Brain age estimation and FreeSurfer - FastSurfer comparison

Maddalena Cavallo

Pattern Recognition

1 Introduction

Human brain changes across the adult lifespan and undergoes complex aging effects that can change its structure and functioning. Normal brain development and healthy aging have been found to follow a specific pattern, consistent between different people, including a decrease of grey matter volume, an increase of white matter volume until the age of 20, when it reaches a plateau, and a complementary pattern in CSF (plateau and then increase) [1]. However, ageing does not affect all people uniformly and deviations from a typical brain ageing trajectory are sometimes observed. For example, from T1-weighted acquisitions it could happen that an individual shows an increased brain atrophy with respect to the majority of people of the same age.
These heterogeneous aging processes have stimulated the onset of many researches that aim to study these aging processes and understand the biological links between ageing and diseases. In particular, researches try to study if, among these heterogeneous aging processes, there is a specific pattern in brains of people with a diagnosed neurological disease such as Alzheimer's Disease (AD) or Parkinson's Disease (PD) ([2] [3]). Furthermore, recent evidences have shown that having an older-appearing brain relates to advanced physiological and cognitive ageing and an increased risk of mortality [4].
Thus the underlying idea that leads these studies is that the extent to which someone deviates from healthy brain ageing trajectories could potentially indicate underlying problems in outwardly healthy people. By measuring how far an individual is from the healthy brain ageing trajectory, researchers hope to be able to quantify advanced and decelerated brain ageing and use this to predict individuals' future trajectories and subsequent risk of age associated health deterioration. This has important consequences for future clinical applications as it will allow to make individualized predictions on patients, identifying people at greater risk of developing diseases and eventually of mortality during ageing. This could also have a great utility for testing potential treatments aimed to stop or at least slow down neurological disease progression.

Many different brain-based biological ageing biomarkers have been proposed, that may help to identify individuals with patterns that deviate from healthy brain ageing trajectories.
One of these is the Brain Predicted Age Difference (Brain PAD), a biological marker that could be extracted from structural neuroimaging data. This biomarker relies on the distinction between an individual chronological age and biological age, with the latter being the hypothetical age that could be defined by measuring some aspects of the organism's biology. These two ages can differ: individuals can appear to be younger or older than their chronological age. Deviations between predicted and chronological age are known to occur in several neurodegenerative diseases. Brain Predicted Age Difference (or simply Brain Age) is the metric introduced to take into account their differences.
Brain Age can be predicted in individuals from magnetic resonance imaging (MRI) data using Machine Learning approaches, that can make statistical analysis from MRI scans and model trajectories of healthy brain ageing. By learning the relationship between chronological age and patterns of data in a training dataset of healthy people, age predictions can be made in test datasets i.e. from people not included in the initial training. The design and development of fast and accurate age prediction models measuring brain age is crucial to provide systems that can be used in the clinical context.

Some of the most common Machine Learning methods are:

- Gaussian Process Regression, that combines different multivariate gaussian distributions in order to model non-linear relationships
- Support Vector Regression (SVR) a method that aim to determine a cost function with deviations

not larger than ϵ from each target point and each training point

- Relevance Vector Regression (RVR), a bayesian alternative to SVR
- Random Forest (RF), an accurate and robust regression method that uses ensemble learning
- Deep Neural Networks, designed to learn accurate representations of provided observations
- Ridge Regression, basically a least square method modified to be used in the case where independent variables suffer from multicollinearity
- Lasso Regression, a method similar to Ridge regression that is able to retain only important features and exclude those yielding a negligible contribution to the model (feature selection)

The general analytic pipeline for brain age prediction with these multivariate methods is the following:

a) Pre-processing: data transformation on training set images e.g. brain segmentation, normalization and registration to a common space

b) Training: Pre-processed neuroimaging data, labelled with chronological age, are given to one of the machine learning regression models mentioned above to predict age. Each model defines some features (i.e. variables that has some relevance) that are extracted from images and used as predictors (independent variables) in the regression with the chronological age as the outcome (dependent variable). At the end of this procedure the Machine Learning algorithm has defined a statistical model of healthy brain ageing.

c) Validation: usually a cross validation procedure to assess the accuracy of the model. Accuracy metrics as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and Pearson's correlation coefficient R are then computed to evaluate the performance of the age prediction model on the training dataset

d) Testing: Once the brain age prediction model reaches a desired level of accuracy, model coefficients can be used to predict brain age in a test dataset i.e. completely new samples

e) Results: A metric such as brain predicted age difference is defined to quantify the discrepancy between the predicted age and the chronological age. This metric can help to identify acceleration or deceleration of individual brain ageing (see also Fig. 1) and could be used to evaluate the presence of diseases. If the analysis was performed on a test dataset of healthy people this could be another way to assess the accuracy of the model, again providing different metrics as MAE, RMSE and Pearson's correlation coefficient R.

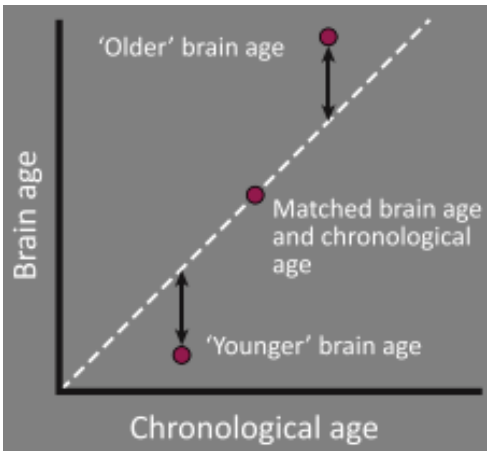

Figure 1: *Brain age versus chronological age [5]*

Mean Absolute Error (MAE) is computed as:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{1}$$

with N sample size, y_i chronological age and $\hat{y}_i$ the predicted brain age. The MAE provides a direct way to assess the difference among the various learning methods and it was found to be the most meaningful

measure for assessing the influence of different parameters [1]. Usually also the **Root Mean Squared Error** is computed for comparison as the following:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{2}$$

Finally, the formula for the computation of **Pearson's correlation coefficient** is:

$$R = \frac{\sum_{i=1}^{N}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{N}(\hat{y}_i - \bar{\hat{y}})^2}} \tag{3}$$

where y_i is the chronological age, $\hat{y}_i$ is the predicted brain age and $\bar{y}$ and $\bar{\hat{y}}$ are their means.

Metrics as MAE and RMSE take into account only the relative difference between observed and predicted values i.e. they tend to reproduce in age sub-samples the same behaviour they have on the entire dataset. This is not true for correlation coefficient, that is heavily affected by the overall range: when considering age sub-samples this range decreases and the resulting correlation became worse [6].

In the recent years Machine Learning approaches have demonstrated the possibility to accurately predict age from brain MRI scans in healthy subjects.

Cole et al. in [7], for example, used Convolutional Neural Networks (CNN) on BAHC dataset (Brain Age Healthy Control). One hindrance to clinical applications of brain age estimation is the time needed for doing data pre-processing before providing them to the machine learning algorithm. Furthermore, often assumptions not met by data are made in pre-processing to get more compact data. Thus they tried to use CNN not only on pre-processed data but also on raw data. They chose to perform the analysis also with a Gaussian Process Regression model, in order to have a performance comparison with a well-known high accuracy model for brain age prediction. Results of these analysis are shown in Fig. 2, where it could be seen that lowest MAE is achieved using GM data as input data and CNN method.

Method	Input data	MAE (years)	r	R^2	RMSE
CNN	GM	4.16	0.96	0.92	5.31
	WM	5.14	0.94	0.88	6.54
	GM+WM	4.34	0.96	0.91	5.67
	Raw	4.65	0.94	0.88	6.46
GPR	GM	4.66	0.95	0.89	6.01
	WM	5.88	0.92	0.84	7.25
	GM+WM	4.41	0.96	0.91	5.43
	Raw	11.81	0.57	0.32	15.10

Figure 2: *MAE, R, R^2 and RMSE from [7]*

Doing test-retest trials they found also that brain predicted age was highly reproducible, finding an intra-class correlation coefficient $ICC > 0.90$ for all analysis. This is crucial for any measure to be used in longitudinal and multi-centre studies and, potentially, in future clinical settings.
Actually, there is still a limitation for clinical usage: a MAE = 4-5 years is clearly an insufficient precision for doing estimates in individual case and making clinically-meaningful decisions. Thus further researches are necessary to reduce MAE in brain age estimation in healthy subjects.
Another limitation of this method is that the complexity of CNN models prevents the possibility to provide a neuroanatomical interpretation of features used in brain age prediction.

Neuroimaging-derived age predictions have also been explored in the context of different brain diseases. Using these models, deviations from healthy brain ageing have been identified for example in Down's Syndrome, Alzheimer's disease, Mild Cognitive Impairment and other diseases as HIV.
In [8] they successfully employed a Gaussian Process Regression model to predict age in people with Down Syndrome and typically developing controls. Chronological age was then subtracted from brain-predicted age to generate a brain predicted age difference. On the training set they found R = 0.94, MAE

= 5.02 years and RMSE = 6.31 years. Analyzing data from the DS group they found a mean brain-PAD significantly greater than controls, supporting the idea that DS is associated with premature structural brain aging.

Gaussian Process Regression method was exploited also in [9], where it was used to establish whether HIV disease as well is associated with abnormal levels of age-related brain atrophy. Prediction accuracy results very similar to the previous study with R = 0.94, MAE = 5.01 and RMSE = 6.31. HIV was indeed found to accentuate brain aging: mean brain PAD score in HIV-positive individuals was 2.15 years, while in HIV-negative individuals it was -0.87 years.

A lot of studies were performed on Alzheimer's Disease patients, relying on the hypothesis that pathological atrophy in AD reflects an accelerated aging process.
Franke et al. in [1] analyzed Alzheimer's Disease Neuroimaging Initiative (ADNI) database with both Relevant Vector Regression and Support Vector Regression methods. Accuracy on an healthy dataset (IXI database) was found to be better when using RVR, obtaining R = 0.92 and MAE = 5 years. Size of the training dataset was found to have a strong effect on age estimation accuracy, finding MAE = 5.2 for half of the dataset and MAE = 5.6 years taking a quarter of the dataset. For the AD group, the mean Brain Age Gap was + 10 years, implying a systematically higher estimated brain age than true age.
Brain predicted age measures were found again to be stable and reproducible, even across scanners. This allows to use Relevance Vector Regression algorithm also in longitudinal studies as presented by the same authors in [2]. They included all subjects from the ADNI database for whom at least the baseline scan and one follow-up scan were available and explored the patterns of longitudinal changes in individual brain age within a follow-up period of up to 4 years.
Subjects were grouped as: NO (healthy subjects), sMCI (stable MCI), pMCI (progressive MCI, turned in AD after a certain amount of time) and AD (Alzheimer's Disease patients).
Regarding NO and sMCI subjects, the estimated brain age at baseline did not differ significantly from the chronological age and the brain age score remained stable across the follow up period of up to 4 years, showing only normal age-related atrophy. Patients with AD and subjects who had converted to AD within 3 years (pMCI) showed instead accelerated brain atrophy by +6 years at baseline, and an additional increase with a score of about +9 years during the follow-up. This shows how accelerated age-related changes in brain atrophy are already evident also at the prodromal stage of AD i.e. MCI.
The fit of the longitudinal changes in brain age results in the following changing rates (brain age years per follow up year): NO = 0.12, sMCI = 0.07, pMCI = 1.05, AD = 1.51. These results suggest that the acceleration in brain aging in pMCI and AD found at baseline becomes even more accelerated during the next months and years. Furthermore, accelerated brain aging was found to be related to prospective cognitive decline and disease severity.

