbour interpolation) of the segmentation.

Nonetheless, this segmentation did not include cerebellum and brainstem and was not binarized. It was then needed to threshold the segmentation to produce

a white-matter mask (by setting all the voxels with values out of a given interval around the white matter peak to 0 and those inside the interval to 1). As there is a slight overlap between the white and grey matter peaks, any chosen threshold lead to white matter labelled as grey matter or inversely, resulting in poorly segmented white matter at the white-matter/grey-matter interface. As shown on figure 3.1, a large interval lead to grey matter identified as white matter and a small interval lead to white-matter voxels identified as non white-matter ones.

On the other hand, the Freesurfer parcellation seemed to fit the anatomy of the patients better than any binarized white-matter mask, and the cerebellum and brainstem was included in it. For this reason a white-matter mask was extracted from the parcellation with MRtrix3's 5TTgen [64] command: it reads the parcellation and creates five images according to its labels: white-matter, cortical grey-matter, sub-cortical grey-matter, CSF and injured tissues (this last one was empty in our case)[1]. This allowed to produce accurate white- and grey-matter mask, as shown on figure 3.2.

[1] As the command did not recognized the Thalamus-Proper* label, a copy of the parcellation was used with thalami relabelled as putamen so they are included in the subcortical grey-matter image.

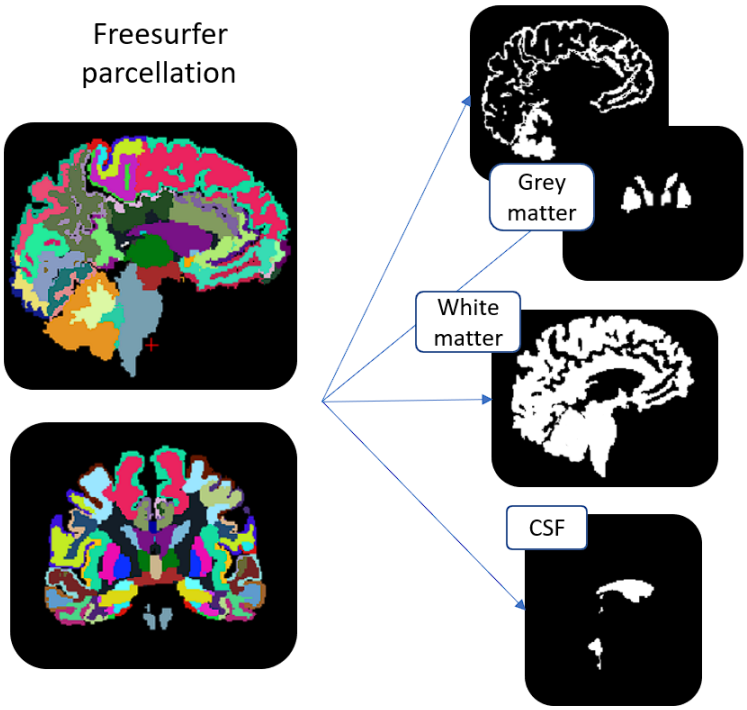

Figure 3.2: Freesurfer parcellation-based masks.

Though, the cerebellum segmentation was imprecise (only white/grey matter based on the segmentation) and was replaced with one produced by Volbrain[65] for microstructure metrics sampling, and another mask was produced by setting all cerebellum labels of the Freesurfer parcellation to cerebellum white-matter in order to have the less restrictive white matter mask as possible for the tractography step. Indeed, cerebellum gyri are much thinner and many streamlines would be discarded by using the white-matter mask for the cerebellum. Moreover, Volbrain does not segment the Brainstem and is not in exact agreement with Freesurfer's cerebellum/brainstem boundary, it was then needed to extend the white-matter labels of Volbrain with the ones adjacent to this boundary to not have holes in the parcellation.

Though, Volbrain crops the images around the cerebellum during their processing pipeline, so that output segmentation is in the space of the box used for cropping, and not in the initial space of the images. A rigid registration was then applied with Dipy to make correspond the binarized cerebellum segmentations of Volbrain and Freesurfer before changing the labels. The final result can be seen on figure 3.3 for one patient.

Table 3.1: Volbrain to Freesurfer translation of left cerebellum, the equivalent labels were used for the right cerebellum.

Volbrain region (label)	Freesurfer region (label)
Left lobule I-II (1)	Cbm_Left_I_IV (601)
Left lobule III (2)	Cbm_Vermis_VI (606)
Left lobule IV (3)	Cbm_Right_CrusII (612)
Left lobule V (4)	Cbm_Left_V (603)
Left lobule VI (5)	Cbm_Left_VI (605)
Left lobule Crus I (6)	Cbm_Left_CrusI (608)
Left lobule Crus II (7)	Cbm_Left_CrusII (611)
Left lobule VIIB (8)	Cbm_Left_VIIb (614)
Left lobule VIIIA (9)	Cbm_Left_VIIIa (617)
Left lobule VIIIB (10)	Cbm_Left_VIIIb (620)
Left lobule IX (11)	Cbm_Left_IX (623)
Left lobule X (12)	Cbm_Left_X (626)

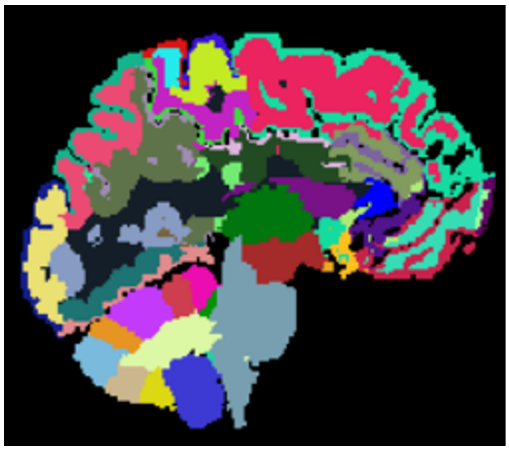

Figure 3.3: Freesurfer parcellation mask with Volbrain's cerebellum segmentation.

3.2 Tractography

Both deterministic and probabilistic tracking were performed iteratively (because of the size of produced files) with Dipy: tracking is performed once, intermediate results are saved and added to results of previous iterations. Deterministic tracking was stopped either after 12 iterations or if the file size exceeded 450Mo (almost always the second case) while probabilistic tracking was stopped either after 25 iterations or after reaching a file size of 900Mo. Indeed, deterministic tracking

lead to approximately 3 times more streamlines selected after filtering with AFQ than with probabilistic tracking for equivalent number of streamlines before filtering.

At each iteration, tracking was done using 1 seeding point per voxel of a dilated white-matter mask (dilated to enhance track density at the grey-matter/white-matter interface and especially in the gyri). Each seeding point was placed at random in its assigned voxel to not generate the same streamlines at each iteration with deterministic tracking.

The white-matter response was estimated with the FA-threshold based method presented in 2.2: voxels of the white matter mask that exceeded a FA threshold of 0.8 were used to compute the response. ODF was then produced with Dipy's CSD with Dipy's default 362-vertices sphere. For deterministic tracking, orientation peaks were obtained from the CSD model with a random peak threshold between 0.75 and 0.9 for deterministic tracking and between 0.8 and 0.9 for probabilistic tracking.

This was done for two reason: first, acceptable and comparable results were obtained with values in those intervals for two patients, and moreover it is to be expected that the optimal value (if there is one) would change from one patient to another. Knowing a higher value imply simpler trajectories and too low values imply a higher number of false-positive tracts and a higher noise, it is then better to use different thresholds from one iteration to another to ensure we use coherent threshold for some iterations for all patients.

Similarly, the maximum allowed angle at each step for probabilistic tracking was chosen at random between 20 and 25 degrees. According to Dipy's default, both direction peaks for deterministic tracking and ODF for probabilistic were normalized between 0 and 1 before filtering the directions with the aforementioned thresholds.

After tracking, obtained streamlines were then cropped to the white-matter mask. To reduce the number of invalid streamlines, those that did not cross the gray-matter mask were removed, and those that crossed the eroded CSF mask were discarded too. Finally, only those longer than 20 mm were kept. These 3 final steps have for purpose to reduce the size of the saved files and both accelerate and improve tract segmentation as only long tracts ($\simeq$ 50mm) are to be selected, and many short streamlines are produced when cropping them to the white-matter mask (because of tracts connecting 2 gyri but getting out of the white-matter mask, at the limit of the dilated white-matter mask used for tracking).

Dilation of the white mask was done with the same procedure as in *connectomemapper3* [66]: with a binary dilation of followed by a gaussian smoothing and a thresholding to smooth then binarize the mask. All of these steps were done with FSL's command line tools. However, using a mask without the subcortical regions (to produce thalamic projections) yield to unsatisfying results (no streamline found for the CST in some cases), so that three white-matter masks were used instead: (see figure 3.4)

- subcortical regions were added to the dilated white-matter mask and used for seeding (3.4C)
- the non-dilated white-matter mask where an erosion of the subcortical regions was removed was used during odd iterations to crop streamlines exiting the white-matter mask (3.4B)
- the non-dilated white-matter mask with the subcortical regions was used during pair iterations to crop streamlines exiting the white-matter mask (3.4A)

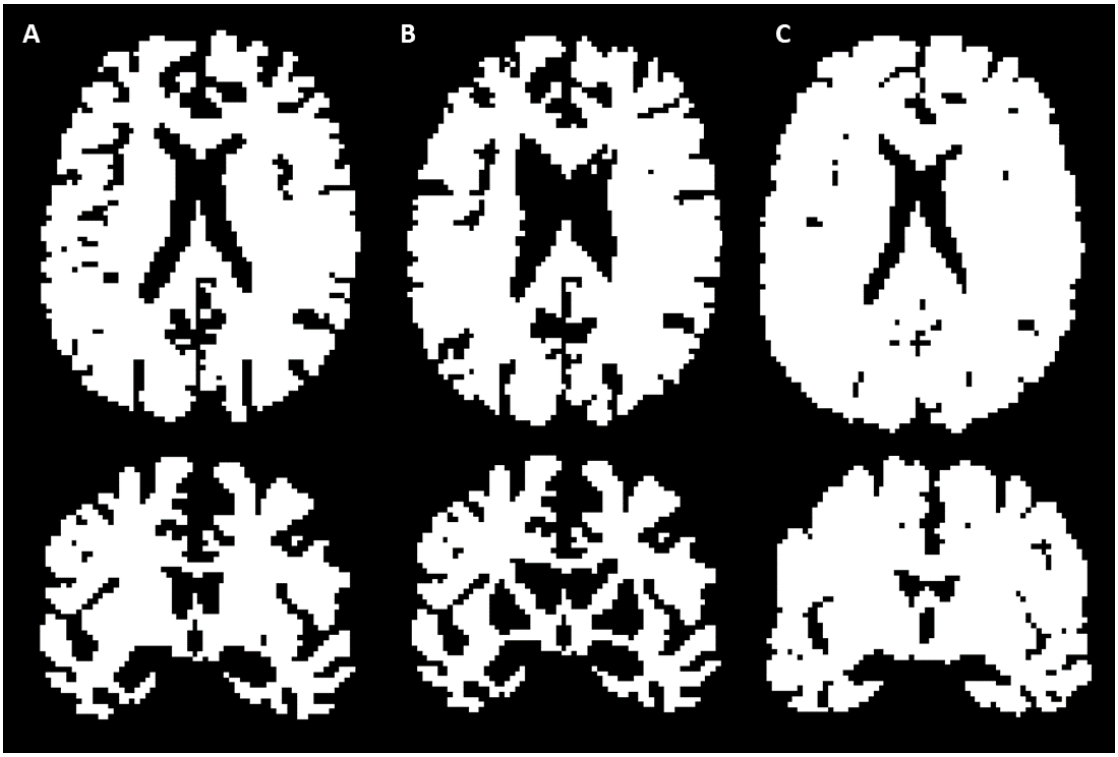

Figure 3.4: White-matter masks for one patient, after registration to the dMRI space. Original (A), without eroded subcortical gray matter (B), and dilated mask (C).

All masks were processed based on the Freesurfer parcellation-based masked produced with 5TT, modified (dilations, erosions, ...) in the resolution of the T1, and finally registered into the dMRI space. To do so, the T1 was registered with FSL on the T1 already registered on the dMRI by Elikopy, using a rigid affine transform (6 degrees of freedom). The affine transform was then applied on the masks with the GenericLabel interpolation of ANTs.

3.3 Tractometry

Tractometry was then done based on AFQ procedure: segmentation, cleaning and splitting into segment. AFQ was chosen because it has no parameters for the user to optimize (if default registration method is kept to register the ROIs), compared to RecoBundles (RecoBundles). However, AFQ has some disadvantages for our particular application:

- 100 segments are used for all bundles, whatever their length, and resamples streamlines to 100 points to do so, which cause two problems: spatial sampling frequency is not constant from one bundle to another, and for short bundles such as the Anterior Thalamic Radiation (ATR) it will resample streamlines to a number of point greater than before resampling.
- include ROIs and probability maps are defined in MNI space and needs to be registered, the quality of the registration then determines the quality of the segmentation.
- there is only 2 include ROIs and no exclude ROIs for most bundles, meaning include ROIs must be very restrictive to be efficient.
- a streamline is selected only if it respects all include and exclude conditions, so that many streamlines are discarded as the ROIs are often placed at the end and beginning of the bundles and streamlines belonging to the bundle but with one end between the ROIs are discarded. Though it is possible to increase the tolerated distance between streamlines end and ROIs, this would also increase the number of wrong streamlines that are selected.

It was then chosen two bring some improvements to this workflow (as explained further in **??**) and to compare it with WMQL. When there were not enough streamlines in a given bundle, another round of tractography was performed using a binarized version of the probability density maps of NatbrainLab [67] as seeding masks. NatbrainLab's probability density maps [68] were also used to add cerebellar bundles that were not included in the default AFQ bundles.

To use WMQL, it is needed to define each bundle according to the regions of Freesurfer's parcellisation. As those definitions needs to be both anatomically coherent and robust throughout hemispheres (for non-interhemispheric connection within a given patient) and patients. That's why default definitions provided in WMQL's code were used as baseline.

From those definitions, new definitions were produced by merging bundles having similar spatial trajectories (such as the thalamo-cortical connections with the pars-opercularis, the pars-triangularis and the lower part of the precentral gyrus that were merged into a *thalamo-broca* pathway) to reduce the number of thalamo-cortical connections considered. It was also needed to split some bundles as the corpus callosum: removing the tapetum from its posterior segment, as the tapetum does not have the same trajectory as the rest of this segment of the corpus callosum.

By visual inspection of the results on a subset of 5 patients, these definitions were slightly refined to ensure no overlap between bundles and similar shaped bundles throughout patients and hemispheres. For instance, the initial definitions of corpus callosum's segments kept streamlines ending in the cingulate cortex, which is not suitable here as it would impair the construction of the mean bundle. There were also the thalamic projections that were not truly defined [2]. Hence, the following modifications were done, as detailed in the Annexes (6.1). Exact definitions can be found on Github[3].

Using AFQ on a subset of patients led to unusable results as most of the bundle were not segmented (or poorly, less than 50 tracts) in all patients, especially for probabilistic tracking. To give an order of magnitude, we got from a starting file of 1Go to an output file of $\simeq$ 2Mo for probabilistic tracking and from 500Mo to 5Mo for deterministic tracking.

However, WMQL led to much better result: $\simeq$ 65Mo files for both deterministic and probabilistic tracking because of a greater amount of selected streamlines as depicted on figure 3.5. Here it is shown for only one patient (not the best nor the worst results) but the same conclusion was drawn on other patients. WMQL is moreover much faster than AFQ: a tenth of minutes for WMQL against more than an hour for AFQ.

