Automatic labeling of cortical sulci using patch- or CNN-based segmentation techniques combined with bottom-up geometric constraints

Léonie Borne, Denis Rivière, Martial Mancip, Jean-François Mangin

Medical Image Analysis 62 (2020) 101651

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Automatic labeling of cortical sulci using patch- or CNN-based segmentation techniques combined with bottom-up geometric constraints

Léonie Borne[a,*], Denis Rivière[a], Martial Mancip[b], Jean-François Mangin[a]

[a] Université Paris-Saclay, CEA, CNRS, Neurospin, Baobab, Gif-sur-Yvette, 91191, France
[b] Maison de la Simulation, CNRS, CEA Saclay, Gif-sur-Yvette, 91191, France

ARTICLE INFO

Article history:
Received 28 May 2019
Revised 14 January 2020
Accepted 16 January 2020
Available online 28 February 2020

Keywords:
Convolutional neural network
Multi-atlas segmentation
Cortical sulci labeling

ABSTRACT

The extreme variability of the folding pattern of the human cortex makes the recognition of cortical sulci, both automatic and manual, particularly challenging. Reliable identification of the human cortical sulci in its entirety, is extremely difficult and is practiced by only a few experts. Moreover, these sulci correspond to more than a hundred different structures, which makes manual labeling long and fastidious and therefore limits access to large labeled databases to train machine learning. Here, we seek to improve the current model proposed in the Morphologist toolbox, a widely used sulcus recognition toolbox included in the BrainVISA package. Two novel approaches are proposed: patch-based multi-atlas segmentation (MAS) techniques and convolutional neural network (CNN)-based approaches. Both are currently applied for anatomical segmentations because they embed much better representations of inter-subject variability than approaches based on a single template atlas. However, these methods typically focus on voxel-wise labeling, disregarding certain geometrical and topological properties of interest for sulcus morphometry. Therefore, we propose to refine these approaches with domain specific bottom-up geometric constraints provided by the Morphologist toolbox. These constraints are utilized to provide a single sulcus label to each topologically elementary fold, the building blocks of the pattern recognition problem. To eliminate the shortcomings associated with the Morphologist's pre-segmentation into elementary folds, we complement this regularization scheme using a top-down perspective which triggers an additional cleavage of the elementary folds when required. All the newly proposed models outperform the current Morphologist model, the most efficient being a CNN U-Net-based approach which carries out sulcus recognition within a few seconds.

1. Introduction

The surface of the brain is divided into many convolutions, called gyri, delimited by folds, called sulci. The main sulci are considered as the limits between functionally and architecturally different regions. Additionally, cortex morphometry is used to quantify brain development and degenerative diseases. Despite the many tools available for 3D visualization of sulci, sulci labeling is a long and fastidious process. It takes several hours for an expert to label all sulci in a single brain and reliable labeling requires the opinion of several experts. However, because of the large variability of the folding pattern in the general population, inferring developmental biomarkers requires the mining of data from a large number of brains. These biomarkers may correspond to characteristics of the sulci, such as size, depth or opening. However, these measures require the prior labeling of sulci. Therefore, automation of the sulcus recognition is essential.

Nevertheless, learning to label sulci is an extremely complex challenge for several reasons. First, as illustrated in Fig. 1, sulci are highly variable structures, some sulci are even absent in more than 70% of brains and some subjects have up to 8 sulci missing. Additionally, each brain contains more than 120 different sulci and only a small number of segmentation algorithms are made for as many structures. Finally, the number of manually labeled subjects which can be used for supervised learning is limited.

* Corresponding author.
E-mail address: leonie.borne@gmail.com (L. Borne).

https://doi.org/10.1016/j.media.2020.101651

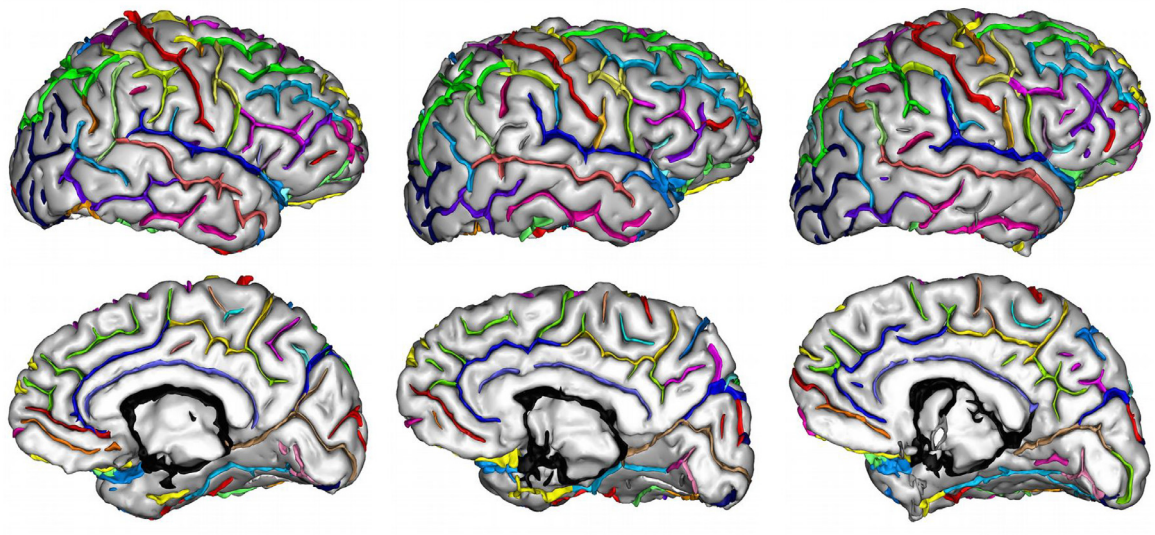

Fig. 1. Illustration of cortical folds variability. The manual labeling of the three right hemispheres represented here shows the variability of cortical sulci by their shape, size, and position.

1.1. Overview of automatic sulci recognition methods

Algorithms dedicated to automatic sulci recognition are primarily based on graphical representations, which represents the relative positions of the sulci with respect to each other, as well as their position and their location in a standardized space (Royackkers et al., 1998; Riviere et al., 2002; Vivodtzev et al., 2006; Shi et al., 2007; Yang and Kruggel, 2009; Belaggoune et al., 2014). To ensure their robust recognition, other methods have previously been experimented with to model inter-subject variability using several frameworks ranging from principal component analysis to Bayesian approaches (Lohmann and von Cramon, 2000; Behnke et al., 2003; Fischl et al., 2004; Perrot et al., 2011). All of these methods are based on a segmentation algorithm followed by a classification algorithm, in which the sulci are first extracted, according to different representations, then labeled.

In this paper, the objective is to improve the model proposed in the BrainVISA/Morphologist package (Perrot et al., 2011). To do this, we focused on two aspects of the pipeline: on the one hand, the sulci labeling algorithm and, on the other hand, the regularization of the results. Note that we did not try to improve the sulci extraction algorithm.

1.2. New sulci labeling approaches: MAS and CNN

Currently, the sulci labeling model proposed in the BrainVISA/Morphologist package, referred as the Statistical Probabilistic Anatomy Map (SPAM) model in this paper, is based on a Bayesian approach. As this labeling model has shown significant weaknesses, we have been inspired by two segmentation approaches for biomedical applications that are among the most widely used today, multi-atlas segmentation (MAS) and convolutional neural networks (CNNs).

MAS techniques, initially introduced by Rohlfing et al. (2004), use each manually segmented image as an atlas: the atlases are adjusted to the image to be segmented and the best matches are used to participate in the segmentation. Thus, MAS techniques make it possible to more accurately represent anatomical variability by not attempting to model a segmentation problem using an average model. These techniques are now widely used, but have a major disadvantage: the registration of the atlases to the images is particularly expensive.

Among the many variations of these techniques, the patch-based approach introduced by Coupé et al. (2011) and Rousseau et al. (2011) have particularly attracted our attention. By using a patch-based search strategy to identify matches with the atlases, the image no longer needs to be aligned globally with all the atlases via expensive non-linear registration. Thus, the registration and selection of matching patches can be particularly accelerated thanks to the Optimized PatchMatch algorithm proposed by Ta et al. (2014). This algorithm is an adaptation to segmentation of 3D images of the PatchMatch algorithm (Barnes et al., 2009) that aims to assign to each patch of an image, a patch similar to it in another image.

Inspired by these approaches, we propose two algorithms for cortical sulci recognition. The first is directly inspired by Romero et al. (2017), that proposes a cerebellum lobule segmentation method using an approach similar to the one originally proposed by Coupé et al. (2011); Rousseau et al. (2011) with some improvements. In the second algorithm, we propose a new patch generation strategy based on a high level representation of the sulci, as the standard way of extracting cubic patches does not seem capable optimally exploiting the sulci geometry and the relations between them, which we believe to be the discriminative features for their recognition. These two algorithms will be designated respectively by PMAS (for Patch-based MAS) and HPMAS (for Patch-based MAS with High level representation of the data).

The CNNs were initially developed to address problems in image classification and are now renowned for their formidable effectiveness in dealing with numerous computer vision problems. These techniques allow effective image analysis by learning an abstract representation of the image. Concerning segmentation problems, the first approach was proposed approximately ten years ago by Ciresan et al. (2012) where a neural network was trained to classify each voxel of the image to be segmented from its surrounding patch. Since then, new approaches allow the entire image segmentation using fully convolutional neural networks, such as the one initially proposed by Long et al. (2015) and dedicated to semantic segmentation. Concerning segmentation problems in medical imaging, the most commonly used architecture is the U-Net, a fully convolutional neural network which was initially proposed by Ronneberger et al. (2015) and whose adaptation to 3D images was proposed in (Çiçek et al., 2016; Milletari et al., 2016). Here, we propose to compare two approaches based on CNNs. The

first is inspired by Ciresan et al. (2012), adapted to address problems associated with 3D imaging. The second uses the 3D U-Net architecture proposed in (Çiçek et al., 2016). These two approaches will be called PCNN (for Patch-based CNN) and UNET, respectively.