One important note should now be done on accuracy of different regression methods. As exposed in different studies such as in [6] [7] [10] , it turns out that the prediction error is not constant but is subject to significant changes when considering specific age ranges. As an example, authors in [6] divided the all data set (484 subjects, 7-80 years) in 4 age range subsamples and evaluated the regression accuracy as MAE and RMSE in each of them as shown in Fig. 3. Best results were obtained for younger subjects (age range 7-20) while the performance has a significant drop when considering groups with older subjects. Worst prediction accuracy was obtained in the age range 40-60, reflecting the high specificity and variability characterizing brain atrophy in these years.

Age range	MAE	RMSE
7 – 20	**3.7 ± 0.2**	**3.9 ± 0.1**
20 – 40	5.1 ± 0.2	6.6 ± 0.1
40 – 60	6.5 ± 0.2	8.2 ± 0.2
60 – 80	4.4 ± 0.2	6.6 ± 0.3

Figure 3: *MAE and RMSE in different age ranges [6]*

This means that, although using analogous predictions models, results can severely be affected by the age range of the samples. As a consequence it would be better to make comparisons of performance accuracy between different methods considering various database with consistent age distribution. For these reasons, there are different studies that make a comparison of different Machine Learning algorithms on the same database.

For example, authors in [6] introduced a method based on multiplex networks combined with deep learning regression and made a comparison of its performance on the same dataset with state-of-the-art regression strategies such as Lasso Regression, Ridge Regression, Support Vector Machine and Random Forest regressions. Results are reported in Fig. 4. They found out that deep learning provides the most accurate model with respect to all the considered metrics.

Model	MAE	RMSE	ρ
Deep learning	**4.7 ± 0.1**	**6.2 ± 1.1**	**0.95 ± 0.02**
Ridge regression	6.0 ± 0.7	7.8 ± 1.3	0.92 ± 0.03
Lasso regression	6.4 ± 0.7	8.2 ± 1.3	0.92 ± 0.03
Random forest	5.9 ± 0.7	7.6 ± 0.9	0.94 ± 0.02
Support vector machine	5.6 ± 0.7	7.2 ± 0.9	0.94 ± 0.01

Figure 4: *MAE, RMSE and R for different Machine Learning methods [6]*

They studied also the sample size effect on the accuracy of the model, finding that as the sample size increases, predicting models tend to be more accurate (both in terms of MAE and R); when using 80 % of data, the DNN model performance reaches a robust plateau.
Furthermore, their DNN model allows also to identify brain regions which seem to majorally affect the age prediction, identifying features that had a strategic role in the age prediction.

Deep neural network learning method was exploited also in Bellantuono et al. [11], considering a simpler network model and using a database of 1112 individuals with age range 7-64 years. Results of performance comparison on the same database is shown in Fig. 5. Also in this case DNN is the best algorithm, supporting the hypothesis that Deep Neural Networks are an appropriate option to manage the intrinsic complexity of the brain and identify features which accurately predict its age.

Algorithm	MAE	RMSE	ρ
Deep learning	**2.19 ± 0.02**	**2.91 ± 0.03**	**0.890 ± 0.003**
Random Forests	3.09 ± 0.02	4.13 ± 0.03	0.770 ± 0.006
Lasso	2.54 ± 0.02	3.34 ± 0.02	0.850 ± 0.002
Ridge	2.49 ± 0.01	3.29 ± 0.02	0.858 ± 0.002
Elastic net	2.50 ± 0.02	3.30 ± 0.02	0.855 ± 0.002
Support Vector Machine	2.40 ± 0.03	3.19 ± 0.04	0.870 ± 0.004
Relevance Vector Machine	2.48 ± 0.03	3.25 ± 0.03	0.859 ± 0.003

Figure 5: *MAE, RMSE and R for different ML algorithms [11]*

In the same study the performance of the model in the development age was analyzed, considering a subset of subjects within the 7-20 age range, reporting MAE = 1.53 ± 0.02, RMSE = 1.94 ± 0.02 and R = 0.787 ± 0.006. As expected, MAE and RMSE showed an improvement while correlation worsened, as it suffers more from sample size reduction. This behaviour confirms the necessity to compare age prediction accuracy declared in different studies paying attention to the age distribution of examined cohort.

Finally, feed-forward DNN on raw data was used by Lombardi et al. in [10]. The database used in this article comes from Predictive Competition 2019 and is made of data of 2638 individuals from 17 sites. Competition aimed to achieve the lowest MAE for brain age prediction, while keeping the Spearman correlation between the brain age delta and the chronological age under 0.1. Performance results in training set are shown in Fig. 6, where they are compared with the ones of other common techniques. Different performances are due to different classification approaches that identify different features as important descriptors.

DNN was found to have the greatest homogeneity across different acquisition sites and the greatest stability over sample size variations and sample heterogeneous age distribution.
The aim to minimize the Spearman correlation coefficient between age gap and chronological age was

Model	MAE	R
RF	6.71 ± 0.39	0.83 ± 0.02
SVR	6.25 ± 0.41	0.85 ± 0.02
LASSO	5.99 ± 0.36	0.86 ± 0.02
DNN	5.39 ± 0.34	0.84 ± 0.03

Figure 6: *MAE and R for different ML models [6]*

made to achieve an unbiased algorithm for brain age prediction. Although DNN models exhibit the lowest correlation values (R = 0.38), a systematic age underestimation in the age range 60-90 and overestimation in the age range 20-35 can be noticed. This variation of brain PAD as a function of age was found also in [7]. Thus, age bias correction techniques need to be further applied to achieve less biased models. Reporting a constant behaviour for different age bins is important to ensure the reliability and generalization of ML models.

2 Segmentation pipelines - FreeSurfer and FastSurfer

Many machine learning methods use automatic segmentation pipelines as preprocessing steps for further analysis. Brain identification and segmentation can be made through software packages such as FreeSurfer and FastSurfer.

FreeSurfer is a toolkit for the analysis and visualization of structural, functional, and diffusion neuroimaging data. Among different things, it allows the user to reconstruct cortical and subcortical volumes from 3D MRI volumes.

The main FreeSurfer pipeline's drawback is that processing a single image is strongly time-consuming. To overcome this issue, FastSurfer was introduced. FastSurfer is a fast and extensively validated deep-learning pipeline for fully automated processing of structural human brain MRIs. It consists of two main parts building upon each other:

- FastSurferCNN: an advanced deep learning architecture capable of whole brain segmentation into 95 classes in under 1 minute, mimicking FreeSurfer's anatomical segmentation and cortical parcellation
- Recon-surf: full FreeSurfer alternative for cortical surface reconstruction, mapping of cortical labels and traditional point-wise and ROI thickness analysis in approximately 60 minutes

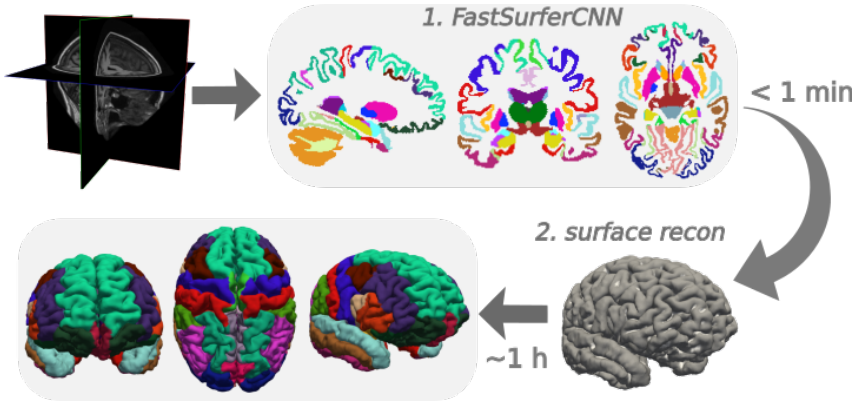

Figure 7: *FastSurfer pipeline [12]*

Different studies were made in order to perform the comparison between the two software, regarding both runtime and quality of results.

For example, in [13] authors used FreeSurfer and FastSurfer pipelines in order to extract volumetric features from MRI scans and compared execution times, finding that FastSurfer one was substantially smaller than FreeSurfer's (Fig. 8).

Execution time	FreeSurfer CPU	FreeSurfer GPU	FastSurfer GPU
Mean	04:11:55.7	04:33:34.2	**00:05:49.3**
Standard deviation	02:15.04.9	05:14:38.8	**00:01:09.8**
Minimum	03:04:42.0	02:53:26.0	**00:05:09.0**
Maximum	15:45:01.0	31:44:59.0	**00:11:46.0**
Ratio of failed MRI scans	1/30	1/30	**0/30**

Figure 8: *FreeSurfer - FastSurfer execution times comparison [13]*

Concerning the quality of results, in [12] authors tested accuracy and reliability of FastSurfer vs FreeSurfer on different datasets, obtaining results in Fig. 9.