[2]The available definitions only provide definitions for connections between the thalamus and any existing gyri and the anterior thalamic projections that were defined as connections between the thalamus and the entire frontal lobe.

[3]`https://github.com/Dany546/TFE_vertigo.git`

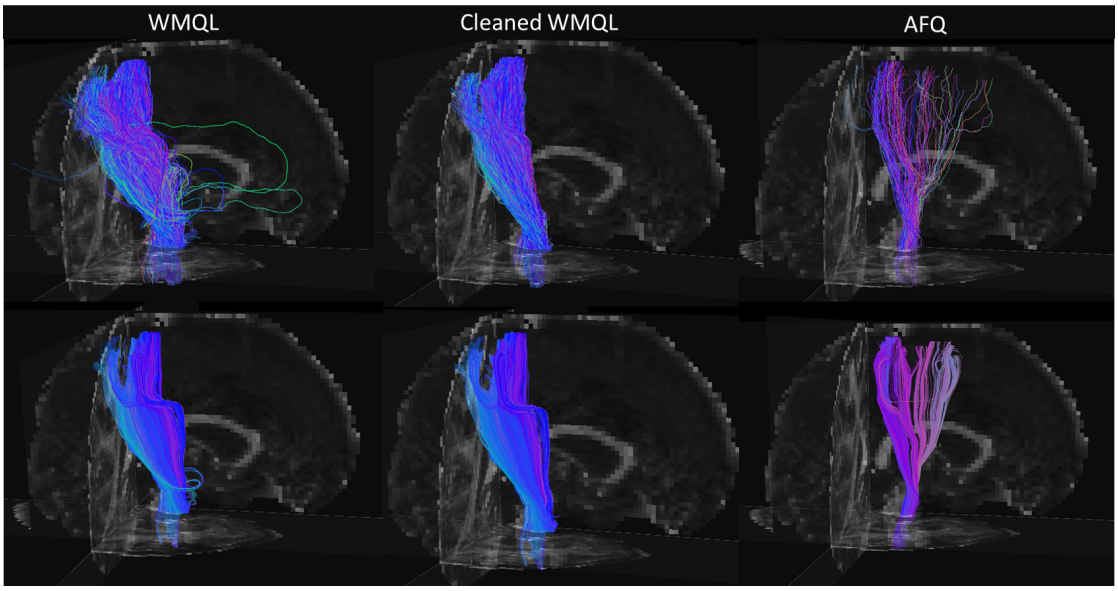

Figure 3.5: AFQ and WMQL comparison for the CST for deterministic tracking (bottom) and probabilistic tracking (top).

As pictured on figure 3.5, and as just mentioned, WMQL led to much better results than AFQ, by keeping more valid streamlines, which is preferable for the consequent analysis.

While being much more accurate on this dataset, WMQL only preforms bundle segmentation, and does not split them into segments. Nor does it include any filtering of the selected streamlines to remove fibres that do not belong to the mean bundle. It was then complemented by cleaning methods similar to AFQ.

It was tried to compensate for the lower number of streamlines selected by AFQ by repeating tracking for all bundle that showed low number of fibres, using Natbrainlab's probability density maps as seeding masks to track within this bundle only. However, the results were not convincing for most bundles and the segmentation remained poorer than with WMQL.

Here, WMQL was used to segment the tractograms into bundles before using AFQ-inspired methods to clean the bundles and split them into segments. This was compared with AFQ's segmentation, using the same steps after segmentation in both cases.

As first round of cleaning was done with the cleaning function of AFQ to keep only tracts that follows the same spatial patterns as the majority of the tracts within the bundle. For deterministic tracking, the cleaning step ends here end we

get to subdividing into segments. For this step a procedure similar as in AFQ was used:

- streamlines are resampled into 2∗n +1 points to form the AFQ segments
- the mahalanobis distance is evaluated at each of these segments
- the mean tract is computed at each segment as the average streamline weighted by the mahalanobis distance

Indeed, it is important for interpretability of the data that differences that are identified between patients are not due to a mismatching of the segments position. Though it could be interesting to know if there is a significant difference in bundle shapes between healthy and ill patients, it is an analysis that should be done separately from the microstructure analysis to be able to appreciate the contribution of both variables to the patients' differences.

Last but not least, this modification should only impact with spread endpoints like the arcuate fasciculus, and not/less the other ones, for which it would only fasten the computations if anything.

Once the mean tract is defined, it is needed to assign streamlines points to the correct segment. Indeed, here it was chosen to use a method that takes advantages of all the points generated by the tractography algorithm rather than sampling microstructure at only one point per streamline in each segment as in AFQ. The method used was then more inspired by RecoBundles's methodology:

- n centroids were defined from the 2n + 1 points of the mean tract (there are n segments, n points in their centre and n + 1 points at their boundaries)
- all points of each streamline were assigned to the segment for which the Euclidean distance with its centroid was minimal
- microstructure metrics were sampled at each point using trilinear interpolation

From there, the methods differ a bit from RecoBundles's. By taking the mean of all points belonging to a given segment, streamline segments with more points would have a greater impact than shorter ones. Similarly, streamline points corresponding to extreme metrics values or noisy streamline would strongly impact the mean value.

To resolve the first problem, it was chosen to use the mean of the sampled metrics for each points of the same streamline belonging to the segment before

taking the mean of the obtained values in the segment (which is equivalent to using a weighted average to take into account the number of points in each streamline).

Concerning the second problem, it was chosen to use a bootstrapped mean rather than a simple average. AFQ uses a weighted average according to the Mahalanobis distance to address this problem, but this implies making the hypothesis that streamlines are normally distributed around the mean tract.

Bootstrapping, on the other hand, will equivalently make a weighted average but without assuming the normality of streamlines' spatial distribution (which could be wrong in long bundles that could be organized in sub-bundles), and will not only take into account spatial distribution of the streamlines, but also the distribution of the sampled metrics. As bootstrapping is done on the metrics, so values that will be taken into account the most will be the most sampled ones, corresponding either to the metrics sampled where there are a lot of streamlines, or to the metrics sampled where there are less of them but the sampled metric is not different.

This should also reduce the impact of misaligned streamlines in the segment as they would be longer than aligned ones.

As the metrics images had low resolution (128x128x56) and some streamlines travels near the CSF or the image background (near cortical areas that are thinner than the voxels dimensions), it was needed to discard the points where the sampled FA is lower than 0.1. However, it would be better to not remove the contributions of those points but rather adapt the trilinear interpolation algorithm so that it only takes into account voxels where the FA is higher than 0.1.

To define the number n of segments, mean tracts were first computed on a randomly chosen patient by fixing the segment length to 4mm (to have segments with length greater than the voxel size). Then, to standardize the result, the number of segments of the left and right bundles were replaced by the mean of the two, and the number of segments of the callosum bundles were chosen to be equal. The number of segments of each bundle were then adjusted to be consistent with one another (2 times more segments for the IFOF than for the anterior thalamic radiations, for instance). A last adjustment was performed by computing the mean length of the streamlines segments, for each segment [4] and comparing between 5 patients to try keeping a 4mm thickness in average for all segments.

These are the only steps for deterministic tracking, but probabilistic tracking

[4] which would be overestimated with probabilistic tracking and for highly curved bundles

needed a further round of cleaning. It was done using a mixed method implying the Mahalanobis distance and the Relative Fiber to Bundle Coherence (RFBC) to remove streamlines that were badly aligned because they deviate from the mean bundle in a restricted location, while preserving streamlines that are in the core of the bundle but not perfectly aligned with it. It was possible to use this method to clean bundles quite well but not without removing a lot of streamlines while it is not needed to clean that much. It was then chosen to use less strict parameters to keep a certain compromise between the amount of noise in streamlines trajectory, and the amount of streamlines that allows to compensate for the noise in the sampled metrics images.

3.4 Analysis

Several analyses were conducted to study the impact of vertigo on cerebral microstructure. First, statistical analysis was performed to find any significant dependency of the vestibular tests with age or structural lateralization.

Second, Machine learning analysis was performed on several datasets: the values sampled with the Freesurfer white-matter parcellation, along the bundles obtained with deterministic tracking, along the bundles obtained with probabilistic tracking and finally on the structural lateralization data.

Finally, statistical analysis of the FA maps was performed using Elikopy [57] (with a similar procedure to FSL's TBSS [69, 70]) to compare the results with machine learning analysis.

3.4.1 Structural lateralization

To study structural lateralization, the numbers of streamlines in left and right bundles were measured for each patient and used to compute a lateralization index [37]:

$$I = \frac{R - L}{R + L}$$

Where R is the number of streamlines in the right hemisphere and L the number of streamlines in the left hemisphere. This index goes from -1 (fully lateralized to the left) to 1 (fully lateralized to the right). As bundles are resampled to 3000 streamlines before cleaning and knowing that cleaning removes a given percentage of the streamlines, it was chosen to use the number of streamlines in the not-cleaned

bundles.

As some bundles were not found in some patients (especially for deterministic tracking), some lateralization index had incorrect values of 0 (the bundle was not found in both hemispheres) or $\pm$ 1 (because the bundle was not found in one of the hemisphere). In the last case, it obviously impairs the estimation of the lateralization index as finding 10, 100 or 1000 streamlines in the right hemisphere bundle would lead to a lateralization index of 1 if no streamlines are found in the left hemisphere. That is why when performing the statistical tests, only patients with streamlines in both left and right bundles were kept for each bundle.

Similarly, as some vestibular tests were not carried out on all subjects, only patients that have undergone a given vestibular test were used for the statistical tests corresponding to this vestibular test.

3.4.2 Machine learning

Data analysis was conducted by fitting several models (Logistic regression with elasticnet, lasso, ridge or sparse-group lasso penalty, SVM and RDF) in a nested cross-validation loop to exhaustively evaluate the ability of the models to predict vertigo for all tested combination of hyperparameters. Nested cross-validation was performed using *afqinsight*'s library that extends *scikit-learn*'s, and offers facilities for models using groups and saving models used during cross-validation.

As there were a limited amount of patients in the least populated class (6 healthy patients for 50 patients with non-compensated vertigo and 12 with compensated vertigo), Repeated Stratified 5-fold cross-validation was used for the outer cross-validation loop to limit the most the impact of class representation unbalance. As the cross-validation is stratified it maintains the ratio of class populations in the folds and each fold contains 1-2 healthy patients, 2- patients with compensated vertigo and 8 patients with non-compensated vertigo. But leave-one-out is known to have a great bias and variance in performance estimation, and a "repeated" cross-validation was used to improve the estimation of models performances by rehearsing the cross-validation 7 times (for a total of 49 folds).

Hyperparameters as well as pre-processing steps (choice of scaling and filtering methods) were tested using the GridSearch approach: by fitting all possible combination of the chosen parameters. Here a Stratified 4-fold cross-validation was used as the inner loop of nested cross-validation, with no repeats.

A quite broad range of values were tested for all parameters to apprehend the

behaviour of models on this dataset to find the best parameters for each model in a reasonable amount of time, and tested hyperparameters were adjusted afterward if necessary.

Several steps were done before running the models on the data:

- a first feature selection to remove feature that were either constant across subjects or missing in more than half patients of one class
- missing values were replaced by 0
- PCA was applied in some cases to reduce the dimension of data, as explained later
- data was pruned by removing features that were strongly correlated together, keeping only the one that is the more correlated with the target (vertigo or the result to the vestibular tests)
- scaling of the data
- selection of the features using F-test or mutual information

Scaling For the scaling step, standardization was used as baseline and compared that are more conservative of features singularity (for example if the precise value of the features conveys information):

- Min-max scaling: minimum value is subtracted before dividing by the difference between the maximal and minimal values. All features are then between 0 and 1 but have their own means and standard deviation.
- Max-abs scaling: it rescales the data so that their maximum absolute value is 1. For positive features, it will lead to a stronger difference of mean and standard deviation than Min-max scaling[5].

Those two approaches then tends to preserve more information about the mean and standard deviation of the features. The choice of the right scaling method strongly depends on both the data and the models, it is then often hard to predict the impact on the final results.

[5]For features with negatives values, each feature is then comprised between -1 and 1, and have its own mean and standard deviation, but its sign is conserved.

PCA The results obtained with the base data were then compared with results obtained on the data transformed by PCA. Concerning the along-tract sampled data, previous work has shown a reduced performance when using PCA to reduce the number of studied segments [47], while using PCA on data sampled along a given tract to reduce the number of metrics considered yield interpretable results [71, 72]. It was then chosen to apply the PCA independently for each tract to find components expressed by the DTI.

The procedure was the same with the white-matter parcellation, except it was applied for each region instead of each tract, and it was compared with applying the PCA on corresponding grey and white matter regions of both right and left hemispheres together. In both cases, only 2 components were kept.

Age influence As there were a significant difference (p-value < 0.05 to an ANOVA test) between the mean age of the patients in each class ("healthy"<"compensated"<"vertigo"), several results were compared:

- using the raw data
- including the age as a new feature in the data in the hope it would be selected in place of variables that depends more on the age than on the vertigo, so that only features that are poorly (on their own) related to the age would be selected
- using linear regression to remove the **linear** link that exists between the age and the microstructural metric

A drawback of the second approach is that it would also bias the performances of the model: if we add the age to the parameters and the age is predictive of the patients condition, then the performances should increase if the age was selected and there were no links between the age and the microstructure and the age. But if the age is selected and there is no increase in performances, it means it has replaced features that were only selected for their link with age[6]. And if the age is not selected, it does not mean there is no relation with the age, but rather that it was judged (by the model) as less significant than the link with vertigo.

The first approach was only run with the white-matter parcellation to compare with the second approach, as the age was not often selected as one of the most important features using the second approach. For the second approach, a model that uses the age as one of the most important variable for its prediction was not

[6]Or at least that the link with the age is more important than the link with vertigo.

considered as a viable model for vertigo prediction, but still studied for its ability to extract age dependencies in the data. This is the reason for the use of the third method, to see if applying a correction for the age lead to significantly different results for the previously selected models, and if the model selecting the age as a core variable would be able to compensate for its absence by selecting other variables.

This last method is not ideal as the evolution of microstructure with age is not linear, as suggested by some paper it is more probable to be a quadratic relationship [73, 74], but *Kochunov et al.* showed a linear relationship could still capture most of the link above 30 years [74] (which is inferior to most of the patients' age here), and using a polynomial regression would increase the risk of a misestimation of the metrics evolution with age, especially knowing that the exact relation seems to differ from men to women [73]. But a simple linear regression would still be impaired by the much higher number of older patients than younger ones, and the correction would be good for older patients but not for young ones.

Moreover, microstructure evolution with age varies with brain regions and sex [73, 75]. We then used a weighted linear regression for each region: by dividing patients in 10 groups of similar ages and weighting the contribution of each patient to the linear regression by the inverse of the number of patient in the same age group[7].

Model evaluation Models performances were assessed with several metrics of interest. Apart from accuracy, precision and recall, were used:

- balanced accuracy: the mean of accuracy for each class, so that each class weight for 50% in the score in a binary classification scheme, or 33% if there are 3 classes
- top 2 accuracy: it counts not only the number of correct predictions, but also the number of incorrect ones for which the correct prediction was the second choice of the model. It was adjusted for class imbalance by taking the same weighted average used in *scikit-learn* models implementation
- "corrected" top 2 accuracy: an alternative implementation of the top 2 accuracy based on *scikit-learn*'s version. The difference is that we only count the second choice as valid if its associated probability is closer to the one of the first choice than the third choice (ie the model hesitate between the two first choices but is sure of what the patient is not: the third choice), and if

[7]The weights were normalized by the mean of the weights, to ensure the mean would be equal to 1.

the probability associated to the third choice is below 20% (ie the model is at least sure at 80% that the patient does not belong to the class of its third choice). It was adjusted for class imbalance by taking the same weighted average used in *scikit-learn* models implementation

- the "unclassable" patients were defined as the patient for which all model's predicted probabilities were above 30%, so that the maximum gap between the predicted probability for 2 classes in a 3 classes classification scheme is 10% (30%, 30% and 40% for instance). It was adjusted for class imbalance by taking the same weighted average used in *scikit-learn* models implementation

To gain time in the cross-validation process, after the first inner loop, models with mean performances across all inner loops lower than what is expected by a random model were discarded (with 3 classes, accuracy of 33%, same value for the balanced accuracy). After each inner loop, the model was tested on the test-set data of the outer loop to estimate generalization abilities of the model. If mean scores on the test set were significantly lower than mean scores of the inner loop, the model was considered to overfit, and consequently discarded.