To the best of our knowledge, despite their current popularity, no MAS or CNN-based approach has yet been proposed for cortical sulci recognition. Note that these two approaches are generally used to segment the entire image while in this study only the presegmented folds need to be labeled, requiring several adjustments in the proposed models.

1.3. Bottom-up geometric constraints

There is no guarantee that the geometric definition of a sulcus, as a set of topologically simple surfaces, is respected in the case of MAS and CNN-based methods described above. This is particularly disadvantageous for morphometric studies whose measurements are based on the definition of sulci. To remedy this, the BrainVISA/Morphologist pipeline provides an algorithm for bottom-up aggregation of voxels into elementary folds, which are the geometric building blocks of the problem. Once the voxels have been labeled by one of the methods proposed above, it is possible to regularize the results at the scale of the elementary folds. However, the upstream extraction of the elementary folds may sometimes be inaccurate. Although from the same MRI, vastly different fragmentations can be obtained because of stochastic optimizations embedded in the pipeline. This was previously a problem in the model proposed in (Perrot et al., 2011), which uses the same geometric entities to perform recognition, but is not capable of automatically re-dividing the elementary folds.

In this paper, we propose to use voxel-wise labeling to give a top-down perspective to a traditional bottom-up pattern recognition system. Thus, the initial cutting into elementary folds proposed by BrainVISA/Morphologist is challenged by voxel-wise labeling, eliminating under-segmentation errors in the model. The proposed approach is particularly robust to the spatial inconsistencies that can occur during voxel labeling and to the potential incorrect definition of upstream geometric entities.

2. Database

The training base is composed of 62 healthy brains selected from different heterogeneous databases and labeled with a model containing 63 sulci for the right hemisphere and 64 for the left hemisphere. The “unknown” label is used to designate unidentified structures (usually small sulci). The two ventricles are labeled but not considered as sulci. Most of the subjects are right-handed men, aged 25 to 35 years old.

Unfortunately, there is no gold standard definition of sulci morphology. Even the boundaries of the well-known central sulcus can be difficult to define (Fig. 2). Moreover, Fig. 2 shows that the definition of sulci morphology impacts the level of granularity of the nomenclature. Therefore, for this study, the elementary folds of each brain were manually labeled according to a sulcus nomenclature following a long iterative process to achieve a consensus across a panel of several experts on cortex morphology. The last iteration of the database labeling was performed using the TileViz visualization tool (Mancip et al., 2018). This tool allows the entire database to be visualized and labeled simultaneously on a wall of screens (See Fig. 19 in supplementary material). Until now it was only possible to label and simultaneously evaluate a limited number of hemispheres, generally four, on a standard screen. Thus, this tool helps to limit the bias of labeling induced by a restricted view of the database. To support this new iteration, the elementary folds were manually cut when necessary, which was not possible during the study of Perrot et al. (2011).

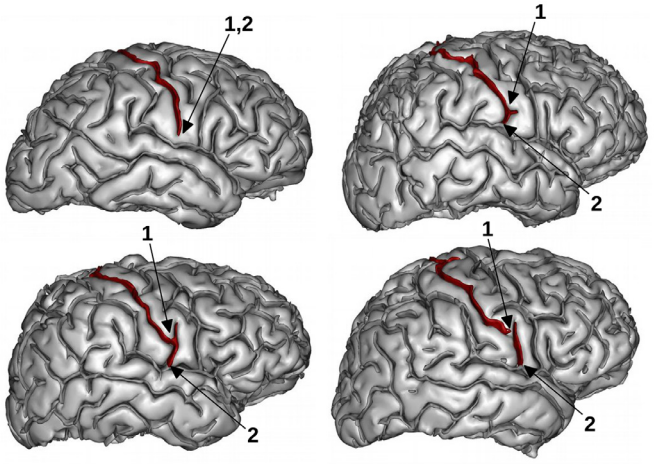

Fig. 2. Where should the central sulcus end? The folds that may belong to the central sulcus are shown in red. Limits 1 or 2 can be chosen according to the morphological definition of the central sulcus used. Note that depending on the definition chosen, the question then arises of adding a label to the nomenclature to identify the sulcus located between boundaries 1 and 2. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Note that compared to traditional labeling approaches where only one expert can label images, this database has been progressively labeled by several experts, both successively and simultaneously. This consensus-based labeling has sometimes led to the introduction of new sulci labels when it was considered necessary, making it essential to use the video wall. However, the different experts have thus not produced independent labelings, which prevents us from assessing human-level performance on this dataset.

Compared to (Perrot et al., 2011), the same MRI acquisitions were used but a new iteration of labeling was performed, resulting in the introduction of four new sulci in the nomenclature used. The new nomenclature is described in the Fig. 3. A more detailed description is provided in the Fig. 23 of the supplementary materials subsection. The manually labeled database is now available on the BrainVISA website (http://brainvisa.info/data/sulci_database/base_62/2019).

3. Method

The Morphologist/BrainVISA pipeline presented in (Perrot et al., 2011) has two major deficiencies. First, the SPAM model of sulci labeling makes obvious labeling errors that are problematic in practice. Typically, it tends to duplicate the central sulcus, which is an aberration. Then, the model uses bottom-up geometric constraints to group the voxels to be labeled in elementary folds, and this step is subject to errors. In this article, we therefore seek to improve the performance of the sulci labeling model and its robustness to sub-segmentation errors in elementary folds.

In this section, sulci labeling from an MRI is described in three steps (Fig. 4). First, the folds are segmented from the MRI using the BrainVISA/Morphologist pipeline (3.1.). Then, they are labeled using different algorithms (3.2.). Finally, the agglomeration of the voxels into elementary folds proposed by the BrainVISA/Morphologist pipeline is used to regularize the results (3.3.).

Note that the strategies used to set the method hyperparameters are detailed in the supplementary material.

3.1. Folds representation

The Morphologist pipeline of the BrainVISA software (www.brainvisa.info), a widely used resource for studying cortical

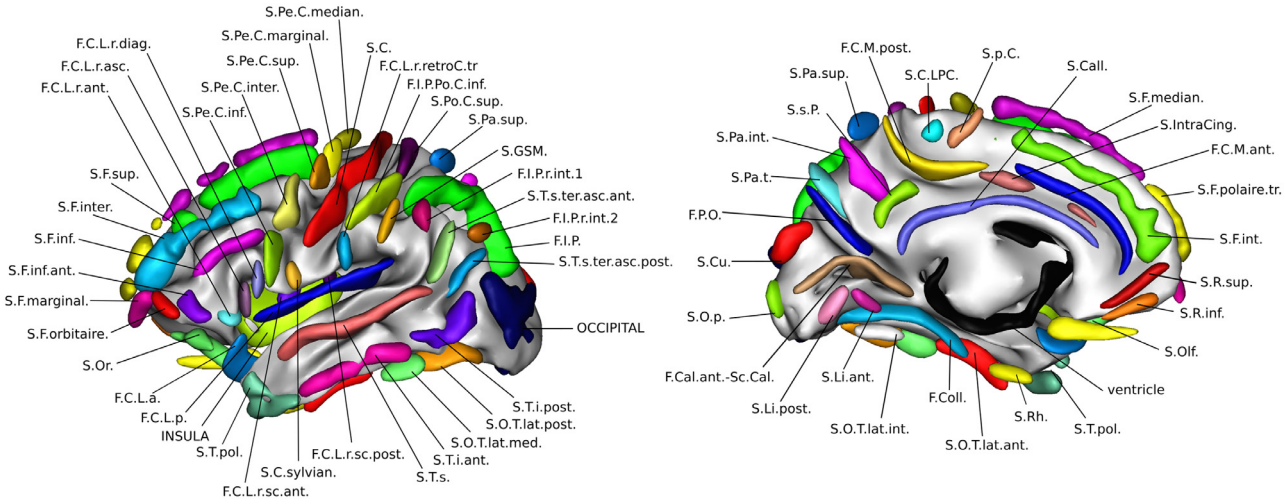

Fig. 3. New nomenclature used to label sulci. The visualization of the sulci labels is done thanks to the SPAM representation used by Perrot et al. (2011) which averages the position of the sulci as probability maps that are thresholded for this image. The new nomenclature includes 63 labels for the right hemisphere and 64 for the left hemisphere. Only the left hemisphere is represented in this figure. The right hemisphere has the same labels except the S.GSM. label. Compared to Perrot et al. (2011), two new sulci are labeled (S.intraCing. and S.R.sup.). The ventricle label does not correspond to a sulcus label, but belongs to the fold skeleton extracted by the BrainVISA/Morphologist toolbox. Only the "unknown" label is not shown in this figure. Please refer to the Fig. 23 of the supplementary material section for English translations of each label.