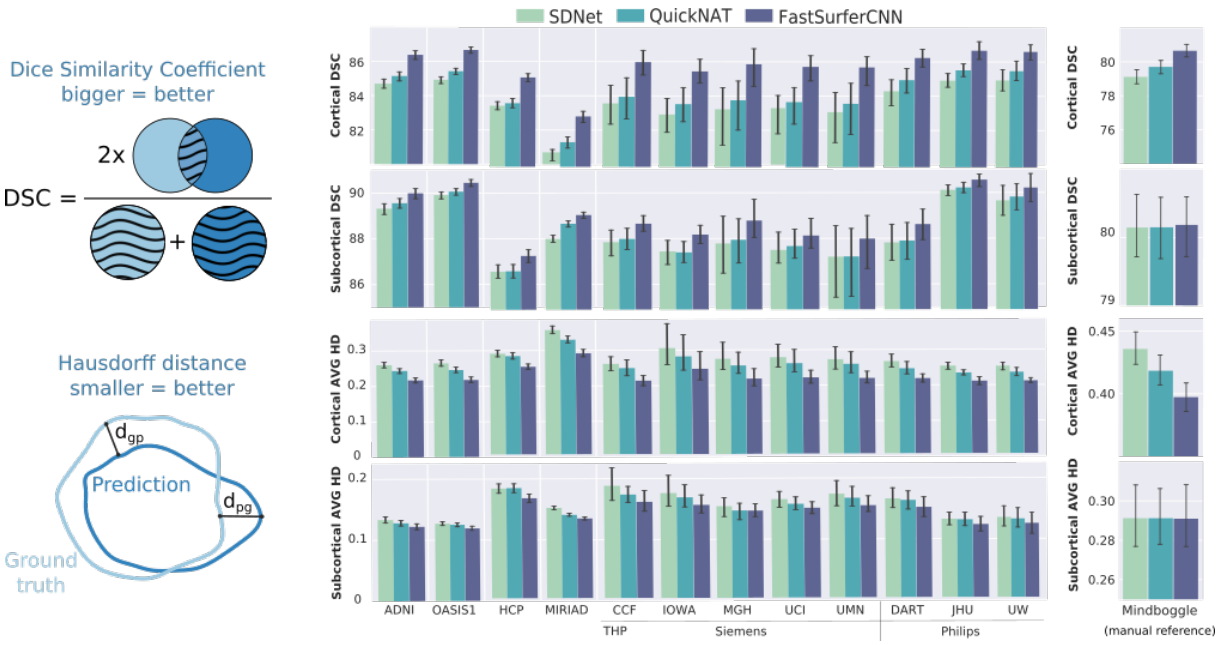

Figure 9: *FastSurfer performances [12]*

Therefore FastSurfer is a valid and preferable alternative to FreeSurfer, as outputs are comparable (or even better) but runtimes are lower. These features make it possible to perform big data analysis and eventually to develop clinical applications. For these reasons FastSurfer's usage as a preprocessing step in ML methods is increasing throughout years.

In the following work, FreeSurfer and FastSurfer outputs will be compared, in order to verify the similarity and find critical points and main segmentation differences between the two pipelines.

2.1 Dataset

Segmentation comparison was performed on data extracted from an MRI dataset of pediatric patients affected by sickle cell disease. In particular, three patient volumes were used as samples to perform the comparisons and analyze the results. In the following, these volumes will be referred as volume 1, 2 and 3.

For both software *aseg.mgz* (automatic segmentation of the entire brain) outputs were used. For a further comparison, also *aseg.auto.mgz* FreeSurfer output was taken into account.

For each volume, comparisons were performed on these outputs taking two of them at time. In particular, in the following we will name comparisons as following:

- *Case A*: FreeSurfer *aseg* $\iff$ FreeSurfer *aseg.auto*
- *Case B*: FreeSurfer *aseg* $\iff$ FastSurfer
- *Case C*: FreeSurfer *aseg.auto* $\iff$ FastSurfer

3 Data Analysis and Results

3.1 Statistics

Some statistics were evaluated in order to make the comparison between different segmentations. These are Dice Similarity Coefficient, Jaccard Index and Hausdorff Distance.

Dice Similarity Coefficient

The Dice Similarity Coefficient (DSC) is a metric that measures the similarity between two sets of data; in this contest it is used to evaluate the segmentation performance of the deep learning networks. The Dice Coefficient ranges from 0 to 1, where 0 indicates no similarity and 1 indicates a perfect similarity. Thus in our case the higher the Dice coefficient, the more similar the two segmentations are.
The Dice coefficient can mathematically be expressed as in Eq. 4, where $|A \cap B|$ represents the common elements between sets A and B, and $|A|$ represent the number of elements in set A (and likewise for set B). In our case, we can approximate $|A \cap B|$ computing the element-wise multiplication between the two segmentations and then summing the resulting matrix.

$$DSC(A, B) = \frac{2|A \cap B|}{|A| + |B|} \tag{4}$$

Jaccard Index

The Jaccard Index, also called Intersection over the Union (IoU), is essentially a method to quantify the percent overlap between two images. This metric is closely related to the Dice coefficient and similarly it ranges from 0 to 1. It is computed as the following:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{5}$$

Hausdorff Distance

The Hausdorff Distance is a metric for measuring the similarity between two sets of points and is often used to evaluate the quality of segmentation boundaries. Having two binary maps, A and B, it is defined as:

$$d_H(A, B) := \max \left\{ \sup_{a \in A} d(a, B), \sup_{b \in B} d(A, b) \right\} \tag{6}$$

Taking into account the maximum distance, the standard Hausdorff distance considers the "worst-case scenario" and this makes it sensitive to outliers. To avoid this, Average Hausdorff Distance was introduced, computed as following:

$$AVG\ d_H(A, B) = \frac{1}{|A|} \sum_{a \in A} min_{b \in B} d(a, b) + \frac{1}{|B|} \sum_{b \in B} min_{a \in A} d(b, a) \tag{7}$$

with $|A|$ and $|B|$ representing the number of voxels in A and B, respectively. In contrast to the DSC, a smaller AVG HD indicates a better capture of the segmentation boundaries with a value of zero being the minimum (perfect match).

3.1.1 Results

Previously cited statistics were computed for the three volumes. In addition to these, also the volumetric difference was computed as the difference between the two volume total sums. Results are shown in Tab. 1.

	1			***2***			***3***		
	Case A	*Case B*	*Case C*	*Case A*	*Case B*	*Case C*	*Case A*	*Case B*	*Case C*
Dice Similarity Coefficient	0.963	0.987	0.958	0.956	0.987	0.955	0.955	0.987	0.954
Jaccard index	0.928	0.974	0.919	0.916	0.975	0.914	0.914	0.973	0.912
Volumetric diff.	-64988	8439	73427	-97433	-1582	95851	-105436	-1691	103745
Hausdorff distance	0.10	0.02	0.12	0.11	0.02	0.12	0.12	0.02	0.12

Table 1: *Dice Similarity Coefficient, Jaccard Index, Volumetric difference and Hausdorff distance for each case for each different volume (1, 2, 3)*

Considering the results for each single comparison, large similarity was found, having high Dice Similarity Coefficients and Jaccard indexes (all above 0.9) and low Hausdorff distances.

Concerning the comparison between case A, B and C, in all three images the best statistics were reached in case B, comparing FastSurfer with FreeSurfer *aseg* volume. In particular this case has the largest Dice Similarity coefficient and Jaccard index as well as the lowest Hausdorff distance. The result given by the volumetric distance confirms these observations, as in case B is one/two order of magnitude lower than the other cases. In addition to this, volumetric distance was found to be always negative in case A and positive in case C. This induces the idea that FreeSurfer *aseg.auto* segments more regions compared to both FreeSurfer *aseg* and FastSurfer. This result will be found also in the following analysis.

Comparing different images, in case B a Dice Similarity Coefficient equal to 0.987 was found in all of them. Also Jaccard index of the three volumes was pretty similar (0.974, 0.975, 0.973).

3.2 Difference matrix

Difference matrix can be computed making the difference pixel by pixel between two segmented volumes. This was performed in all three cases (A, B and C) for the three volumes available (volume 1, volume 2 and volume 3). In the following, results for the first volume will mainly be examinated, but they should be considered as representative also of the others, as conclusions found were generally comparable.

3.2.1 Slice by slice

Difference matrix can be analyzed slice by slice. In particular, the sum of differences per slice and the percentage of different pixels per slice were plotted. These features could be visualized in the three common brain planes: axial, coronal and sagittal.

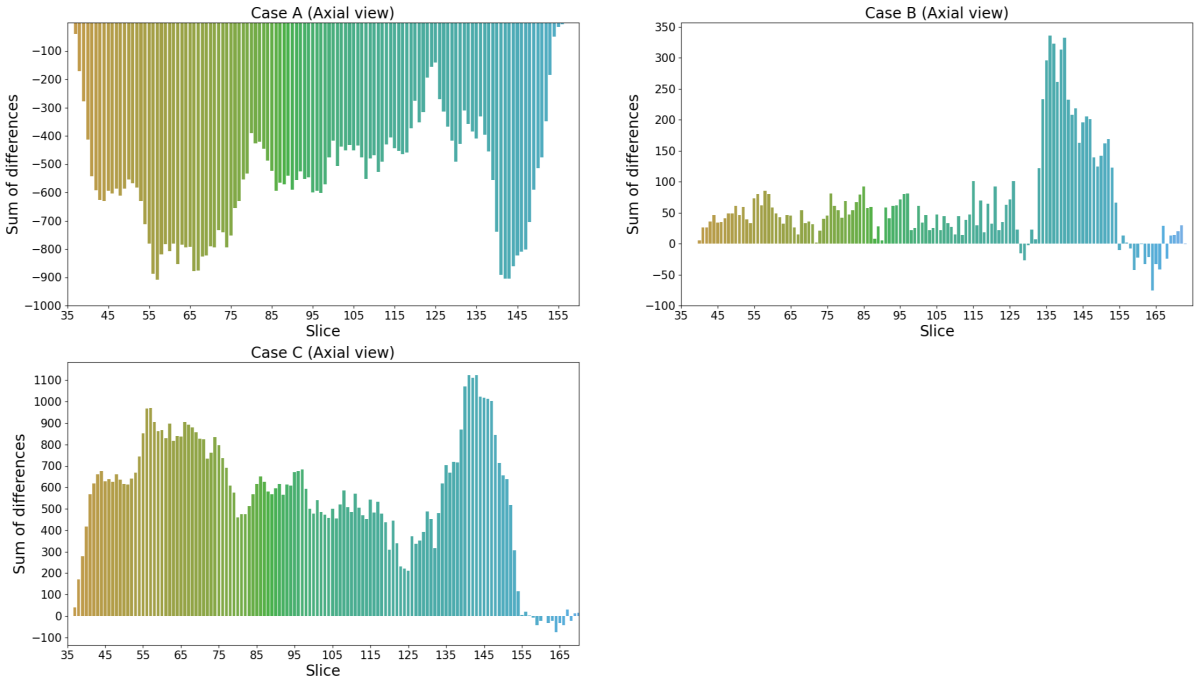

Figure 10: *Sum of the values of the difference matrix in each slice (axial view)*