Features' impact on the models' predictions was estimated with the permutation framework presented earlier in 2.5.3, and the computed feature importances were averaged across the cross-validations. Similarly, the permutation test score was used with a 2-repeats-5-folds cross-validation with 399 permutations [8] and a model was considered reliable if the cross-validated p-value obtained by the test was lower than 0.05.

Figure 3.6: Caption

Ensemble models Base *scikit-learn*'s implementation of both BaggingClassifier and RandomForest does not entirely take class imbalance into account: for each sub-model it randomly draws the subjects on which it will train itself. But when randomly drawing subjects with the same weights, it can happen there are no "healthy" patients in the drawn samples, and even when it does not the class imbalance will be drastically increased and have a stronger impact on model's performances evaluation than the sub-models that had drawn (by chance, it will rarely happen) more "healthy" subjects than there are in the initial data.

[8] for a total of 3990 permutations and a minimum p-value of 0.0025 for each cross-validation fold

Those base ensemble models were then compared with another implementation, where the sampling function of *scikit-learn* was modified to either maintain the class ratios in the sampled subjects, or to reduce it.

Chapter 4

Results

4.1 Age influence

Difference of age between patients groups was studied with a False Discovery Rate (FDR) corrected t-test that revealed significant age differences between patients with normal and abnormal results to the following vestibular tests: bilateral cVEMP tests and left oVEMP test with after correction, and between the left caloric VNG, and the right oVEMP test.

An ANOVA test revealed the 3 classes "healthy/compensated/vertigo" were nearly significantly different (p-value of 0.053). The difference between the 3 classes is displayed figure 4.1. We can see that Vertigo and Compensated patients are on average older than Healthy ones, and the difference is stronger for Vertigo patients.

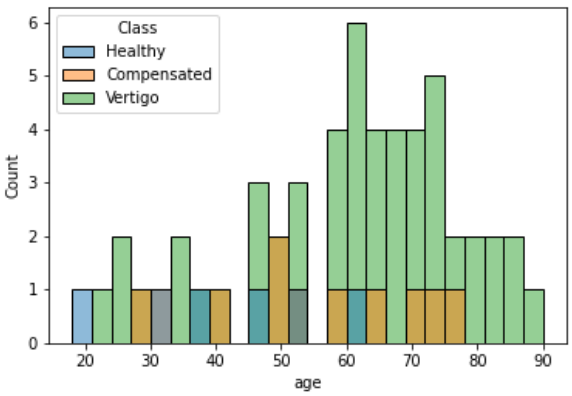

Figure 4.1: Age differences between patients

Correction of the microstructural metrics using linear regression with the par-

cellation data was quite incoherent with age evolution of FA that can be found in the literature: there were an increase of FA with age in many regions, while most brain structures are known to have reduced FA with age in healthy subjects [76]. This would suggest there may be an increase in FA with vertigo (as healthy patients were younger) that was sufficient to noise the estimation of age dependence.

The same was observed using the tract-based data for the linear regression. However, regions with incoherent age evolution were also regions with poorer correlation with age. This could explain why evolution with age in those regions was miscomputed.

However,

4.2 Elikopy statistical analysis

The results of statistical analysis using Elikopy led to no significant p-values after correction so that the results presented here are only related to regions that have low p-values before correction. For sake of brevity, only 3 sections of different anatomical orientations are presented here, but more complete images can be found on the Github. Significant regions of the FA skeleton were filled (using tbss_fill script of FSL) to make the presentation easy (skeleton appears clearer than surrounding white matter).

As Elikopy uses the FSL randomise analysis, which relies on bootstrapping, the selected regions should be the ones that do not variates much with age or that is more influenced by vertigo than age. For instance, if Vertigo patients have lower FA, even younger ones, the difference with other patients will be even stronger with older vertigo patients. On the contrary, if the age have a stronger impact than Vertigo the younger patients would typically have higher FA than Healthy or Compensating patients and the difference of FA would only be significant for older patients and the regions are less likely to be selected.

Healthy vs Vertigo The regions near the selected regions were coherent with previously reported regions related to the vestibular system:

- Posterior Insula (left)
- Occipital lobe
- Inferior Parietal lobe (mostly right)

- Prefrontal cortex (mostly right)
- Brainstem (mostly right)
- higher cerebellum

However, those regions were quite small as can be seen on figure 4.2, so that the only large region that is selected is the right brainstem. Nearly all regions were selected because FA of Vertigo patients was lower than Healthy patients, which is opposite to the natural evolution with age and could explain regions were quite limited.

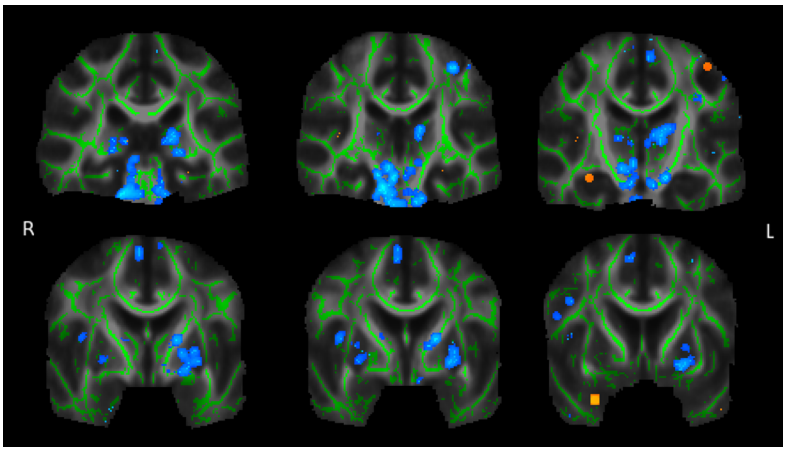

Figure 4.2: Healthy - Vertigo FA analysis. Blue regions means FA is lower for Vertigo patients than Healthy ones, Red means FA is higher for Vertigo patients than Healthy ones.

Healthy vs Compensated There were selected regions similar to the ones discriminating the Healthy and Vertigo patients (brainstem and), but other regions were selected too: temporal (apparently along the MdLF and ILF, mostly in the right hemisphere), occipital (mostly in the right hemisphere) and frontal (near the prefrontal cortex, mostly in the left hemisphere).

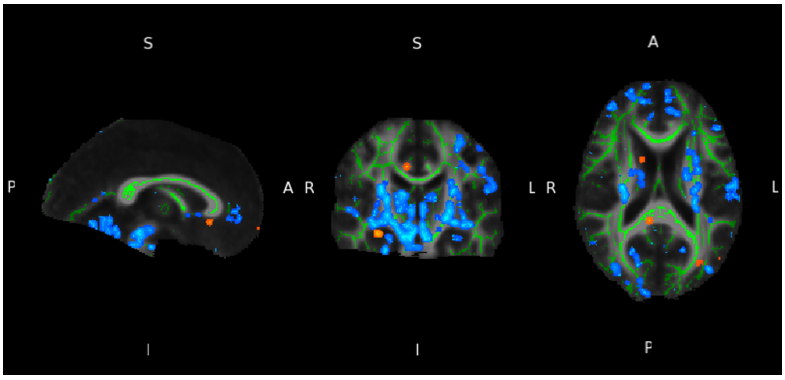

Figure 4.3: Healthy - Compensated FA analysis. Blue regions means FA is lower for Compensating patients than Healthy ones, Red means FA is higher for Compensating patients than Healthy ones.

Compensated vs Vertigo As expected from the comparison Healthy/Compensated, frontal regions were found as cores of differences between the compensating and non-compensating patients. However, frontal regions were significant in both hemispheres for the Healthy/Compensated comparison and in the right hemisphere for the Healthy/Vertigo comparison. Therefore, the fact the left hemisphere is selected in this case seems coherent.

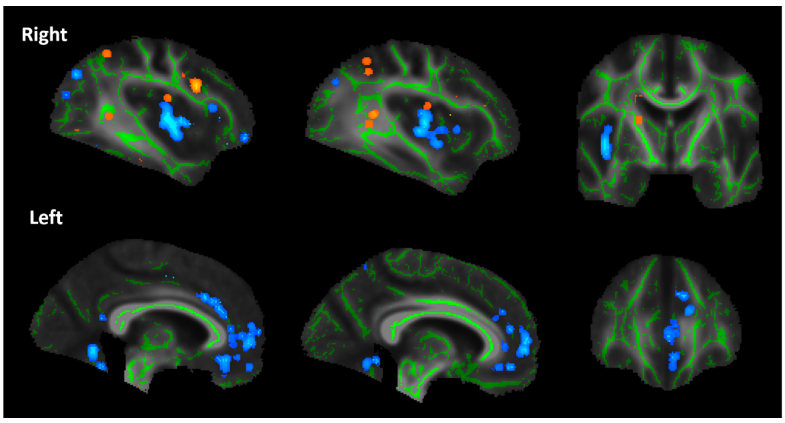

Figure 4.4: Compensated - Vertigo FA analysis. Right hemisphere on the top, left on the bottom. Blue regions means FA is lower for Compensating patients than Healthy ones, Red means FA is higher for Compensating patients than Healthy ones.

A point of concern with those results is the age difference between Compensating and Vertigo patients, as FA is lower in older adults and selected regions are found with higher FA for Vertigo patients (who are younger). The results can then be linked to compensation, or only to age. But as those regions were also selected as

having lower FA than Healthy patients, the latter case is less likely as Compensating and Healthy patients have similar ages.

4.3 Machine learning analysis

This section will expose the main results obtained from machine learning analysis, starting from basic analysis of the interactions between the features and estimating the importance of each vestibular test in the "healthy/compensated/vertigo" classification.

For brevity's sake, "ElasticNet" will designate the logistic regression classification model with ElasticNet penalty. The vestibular tests were simplified to "normal" and "abnormal" to ease classification when predicting the results to those tests.

SVM model trained on the vestibular tests had balanced accuracy between 49 and 70%. Some models designated the left caloric VNG test as the most reliable parameter for the prediction followed by the VEMPs and other VNG tests, the age, and finally the VHIT test that were always left as unimportant, while other models took the VHIT into account in place of the age, VEMPs or VNG tests but the left caloric VNG was always in the top important features.

A last note before going into details for the different datasets is that RDF never led to satisfying results. The performances were actually most of the time worse than a random model, even with its updated implementation to maintain class ratios. It could be hypothesized that the high cardinality of data is responsible for it, but this is probably not the only reason. Indeed, random performances were obtained with Bagging too, which means that predictions strongly varies between the sub-models, and taking their average leads to random predictions. That would sustain the fact there are different patterns of differences between patients and that another classification of the patients could be studied to group Vertigo patients with similar symptoms together, for example.

Similarly, no SGL model survived cross-validation, either because the correction for class imbalance was not properly implemented[1], or because there are no sufficient bundle-wise differences between classes: no bundles shows a significant difference between classes across all of its segments and the "best" models use only Lasso penalty and no Ridge penalty.

[1]SGL had never a balanced accuracy above 33% without the correction, and get untill 38% with it

4.3.1 White-matter parcellation

In this part, we will summarize the main results obtained with the data sampled from the white-matter parcellation, starting from the best results that were obtained with the entire data (white-matter, grey-matter, subcortical grey-matter and cerebellum regions).

Entire data As can be seen on figure 4.5, values sampled with the white-matter parcellation were strongly correlated together, but applying the PCA allowed to decorrelate the data on top of reducing its dimensions. Obviously, the PCA will inherently decorrelate features as it projects data onto the directions of maximum covariance, but here PCA was not applied on the whole data. It was applied region by region, by grouping analog cortical and white-matter regions of both hemispheres together. It is then normal that features decorrelate within the same group, but it is interesting to see that the correlation between principal components of different groups is much smaller than between the initial features.

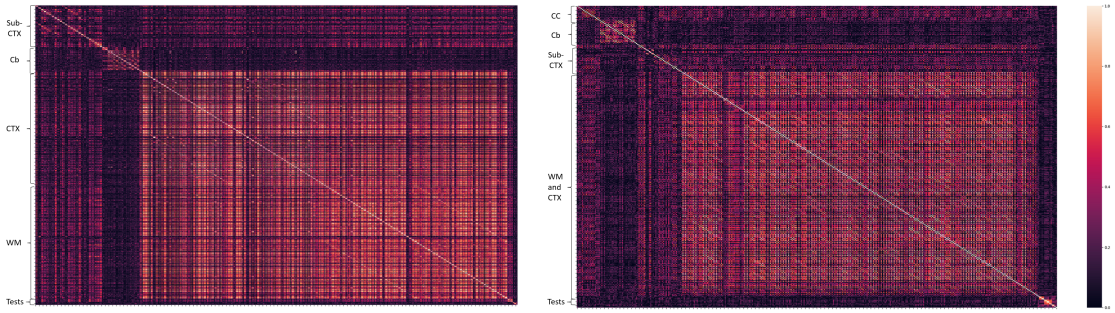

Figure 4.5: Correlation maps for the parcellation dataset, for the original data (left) and after PCA (right). Columns and ligns follow the same order: CC stands for corpus callosum, Cb for cerebellum, CTX for cortex, WM for white matter, subCTX for subcortical grey matter and Test for the vestibular tests.

Even if those components quite varies from one group to another - not only from subcortical regions to white-matter but also from one lobe to another - there are always several groups with similar components (for example equivalent contribution of FA and all diffusion metrics for the first component in some regions, or equivalent contribution of all diffusion metrics and a negative contribution of FA for other regions), but the correlation between their components is still poor. This means that even if one metric in one region is correlated with the same metric in other regions, the variations of the four metrics as a whole is specific to each region.

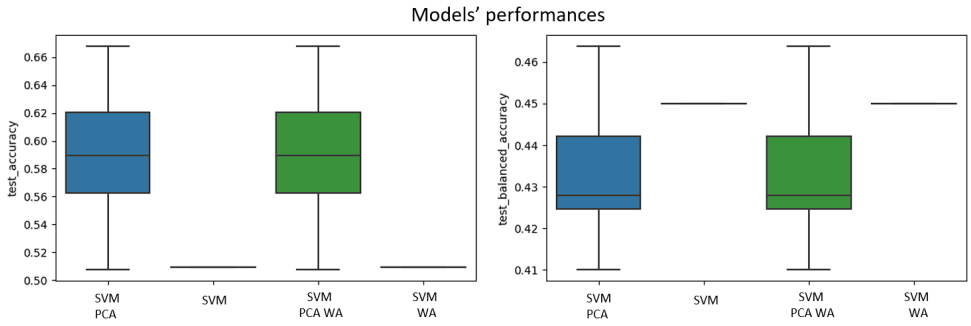

Figure 4.6: Models performances on the parcellation data. WA means the age correction was used.

Figure 4.6 show the performances of the models for several metrics. We can see only SVM models survived the cross-validation with a positive result to the permutation-test, especially when PCA was used. SVM tends to have much higher accuracy than balanced accuracy, meaning Vertigo patients are well classified (it is the overrepresented class), while Compensated and Healthy patients are less. As the balanced accuracy remains quite low (below 55%), it then means that in some folds of the cross-validation, no Healthy and/or Compensated patients were correctly classified.