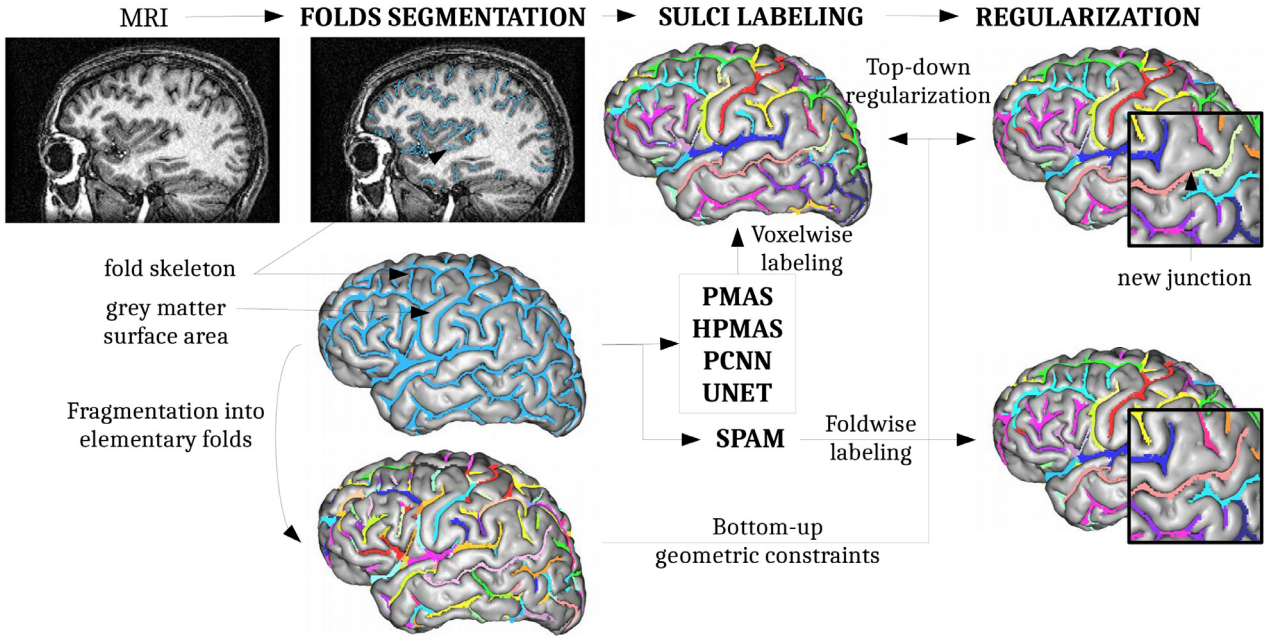

Fig. 4. MRI to labeled cortical sulci: a three-step pipeline. First, the fold skeleton is extracted using the BrainVISA/Morphologist toolbox. This toolbox also makes it possible to fragment the skeleton into elementary folds. Second, skeleton voxels are labeled by different algorithms. Algorithms based on MAS techniques (PMAS, HPMAS) and CNN-based algorithms (PCNN, UNET) label each skeleton voxel while the SPAM algorithm directly labels the elementary folds. Third, voxelwise labeling is regularized through the elementary folds, while automatically re-dividing them when the labeling indicates that it is required.

anatomy, allows first to represent the folds as a set of voxels corresponding to a skeleton of the cerebrospinal fluid filling the fold and then to label them using the SPAM model (Fig. 5). This first step of fold segmentation is common to all the models presented in this article. It consists of three major steps: first, the segmentation of white and grey matter from MRI, then the extraction of the skeleton of cortical folds, followed by its division into elementary folds. The fragmentation into elementary folds satisfies topological and geometric constraints specific to the sulci's definition. It is first based on the topological characterization of a simple surface proposed by Malandain et al. (1993) which isolates surface pieces that do not include any junction. The skeleton is also fragmented at the level of the buried gyri (Fig. 6).

The skeleton representation has three main advantages. First, this 3D representation is essential during manual labeling because it allows the visualization of the relative position of the sulci between each other and the evaluation of their depth, size, etc. Additionally, the agglomeration of the voxels into elementary folds makes it possible to speed up labeling by giving a label to a set of voxels rather than individually. Second, as the data are particularly influenced by the type of MRI sequence, the age of the subjects (which has a significant impact on the opening of the sulci) or even their pathologies, this pre-processing enables optimal normalization of the data. Moreover, the algorithm can focus on labeling only after its segmentation. Finally, this representation has previously been used in other pipelines, making it possible to

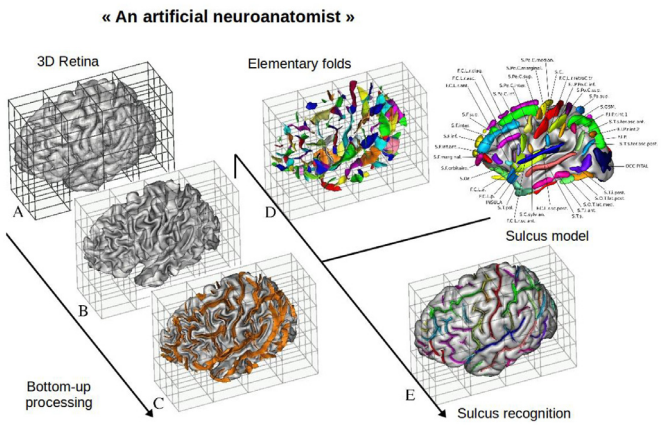

Fig. 5. A computer vision pipeline mimicking a human anatomist (Mangin et al., 2015). A: interface between the cerebral envelope and the cortex. B: interface between white matter and grey matter. C: extraction of the fold skeleton. D: cutting of the skeleton into elementary folds. E: Folds labeling using the SPAM model of Perrot et al. (2011)

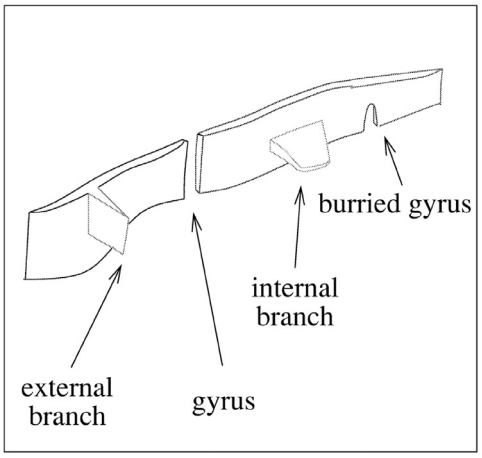

Fig. 6. Schematic representation of the fold skeleton. The fragmentation into elementary folds isolates the internal and external branches and cuts the skeleton at the level of the buried gyri. Image taken from Riviere et al. (2002).

automate the calculation of measurements (depth, length, connectivity, etc.) used in morphometric studies or to realign the brains according to the major sulci (Auzias et al., 2011; 2013), which is why we have chosen to keep it. However, if we had chosen to construct a model to recognize the sulci, that carries out both their extraction and labeling without relying on this representation, it was highly probable that the results obtained would not conform to the representation used by these pipelines and that some significant post processing steps would be necessary.

Although the extraction of the fold skeleton is robust, its fragmentation into elementary folds demonstrates certain significant instabilities, such as vastly different fragmentations can be observed from the same MRI (Fig. 7). Several stochastic optimizations were included in the segmentation pipeline (e. g. for bias correction, brain masking, skeletonization, etc.). These optimizations only have a slight impact on the shape of the resulting fold skeleton. However, for the topological fragmentation into elementary folds, a single voxel can then make the difference. Thus, these stochastic optimizations can have important consequences on the fragmentation of large simple surfaces. To remedy this, during manual labeling, the folds were cut manually when necessary. During automatic labeling, we propose a technique, based on a clustering algorithm, to automatically redivide the elementary folds from a voxelwise labeling during the regularization step. This technique is described in Section 3.3.

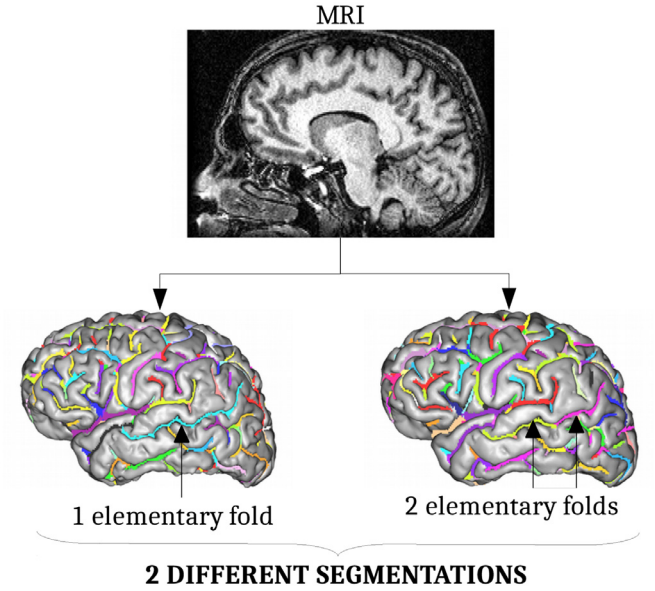

Fig. 7. Extraction of the elementary folds from the same MRI. In the two lower brains, each color represents a different elementary fold. We observed that the skeleton extraction is visually stable, but its division into elementary folds can produce very different results. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.2. Labeling methods

The methods described below seek to automatically label the voxels of the fold skeleton. Among the possible labels, while most correspond to cortical sulci, three other labels are used: those corresponding to the right and left ventricles and the “unknown” label. According to the methods presented here, the ventricles are treated as sulci, as they are relatively stable anatomical structures of the brain negative mold. However, the “unknown” label, corresponding to voxels that do not belong to any of the other labeled structures, must be treated differently in some cases.

3.2.1. Statistical probabilistic anatomy map (SPAM) models

In this comparative study, the reference method corresponds to the one described in (Perrot et al., 2011), where they propose a coherent Bayesian framework to automatically identify sulci based on a probabilistic atlas (a mixture of SPAM models) estimating simultaneously normalization parameters. This method, currently available in the BrainVISA/Morphologist pipeline, has been widely used on very large databases for large-scale morphometric studies (Le Guen et al., 2019). However, the model is still making obvious errors and we believe that this is due to the fact that the SPAM approach is based on a single template atlas, which prevents it from fully representing the high variability of folding patterns. Each sulcus can have several configurations, which may prove difficult to represent with a single average model.