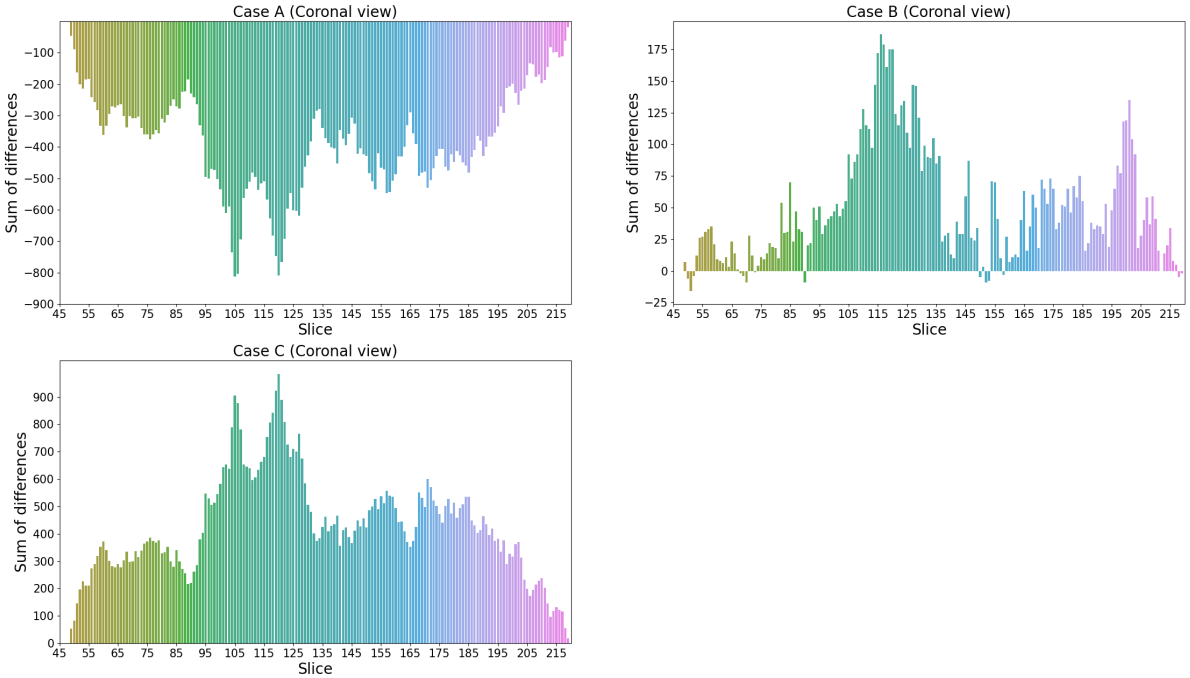

Figure 11: *Sum of the values of the difference matrix in each slice (coronal view)*

First of all one could compute the sum of the difference matrix in each slice for each plane, to precisely localize the differences. Histograms in Fig. 10, 11, 12 show the results of this analysis for the axial, coronal and sagittal planes respectively.

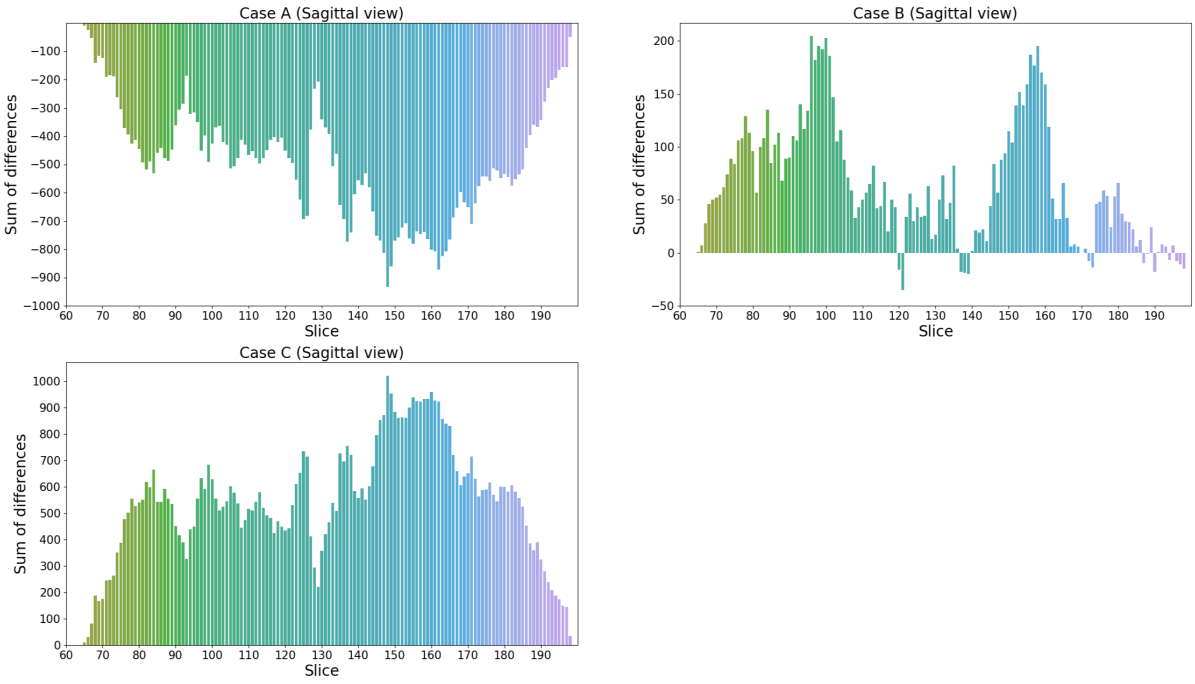

Figure 12: *Sum of the values of the difference matrix in each slice (sagittal view)*

Generally speaking, among the different planes, the first conclusion one could derive from these graphs is that the differences in case B are clearly lower than the ones in case A and case C. Negative sign in case A and positive sign in case C confirm that FreeSurfer *aseg.auto* output segments more regions.

For further analysis, one could look at the slices where there are larger values (i.e. peaks, where differences are localized) in order to visualize them.

One note should be made at this point. As previously said, the difference matrix is computed as a difference in each voxel. Original images have binary values, 0 or 1, therefore the difference matrix will have the following values: 0, ± 1. Adding all the values on a certain slice, one could have that a lot of +1 and -1 cancel off. As a consequence, strong differences of opposite sign located in different regions of the same slice could appear as small peak in the sum of differences histogram. Strong peaks on the histograms are anyway valid, and represent very strong differences as they "survive" the sum.

Lineplots of the percentage of different pixels per slice have been calculated to overcome this issue and have a more clear and reliable visualization of the differences. This feature turns out to be really small, especially for case B, but anyway useful to localize the differences. Results are shown in Fig. 13, 14, 15 for the first volume (but generalizable also for the other two volumes).

From these pictures we can again appreciate as case B has the lowest difference among the comparisons, while case A and case C have a comparable behaviour across the slices.

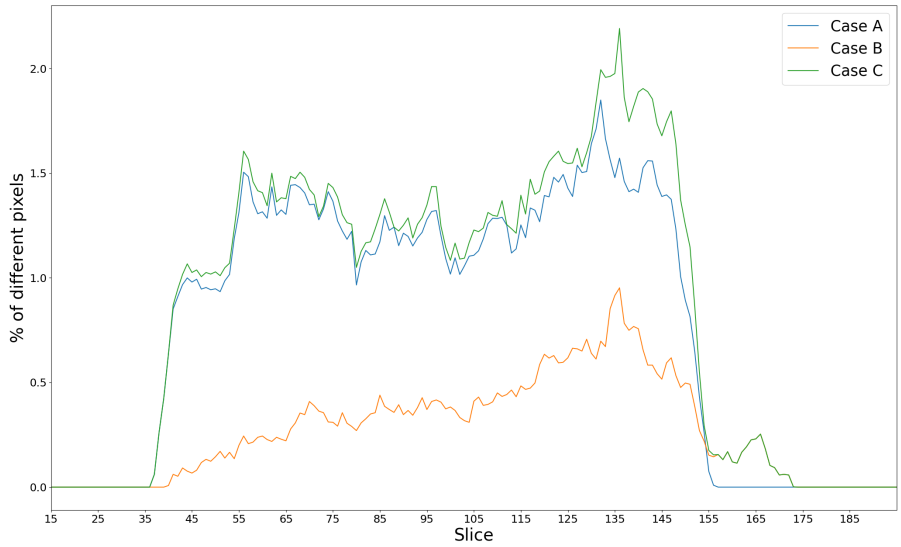

Figure 13: *Plots of the percentage of different pixels per slice for the first patient (axial projection)*

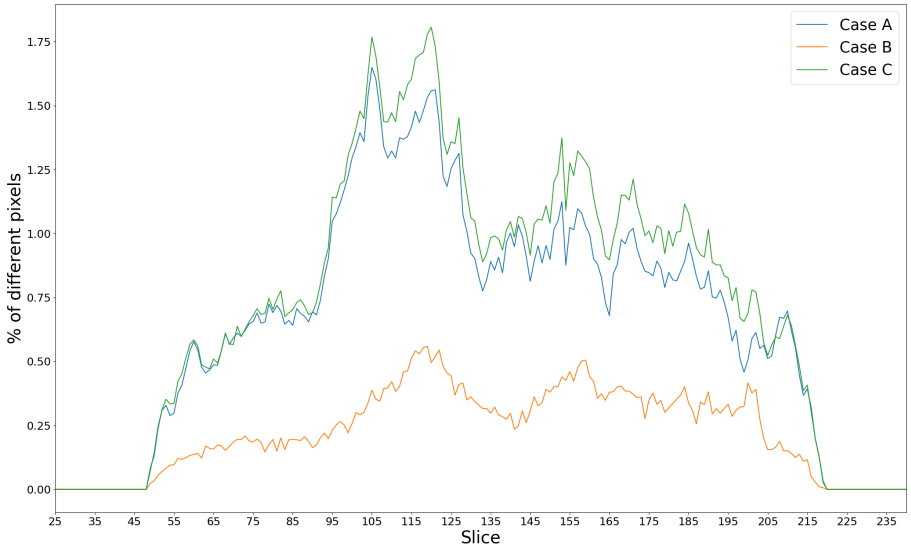

Figure 14: *Plots of the percentage of different pixels per slice for the first patient (coronal projection)*

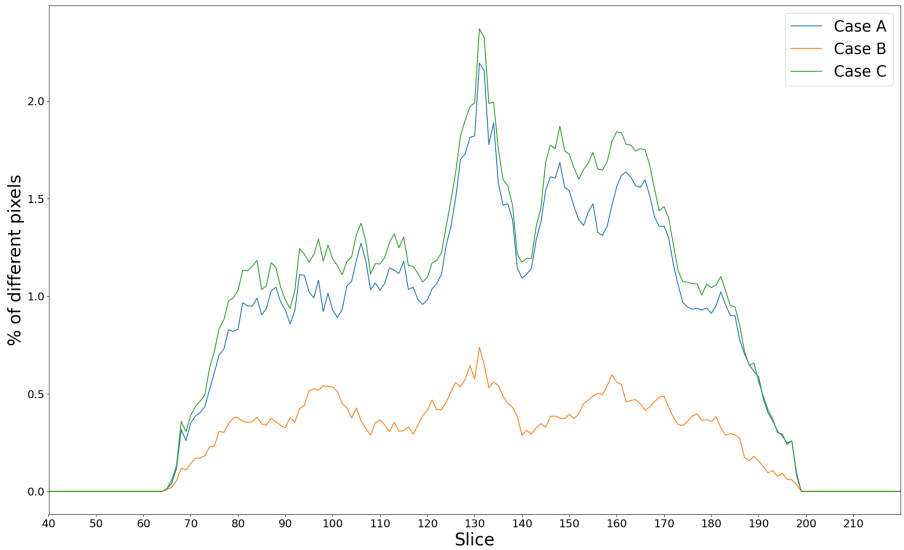

Figure 15: *Plots of the percentage of different pixels per slice for the first patient (sagittal projection)*

Case B

As we can easily deduce from histograms and line plots, case B is the one with lower differences. Indeed, most of them are visualized as single spots each one far from the others.
We can further study this case distinguishing among the three different planes.