White-matter and grey-matter If we remove the cerebellum and subcortical regions, the models select more or less the same regions, but the performances are reduced: fewer models survive cross-validation and the models that does not reach the maximum performance obtained with the entire data. This confirms that cerebellum and subcortical regions contains unique information that is relevant for vertigo identification, and that most models that were selected before are not able to compensate for this loss of information with other variables.

White-matter only Here instead of keeping only white-matter and cortical grey-matter, the only regions that were used were the white-matter ones. Similarly as in the previous paragraph: performances were reduced, but this was emphasized for PCA as here the performances using PCA were inferior to the performances using the microstructural metrics.

With this dataset, PCA could obviously not be applied the way explained in the *Methods* section as the grey-matter was not used any more. It was then applied with the white-matter only, and the components were different from before, which could explain the reduced performances of models.

PCA The PCA components quite varied across regions: they were similar in subcortical and deep white-matter regions (such as the corpus callosum and the brainstem), but different in cerebellum regions, and again different in the cortical regions and their corresponding white-matter regions. The main observations for the first component are

- deep white-matter[2]: strong and equivalent contribution of AD, MD, RD and smaller (half to third) and negative contribution of FA. The exceptions were the central corpus callosum for which the FA contribution was positive, and the mid-posterior and posterior corpus callosum for which it had nearly the same absolute contribution as MD and RD, and the contribution of AD was lower.

- subcortical grey-matter regions: strong and equivalent contribution of AD, MD, RD and smaller (third to sixth) and negative contribution of FA. The only exception were the Amygdala for which the FA contribution was positive, and the pallidum for which it was half the contribution of AD, MD or RD.

- cerebellum: strong and equivalent contribution of AD, MD, RD and smaller (half to fourth) contribution of FA.

- cortical and white-matter: strong and equivalent contribution of AD, MD, RD and smaller though similar (more than half) and negative contribution of FA.

Concerning the second component, the main conclusions are:

- deep white-matter: strong contribution of FA, and a lower (half) contribution of either AD (positive) or RD (negative).

- subcortical regions: strong contribution of FA, lower (fourth) contribution of AD, low and negative contribution of RD or no contribution of it.

- cerebellum: strong contribution of FA, lower (fourth) and negative contribution of AD and RD.

- cortical and white-matter: strong contribution of FA and AD (half FA's), lower (fourth) contribution of MD and RD.

When correcting for the age, most principal components were similar for deep white-matter, subcortical regions and cerebellum, with a slightly decreased FA contribution (as FA is increased for aged patients after the correction) and increased

[2]corpus callosum, optic chiasm, brainstem and unsegmented white-matter

diffusion contribution (it is decreased for aged patients after the correction). This interpretation is however less pertinent for the cerebellum where the FA contribution is positive, but it may be due to cerebellum segmentation that was used: white matter of cerebellum's gyri was taken as belonging to the lobe instead of being separated into cortical grey-matter and subcortical white-matter. This would obviously increase the FA measured and could explain the difference of principal components with the other regions.

For cortical regions and their associated white-matter, there were a strong difference with the correction for most regions. It led to principal components with lower loadings for the first components and higher for the second components: the loadings were typically 90% and 9% before correction, while 70% and 29% after the correction.

The second component of most regions were similar after the correction: strong contribution of FA, lower (third to fifth) contribution of AD (positive) and RD (negative). Concerning the first component, the absolute contribution of FA was lower in most regions (barely or not contributing), and opposite in some regions (from a negative contribution to a positive one), or even gets from a negative to a positive one. This component is then much harder to interpret, as it varies more between regions.

From those observations, we could hypothesize that the first component is mainly indicative of the isotropic diffusion (high diffusion, low FA) while the first one would be more indicative of white-matter integrity (high FA and AD, low RD)[77] for most regions. This hypothesis is much more relevant after age correction.

Graphical summary Figure 4.7 depicts the regions selected by the models. More detailed images can be found on Github[3]. Many regions are selected, but left parietal is more strongly selected for the first component of PCA, along with the white matter underlying the primary and supplementary motor cortex and cingular cortex (bilateraly). Regions selected for the second component were right temporal and frontal regions, and left medialorbitofrontal.

[3] `https://github.com/Dany546/TFE_vertigo.git`

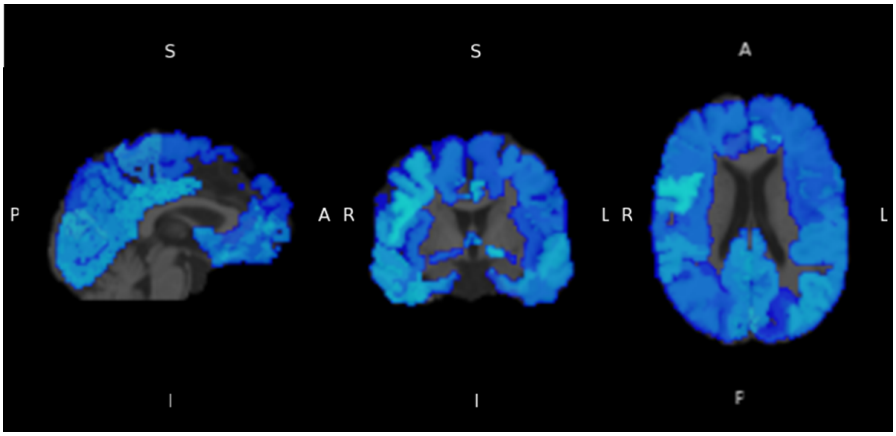

Figure 4.7: Selected regions of Freesurfer's parcellation shown on one patient's anatomy. The colour map scale was arbitrarily chosen for visualization purposes, averaging the feature importances across models. Brighter colours mean a greater impact on models' performances.

However, the fact the results are lateralized must be mitigated by the correlation between left and right hemisphere homologous. The regions selected with the age correction or without PCA were the same.

Models trained to predict the results of patients to the vestibular tests were less efficient (no more than 60% of balanced accuracy for binary classification) and much fewer models survived cross-validation. However, the models that pass cross-validation selected similar regions than the ones selected in the Healthy-Vertigo-Compensated classification. More precisely, VHIT test were related to brainstem, corpus callosum and motor cortex, while VEMP tests were more related to frontal regions, motor cortex, and parietal cortex.

4.3.2 Tract-based analysis

Here, the results were obtained from with different methods: either with the original tract-based dataset (with or without age correction), or with the same data transformed by PCA. As explained in the Methods (**??**), there were several ways of applying PCA: on each bundle independently or using segments of both hemispheres as samples for the PCA.

Results obtained with probabilistic tractograms were compared with results obtained with deterministic tractograms. As deterministic tracking did not allow finding biomarkers for vertigo (no models getting through the cross-validation), only the results obtained with probabilistic tracking will be detailed.

The first observations that can be drawn from these datasets are related to the correlations between features. As pictured on figure 4.8, features are much less correlated compared to the parcellation data. The features that are correlated together are from the same bundle but from different segments when there are only slight evolution of the microstructure along the bundle, or metrics sampled from segments of similar locations (for example crossings of the CC, CST and thalamo-sensorimotor bundles).

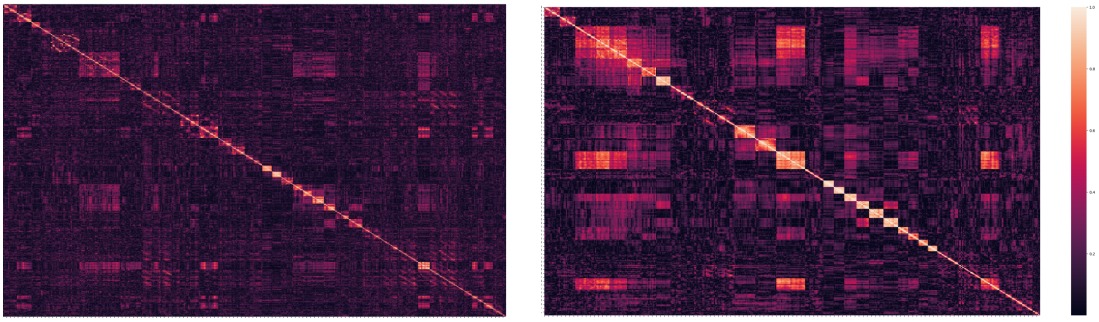

Figure 4.8: Correlation maps for the tract-based dataset, for the original data (left) and after PCA (right). Ligns and columns are organized in alphabetical order: AF, ATR, etc ...

As we can see on figure 4.9, models have higher performances than what was obtained with the parcellation dataset, especially when correcting for the age. There were much more models that pass the performance permutation-test, and less overfitting.

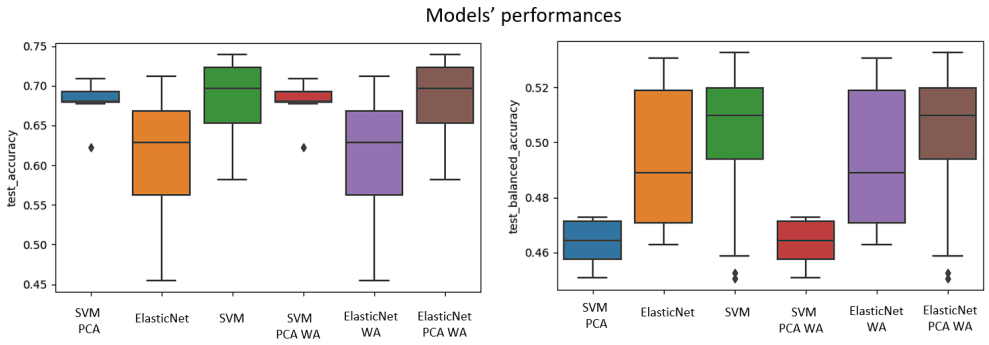

Figure 4.9: Models performances on the tract-based data, on the test sets. Accuracy (left) and balanced accuracy (right). WA means the age correction was used.

Performances were lower with PCA on SVM, but had a positive impact on ElasticNet's performances as no models survived cross validation without it.

PCA on original data In contrast to the parcellation data, PCA applied along bundles yield to similar results for all bundles. For most of them, the 2 main components ($\simeq$99% of cumulated explained variance ratio) were the following:

1st component : it had an equivalent contribution of AD, MD and RD, and smaller (third to 2/3) contribution of FA

2nd component : it had a strong contribution of FA, a smaller (half) and negative contribution of RD, and a smaller (third) to no contribution of AD and MD

The first component could be interpreted as indicative of the isotropic diffusion, with a FA contribution because of the small correlation of FA with AD, MD and RD in white matter. On the other hand, the second component could be indicative of white matter integrity because it increases with FA, AD and MD but decreases with RD[77].

However, there were some exceptions: the first component had a negative contribution of FA for the CST, and thalamic projections with the superiorparietal and inferiorparietal gyri. This remains coherent with the previous interpretation for the first component as related to isotropic diffusion, except here there is an increased difference between patients in segments where the FA is low, as near the cortex.

PCA on data corrected by age As for the parcellation data, the components after correction by age were similar for most bundles, with slightly increased contributions of AD, MD and RD and a slightly lower contribution of FA.

The performances were better than the ones obtained without the correction: a higher balanced accuracy and lower accuracy (less favouritism of one class), but also a higher top-2-accuracy (the right class is less often taken as the last choice).

PCA applied unilaterally When the PCA was applied on each bundle separately, instead of both right and left hemisphere simultaneously, there were obviously small differences between right and left components contribution and loadings. However, these small differences were sufficient to reduce the performances of the models, so that no models were selected as reliable.

Graphical summary Figure 4.10 depicts the regions selected by the models. More detailed images can be found on Github[4]. Many regions are selected, but left

[4]https://github.com/Dany546/TFE_vertigo.git

parietal is more strongly selected for the first component of PCA, along with the white matter underlying the primary and supplementary motor cortex (bilateraly). Regions selected for the second component were right temporal and frontal regions, and left medialorbitofrontal.

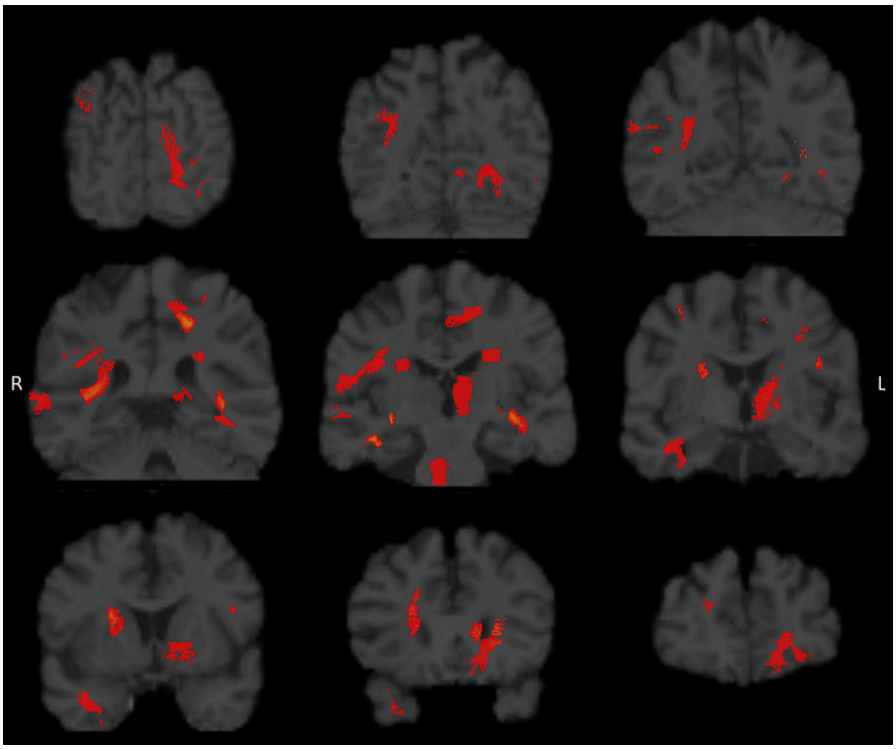

Figure 4.10: Selected segments shown on one patient's anatomy. The colour map scale was arbitrarily chosen for visualization purposes, averaging the feature importances across models. Brighter colours mean a greater impact on models' performances.

However, the fact the results are lateralized must be mitigated by the correlation between left and right hemisphere homologous.
Interestingly, more regions are fond when identifying those that have a negative impact on the predictions when the importance of these features is miscomputed by the model. Occipital, parietal, temporal and frontal region are more strongly selected, especially in the right hemisphere for occipital and inferior-temporal areas for the first component of PCA, and right parietal and superior-frontal and left temporal regions for the second component.

Models trained to predict the results of patients to the vestibular tests showed similar performances and results than with the parcellation, but selected fewer regions: VHIT test were related to brainstem, corpus callosum and motor cortex,

while VEMP tests were only related with segments of the AF and SLF in the opercular regions, near the posterior insula (see figure 4.11).

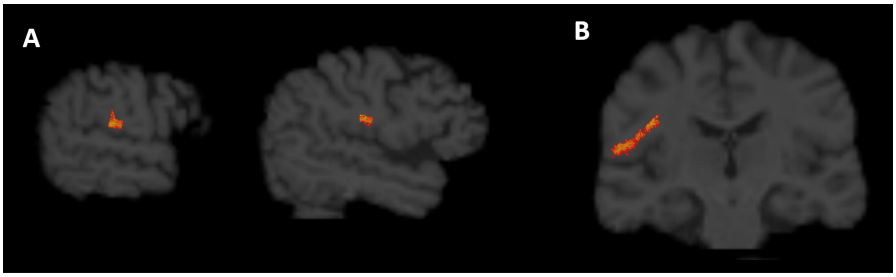

Figure 4.11: Selected segments shown on one patient's anatomy for right oVEMP test: right hemisphere sagittal section (A), frontal section (B). The colour map scale was arbitrarily chosen for visualization purposes, averaging the feature importances across models. Brighter colours mean a greater impact on models' performances.