3.2.2. MAS approaches

Two MAS approaches, PMAS and HPMAS, are compared in this section. The first approach is largely inspired by the one proposed in (Romero et al., 2017) in which, unlike most MAS approaches, similar atlases are searched between two cubic patches, instead of two full images. The second MAS algorithm presented here, and described in Borne et al. (2018), aims to define a library of local patches embedding enough geometrical information to minimize

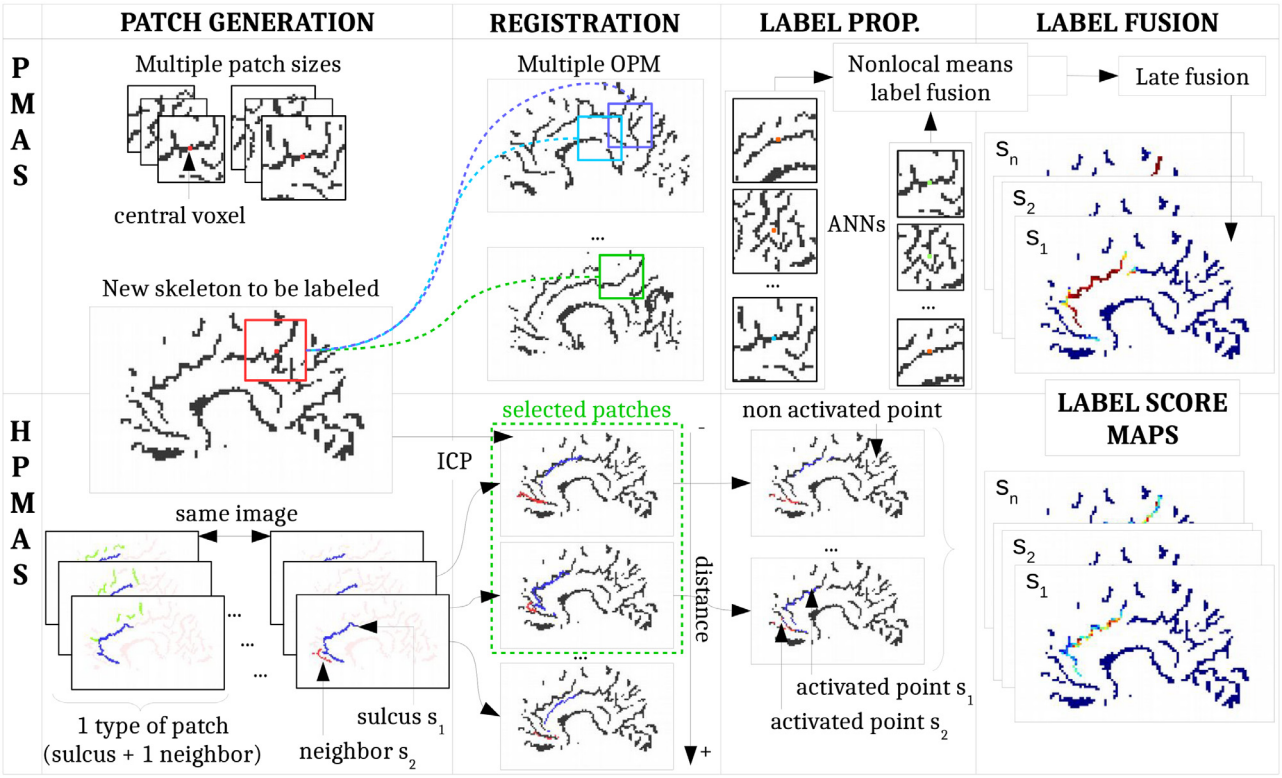

Fig. 8. Comparison of MAS approaches: PMAS vs. HPMAS. First of all, the patches are designed. Second, they are transferred to a new image to be labeled, where the fold skeleton has been extracted. Third, the best matches were selected and patch labels were propagated on the image to be labeled. Finally, the propagated labels are used to calculate the label score maps. In order to make the figures as readable as possible, we have chosen to represent the images in 2D while they are processed in 3D. All images are represented in 2*2*2 mm resolution, while for HPMAS, images are processed with the acquisition resolution. The acronym ANNs refers to the Approximate Nearest Neighbors patches obtained by the multiple run of the Optimized PatchMatch (OPM) algorithm.

ambiguities when searching for a high similarity hit in the unknown subject morphology. Therefore, instead of taking native cubic patches, this algorithm builds virtual patches containing whole sulci.

These two approaches are described in four steps: first, the design of the patches (patch generation), second, the strategy of realigning the patches between them and selecting the best matches (distance calculation), third, the strategy of propagating the labels from the patch to the brain to be labeled (label propagation) and finally the combination of the labels of the propagated patches (label fusion) (Fig. 8).

Patch-based MAS approach (PMAS)

Patch generation. The patches are cubes containing the fold skeleton. They are extracted from images with a resolution of 2*2*2 mm, that has been automatically relocated thanks to the BrainVISA/Morphologist pipeline in the well-known MNI space (Collins et al., 1994), which aligns the rough shapes of the brains through an affine transformation. We chose to harmonize the resolution of the images at 2*2*2mm, because it seemed sufficient to us to visually recognize the sulci.

We chose to take into account only the patches with the central voxel belonging to the fold skeleton for two main reasons. First, it limits the number of patch matches that require optimization as the voxels belonging to the skeleton represent only a small part of the image's voxels. Second, since the patches are extracted from binarized images, the calculation of the distance between two patches can be successful only if the patches contain a minimum number of skeleton voxels.

As proposed in (Giraud et al., 2016), we adopted a multi-scale approach, which involves the independent use of several patch sizes (determined by inner cross validation), to produce several score maps per label, which are then averaged.

Distance calculation. In order to find the most similar set of patches, we aimed to optimize the following distance d between two patches $P(S_A)$ and $P(S_B)$, respectively belonging to the fold skeletons S_A and S_B (superimposed by a simple translation):

$$d(P(S_A), P(S_B)) = \frac{d(P(S_A) \rightarrow S_B) + d(P(S_B) \rightarrow S_A)}{2} \quad (1)$$

The measurement from a patch $P(S_A)$ to a fold skeleton S_B corresponds to the average of quadratic Euclidean distances d_E of the skeleton voxels $p_A \in P(S_A)$ and their nearest neighbor in the fold skeleton S_B (Fig. 9):

$$d(P(S_A) \rightarrow S_B) = \frac{1}{|P(S_A)|} \sum_{p_A \in P(S_A)} \min_{p_B \in S_B} [d_E^2(p_A, p_B)] \quad (2)$$

Note that, in order to avoid border effects, the closest neighbor of p_A is searched in the entire skeleton S_B and not only among the skeleton voxels contained in the patch $P(S_B)$.

Realigning and comparing all the patches in the database for each skeleton voxel to be labeled would be extremely expensive, making it impossible to label within a reasonable time. Additionally, it would increase the probability of spurious matching between remote areas in the brain while the images are already roughly aligned with each other. It is important to note that because we use binarized images, the risk of obtaining false positives is higher than usual.

In (Romero et al., 2017), the Optimized Patch Match Label fusion (OPAL) (Ta et al., 2014; Giraud et al., 2016) was used. This segmentation method is based on the Optimized PatchMatch (OPM) algorithm which uses a cooperative and random strategy resulting in a very low computational burden. Compared to the PatchMatch algorithm (Barnes et al., 2009) from which it is inspired, OPM is adapted to 3D anatomical segmentation by taking into account the

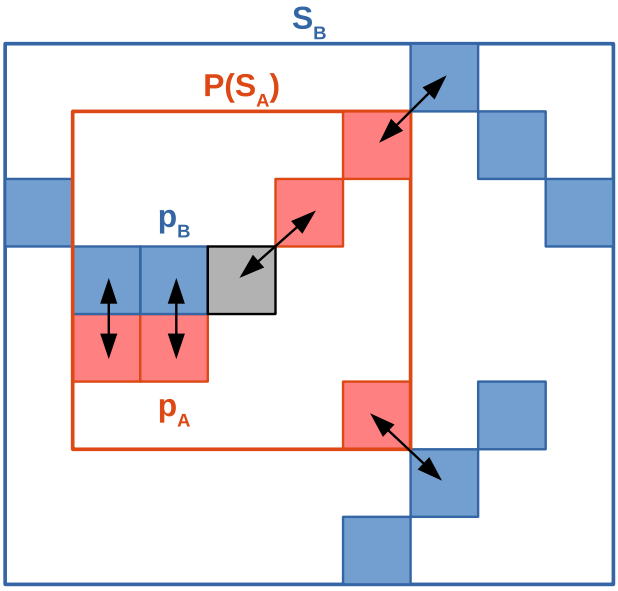

Fig. 9. Calculation of the distance from the patch $P(S_A)$ to the skeleton S_B for the PMAS method. The grey voxel represents the central voxel of the patch $P(S_A)$ which is superposed with a voxel of the skeleton S_B. For each voxel $p_A \in P(S_A)$, we look for its closest neighbor among the voxels of the skeleton S_B. The Euclidean distance between these two voxels is calculated. The distances over all the points $p_A \in P(S_A)$ and their nearest neighbors are then averaged to obtain $d(P(S_A) \rightarrow S_B)$.

rough alignment of images. Here, as only patches with the central voxel belonging to the fold skeleton are considered, an adapted version of the OPM algorithm has been implemented. Please refer to supplementary material for more details.

Label propagation. In order to select several Approximate Nearest Neighbors (ANNs) patch per skeleton voxel for a given patch size, multiple independent OPM were launched. The number of ANNs to be selected is determined by inner cross-validation. Once the ANNs have been selected, all the voxels of each ANN patch participates in the labeling, as done in (Rousseau et al., 2011; Giraud et al., 2016). However, there are only a few voxels belonging to the skeleton of the patch that overlap with the skeleton voxels to be labeled. Thus, we propose to propagate the label of each skeleton voxel of the patch to its nearest neighbor in the skeleton to be labeled.

Label fusion. For this method, we have implemented the non-local patch-based label fusion used in (Romero et al., 2017). In this strategy, the distance between patches is used to perform a robust weighted average of the labels. The label fusion strategy corresponds to the multipoint estimation described in (Rousseau et al., 2011). Once the non-local means estimator has been calculated for all patch sizes, the final estimation is obtained by averaging these estimations thanks to a late fusion (Snoek et al., 2005). Thus, a score map is estimated for each label in the database.