Sagittal plane

A clear peak in both graphs can be found around slices 98 and 160, that is clearly localized in the lower part of the temporal lobe and a bit in the upper part of the cerebellum, as shown in Fig. 16.

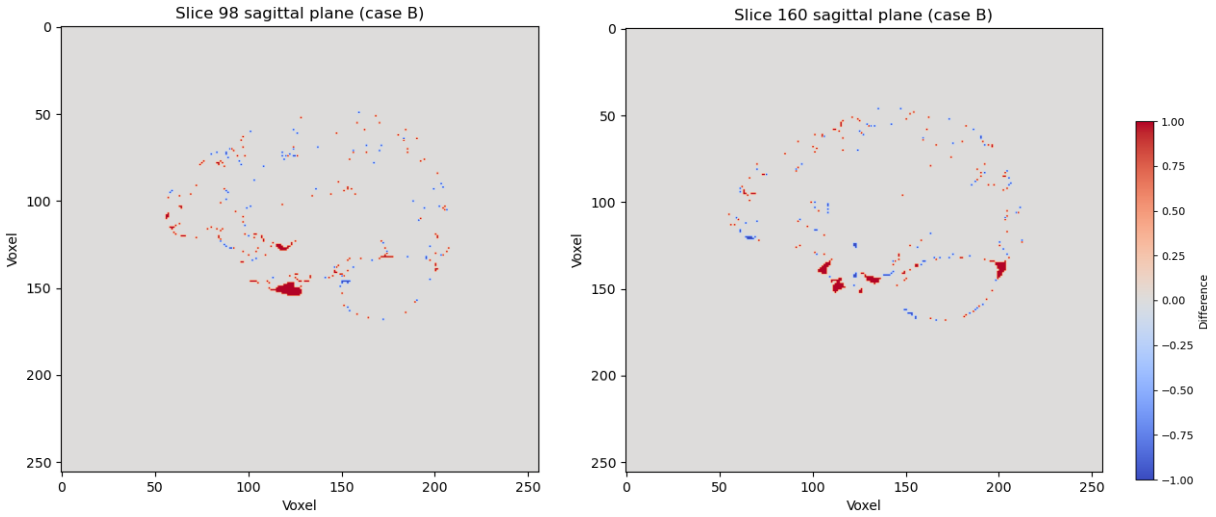

Figure 16: *Slices 98 and 160 on sagittal view. In both cases there are some differences localized in the upper part of the cerebellum and the lower part of the temporal lobe (red areas)*

In addition to these slices, the line plot also shows a peak around slice 130 that the histogram doesn't show. The reason for this is that, even if there aren't clearly localized regions, a lot of diffused blue and red spots occur in the difference matrix: these spots cancel off in the histogram but are shown in the line plot. In the following, high peaks due to spots randomly spread in the slices will not be mentioned as they don't provide any meaningful information on the differences localization.

Axial plane

A clear peak appears around slices 137-141, that could be recognized in axial view in light differences in both the external part of the cerebellum and the lower part of the temporal lobe (in the right side in Fig. 17).

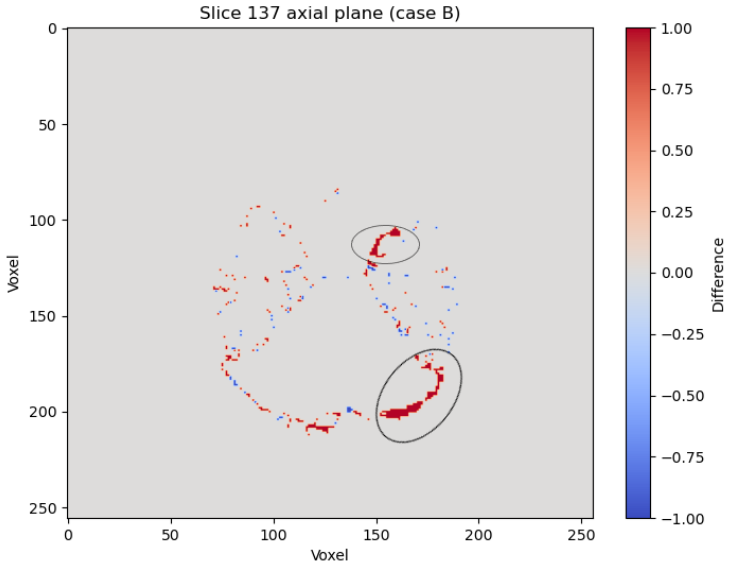

Figure 17: *Difference matrix on slice 137 shows light differences in both areas of interest (axial view)*

Coronal plane
In the coronal view, peaks around slices 117 show differences in the inferior part of temporal lobe (Fig. 18). This could not be cerebellum as the latter is present only in deeper slices (from slice 151 forward).

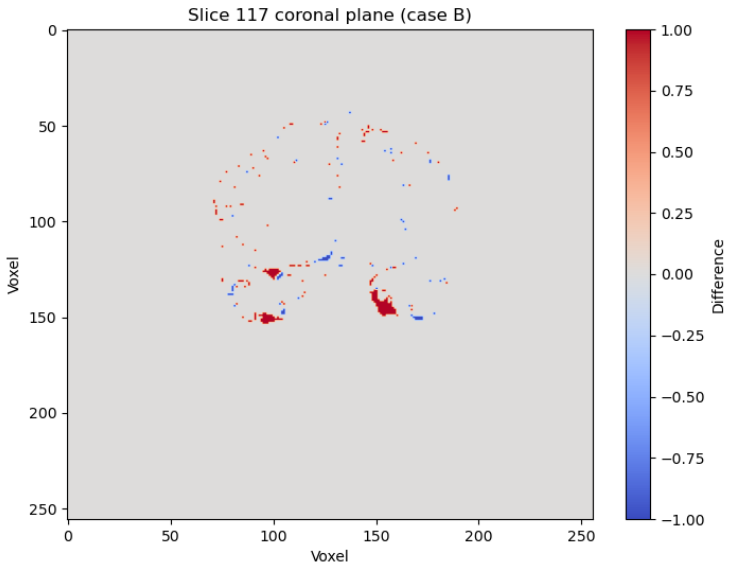

Figure 18: *Difference matrix on slice 117 shows differences in temporal lobe (coronal view)*

Case A and Case C
As we have already seen, differences are widely spread in the whole volume in cases A and C. However, one can anyway try to analyse and localize the main differences, distinguishing among the three planes (axial, coronal and sagittal).

Sagittal plane
Peaks at slices 107, 148 and 163 show that main differences are located in the inferior part of the temporal lobe and in the superior external part of the cortex, as shown in Fig. 19. Having that it is a negative difference in case A and positive in case C, that region is only segmented by FreeSurfer *aseg.auto*. From these images is also clear that cerebellum is segmented in the same way in case A, as it does not appear in the difference, while this happen in case C.

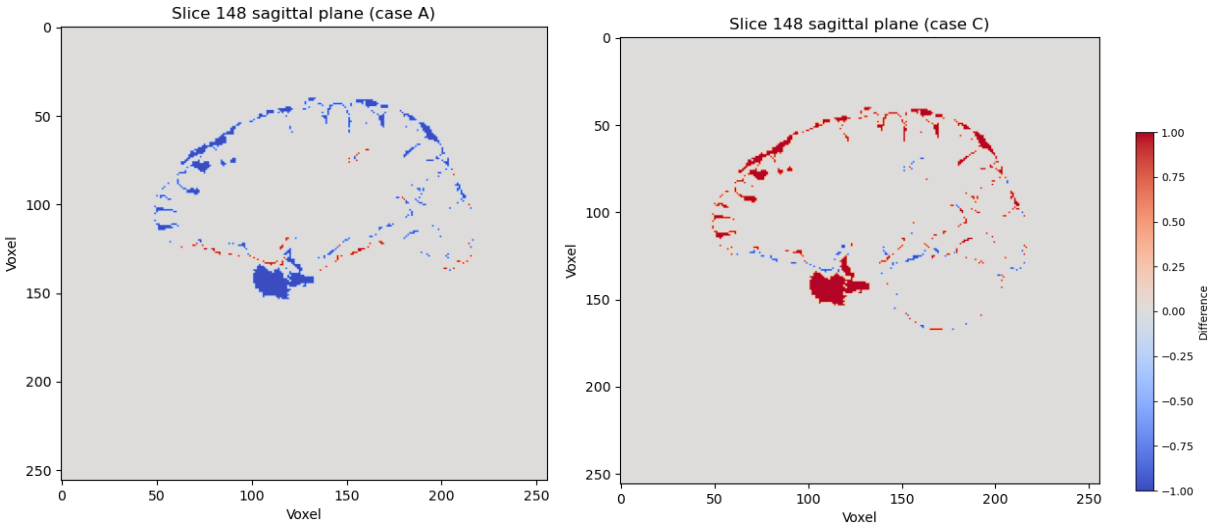

Figure 19: *Difference matrix on slice 148 (sagittal view) shows differences in the temporal lobe and the external part of the cortex*

An interesting note should be made about slices 134-142, shown in Fig. 20. Differences in the back part of the brain, above the cerebellum, are present in both images. These differences are not subject-specific for the first volume, as similar-located difference regions can also be found in the two other brain volumes under study, as shown in Fig. 21 and Fig. 22. This is clearly only a first evidence, that should be further verified on a larger number of samples.