It was shown on figure 4.10 that the right brainstem was relevant for prediction of vertigo. To dig a bit further, a T-test (without correction) was performed on the data sampled along the right CST, and revealed the second component of PCA (pc1) in the first segments of this bundle had a significant link with the results to the VHIT test. In these segments, the results to the VHIT test were worse for patients with higher pca1. Figure 4.12 shows the evolution of this PCA component for the three Healthy/Compensated/Vertigo classes.

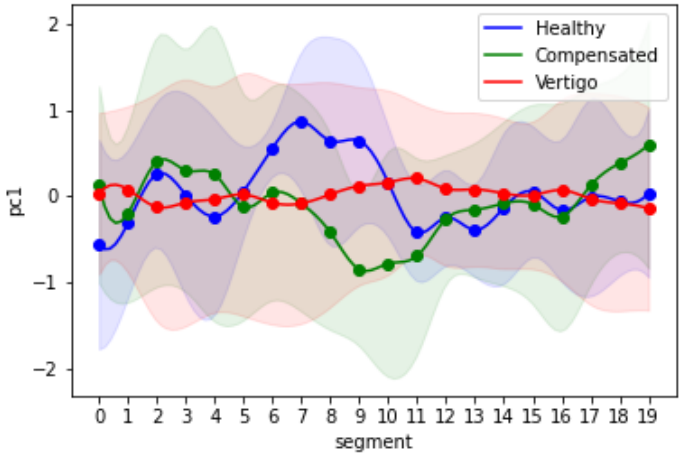

Figure 4.12: Evolution of the second component of PCA along the right CST, from the lower brainstem (segment 0) to the motor cortex (segment 19).

This would mean the results to the VHIT test worsen with a higher FA and AD, and lower RD in the right brainstem. One hypothesis could be there is a lower structural complexity, for instance because of a reduced communication within the right vestibular nuclei and the cerebellum.

4.4 Structural lateralization

Before presenting the results themselves, it is important to first give an insight on what could be the source of structural lateralization. Indeed, the first impulse we could have when seeing there is a lateralization is to think there are really more tracts in one hemisphere than in the other. But the thing is that the lateralization computed here reflects more the capacity of the tracking algorithm to reconstruct the tract than the lateralization of the tract itself.

To be more specific, let's think about an example: if there is a microstructural pattern that would impair tracking efficiency (like low FA and diffusion) in a local region of the brain, as the internal capsule, and that this pattern is present in the left hemisphere only, it could impact the lateralization of several bundles (the CST and thalamic projections with the sensorimotor cortex for example, as both are intertwined). But in reality there could be either lateralization of both bundles, lateralization of one of the 2, or no lateralization at all depending on the source of the microstructural difference between left and right hemispheres (for instance, lateralization of another bundle which cross their trajectory).

It could also be the result of the lateralization of a non-identified bundle: if a bundle that is not segmented is lateralized, it could locally noise the sampled microstructure along the bundle it crosses. So that only local microstructure differences between left and right would be identified along the segmented bundles.

On the other hand, if there really is a lateralization of a given bundle, we could think it would result in an observed microstructural difference between left and right hemispheres along the main part of the bundle. But local differences could still be a sign of lateralization if there is lateralized microstructural differences along bundles that crosses the one of interest such that those differences along the other bundles compensate the differences along the bundle of interest.

Figure 4.13 shows the results for structural lateralization index. Results obtained with deterministic tracking were more accentuated than with probabilistic tracking, and there were more missing data, making deterministic tracking results less reliable for most tracts. Another source of differences is the link between the lateralization

and the tracking algorithm itself: deterministic tracking tend to favour some trajectories against others (typically simple and short ones are preferred to long and complex) so that if a microstructural mismatch exists between the left and right hemisphere, it will have a stronger impact on favoured trajectories and lateralization indexes are poorly correlated together.

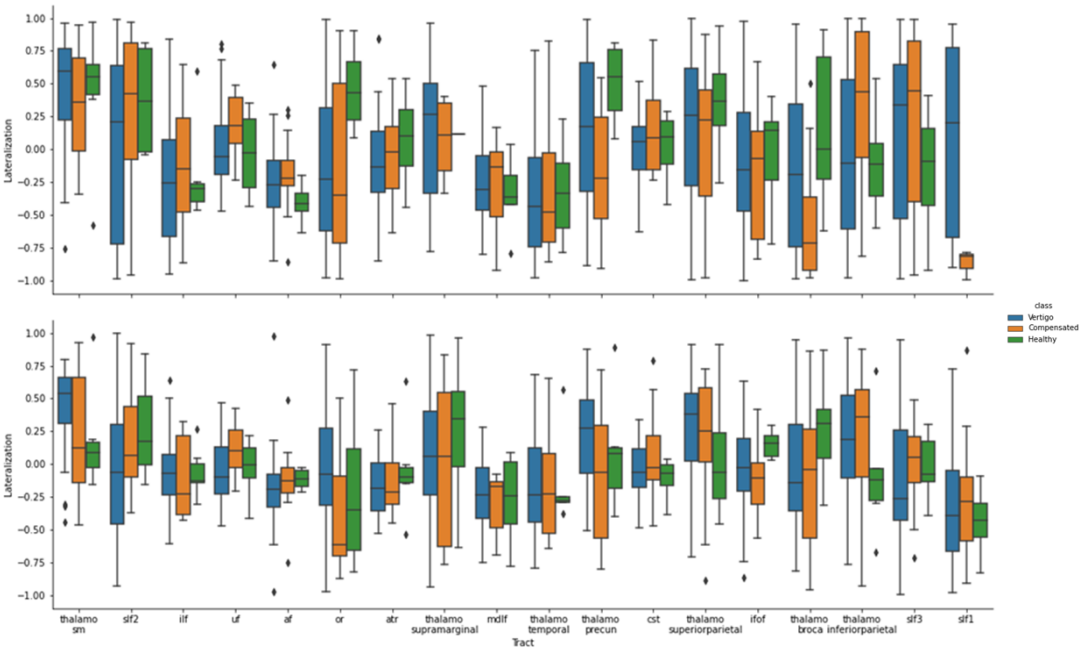

Figure 4.13: Lateralisation index for all segmented tracts obtained with deterministic tracking (top) and probabilistic tracking (bottom).

On the other hand, probabilistic tracking, when facing the same event, will distribute the impact over all possible tracts as it does not favour simple trajectories as much as deterministic tracking, and the lateralization indexes are more correlated together than with deterministic tracking (especially for thalamic projections).

It will then be needed to compare deterministic and probabilistic results with microstructural and lateralization analysis. This is crucial to be able to put the results into the right context and not to come up with misleading interpretation, but also to understand possible contradicting results.

Probabilistic tracking Neither Mann–Whitney U test nor T-test yield to significant result for vertigo (p-value < 0.05) once FDR correction applied. However, the IFOF was the only one with a significant p-value for vertigo before the correction.

The only test that passed FDR correction was concerning the thalamic projections with the broca's area: it is significantly correlated with right cVEMP.

Deterministic tracking Here the Optic Radiation (OR) was significantly (p-value < 0.05) more lateralized to the right in healthy patients than in patients with (compensated-)vertigo

Machine learning When using models to predict the results of vestibular tests based on the structural lateralization, the performances remains quite modest, as pictured on figure **??**. Both accuracy and balanced accuracy stick around 45%, while top 2 accuracy reach 95%. This is revealing of the fact the models are often confronted to a subject for which it is not clear to which class it belongs, and the model makes the wrong prediction even if the right one was its second choice. The models are then quite sure of the class the patients do not belong to, even if it does not have good performances for prediction of the right class itself.

If we look at the features that are selected by the models, the most often selected features were posterior and sensorimotor thalamic projections with probabilistic tracking (ATR and thalamo-broca, thalamo-sm, thalamo-superior/inferior parietal). For deterministic tracking, the optic radiations SLFII and SLFIII were the most significant.

4.5 Results summary and interpretation

Elikopy's analysis As Compensated patients are on average older than Healthy ones and FA is known to decrease with age in most brain areas [76], it is possible that the age (and not the vertigo) is responsible for the decrease in FA between Compensated and Healthy patients in some of the regions highlighted in the comparison Healthy/Compensated. The same would apply for the Healthy/Vertigo comparison.

However, Vertigo and Compensated patients have a slighter difference of age, and if there were an impact of this difference, it would be an increase of FA with compensation. But their comparison showed the opposite[5]: a suspicious decrease of FA in Compensated patients in the right medial prefrontal cortex and left insula.

[5]these results can then be noised by the age difference but not due to it

This would also explain why there were not many significant regions in the Healthy/Vertigo comparison: if Vertigo patients have a higher FA than Healthy ones, which is opposite to the natural decrease with age, the "old-vertigo" patients would have similar or lower FA than "young-healthy" ones, while "young-Vertigo" patients have a higher FA. The resulting effect being that those regions shows no significant differences and only differences that would be found are those positively (or not at all) correlated with the age.

It would then be legitimate to wonder if (at least) some Vertigo patients would have an increase in FA due to vertigo. Indeed, another explanation for the few regions in the Healthy/Vertigo comparison would be that all Vertigo patients does not have the same biomarkers and are too different from each other than to be considered as one group from a "brain microstructure" point of view.

Some selected regions (brainstem, left Parietal-Insular Vestibular Cortex (PIVC), broca's area and medial-frontal) similarly showed a lower FA in Vertigo and Compensated, so that the differences measured in these regions could be linked to Vertigo.

A decrease in FA with compensation could be explained by an increased microstructural complexity connections with more scattered regions because of an increased information integration for instance, similarly to the structural reorganization that occurs with age [78], but less in the insula where the correlation between FA and complexity is lower [76]. There could also be a decrease in FA due to a lower white-matter integrity (reduced number of connections or demyelination for example), which would increase the efficiency of short connections, and may decrease it for long-range connections [77].

Structural lateralization An interesting thing to note is that none of the variables passing the corrected statistical tests lead to more significant p-values when using the absolute lateralization rather than the lateralization itself. This could indicate that, as can be hypothesized from figure 4.13, the difference of mean lateralization between classes is not due to a systematic and similar lateralization difference of all Vertigo patients from Healthy ones, but to the fact there is an important difference between some of the Vertigo patients compared to the mean of Healthy patients.

This observation first ensure there were no significant impact of the handedness of patients (as few were left-handed or ambidextrous). This is also another evidence

that the differences between Vertigo and Healthy patients is complex and varies from one patient to another, i.e. there could be several subgroups of Vertigo patients with similar microstructural differences with the Healthy and Compensated patients.

Machine Learning The regions selected were coherent with the ones identified by TBSS, instead cerebellum regions were only selected by unreliable models, which may be due to the noise brought by the age differences. Similarly, Freesurfer parcellation often highlighted cerebellum and subcortical grey matter regions (hippocampus, amygdala, nucleus accumbens, pallidum), but only when models were identified as unreliable by the permutation-test procedure.

Still, machine learning analysis selected regions near parietal and insular cortex (mostly in the right hemisphere) as well as frontal regions (mostly in the left hemisphere). As there is a natural left lateralization of this Broca's because of language lateralization, there is a risk the results are due to this lateralization, and that in reality the differences are present bilaterally. Indeed, this natural structural lateralization could lead to greater the differences in the left hemisphere compared to the right.

However, the structural lateralization of these frontal regions showed an increased right-lateralization (or lower left-lateralization) in Vertigo patients, which could explain the selection of these regions: a greater white matter complexity (for example) in the left frontal lobe than in the right could lead to such a lateralization. This could translate in higher diffusion metrics and lower FA around Broca's area, which was observed in the TBSS's Healthy/Vertigo comparison.

Finally, the metrics measuring hesitation of the models showed a greater hesitation between Healthy and Compensated patients (up to 45%), followed by the hesitation between Compensated and Vertigo patients (up to 35%), the last being hesitation between Healthy and Vertigo (up to 20%). This further shows the models have

Chapter 5

Discussion and future perspectives

With this dataset, it was shown that no results were significant when a correction was applied on the p-values produced by the statistical test. This depicts the fact the differences between patients are small when studying one feature at a time, but also implies a high rate of false positives. Especially when considering the difference of age between the classes, it is then probable that some of the present findings are not biomarkers of Vertigo in general, but only in this particular dataset. However, it still allows pointing out regions of interest and hypothesis for future investigations that could verify the validity of those results.

Another thing that would impair the ability to generalize the results is the fact all patients are deaf. It would indeed be better to be able to compare with Healthy patients without hearing impairment, and (compensated-)Vertigo -again without hearing impairment - to be able to disentangle the impact of deaf on microstructure evolution with age and vertigo.

As the regions selected using machine-learning algorithm and TBSS statistical analysis were consistent with each other, but the differences between classes remains small, it could be interesting to compare this results with analysis of the functional MRI data, including connectomes-based approches.

5.1 Tractography

FOD The random threshold used to prune the before tracking was chosen to reduce the possibility to favour one bundle to the others, especially as it was not possible to include these parameters in the cross-validation process. But other ways of choosing this parameter could lead to better results, and the impact of this choice on the models' performances and predictions could be studied.

White matter response The functions of Dipy's library were used to generate the from the preprocessed dMRI. However, it is suggested in Mrtrix's documentation to upsample dMRI images to improve the white-matter response estimation and ODF computation [79].

5.2 Tract segmentation and Tractometry

AFQ was designed to segment tracts obtained with deterministic tracking, and the probability maps used are defined by such tracts, so that it's not as appropriate for probabilistic tracts than it is for deterministic ones. It was indeed more difficult to obtain more than a few tracts for most of the bundles with probabilistic tracking.

Segments Another point that could be discussed is the segments production procedure: with predefined number of segments to tend towards the wanted length. First, this length (4mm) could be discussed, as it was chosen to have segments thicker than a voxel dimension and with enough points in each streamline segments (1 point each 0.5mm), but thinner segments could be used as in the base methodology of AFQ and BUAN.

Moreover, once used on all patients, the imposed number of segments led to segments with thickness between 3.5 and 7 mm, indicating the anatomy of patients could strongly variate as the shape of the bundles. But this could bring noise in the measurements, by having segments that are not placed around the same cortical areas from one patient to the other (because one patient has a large frontal lobe compared to another, for example).

It could then be interesting to compare the results with an alternative method, where the centroids of the segments would be defined in a standard space (like an average brain) using streamline-based linear registration, so that segments would have comparable anatomical placement in all patients by adapting the length of the segments to the patients' anatomy.

Cleaning In the process of bundle segmentation, it was chosen to add another round of cleaning for tracts obtained in probabilistic tracking to reduce the impact of spurious fibres on further analysis. Indeed, it was expected such fibres that are misaligned with the mean tract could noise microstructure sampling and subdivision

of the bundle into segments. The second round of cleaning than had the purpose of limiting such phenomenon, but it should be needed to verify if there really is an impact of spurious fibres on the results and if improvements should be done for this step.

Bundles Here, some main white matter bundles were used, as the one segmented by default in AFQ, and some thalamic projections were added. However, it would be needed to take other bundles into account, that could have a link with the vestibular system. An exhaustive analysis would then include the cingulum, connections between the insula or limbic cortex and other regions, connections of the striatum with cortical areas and thalamus, or spinal tract connecting with the insula for instance.