Concerning the "unknown" label, present in the manually labeled database, it is treated like a sulcus label.

Patch-based MAS approach with High level representation of the data (HPMAS)

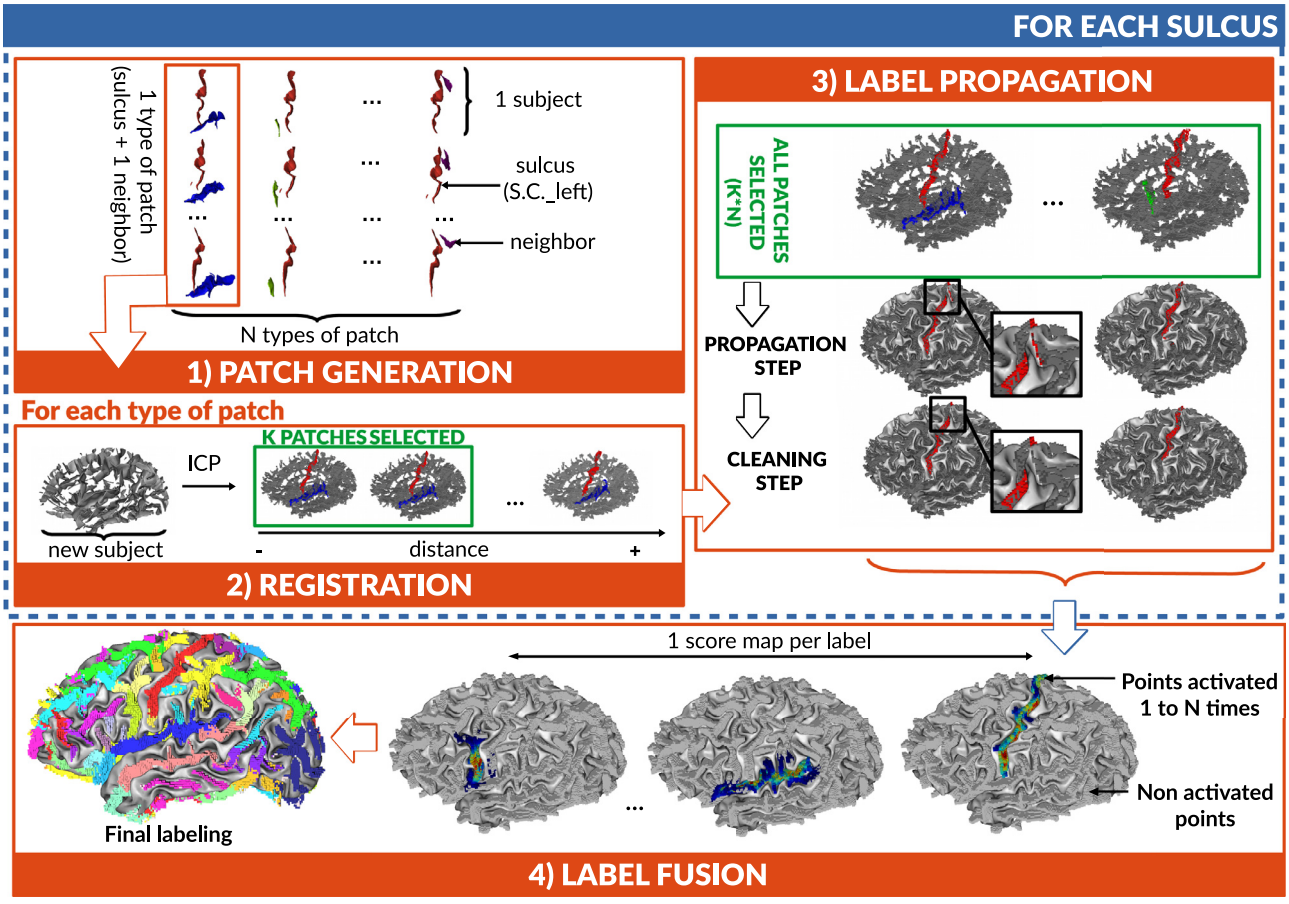

Fig. 10. 3D representation of the HPMAS method. As for the Fig. 8 which represents the method in 2D, the approach is described in four steps: generating the virtual patches, registering them on the image to be labeled, propagating the labels of the selected virtual patches and finally merging the propagated labels to obtain the final labeling.

As the standard way of extracting patches does not seem capable of exploiting the sulci geometry and the relations between them, which we believe to be the distinguishing features necessary for recognition, we have proposed a new virtual patch generation strategy based on a high level representation of the sulci (Borne et al., 2018). This framework is well adapted to leverage more information about the different folding configurations in the training dataset.

Note that this method is the only one of the proposed new methods to have been specifically developed for the recognition of cortical sulci. It includes many arrangements specific to this application. Its complex design gives an idea of the scores that can be obtained by pushing as far as possible in this direction. To facilitate the understanding of this ad-hoc method, Fig. 10 represents the pipeline in 3D, which complements the 2D representation provided in Fig. 8.

Patch generation. In order to take into account as much geometric information as possible, the idea was to define virtual patches containing whole sulci. These virtual patches correspond to a voxel cloud representing a pair of sulci, extracted from MNI space at the image resolution. By defining patches as clouds of voxels and not as cubes, it allows to take into account the sulcus in its entirety without parasitizing the patch with all its surrounding sulci. Note that the shape of small sulci is not specific enough to prevent spurious hits. That is why we have chosen to aggregate two sulci to create discriminative local shapes. In the following, we define a type of virtual patches for each pair of sulci that are neighbors in the brain.

In practice, a pair of sulci is selected in the circumstance that the two sulci are neighbors in at least one brain of the atlas dataset, according to the topology provided by the BrainVISA/Morphologist pipeline that produces the folds. This pipeline endows the list of folds with a graph structure corresponding to either direct connections or to the fact that two folds are separated by a piece of gyrus. Finally, each type is made up of the instances of the pair of sulci in the atlas dataset, most of the time as many shapes as atlases (some atlases miss a few small sulci) (Fig. 10.1).

Note that only the unknown sulcus label is not selected to form virtual patches, as it does not constitute a coherent structure like the other labels. Thus, unlike the previous PMAS method, the unknown label is not treated like other sulcus labels.

Distance calculation. For the distance calculation step, the set of folds of the brain to segment and the virtual patches of the library are represented by point clouds. In order to find an optimal alignment of each virtual patch into the skeleton point cloud of the brain to segment, the well-known iterative closest points algorithm (Besl and McKay, 1992) is used, with the robust implementation of Holz et al. (2015). This algorithm iteratively adjusts the transformations (translation and rotation) in order to minimize the distance between two set of points. Note that compared to the PMAS approach which only uses translations to superimpose patches, the registration here allows rotations.

To build the measure used to rank the matches, the nearest voxels in the new fold skeleton S_B of each skeleton voxel $p_A \in P(S_A)$ are saved as activated voxels $p_B^* \in S^*_{B,P(S_A)}$. Then, the measure corresponds to the sum of the quadratic distances of the skeleton voxels and their corresponding activated voxels, divided by the number of different activated voxels:

$$d(P(S_A) \rightarrow S_B) = \frac{1}{|S^*_{B,P(S_A)}|} \sum_{p_A \in P(S_A)} \min_{p_B \in S_B} [d_E^2(p_A, p_B)] \tag{3}$$

Note that by dividing by $|S^*_{B,P(S_A)}|$, we take into account the number of different activated points. This allows the penalization of virtual patches where several points activate the same point of the skeleton to be labeled (Fig. 11).

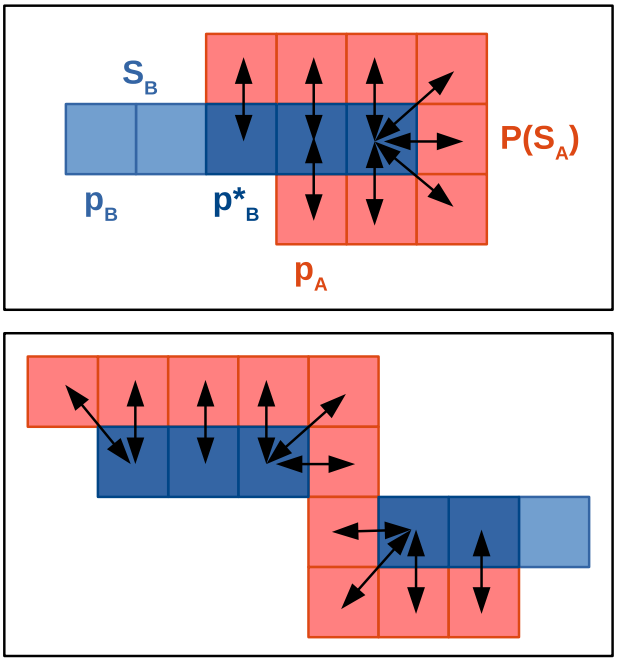

Fig. 11. Calculation of the distance from the virtual patch $P(S_A)$ to the skeleton S_B for the HPMAS method. For the sake of clarity, the skeletons S_A and S_B represented do not overlap in this Figure. For each voxel $p_A \in P(S_A)$, we look for its closest neighbor among the voxels of the skeleton S_B. The Euclidean distance between these two voxels is calculated. The distances over all the points $p_A \in P(S_A)$ and their nearest neighbor are then summed and divided by the number of different activated points p_B^* to obtain $d(P(S_A) \rightarrow S_B)$. The two configurations represented are penalized by the division by the number of different activated points rather than by the number of points in $P(S_A)$ as for a classical average. On the first configuration, we observe that the proposed distance penalizes the virtual patch more if its shape is more complex or if its size is larger than the structure on which it has been registered. On the second configuration, we observe a greater penalization of the virtual patch if it has only one connected component and if it is registered on two different components.