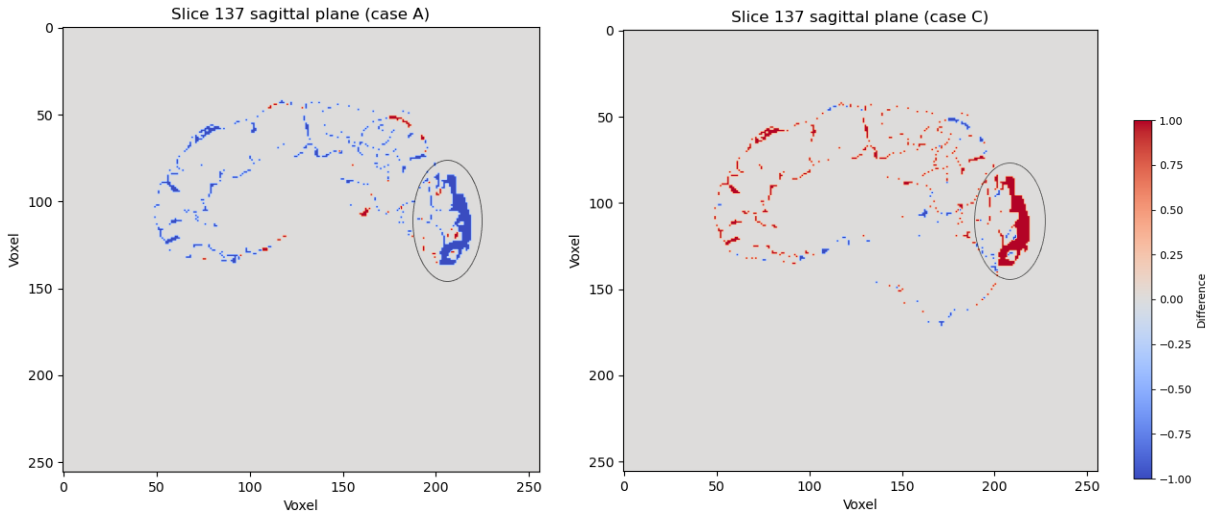

Figure 20: *Difference matrix on slice 137 in case A and C (sagittal view) shows differences in the posterior part of the brain, above the cerebellum, for the first volume*

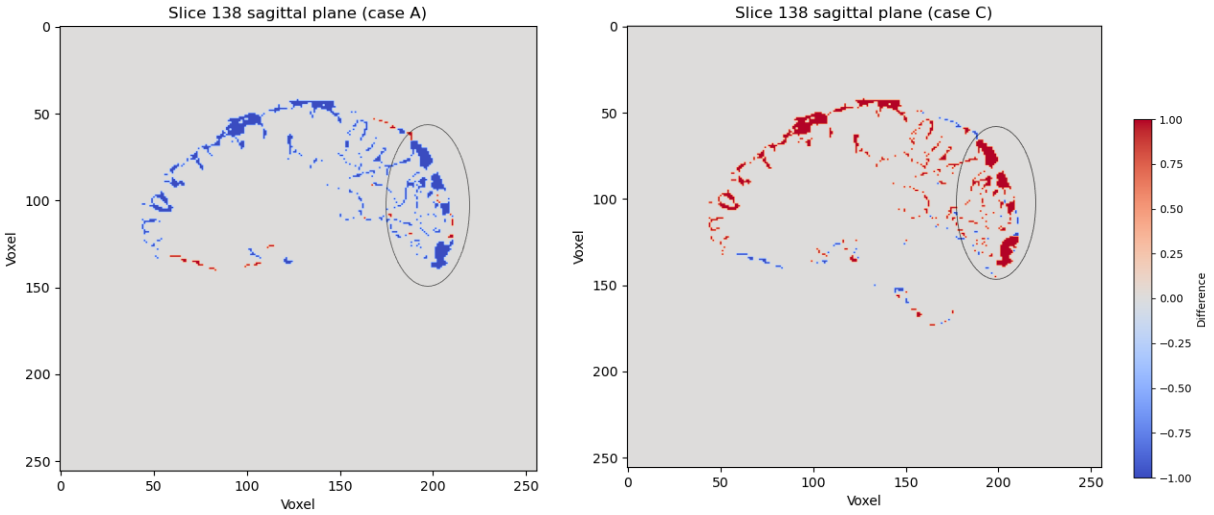

Figure 21: *Difference matrix on slice 138 (sagittal view) shows differences in the posterior part of the brain, above the cerebellum, also for the second volume*

Axial plane
A noteworthy difference is present at slice 141; this is again mainly localized in the inferior part of the temporal lobe, as shown in Fig. 23. Case C also shows some differences in the external part of the cerebellum (highlighted by a circle in the image).

Coronal plane
Once again in both case A and case C peaks from 105 to 128 are localized in the inferior part of the temporal lobe, as shown in Fig. 24.

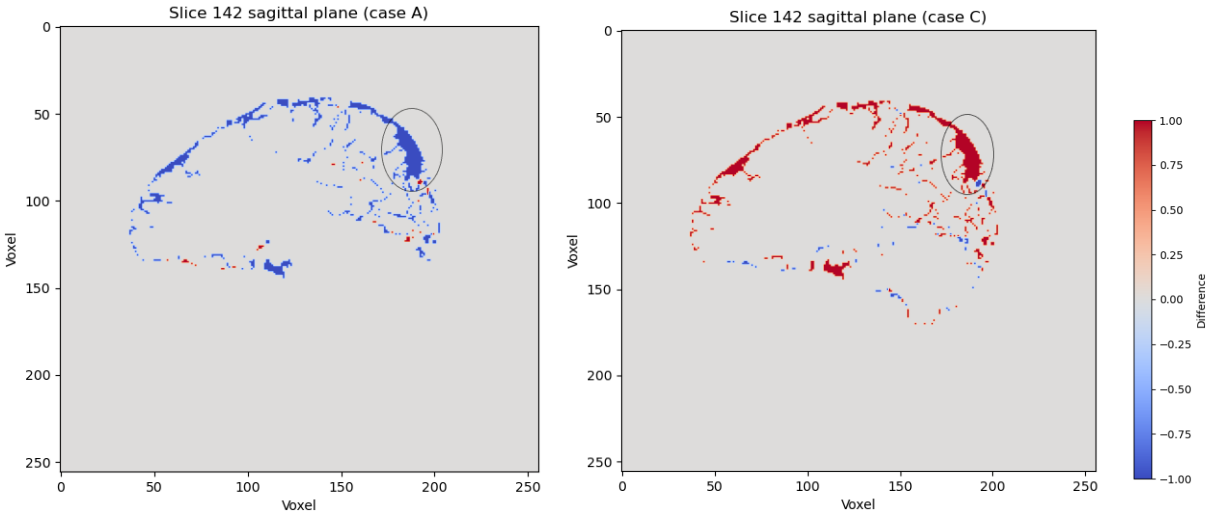

Figure 22: *Difference matrix on slice 142 (sagittal view) shows differences in the posterior part of the brain, above the cerebellum, also for the third patient*

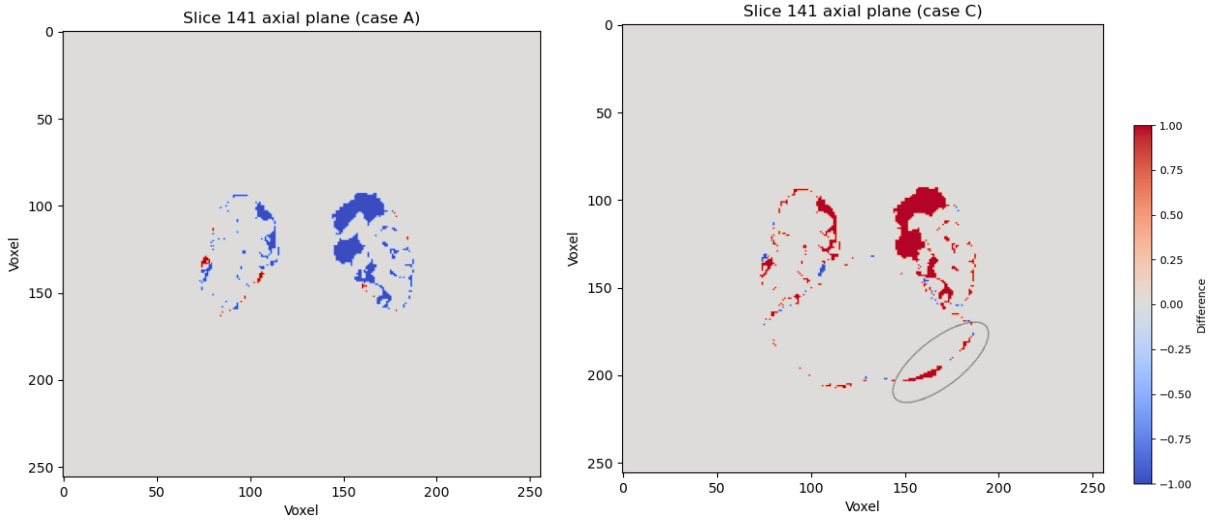

Figure 23: *Slice 141 in case A and C in axial view. In both cases there are some differences localized in the lower part of the temporal lobe. In case C light differences in the cerebellum have also been found*

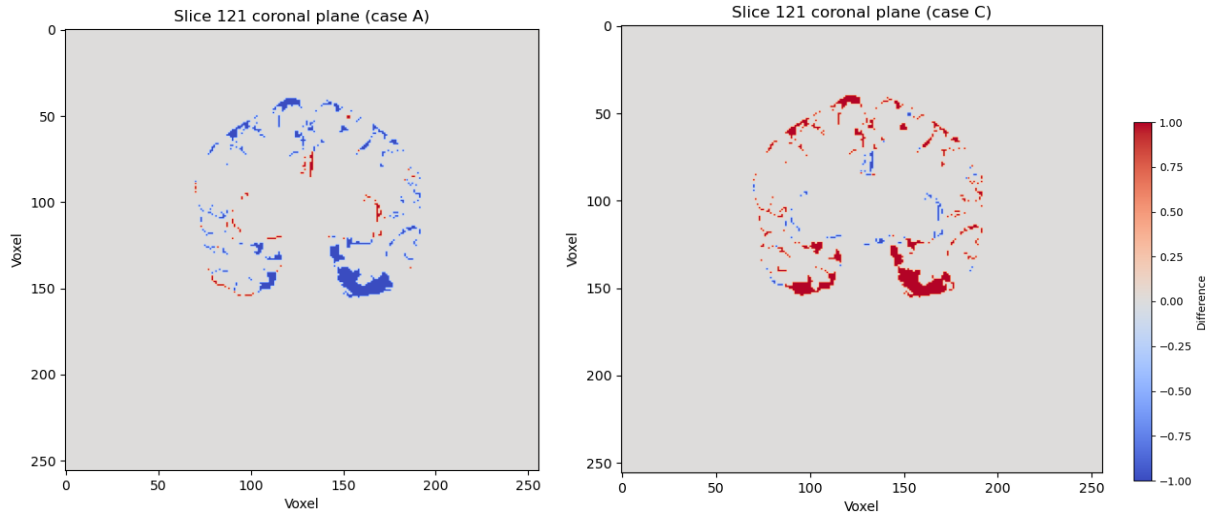

Figure 24: *Slice 121 in case A and C in coronal view. Main differences are in both case localized in the lower part of the temporal lobe*

3.2.2 On the entire volume

Results obtained in the previous paragraph are supported also by the 3D visualization of the difference matrix brain volume .

Case B
Brain volumes are shown in Fig. 25 (first volume), Fig. 26 (second volume) and Fig. 27 (third volume). In all three cases one could notice that the main differences are located in the upper part of the cerebellum and in the lower part of the temporal lobe, in agreement with what has been previously found.