Diffusion metrics sampling Diffusion metrics were sampled directly on the DTI images of the patients using trilinear interpolation. The images then had the same resolution as the dMRI. It could then be better to upsample the metrics images so that they have a higher resolution. The trilinear interpolation would be more accurate and less susceptible to noisy voxels as the upsampling would be done with an interpolation method that will smooth the data.

Microstructure metrics DTI model was used here, but it could be misleading when interpreting the results as this model is sensitive to several noise sources (myelinization or fibre orientation, for instance, could both influence the diffusion and FA metrics). It could also be interesting to compare the results with the ones obtained by adding NODDI measures to the diffusion metrics.

5.3 Machine Learning

Missing values When dealing with tractography data, there were a lot of features where some values were missing. In most cases, this was because a whole bundle was not reconstructed during tractography for one or a few patients and was then missing for those patients.

As explained in the *Methods*, these cases were dealt with by removing from the dataset all features that were missing in at least half of the patients of one class , and replacing other missing data by 0.

But it may be better to replace missing values by the mean value of all patients, or of the patients belonging to the same class. Though this could bias the model

towards good predictions and overestimate models performances, this may lead to better results than the used method as microstructure data are not categorical.

Age Here, only the impact of age on the predictions was studied as a third party-variable because it was the only one with a significative difference of values between the classes and was expected to either (or both): impair the results by overwhelming the features' variation with vertigo because of the evolution with age, or impair the interpretation by overestimating the models performances because of the increased difference between patients (in this case we would have a classification "young-healthy" / "old-compensated-vertigo" / "old-uncompensated-vertigo" rather than "healthy" / "compensated-vertigo" / "uncompensated-vertigo").

But a more extensive study of the impact of other third party variables - such as sex, handedness, white-matter volume or age of onset of the symptoms - could be interesting. Indeed, there are other factors that can affect brain microstructure as well as its evolution with age, as shown on figure 5.1 [80].

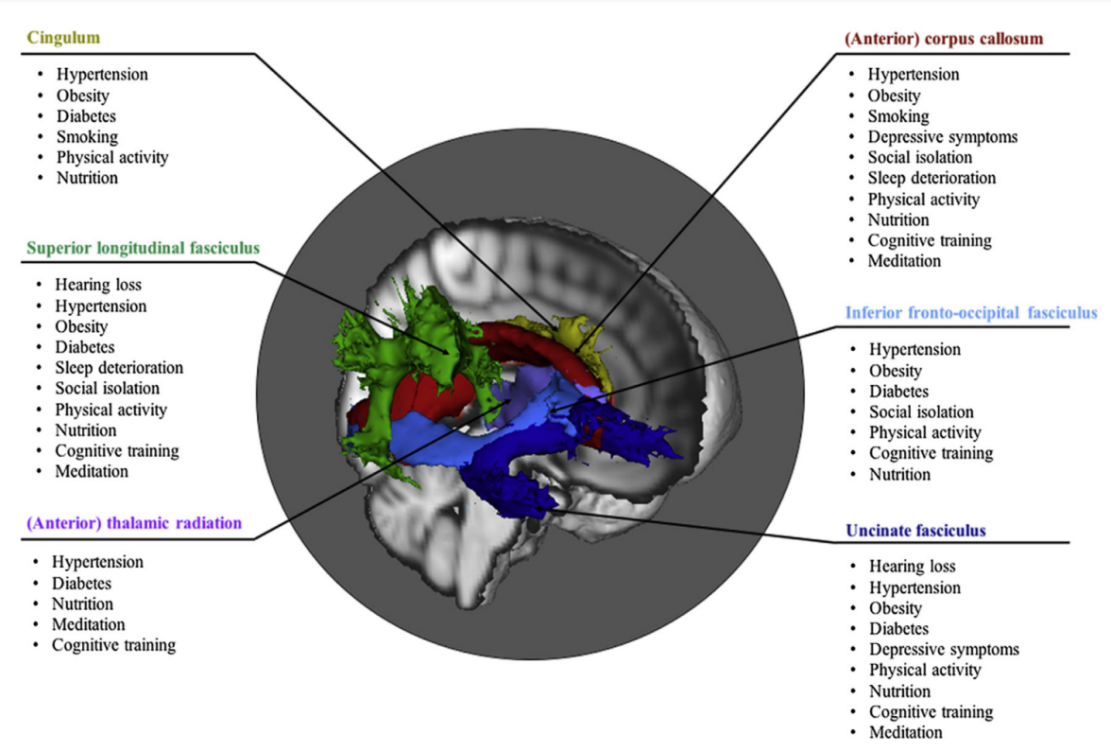

Figure 5.1: Impact of diverse risk factors on the microstructure of important bundles, and on its evolution with age [80].

Another thought about the method used here to correct for the age is that some biomarkers for vertigo could be correlated with age in this dataset, so that correcting for the age would also correct the main linear differences related to vertigo because of the age differences between Healthy and Vertigo patients. For example, if the FA is higher in Vertigo patients than in Healthy ones, the correction would decrease the FA of Vertigo patients instead of increasing it to correct for the natural decrease with age. It results that parameters correlated with the age in this dataset would be less often selected and have lower impact on models predictions (as differences between the classes is smaller) than biomarkers that are not strongly correlated with the age.

PCA In this work, PCA was applied on all DTI microstructural metrics to benefit from its capacity of dimensionality reduction of data with multi-colinear features. However, it is known PCA can be biased if many features are strongly correlated together, as AD, MD and RD in this case. This is why it could be interesting to compare the results with Sparse PCA[1] and with the method described in [71] (removing some of the variables that are strongly correlated).

It could also increase models performances to apply PCA on each segment separately, instead of applying it simultaneously on all segments from the same bundle. But this would also decrease interpretability of the results, as the components would vary from one segment to the other.

Results interpretation The used methods here only give the possibility to identify features that are biomarkers for vertigo on their own, or features that are often selected together (with a lower individual importance) because of the interplay between them[2]. But it does not provide the real piece of information the model found in each feature (does it discriminate all three classes, one class from the two others, or does it allow a better classification for only a subset of patients?). Moreover, the estimation of feature importance used will necessarily underestimate a feature importance when this feature was selected mainly for its interplay with other features.

It would then be interesting to study more deeply the feature importance, and take features interplay into account, as well as studying the impact of each feature on the classification of each patient. This could be done using Shapley explanation

[1] A variant of PCA with an added sparsity constraint that finds better components when data are organized in subgroups **source**

[2] these features bring significantly more information together than the sum of the information they each bring on their own

[81], amongst other methods, though it was not used here because it needs the construction of a matrix $N_p N_f$x $N_p N_f$, with N_p the number of patients and N_f the number of features, which implies a too great computational cost if used on all models with all features. However, it could be used to study more deeply the impact of each selected features on the classification for a chosen model, for instance.

Feature selection Here, machine learning analysis was preceded with a first feature selection was done to remove the less relevant features (based on an ANOVA test). This allowed to increase the performances of the models, but also increased the metrics measuring hesitations between two classes. Moreover, in many regions the features lead to a decrease of models' performances when the feature importance is miscomputed (like the corresponding coefficient in a linear regression analysis that would be taken as negative instead of positive, for example).

Those two observation implies there may be a possibility to find a better feature selection procedure, preceding the machine learning analysis. This could be done using embed feature selections methods, using information about the models' selected variables when discarding features of low importance.

Chapter 6

Conclusion

The goal of this thesis was to identify biomarkers for vertigo using machine learning and tractography analysis brain microstructure, based on dMRI data.

Deterministic tracking algorithm are most often used for tractometry analysis, and was not able to highlight differences between Healthy and Vertigo patients (with or without compensation) using machine learning analysis.

Here, we showed probabilistic tracking allowed to find such differences, that were comparable with other analysis methods such as TBSS, machine-learning analysis of white matter parcellation, or study of the structural lateralization. However, tract-based analysis based on probabilistic tracking was more performant than parcellation-based analysis.

All methods identified regions consistent with previous findings related to the vestibular system in the literature. Though, the low efficiency and/or reliability of most of these results suggest a need to develop further into this subject to understand why those regions are selected and to which vertigo patterns it is related.

The main limitations that were encountered were linked to the patients' characteristics. First, there were very few Healthy and Compensated-vertigo patients and many Vertigo ones. Then, patients had really different ages, which is a difference that could not be successfully overcome.

Bibliography

[1]

[2] Principe et technique de l' irm.

[3] Yi Cao, Lijun Xu, Ye Kuang, Dangsheng Xiong, and Renjun Pei. Gadolinium-based nanoscale MRI contrast agents for tumor imaging. *Journal of Materials Chemistry B*, 5(19):3431–3461, 2017.

[4] Bruno Kastler, Daniel Vetter, Patay Zolta´´n, Philippe Germain, and Auguste Wackenheim. *Comprendre l'IRM: Manuel d'auto-apprentissage*. Elsevier Masson, 2018.

[5] Neda Gholizadeh, Peter B. Greer, John Simpson, Jim Denham, Peter Lau, Jason Dowling, Hubert Hondermarck, and Saadallah Ramadan. Characterization of prostate cancer using diffusion tensor imaging: A new perspective. *European Journal of Radiology*, 110:112–120, January 2019.

[6] Ben Jeurissen, Maxime Descoteaux, Susumu Mori, and Alexander Leemans. Diffusion MRI fiber tractography of the brain. *NMR in Biomedicine*, 32(4), September 2017.

[7] Thijs Dhollander, David Raffelt, and Alan Connelly. Unsupervised 3-tissue response function estimation from single-shell or multi-shell diffusion mr data without a co-registered t1 image. 09 2016.

[8] Chantal M.W. Tax, Ben Jeurissen, Sjoerd B. Vos, Max A. Viergever, and Alexander Leemans. Recursive calibration of the fiber response function for spherical deconvolution of diffusion MRI data. *NeuroImage*, 86:67–80, February 2014.

[9] Timo Roine, Ben Jeurissen, Daniele Perrone, Jan Aelterman, Wilfried Philips, Alexander Leemans, and Jan Sijbers. Informed constrained spherical deconvolution (iCSD). *Medical Image Analysis*, 24(1):269–281, August 2015.

[10] Robert E. Smith, Jacques-Donald Tournier, Fernando Calamante, and Alan Connelly. SIFT: Spherical-deconvolution informed filtering of tractograms. *NeuroImage*, 67:298–312, February 2013.

[11] Jason D. Yeatman, Robert F. Dougherty, Nathaniel J. Myall, Brian A. Wandell, and Heidi M. Feldman. Tract profiles of white matter properties: Automating fiber-tract quantification. *PLoS ONE*, 7(11):e49790, November 2012.

[12] Eleftherios Garyfallidis, Marc-Alexandre Côté, Francois Rheault, Jasmeen Sidhu, Janice Hau, Laurent Petit, David Fortin, Stephen Cunanne, and Maxime Descoteaux. Recognition of white matter bundles using local and global streamline-based registration and clustering. *NeuroImage*, 170:283–295, April 2018.

[13] Bramsh Qamar Chandio, Shannon Leigh Risacher, Franco Pestilli, Daniel Bullock, Fang-Cheng Yeh, Serge Koudoro, Ariel Rokem, Jaroslaw Harezlak, and Eleftherios Garyfallidis. Bundle analytics, a computational framework for investigating the shapes and profiles of brain pathways across populations. *Scientific Reports*, 10(1), October 2020.

[14] Wikipedia contributors. Lobes of the brain, Apr 2022.

[15] Wikipedia contributors. Limbic lobe, Feb 2022.

[16] Wikipedia contributors. Cerebral cortex, Mar 2022.

[17] Guillaume Herbet, Ilyess Zemmoura, and Hugues Duffau. Functional anatomy of the inferior longitudinal fasciculus: From historical reports to current hypotheses. *Frontiers in Neuroanatomy*, 12, September 2018.

[18] Nikos Makris, David N. Kennedy, Sean McInerney, A. Gregory Sorensen, Ruopeng Wang, Verne S. Caviness, and Deepak N. Pandya. Segmentation of subcomponents within the superior longitudinal fascicle in humans: A quantitative, in vivo, DT-MRI study. *Cerebral Cortex*, 15(6):854–869, December 2004.

[19] C. Maffei, C. Lee, M. Planich, M. Ramprasad, N. Ravi, D. Trainor, Z. Urban, M. Kim, R.J. Jones, A. Henin, S.G. Hofmann, D.A. Pizzagalli, R.P. Auerbach, J.D.E. Gabrieli, S. Whitfield-Gabrieli, D.N. Greve, S.N. Haber, and A. Yendiki. Using diffusion MRI data acquired with ultra-high gradient strength to improve tractography in routine-quality data. *NeuroImage*, 245:118706, December 2021.

[20] Wikipedia contributors. Corpus callosum, Mar 2022.

[21] Thijs Dhollander. Tractography. 06 2018.

[22] Wikipedia contributors. Occipitofrontal fasciculus, Mar 2022.

[23] Claire Rollans. Evidence for the functional and structural differentiation of the inferior fronto-occipital fasciculus using dti tractography. 2016.

[24] Kylie H. Alm, Tyler Rolheiser, and Ingrid R. Olson. Inter-individual variation in fronto-temporal connectivity predicts the ability to learn different types of associations. *NeuroImage*, 132:213–224, May 2016.

[25] Steven J. Granger, Stephanie L. Leal, Myra Saraí Larson, John T. Janecek, Liv McMillan, Hal Stern, and Michael A. Yassa. Integrity of the uncinate fasciculus is associated with emotional pattern separation-related fMRI signals in the hippocampal dentate and CA3. *Neurobiology of Learning and Memory*, 177:107359, January 2021.

[26] Thomas Brandt and Marianne Dieterich. Thalamocortical network: a core structure for integrative multimodal vestibular functions. *Current Opinion in Neurology*, 32(1):154–164, February 2019.

[27] V. Kirsch, R. Boegle, D. Keeser, E. Kierig, B. Ertl-Wagner, T. Brandt, and M. Dieterich. Handedness-dependent functional organizational patterns within the bilateral vestibular cortical network revealed by fMRI connectivity based parcellation. *NeuroImage*, 178:224–237, September 2018.

[28] Jay M. Goldberg. Vestibular inputs: The vestibular system. In *Neuroscience in the 21st Century*, pages 1007–1054. Springer New York, 2016.

[29] Inner ear and balance.

[30] Alexandre Bisdorff, Michael Von Brevern, Thomas Lempert, and David E. Newman-Toker. Classification of vestibular symptoms: Towards an international classification of vestibular disorders. *Journal of Vestibular Research*, 19(1-2):1–13, October 2009.

[31] Sertac Yetiser and Dilay Ince. Caloric analysis of patients with benign paroxysmal positional vertigo. *The Journal of International Advanced Otology*, 13(3):390–393, December 2017.

[32] James A. Nelson and Erik Viirre. The clinical differentiation of cerebellar infarction from common vertigo syndromes. *The western journal of emergency medicine*, 10(4):273–277, Nov 2009. 20046249[pmid].

[33] Enass S. Mohamed. Predictors of central vestibular disorders from videonystagmography tests. *The Egyptian Journal of Otolaryngology*, 32(3):202–209, July 2016.

[34] Terry D. Fife, Saty Satya-Murti, Robert F. Burkard, and John P. Carey. Vestibular evoked myogenic potential testing. *Neurology: Clinical Practice*, 8(2):129–134, March 2018.

[35] Sun-Young Oh, Hyo-Jeong Kim, and Ji-Soo Kim. Vestibular-evoked myogenic potentials in central vestibular disorders. *Journal of Neurology*, 263(2):210–220, August 2015.

[36] Marianne Dieterich and Thomas Brandt. Global orientation in space and the lateralization of brain functions. *Current Opinion in Neurology*, 31(1):96–104, February 2018.