With regards to each type of virtual patch, all matches are ranked according to the distance proposed above. A fixed number of matches (determined by inner cross-validation) leading to the shortest distances is selected to propagate the two parent sulci. All types of virtual patches are selected the same number of times even if they are not all equally informative. It is important to note that some sulcus instances are selected several times, because they win the competition for several virtual patch types, but their multiple contributions will be associated with slightly different alignments. Hence, sulcus instances maximizing regional similarity to the unknown subject get more weight.

Label Propagation. Each selected virtual patch after the optimal alignment to the unknown subject, concomitantly propagates the label of each voxel to its nearest neighbor in the target brain. To consider the virtual patch structure, each connected set of voxels in the virtual patch should correspond to a unique connected set in the target brain: the smallest non-connected sets are excluded (Fig. 10.3).

Label Fusion. Post complete propagation of all the proposed virtual patches $p \in V_l$ that contain the sulcus l, the score map S_l is calculated by averaging the number of times the points of coordinates (x, y, z) are activated by different virtual patches:

$$S_l(x, y, z) = \frac{\sum_{p \in V_l} act_p(x, y, z)}{|V_l|} \tag{4}$$

with $act_p(x, y, z)$ equals to 1 if the voxel of coordinates (x, y, z) is activated by the patch p, and to 0 otherwise.

Compared to PMAS, where patches are weighted by their distance to the patch to be labeled, here each propagated point has

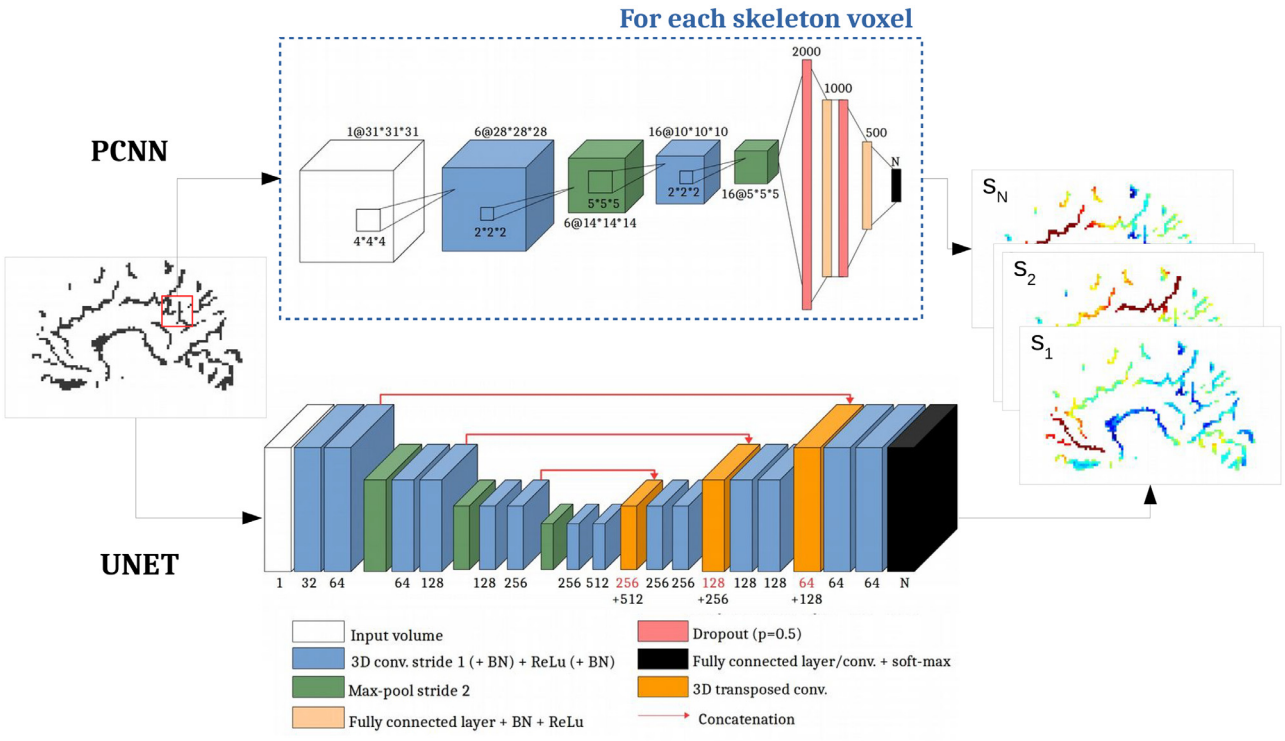

Fig. 12. Comparison of CNN-based approaches: PCNN vs. UNET. Boxes represent feature maps. The number of channels is denoted next to each feature map. The size of the feature map is indicated after the @ when appropriate. N is the number of different labels to be predicted. For clarity sake, input and output are represented in 2D rather than 3D.

the same weight in the label fusion. In order to perform a similar weighting, we have tested the use of distance from the entire virtual patch to the skeleton to be labeled. This did not seem to significantly improve the results. We also tried to weight by the distance from the virtual patch point to the point it has activated, without any further improvements. We think it is essential to combine these two distances when weighting, for example by averaging the two distances. However, our attempts have also been unsuccessful so far, so we chose to avoid weighting.

As the "unknown" label does not belong to any virtual patch, its score map is empty. This label will be selected only if the score maps of all other labels are also empty for a given elementary fold.

3.2.3. CNNs based approaches

As this is the first time that CNNs are used for sulci labeling, we take inspiration from two models that have proven their efficacy in medical image segmentation (Fig. 12): the first being a patch-based approach inspired by (Ciresan et al., 2012) and the second an approach that treats the entire image with a 3D U-Net as in (Çiçek et al., 2016). First the common modalities used during training of these two networks are detailed followed by an individual description of each network. The models presented are implemented using the Pytorch library (Paszke et al., 2017).

Data. All the fold skeletons are registered in MNI space and used as input: they correspond to 3D binary volumes with a common resolution of 2*2*2mm, where the voxels belonging to the skeleton are one and the others are zero. In order to augment the training dataset, a rotation in a random direction with a random angle (following a Gaussian distribution $\mathcal{N}(0, \frac{\pi}{16}^2)$) is applied to the images at each epoch.

At the output of the neural network, a score per label present in the database is obtained per voxel. Concerning the "unknown" label, it is treated like a sulcus label.

Training design. Initialization of the weights of the neural networks was done as in (LeCun et al., 2012). Stochastic gradient descent was used for training, with learning rate and momentum determined by 3-folds inner cross validation. The learning rate was halved when the loss function had not improved for two consecutive epochs. After four consecutive epochs without improvement, training was stopped. The selected trained neural network corresponds to the epoch obtaining the lower error rate E_{SI}, described in (Perrot et al., 2011) and in the following section.

The loss function used is the cross- entropy loss. In most cases, for unbalanced problems, the loss function must be weighted to avoid favoring the labels most involved in backpropagation, due to their higher presence in the database. Although the average size of each sulcus is extremely unbalanced, we have chosen not to weigh this loss function because large sulci are also the most interesting from a neuroanatomical point of view and need to be better recognized than small ones.

Patch-based model with a 3D CNN (PCNN) PCNN method adapts the approach proposed in (Ciresan et al., 2012), addressing a segmentation problem as a classification of each voxel based on its environment contained in a patch. Here, only voxels belonging to the skeleton are selected to participate in the classification.

We designed the architecture of the neural network so that it takes cubic patches of 6.2 cm side in input, which we considered to be large enough to identify its central voxel. During training, the dropout strategy (Srivastava et al., 2014) with a probability of 0.5 is used on fully connected layers. Batch normalization (Ioffe and Szegedy, 2015) was also used on convolutional and fully connected layers. The batch size has been set at 100 to minimize learning time and fit in memory. In order to ensure that the inner cross-validation is not too time-consuming, only three epochs are calculated for each hyperparameter value tested.

3D U-Net based model (UNET) For the UNET method, the network architecture used is the one presented in (Çiçek et al., 2016), with the Pytorch implementation of (Wolny and enfisan, 2019). The particularity of this application of U-Net lies in the fact that all the voxels that do not belong to the fold skeleton, i.e. a large majority

of the voxels in the image, do not need to be classified. Indeed, as the values predicted by U-Net are masked by the segmentation of the fold skeleton made upstream, the background voxels do not need to be predicted and therefore do not need to be learned. Thus, during training, all voxels that do not belong to sulci are not used for gradient backpropagation. The batch size has been set at 1 in order to fit in memory.

3.3. Bottom-up geometric constraints

In order to standardize the results, the voxels were agglomerated into elementary folds. However, these folds are not always sufficiently fragmented, so we propose to use the label score maps to reconsider their fragmentation.

The straightforward approach to regularize the results is to do a weighted majority vote. The scores of each elementary fold were averaged by label and the highest score label was selected. This strategy was used as a reference to evaluate the impact of the automatic re-division of elementary folds.

In this paper, we propose to re-divide the elementary folds with help of the Ward's hierarchical agglomerative clustering method (Ward Jr, 1963). Clustering for each elementary fold was performed based on the label score maps. In order to ensure spatial consistency, a spatial connectivity constraint was imposed during cluster agglomeration. Then, the Calinski-Harabasz index I_{CH} (Caliński and Harabasz, 1974), implemented in the scikit-learn library (Pedregosa et al., 2011), was used to quantify the quality of the proposed clustering. This score corresponds to the ratio of the between clusters dispersion mean B and the within cluster dispersion W:

$$I_{CH} = \frac{Tr(B)}{Tr(W)} * (N-2) \tag{5}$$

$$W = \sum_{k=1}^{2} \sum_{x \in C_k} (x - c_k)(x - c_k)^T \tag{6}$$

$$B = \sum_{k=1}^{2} n_k (c_k - c)(c_k - c)^T \tag{7}$$

with N be the number of voxels in the elementary fold E, C_k be the set of voxels in cluster k, c_k be the center of cluster k, c be the center of E, n_k be the number of points in cluster k.