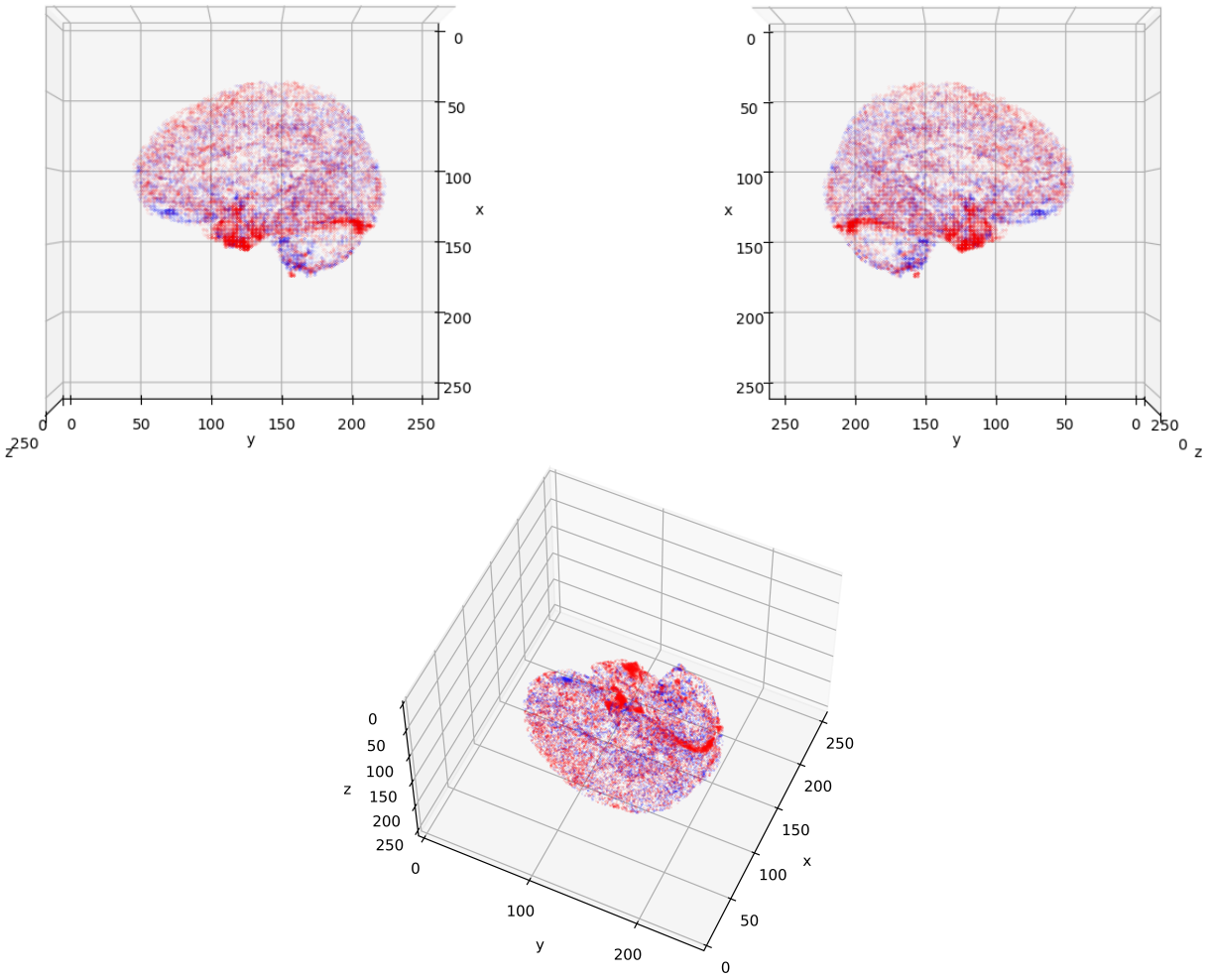

Figure 25: *Tridimensional plots of the difference matrix volume for the first patient. Red points are the ones only found by FreeSurfer. Blue points are the ones only found by FastSurfer*

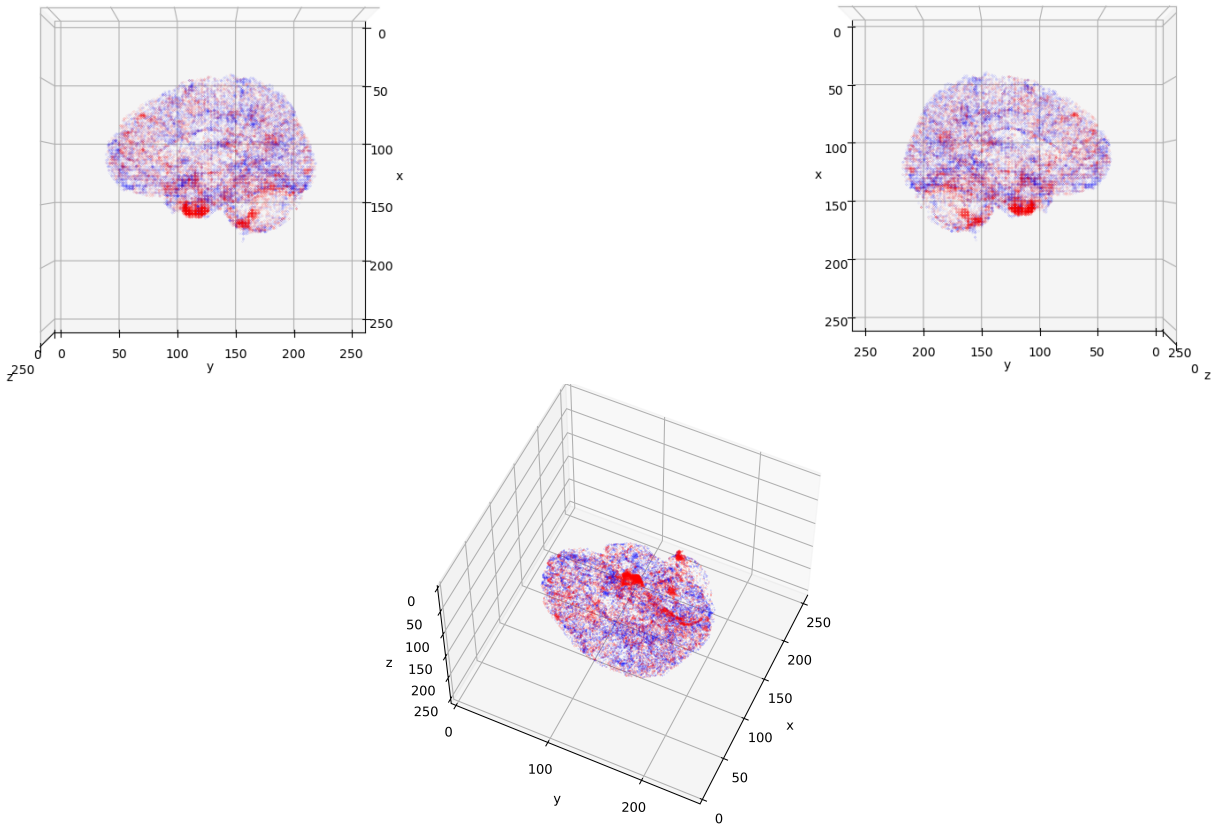

Figure 26: *Tridimensional plots of the difference matrix volume for the second patient. Red points are the ones only found by FreeSurfer. Blue points are the ones only found by FastSurfer*

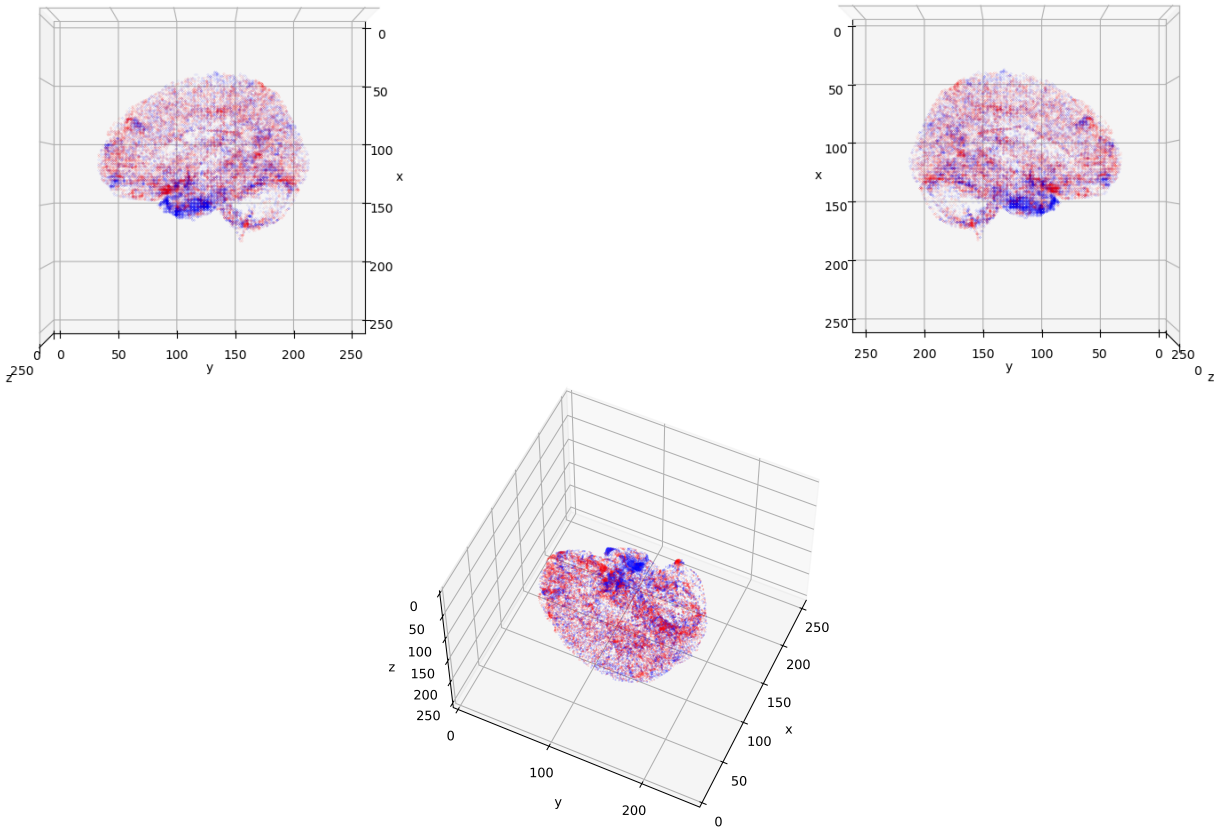

Figure 27: *Tridimensional plots of the difference matrix volume for the third patient. Red points are the ones only found by FreeSurfer. Blue points are the ones only found by FastSurfer*