[37] Iole Indovina, Gianfranco Bosco, Roberta Riccelli, Vincenzo Maffei, Francesco Lacquaniti, Luca Passamonti, and Nicola Toschi. Structural connectome and connectivity lateralization of the multimodal vestibular cortical network. *NeuroImage*, 222:117247, November 2020.

[38] Louise Devantier, Allan K. Hansen, Jens-Jacob Mølby-Henriksen, Christian B. Christensen, Michael Pedersen, Kim V. Hansen, Måns Magnusson, Therese Ovesen, and Per Borghammer. Positron emission tomography visualized stimulation of the vestibular organ is localized in heschl's gyrus. *Human Brain Mapping*, 41(1):185–193, September 2019.

[39] I Indovina, R Riccelli, L Passamonti, V Maffei, G Bosco, F Lacquaniti, and N Toschi. Structural connectome of the human vestibular, pre-motor, and navigation network. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, July 2018.

[40] Xia Zhe, Jie Gao, Li Chen, Dongsheng Zhang, Min Tang, Xuejiao Yan, Fuxia Bai, Xin Zhang, Ze Zou, Weibo Chen, Xiaoyan Lei, and Xiaoling Zhang. Altered structure of the vestibular cortex in patients with vestibular migraine. *Brain and Behavior*, 10(4), March 2020.

[41] Jason Brownlee. *Master machine learning algorithms: Discover how they work and implement them from scratch*. Jason Brownlee, 2016.

[42] Rich Larrabee Software Architect / Data Scientist, Frederic Marechal Data Scientist, Alex Chicote Data Scientist, Ilias Zavitsanos Software Engineer / Researcher, and Diego Ferreño Researcher. Imbalanced classification with python, Oct 2021.

[43] Christopher M. Bishop. *Pattern recognition and machine learning*. Springer Verlag, 2009.

[44] Leonardo Gutiérrez-Gómez, Jakub Vohryzek, Benjamin Chiêm, Philipp S. Baumann, Philippe Conus, Kim Do Cuenod, Patric Hagmann, and Jean-Charles Delvenne. Stable biomarker identification for predicting schizophrenia in the human connectome. *NeuroImage: Clinical*, 27:102316, 2020.

[45] C. Fernandez-Lozano, C. Canto, M. Gestal, J. M. Andrade-Garda, J. R. Rabuñal, J. Dorado, and A. Pazos. Hybrid model based on genetic algorithms and SVM applied to variable selection within fruit juice classification. *The Scientific World Journal*, 2013:1–13, 2013.

[46] Leo Breiman. *Machine Learning*, 45(1):5–32, 2001.

[47] Adam Richie-Halford, Jason D. Yeatman, Noah Simon, and Ariel Rokem. Multidimensional analysis and detection of informative features in human brain white matter. *PLOS Computational Biology*, 17(6):e1009136, June 2021.

[48] Jitendra K. Tugnait. Sparse-group lasso for graph learning from multi-attribute data. *IEEE Transactions on Signal Processing*, 69:1771–1786, 2021.

[49] Jan Hendrik Metzen. Comparison of calibration of classifiers.

[50] Alexandre Gramfort and Lucy Liu. Test with permutations the significance of a classification score.

[51] Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8(1), January 2007.

[52] S. Yousefi, N. Kehtarnavaz, and A. Gholipour. Improved labeling of subcortical brain structures in atlas-based segmentation of magnetic resonance images. *IEEE Transactions on Biomedical Engineering*, 59(7):1808–1817, July 2012.

[53] Joël Schaerer. A generic interpolator for multi-label images. *Insight Journal*, 11 2014.

[54]

[55] Paul McCarthy. Fsleyes, 2022.

[56]

[57] Quentin Dessain and Mathieu Simon. Elikopy documentation.

[58] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, January 1979.

[59] Jelle Veraart, Dmitry S. Novikov, Daan Christiaens, Benjamin Ades-aron, Jan Sijbers, and Els Fieremans. Denoising of diffusion MRI using random matrix theory. *NeuroImage*, 142:394–406, November 2016.

[60] Jesper L.R. Andersson and Stamatios N. Sotiropoulos. An integrated approach to correction for off-resonance effects and subject movement in diffusion MR imaging. *NeuroImage*, 125:1063–1078, January 2016.

[61] Jesper L.R. Andersson, Stefan Skare, and John Ashburner. How to correct susceptibility distortions in spin-echo echo-planar images: application to diffusion tensor imaging. *NeuroImage*, 20(2):870–888, October 2003.

[62] Kurt G. Schilling, Justin Blaber, Yuankai Huo, Allen Newton, Colin Hansen, Vishwesh Nath, Andrea T. Shafer, Owen Williams, Susan M. Resnick, Baxter Rogers, Adam W. Anderson, and Bennett A. Landman. Synthesized b0 for diffusion distortion correction (synb0-DisCo). *Magnetic Resonance Imaging*, 64:62–70, December 2019.

[63] Nicholas J Tustison, Brian B Avants, Philip A Cook, Yuanjie Zheng, Alexander Egan, Paul A Yushkevich, and James C Gee. N4itk: Improved n3 bias correction. *IEEE Transactions on Medical Imaging*, 29(6):1310–1320, June 2010.

[64] Robert E. Smith, Jacques-Donald Tournier, Fernando Calamante, and Alan Connelly. Anatomically-constrained tractography: Improved diffusion MRI streamlines tractography through effective use of anatomical information. *NeuroImage*, 62(3):1924–1938, September 2012.

[65] José V. Manjón and Pierrick Coupé. volBrain: An online MRI brain volumetry system. *Frontiers in Neuroinformatics*, 10, July 2016.

[66] Sebastien Tourbier, Joan Rue Queralt, Katharina Glomb, Yasser Aleman-Gomez, Emeline Mullier, Alessandra Griffa, Mikkel Schöttner, Jonathan Wirsich, Anil Tuncel, Jakub Jancovic, Meritxell Bach Cuadra, and Patric Hagmann. connectomicslab/connectomemapper3: Connectome mapper v3.0.3, 2022.

[67] Atlas maps.

[68] M CATANI and M THIEBAUTDESCHOTTEN. A diffusion tensor imaging tractography atlas for virtual in vivo dissections. *Cortex*, 44(8):1105–1132, September 2008.

[69] Mark Jenkinson. Contents.

[70] Anderson M. Winkler, Gerard R. Ridgway, Matthew A. Webster, Stephen M. Smith, and Thomas E. Nichols. Permutation inference for the general linear model. *NeuroImage*, 92:381–397, May 2014.

[71] Maxime Chamberland, Erika P. Raven, Sila Genc, Kate Duffy, Maxime Descoteaux, Greg D. Parker, Chantal M.W. Tax, and Derek K. Jones. Dimensionality reduction of diffusion MRI measures for improved tractometry of the human brain. *NeuroImage*, 200:89–100, October 2019.

[72] Danka Jandric, Geoff J.M. Parker, Hamied Haroon, Valentina Tomassini, Nils Muhlert, and Ilona Lipp. A tractometry principal component analysis of white matter tract network structure and relationships with cognitive function in relapsing-remitting multiple sclerosis. *NeuroImage: Clinical*, 34:102995, 2022.

[73] Rishu Rathee, V.P. Subramanyam Rallabandi, and Prasun K. Roy. Age-related differences in white matter integrity in healthy human brain: Evidence from structural mri and diffusion tensor imaging. *Magnetic Resonance Insights*, 9:MRI.S39666, January 2016.

[74] P. Kochunov, D.C. Glahn, J. Lancaster, P.M. Thompson, V. Kochunov, B. Rogers, P. Fox, J. Blangero, and D.E. Williamson. Fractional anisotropy of cerebral white matter and thickness of cortical gray matter across the lifespan. *NeuroImage*, 58(1):41–49, September 2011.

[75] Kaoru Amemiya, Eiichi Naito, and Hiromasa Takemura. Age dependency and lateralization in the three branches of the human superior longitudinal fasciculus. *Cortex*, 139:116–133, June 2021.

[76] Shannon Kelley, John Plass, Andrew R Bender, and Thad A Polk. Age-related differences in white matter: Understanding tensor-based results using fixel-based analysis. *Cerebral Cortex*, 31(8):3881–3898, April 2021.

[77] Ian M. McDonough and Jonathan T. Siegel. The relation between white matter microstructure and network complexity: Implications for processing efficiency. *Frontiers in Integrative Neuroscience*, 12, September 2018.

[78] Epifanio Bagarinao, Hirohisa Watanabe, Satoshi Maesawa, Daisuke Mori, Kazuhiro Hara, Kazuya Kawabata, Noritaka Yoneyama, Reiko Ohdake,

Kazunori Imai, Michihito Masuda, Takamasa Yokoi, Aya Ogura, Toshiaki Taoka, Shuji Koyama, Hiroki C. Tanabe, Masahisa Katsuno, Toshihiko Wakabayashi, Masafumi Kuzuya, Norio Ozaki, Minoru Hoshiyama, Haruo Isoda, Shinji Naganawa, and Gen Sobue. Reorganization of brain networks and its association with general cognitive performance over the adult lifespan. *Scientific Reports*, 9(1), August 2019.

[79] Fibre density and cross-section - single-tissue csd.

[80] Catherine Lebel, David Salat, and Jason Yeatman. *Tractography: Applications to Neurodevelopment, Aging, Plasticity*. 03 2022.

[81] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[82] Chantal M.W. Tax, Ben Jeurissen, Sjoerd B. Vos, Max A. Viergever, and Alexander Leemans. Recursive calibration of the fiber response function for spherical deconvolution of diffusion mri data. *NeuroImage*, 86:67–80, 2014.

[83] Zhengwei Chen, Lijie Xiao, Haiyan Liu, Qingxiu Zhang, Quan Wang, You Lv, Yujia Zhai, Jun Zhang, Shanshan Dong, Xiue Wei, and Liangqun Rong. Altered thalamo-cortical functional connectivity in patients with vestibular migraine: a resting-state fMRI study. *Neuroradiology*, 64(1):119–127, August 2021.

[84] Sebastian M. Frank, Maja Pawellek, Lisa Forster, Berthold Langguth, Martin Schecklmann, and Mark W. Greenlee. Attention networks in the parietooccipital cortex modulate activity of the human vestibular cortex during attentive visual processing. *The Journal of Neuroscience*, 40(5):1110–1119, December 2019.

[85] Tabinda Sarwar, Kotagiri Ramamohanarao, and Andrew Zalesky. Mapping connectomes with diffusion MRI: deterministic or probabilistic tractography? *Magnetic Resonance in Medicine*, 81(2):1368–1384, October 2018.

[86] Jason D. Yeatman, Adam Richie-Halford, Josh K. Smith, Anisha Keshavan, and Ariel Rokem. A browser-based tool for visualization and analysis of diffusion MRI data. *Nature Communications*, 9(1), March 2018.

[87] S.P.L. Meesters, G. Sanguinetti, E. Garyfallidis, J.M. Portegies, and R. Duits. Fast implementations of contextual pde's for hardi data processing in dipy.

May 2016. 24th ISMRM Annual Meeting and Exhibition, May 7-13, 2016, Singapore ; Conference date: 07-05-2016 Through 13-05-2016.

[88] Thijs Dhollander and Alan Connelly. A novel iterative approach to reap the benefits of multi-tissue csd from just single-shell (+b=0) diffusion mri data. 05 2016.

[89] Robert E. Smith, Jacques-Donald Tournier, Fernando Calamante, and Alan Connelly. Anatomically-constrained tractography: Improved diffusion MRI streamlines tractography through effective use of anatomical information. *NeuroImage*, 62(3):1924–1938, September 2012.

[90] Simona Schiavi, Mario Ocampo-Pineda, Muhamed Barakovic, Laurent Petit, Maxime Descoteaux, Jean-Philippe Thiran, and Alessandro Daducci. A new method for accurate in vivo mapping of human brain connections using microstructural and anatomical information. *Science Advances*, 6(31), July 2020.

[91] David Raffelt, J.-Donald Tournier, Stephen Rose, Gerard R. Ridgway, Robert Henderson, Stuart Crozier, Olivier Salvado, and Alan Connelly. Apparent fibre density: A novel measure for the analysis of diffusion-weighted magnetic resonance images. *NeuroImage*, 59(4):3976–3994, February 2012.

Annexes

6.1 White Matter Query Language

In summary, the following modifications were done on the base definitions of WMQL:

- the "thalamo-broca" connections were defined as the connections between the thalamus and the parstriangularis, parsoprecuparis, and the part of the precentral gyrus that is not higher than the parsopercularis
- the anterior thalamic projections were restrained to the connections between the thalamus and the fontal lobe for which the endpoints in the frontal lobe were not higher than the caudal-middle-frontal gyrus, nor posterior to the parstriangularis nor included in the "thalamo-broca" connections
- the "thalamo-sm" connections were defined as the connections between the thalamus sensori-motor areas (precentral, paracentral and postcentral gyrus) that were not included in the "thalamo-broca" connections
- the interhemispheric tracts passing through the anterior corpus callosum were subdivided in 3 parts: an orbital bundle connecting the broca's areas of the 2 hemispheres, an anterior callosum bundle connecting the frontal lobes (not higher than the caudal-middle frontal gyrus and not included in the orbital) and a superior anterior callosum bundle connecting the rest of the frontal lobes (the main part of the superiorfrontal gyrus)
- interhemispheric connections between the temporal lobes and insulas going through the corpus callosum were also defined
- definitions of the corpus callosum parts were refined by ensuring no streamline was part of two of them, which was necessary after splitting the anterior callosum bundles and after defining the temporal and insular callosum tracts
- the posterior callosum bundle was refined by removing streamlines with an endpoint that was not posterior to the ipsilateral thalamus, as some

streamlines selected with the original definition were half included in the tapetum

- the original definition for the uncinate fasciculus wrongly selected streamlines half part of the ILF/MdLF and of the IFOF, and streamlines with endpoints in the posterior superiorfrontal gyrus. Two modifications were then introduced: streamlines with endpoints in the superiorfrontal gyrus that were not anterior to the rostral-midlle frontal gyrus were discarded, streamlines with a part posterior to the ipsilateral thalamus were discarded

- the arcuate fasciculus was subdivided in its anterior, posterior and long segments. The anterior segment matched with the definition of the third segment of the SLF and they were merged together. The posterior segment matched with the definition of the MdLF, they were merged together too.

- as cerebellum was at most half in the image in all patients, and as there were no brainstem segmentation available, 2 cerebellum tracts were defined: one connecting the superior lobes to the brainstem, and one connecting cerebellum's white matter to the brainstem

6.2 Cleaning for probabilistic tractogram

As stated in the Methods section 3.3, a second round of cleaning was used for probabilistic tracking as cleaning using the Mahalanobis distance was judged insufficient for noisy probabilistic bundles. A noisy UF bundle was then produced using bad WMQL definitions on purpose, to test several methods using the Mahalanobis distance and/or the Fiber to Bundle Coherence (FBC) [87]. This noisy bundle and the result of the first cleaning round (using Mahalanobis distance only) are shown on figure 6.1. We can see the first round was able to remove the most incoherent fibres, but did keep other streamlines that should be removed.