The ratio was higher when clusters are dense and well separated. If this score was higher than a threshold determined by inner cross validation, the partitioning was performed. When an elementary fold was split in two, each of the two clusters obtained were also challenged with the same manipulation, until all the elementary folds had a Calinski-Harabasz index below the threshold.

3.4. Performance evaluation of labeling models

As in Perrot et al. (2011), two measures were used to compare the different models proposed above: E_{local} at the sulcus scale and E_{SI} at the subject scale. Error rates were assessed by 10-folds cross validation. One model was trained per hemisphere.

3.4.1. Mean/max error rates

To take into account the variability of the fragmentation into elementary folds and therefore the robustness of the labeling methods to this variability, each image was re-segmented ten times (See Fig. 20 in supplementary material). Thus, if the image belonged to the training set, only the segmentation used for manual labeling was considered. However, if the image belonged to the test set, ten other segmentations (whose true labels have been transferred from manual segmentation) were labeled and used to quantify the error rates. Note that manual segmentation was not used to calculate error rates. Using ten different segmentations for each sulcus highlights the weaknesses of the BrainVISA/Morphologist preprocessing since we can compute errors from the worst result, typically associated to an issue of under-segmentation.

To quantify errors, for each new segmentation, the manual labeling on the initial segmentation must be transferred to the new one. Because of the variability of the segmentations obtained and the sparsity of the fold skeleton, the simple superposition of images was insufficient. We have given to skeleton voxels that do not overlap with those of the initial segmentation, the label of the nearest skeleton voxel of the initial segmentation. To do this, a Voronoi diagram of the manual labeling is performed. Note that the elementary folds were not used to transfer the labeling and that the true labeling was on the voxel scale.

For each subject, from the ten segmentations, the average of the errors (E_{SI}^{mean} and E_{local}^{mean}) and the maximum error ($E_{SI}^{\max}$ and $E_{local}^{\max}$) were calculated. Note that the training segmentation used for manual labeling was not used in the error calculation because it would bias our evaluation. By considering the maximum error rates, labeling errors due to model variability were highlighted. These errors in most models were related to an incorrect fragmentation of the fold skeleton into elementary folds. Only the PMAS labeling model was not deterministic and includes stochastic optimizations that can penalize the calculation of maximum error rates.

3.4.2. Error at the sulcus scale: E_{local}

Given a sulcus l,

$$E_{local}(l) = \frac{FP_l + FN_l}{FP_l + FN_l + TP_l} \tag{8}$$

with TP_l, FP_l and FN_l, respectively the number of true positive, false positive and false negative voxels for the sulcus l.

It is important to note that the error rate was one, when the sulcus was absent and labeled by the model. Similarly, for when the sulcus was present but not labeled by the model. As small sulci are frequently absent, this explained why error rates can be highly variable when averaging the error rates per subject.

3.4.3. Error at the subject scale: E_{SI}

Given a set of sulci L,

$$E_{SI} = \sum_{l \in L} w_l * \frac{FP_l + FN_l}{FP_l + FN_l + 2 * TP_l} \tag{9}$$

with $w_l = s_l / \sum s_l$ and $s_l = FN_l + TP_l$, the sulcus l true size.

The error at the subject scale allows local errors to be generated in a single measurement. As explained in Perrot et al. (2011), each component of the sum over labels differs on two points compared to $E_{local}(l)$. First, true positive measures are counted twice as compared to the false positive and negative measures, in order to remove errors shared by several labels, since each extra sulcal piece for a given label is a missing part for another label. Second, each component was weighted according to the sulcus true size so that each local component count as much as its size.

Compared to Perrot et al. (2011), three labels were not included in the set of sulci (unknown and both ventricles). These labels were not particularly considered as sulcus labels, but correspond to other structures, not pertinent to our study. Thus, the scores presented here for the SPAM method are worse than presented in Perrot et al. (2011) for four reasons. First, because removing the two labels considerably improved the scores. Second, because we cut the elementary folds during manual labeling while the SPAM model cannot automatically correct this kind of sub-segmentation errors. Third, because we are interested in the mean/max of the error rates. Finally, because the error rates are estimated by 10-folds

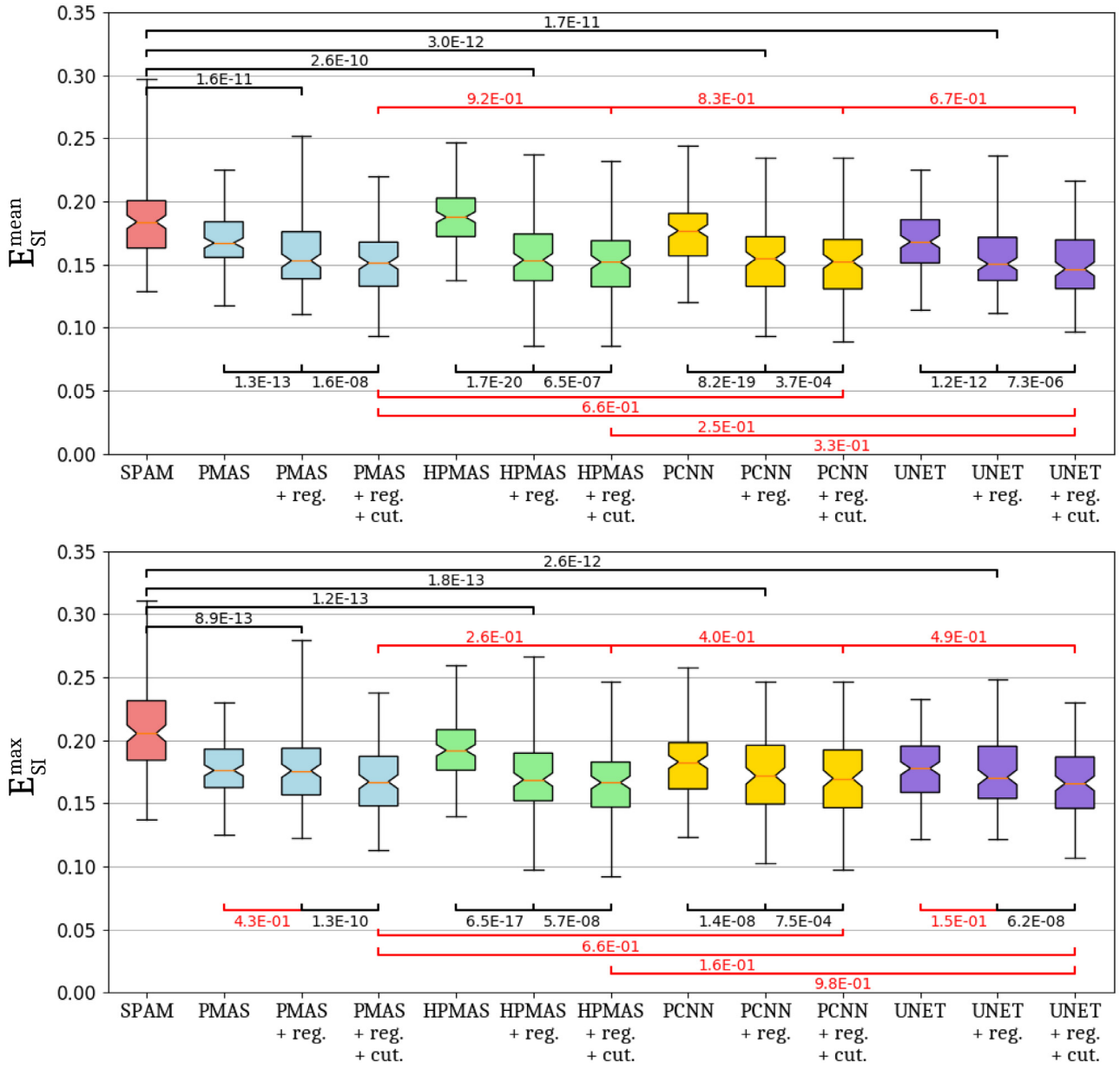

Fig. 13. Comparison of E_{SI} error rates by model. Once the 10 segmentations have been labeled by hemisphere, we consider the average error in the upper chart and the maximum error in the lower chart. The box extends from the lower to upper quartile error values, with a line at the median. The whiskers extend from the box to show the minimum and maximum limits of the error rates. The SPAM model is represented in red, the PMAS model in blue, the HPMAS model in green, the PCNN model in yellow and the UNET model in purple. For the four new models, three modalities are represented: first, labeling at the voxel scale, then labeling after regularization at the elementary fold scale (+ reg.), and finally the labeling obtained after automatic re-division of the elementary folds (+ reg. + cut.). The models are compared by Wilcoxon signed-rank test. The *p*-values of the differences in model performances are written above and below the compared models. The *p*-value is written in black if it is less than 0.05 and in red otherwise. Regularization by elementary folds significantly improves results. Automatic fold re-division also significantly improves results. All regularized models are significantly better than the SPAM model. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

cross-validation and not by leave-one-out cross-validation. Moreover, the addition of the four new sulci labels and our refined labeling of the training dataset may also have impacted the results.

3.4.4. Error rate comparison

During the 10-folds cross validation, each fold contained approximately 6 hemispheres labeled to test the model's performance. Error rates are calculated by hemisphere and then averaged over the entire database to obtain the mean error rates per model. When not specified, the average error rate includes the right and left hemispheres. In order to compare the models in pairs, a Wilcoxon signed-rank test was performed between the error rate lists for each hemisphere. If the *p*-value was less than 0.05, the error rates were considered significantly different.

4. Results

4.1. Which is the best model?

In order to compare the five models presented above, we were interested in the E_{SI}^{mean} and $E_{SI}^{\max}$ for each model, trained separately on each hemisphere (Fig. 13). Please refer to the supplementary materials for the numerical values of the error rates per hemisphere (Table 1).