Case A and case C

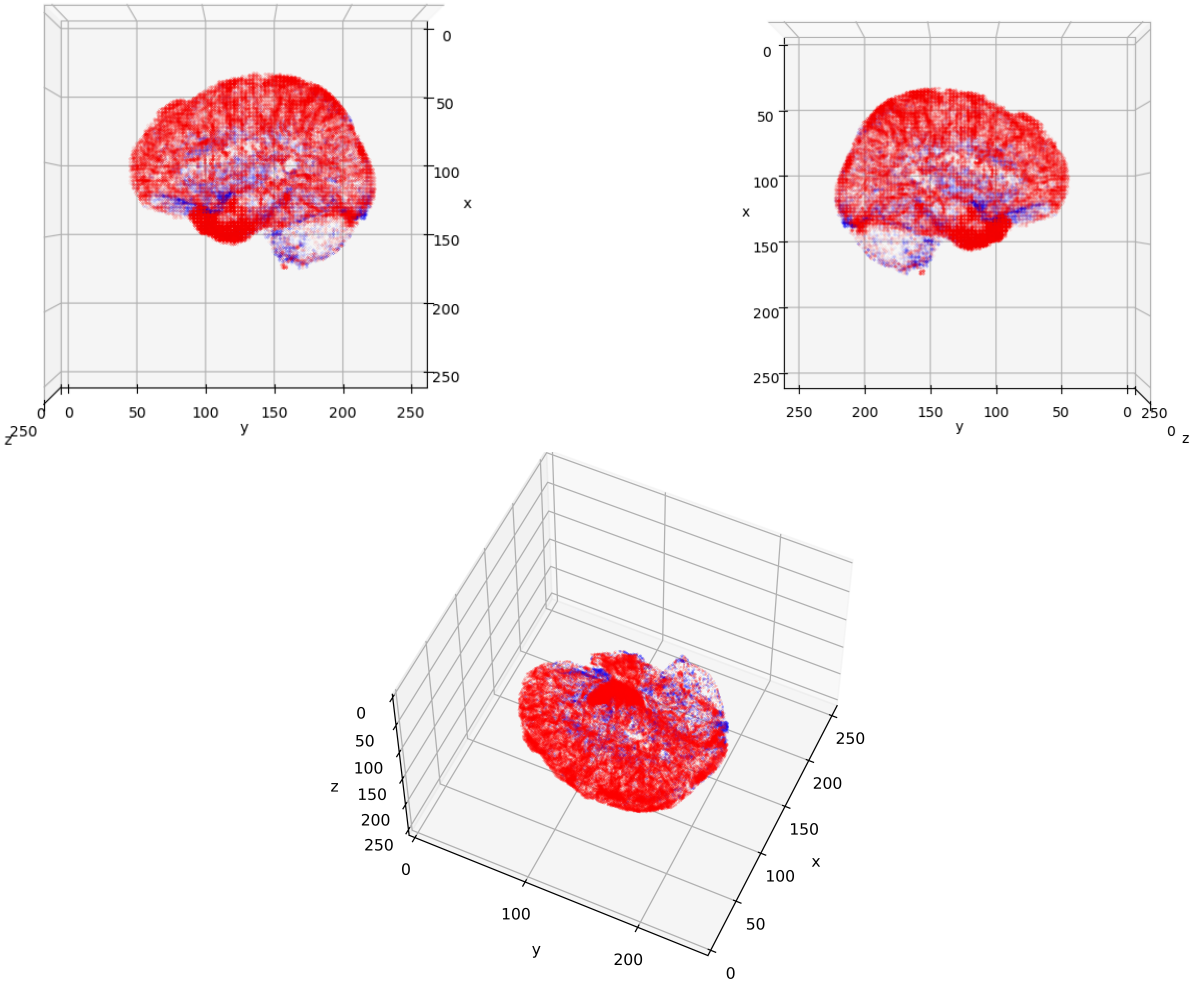

Figure 28: *Difference volume for case C, as an example of widespread differences*

Concerning case A and case C, 3D visualization is more chaotic, as there are more differences and they are not localized but spread all over the volume (especially in the external regions of the cortex). This supports the conclusions previously made. As an example, case C volume difference for the first volume is reported in Fig. 28. Also in this type of comparison one could notice as FreeSurfer *aseg.auto* identify more points than the other two (marked as red points in case C in Fig. 28).

4 Conclusions

This work aimed to perform a comparison between segmentation outputs given by FreeSurfer and FastSurfer, two pipelines commonly used in preprocessing steps of many machine learning algorithms in neuroscience. The main purpose of this comparison was verifying the similarity between outputs and identifying potential recursive differences.

Dice Similarity Index, Jaccard Index, Volumetric Difference and Hausdorff Distance were computed for the three volumes in each case (A, B, C). All comparisons turned out to have good similarity (DSC and Jaccard Index above 0.9, Hausdorff distance equal or lower than 0.1). FreeSurfer *aseg* output and FastSurfer output (named case B in comparisons) proved to have the greatest similarity (Dice Similarity Coefficient = 0.987, Jaccard Index $\simeq$ 0.974, Hausdorff Distance = 0.02). Small differences were found in general also computing the percentage of different pixel per slice in each image (larger differences were found to be of the order of 2%, lower than 1% in case B). These result are in agreement with what can be found in literature [12], [13] that suggest FastSurfer pipeline as a comparable (but faster) and valid alternative to FreeSurfer.

Difference matrix was then computed for all volumes in all three cases. Bidimensional and tridimensional visualizations of these matrices were used to easily identify where main differences could be located. Plots of the percentage of different pixel per slice, together with histograms of the sum of differences, were used to further determine slices where those differences were located. In particular, main peaks in the lineplots were studied. Case B was the easiest case to deal with as differences were less and more spread among the whole volume compared to cases A and C (as previously found from statistics). Indeed, most differences were single and scattered spots. Among these, noteworthy aggregated groups of points were found in the lower part of the temporal lobe and in the upper part of the cerebellum in all three planes (axial, coronal and sagittal). Following on from what has been previously said, differences in cases A and C were widely spread in the whole volume. Noticeable larger different regions were found to be located in the external part of the cortex (especially in the back side of the brain, above the cerebellum) and again in the lower part of the temporal lobe (but in a larger area compared to case B). Greater differences were due to the fact that in all the three volumes FreeSurfer *aseg.auto* output segmented more regions compared to both FreeSurfer *aseg* and FastSurfer output.
To conclude, some potentially systematic differences were identified but further studies on a larger sample are needed in order to characterize them as recurring differences in pipelines segmentation ability.

Results from both statistics and difference matrix analysis have thus shown that FastSurfer and FreeSurfer give comparable results. This suggest to use FastSurfer as a substitute to FreeSurfer, due to its incomparable higher speed. Increasing computational speed is indeed the main goal. We can take brain age estimation as an example. Studies shown in Section 1 described the brain predicted age difference (Brain PAD) as a novel MRI based biomarker, that aggregates the complex, multidimensional ageing pattern across the whole brain into one single value. Preprocessing of brain volumes was the first stage for estimating it. Thus, having a shorter computational time allows to reduce the entire pipeline performance time and to do a first step in the direction to make this technique clinically usable. The ultimate goal will indeed be the usage of MRI as a screening tool to help identifying people at greater risk of general functional decline and mortality during ageing.

References

[1] Katja Franke, Gabriel Ziegler, Stefan Klöppel, and Christian Gaser. Estimating the age of healthy subjects from t1-weighted mri scans using kernel methods: Exploring the influence of various parameters. *NeuroImage*, 50:883–892, 4 2010.

[2] Katja Franke and Christian Gaser. Longitudinal changes in individual brainage in healthy aging, mild cognitive impairment, and alzheimer's disease. *GeroPsych: The Journal of Gerontopsychology and Geriatric Psychiatry*, 25:235–245, 1 2012.

[3] Yashar Zeighami, Seyed Mohammad Fereshtehnejad, Mahsa Dadar, D. Louis Collins, Ronald B. Postuma, Bratislav Mišić, and Alain Dagher. A clinical-anatomical signature of parkinson's disease identified with partial least squares and magnetic resonance imaging. *NeuroImage*, 190:69–78, 4 2019.

[4] James H. Cole, S. J. Ritchie, M. E. Bastin, M. C. Valdés Hernández, S. Muñoz Maniega, N. Royle, J. Corley, A. Pattie, S. E. Harris, Q. Zhang, N. R. Wray, P. Redmond, R. E. Marioni, J. M. Starr, S. R. Cox, J. M. Wardlaw, D. J. Sharp, and I. J. Deary. Brain age predicts mortality. *Molecular Psychiatry*, 23:1385–1392, 5 2018.

[5] James H. Cole and Katja Franke. Predicting age using neuroimaging: Innovative brain ageing biomarkers. *Trends in Neurosciences*, 40:681–690, 12 2017.

[6] Nicola Amoroso, Marianna La Rocca, Loredana Bellantuono, Domenico Diacono, Annarita Fanizzi, Eufemia Lella, Angela Lombardi, Tommaso Maggipinto, Alfonso Monaco, Sabina Tangaro, and Roberto Bellotti. Deep learning and multiplex networks for accurate modeling of brain age. *Frontiers in Aging Neuroscience*, 11, 2019.

[7] James H. Cole, Rudra P.K. Poudel, Dimosthenis Tsagkrasoulis, Matthan W.A. Caan, Claire Steves, Tim D. Spector, and Giovanni Montana. Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker. *NeuroImage*, 163:115–124, 12 2017.

[8] James H. Cole, Tiina Annus, Liam R. Wilson, Ridhaa Remtulla, Young T. Hong, Tim D. Fryer, Julio Acosta-Cabronero, Arturo Cardenas-Blanco, Robert Smith, David K. Menon, Shahid H. Zaman, Peter J. Nestor, and Anthony J. Holland. Brain-predicted age in down syndrome is associated with beta amyloid deposition and cognitive decline. *Neurobiology of Aging*, 56:41–49, 8 2017.

[9] James H. Cole, Jonathan Underwood, Matthan W.A. Caan, Davide De Francesco, Rosan A. Van Zoest, Robert Leech, Ferdinand W.N.M. Wit, Peter Portegies, Gert J. Geurtsen, Ben A. Schmand, Maarten F.Schim Van Der Loeff, Claudio Franceschi, Caroline A. Sabin, Charles B.L.M. Majoie, Alan Winston, Peter Reiss, and David J. Sharp. Increased brain-predicted aging in treated hiv disease. *Neurology*, 88:1349–1357, 4 2017.

[10] Angela Lombardi, Alfonso Monaco, Giacinto Donvito, Nicola Amoroso, Roberto Bellotti, and Sabina Tangaro. Brain age prediction with morphological features using deep neural networks: Results from predictive analytic competition 2019. *Frontiers in Psychiatry*, 11, 1 2021.

[11] Loredana Bellantuono, Luca Marzano, Marianna La Rocca, Dominique Duncan, Angela Lombardi, Tommaso Maggipinto, Alfonso Monaco, Sabina Tangaro, Nicola Amoroso, and Roberto Bellotti. Predicting brain age with complex networks: From adolescence to adulthood. *NeuroImage*, 225, 1 2021.

[12] Leonie Henschel, Sailesh Conjeti, Santiago Estrada, Kersten Diers, Bruce Fischl, and Martin Reuter. Fastsurfer - a fast and accurate deep learning based neuroimaging pipeline. *NeuroImage*, 219:117012, 2020.

[13] Louise Bloch and Christoph M. Friedrich. Comparison of automated volume extraction with freesurfer and fastsurfer for early alzheimer's disease detection with machine learning. pages 113–118, June 2021.