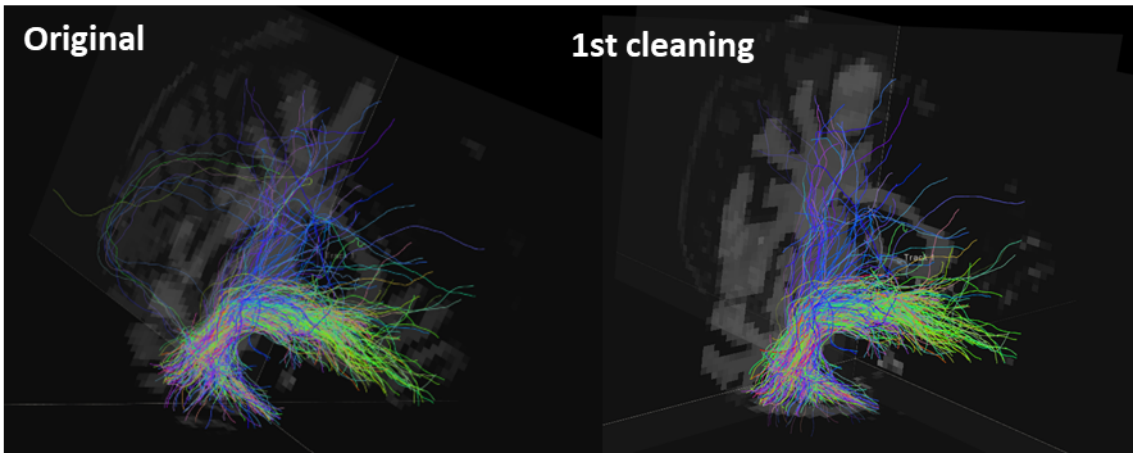

Figure 6.1: Noisy UF before and after the first cleaning.

Figure 6.2 depicts the results of the tested cleaning methods after the first round of cleaning. The first method (figure 6.2.A) used the FBC only to remove spurious fibres, using the default kernel proposed in the paper [87]. In this method, all streamlines belonging to the 5th quantile of RFBC were discarded (removing), as this metric is higher for aligned fibres and lower for misaligned ones. This was performed iteratively (5 times) as the results were comparable with using a higher RFBC threshold but discarding less streamlines in total.

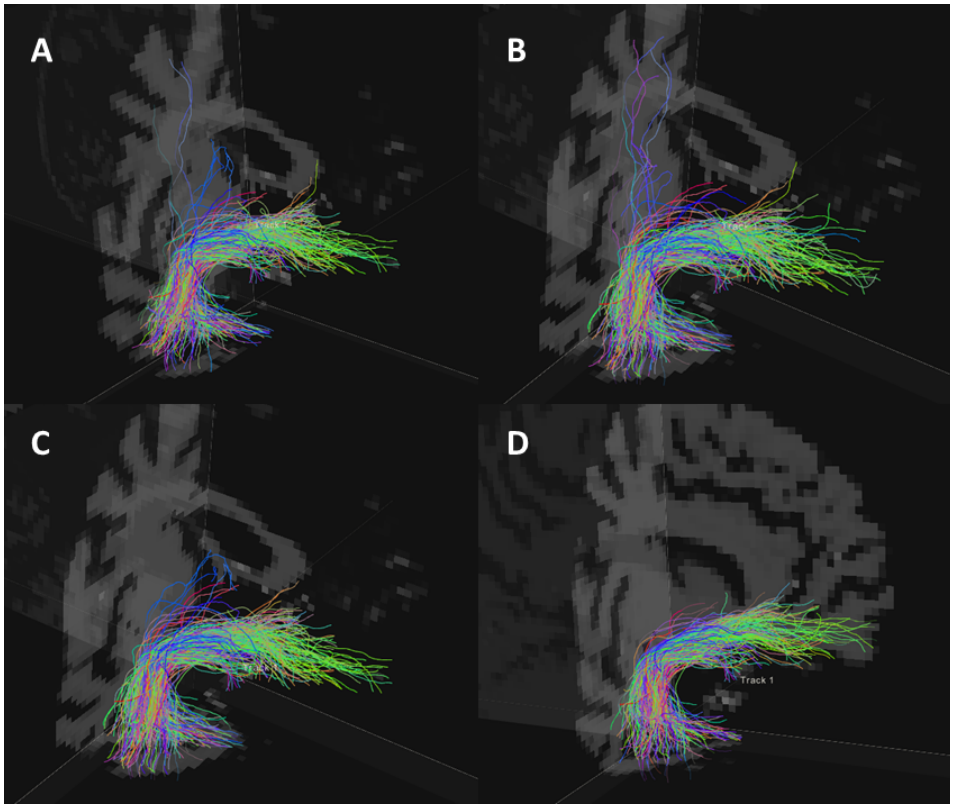

Figure 6.2: The four cleaning methods tested on the UF. A: FBC only. B: FBC and Mahalanobis distance used consecutively. C: FBC corrected by Mahalanobis distance. D: corrected FBC and Mahalanobis distance used consecutively.

The second method (figure 6.2.B) consisted in adding a Mahalanobis distance cleaning at each iteration of the RFBC cleaning, removing the 1% streamlines with the greatest mean Mahalanobis distance. We can see there are still some streamlines with incoherent endpoints in the posterior superior frontal gyrus.

The third method (figure 6.2.C) also used the Mahalanobis distance, but to weight the RFBC of each streamline rather than directly cleaning. This way, the RFBC was increased for streamlines near the mean bundle and streamlines with wrong endpoints were discarded first as their RFBC was decreased. Here the results are quite satisfying, with no more streamlines discarded than with the second method.

However, it was possible to further improve these results by combining both the second and third methods (figure 6.2.D): at each iteration, the bundle is cleaned using the Mahalanobis distance (discarding x% of the streamlines), then

the Mahalanobis distance is recomputed on kept streamlines and used to correct the RFBC (also computed on kept streamlines), and x% of the streamlines are discarded.

The "x" threshold was first chosen as 5 (removing $\simeq$5% at each step), but a higher proportion of streamlines were removed in bundles with much more fibres than here (because of the higher difference between the tails of the RFBC distribution). Moreover, when there were fewer streamlines in the bundle, the threshold had to be increased or no streamlines would be removed. It was then chosen to use an adaptive threshold, from 0.0158 (3000 fibres) to 0.05 (300 fibres), and 0.1225 (50 fibres).

To ensure the second round of cleaning did not remove too many streamlines, cleaning was aborted if more than 75% of the streamlines were discarded. Once the cleaning provided bundle such as on figure 6.2.D, it was possible to obtain cleaner results with a last round of RFBC cleaning.

However, each time we remove some spurious streamlines, valid streamlines are also discarded. It was also judged not necessary to have such a cleaned bundle, as streamlines on figure 6.2.C would be acceptable in our case, especially that those results were obtained on purposely noisy data and should be better with real streamlines. Consequently, only three cleaning iterations were used (instead of 5), and the thresholds were lowered by half (as there are 2 cleaning steps at each iteration).

6.3 Supplementary images

6.3.1 Theoretical background

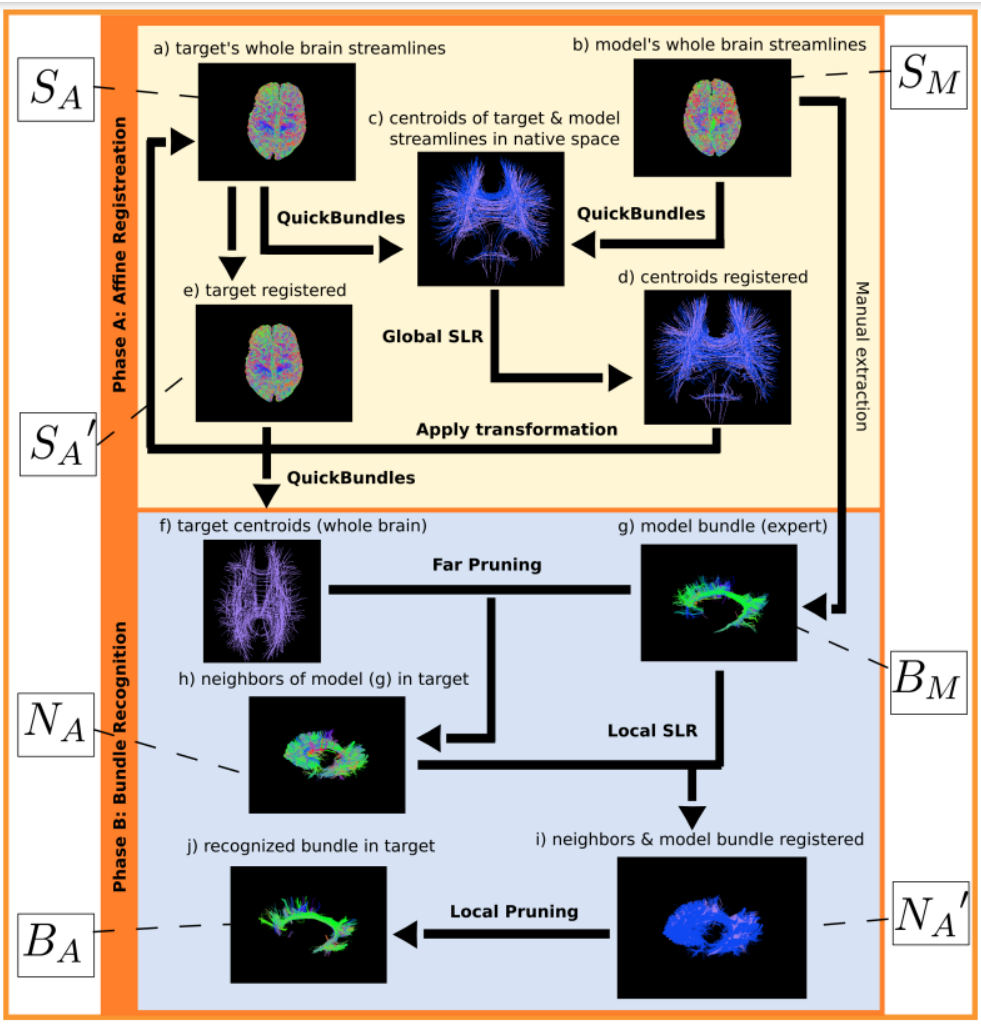

Figure 6.3: Flow chart diagram showing phases A and B of RecoBundles. In Phase A the streamlines of the two subjects are registered using streamline-based linear registration (SLR) and in Phase B a segmented bundle from one participant (model) is used as input to extract a homologous bundle in the other participant. Notice how in Phase B a local SLR is also executed. This SLR is responsible for matching the bundles.

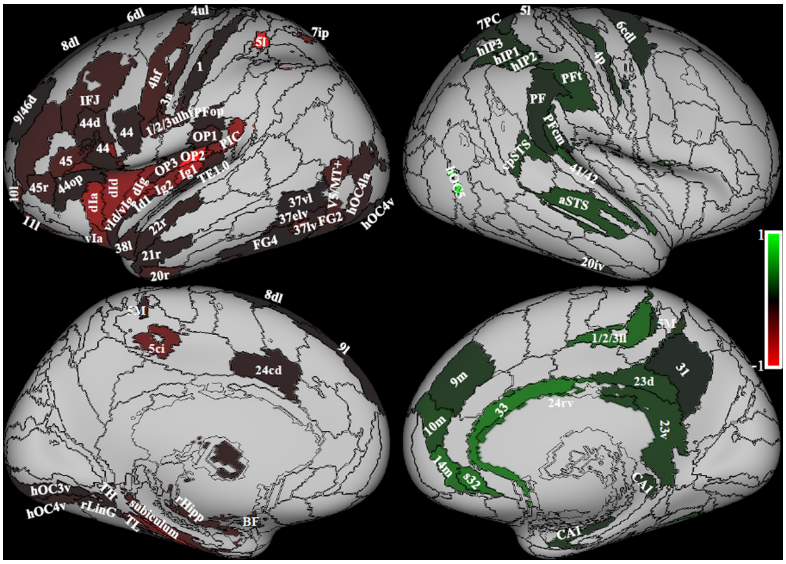

Figure 6.4: *Structural lateralization. On the left in red, areas that show a significant left laterality index; on the right in green areas that show a significant right laterality index; (Wilcoxon signed test, $p < 0.05$ Bonferroni corrected). Data from 794 right handed individuals. The color bars represent the range of laterality index. Regions from the composite atlas are overlapped onto the Conte69 inflated brain in workbench viewer (Glasser et al 2016).*[37]

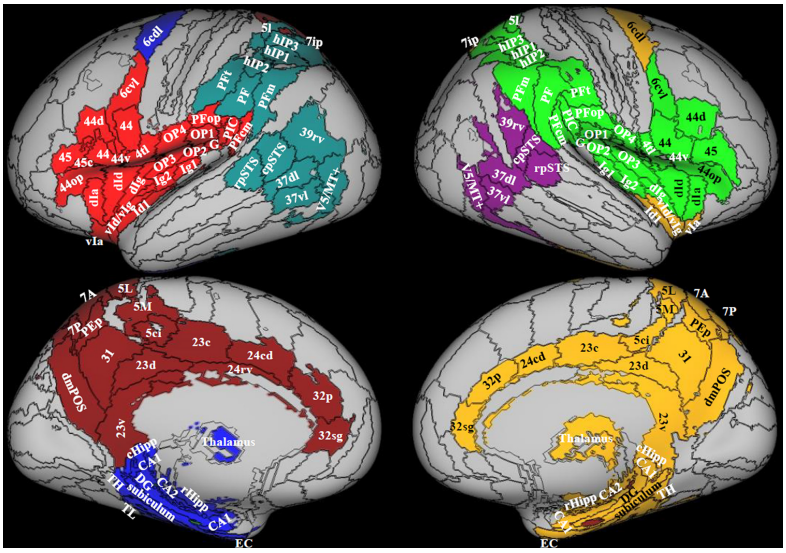

Figure 6.5: *Modularity. Areas in color have been selected as belonging to the multimodal vestibular cortex. Each color represents a module, that is a cluster of areas highly connected. Data from 794 right handed individuals. There are 4 modules on the left and 3 on the right, indicating better integration of the vestibular network on the right. Regions from the composite atlas are overlapped onto the Conte69 inflated brain in workbench viewer (Glasser et al 2016).*[37]

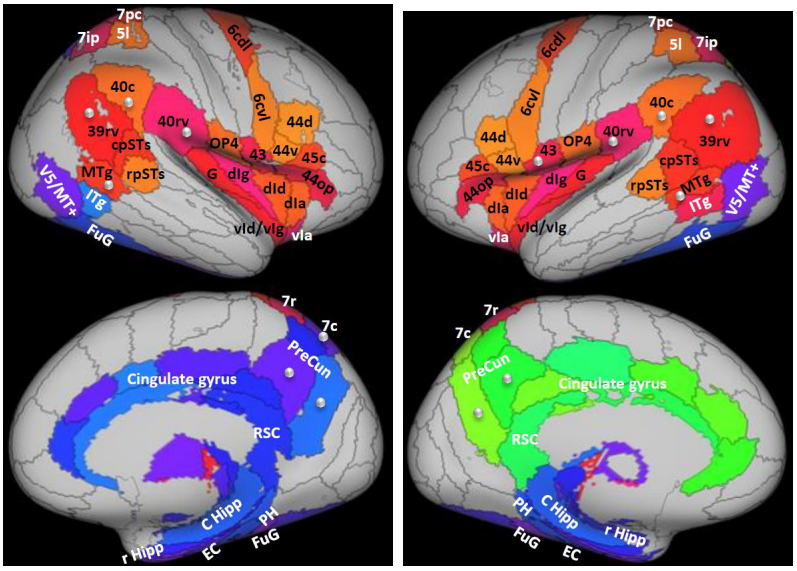

Figure 6.6: *Regions of the vestibular, pre-motor and navigation network are overlapped onto the Conte69 inflated brain in workbench viewer [21]. Different gradations of colors represent different modules. Red, vestibular and premotor modules; blue and green, navigation modules. Hubs are indicated by silver dots.*[39]

UNIVERSITÉ CATHOLIQUE DE LOUVAIN
École polytechnique de Louvain
Rue Archimède, 1 bte L6.11.01, 1348 Louvain-la-Neuve, Belgique | **www.uclouvain.be/epl**