First, we observed that all of the new approaches proposed with regularization per elementary folds were significantly better than the SPAM approach (also based on this regularization), which suggests that a model based on an average template was not the most appropriate to represent the high variability of cortical folds.

Second, with regards to the four proposed methods, regularization by elementary folds of the label score maps significantly improved the results compared to voxel labeling. Most importantly, the automatic re-division of these elementary folds also significantly improved the four methods. Thus, the use of top-down refinement of bottom-up regularization is particularly relevant in this paper.

Third, by comparing the new models in pairs, the models seem to demonstrate equivalent performance.

Concerning the PCNN and UNET models, this paper consequently demonstrated the incredible efficiency of neural networks, even for the recognition of structures as variable as cortical folds. However, it is surprising that the UNET model was not better than the PCNN model due to its deeper architecture.

The fact that these four models do not stand out radically on this dataset suggests that these models may have reached the limit of what can be interpreted from this database, probably due to its insufficient size to represent the high variability of cortical folds. Therefore, the fold variability is such that manual labeling of a brain raises many questions and it may be possible that the models have reached the human-level performances. Unfortunately, since manual labeling is based on consensus among several experts, it is impossible for us to assess human-level performance on this database.

Finally, with regards to the computation time required to label a hemisphere, the SPAM model takes about 5 min, while the UNET model takes about 20 s, PCNN takes slightly more than a minute, PMAS and HPMAS take several hours. Although the PMAS model could be much faster by optimizing the codes as in (Giraud et al., 2016), the UNET model is currently by far the fastest. Thus, since the UNET model has the lowest error rates and is the fastest, we propose to study in more detail the differences between this model and the SPAM model in the following section. In the rest of this study, the UNET model will therefore refer to the model with regularization using elementary folds and automatic redivision of these, if necessary.

4.2. Which sulci are better recognized?

Concerning E_{local}^{mean} and $E_{local}^{\max}$, the SPAM model has average/max error rates from 5% to 77% while the error rates of the UNET model vary between 2% and 68%. Comparing the $E_{local}^{\max}$ of each sulcus (Fig. 14), we can see that the difference between the error rates of both model for a given sulcus reaches up to 25%. Finally, almost all sulci were better recognized by the UNET model, only about twenty sulci are less well recognized. Their comparison with the Wilcoxon signed-rank test, by controlling the false discovery rate with the help of the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995), showed that around 13% of sulci were significantly better recognized by UNET than SPAM, while none were significantly less well recognized. In the figure, we can also see that the sulci with the highest labeling error rates using the UNET model are also the smallest. This is probably due to the fact that small sulci are generally also the most variable and were already less well recognized by the SPAM model. Please refer to the supplementary material for exact values of sulcus error rates (Tables 2 and 3). In order to visualize the location of the sulci better recognized than before, Figs. 15 and 16 give graphical comparisons of sulcus error rates between SPAM and UNET labeling. In Fig. 16, it can be seen that the differences in performance between the SPAM and UNET models are not spatially uniform. This may be due to the fact that some regions have more variable fold patterns than others and the recognition of their sulci was more severely penalized by the use of a mono-template approach. We also noted that the sulci best recognized by the UNET model are also those that were most impacted by sub-segmentation errors in elementary folds.

In the next section, we focus on the impact of the significant improvement in central sulcus recognition, in which the $E_{local}^{\max}$ value has gone from about 8% using the SPAM model to only 3% with the best UNET model.

4.3. Experiment on an external database demonstrating the clinical advantage

Here, the SPAM model and the UNET model were trained on the entire manually labeled database. The hyperparameters of the UNET model were estimated over the entire database, using the same procedures as during inner cross-validation, i.e. by performing a 3-folds cross-validation to select the hyperparameter values that minimize error rates. The database used by Sun et al. (2012) to study the effect of handedness on the shape of the central sulcus was labeled manually and automatically by these two models. This database contains 23 consistent age and sex matched natural dextrals (mean age 34, range 22-59 years; 17 males, 6 females) and 18 similar natural sinistrals (mean age 36, range 25-56 years; 12 males, 6 females). The database used in Sun et al. (2012) also contains a group of 34 forced dextrals that is not studied here.

We propose to investigate the asymmetry index I of the central sulcus length along the brain hull between the left $l_{S.C.\_left}$ and right hemispheres $l_{S.C.\_right}$:

$$I = \frac{l_{S.C.\_left} - l_{S.C.\_right}}{l_{S.C.\_left} + l_{S.C.\_right}} \tag{10}$$

Note that in the nomenclature proposed in this paper, two sulci labels belong to the central sulcus: “S.C.” and “S.C._sylvian.”. Therefore, the lengths of these two “sub-sulci” are added together to obtain $l_{S.C.}$.

With manual labeling, there is a significant difference between left-handed and right-handed people (Fig. 17). Therefore, left-handed people have on average a longer central sulcus in the right hemisphere than in the left, and vice versa for right-handed people. However, when focusing on the asymmetry index with SPAM labeling, no significant difference was found, whereas this difference was significant with UNET labeling.

Considering the worst labeling errors (See Fig. 21 in supplementary material) of each model, we observe that the SPAM model can double the size of the central sulcus, by labeling completely unrelated large structures. However, the UNET model only adds small fragments.

5. Discussion

5.1. PMAS

Considering the hyperparameters selected during the inner cross validation (See Fig. 22 in supplementary material), it seems that this method would benefit from increasing the number of ANNs selected by voxel. Indeed, the number of ANNs is automatically set to 10, which is the upper limit of the values proposed in the inner cross-validation. However, testing a larger number of ANNs would require optimization of the codes currently in use and it is very likely that the model would not gain much in performance. Indeed, the evolution of the scores according to the number of ANNs suggests that a plateau is reached and that increasing this hyperparameter would have little influence on the ranking of the methods obtained.

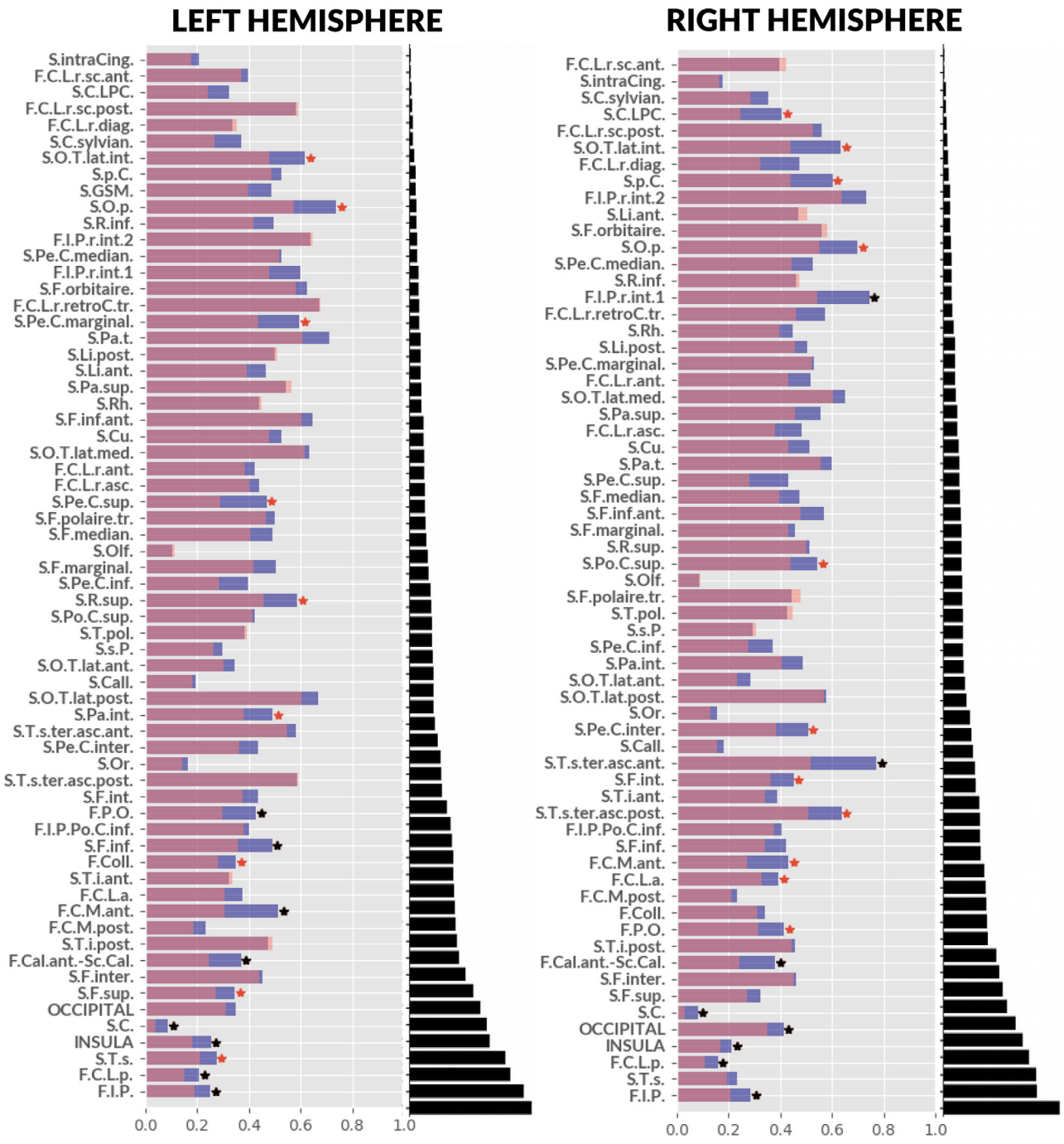

Fig. 14. E_{local}^{max} per sulcus. The graph on the left and the graph on the right present E_{local}^{max} for the sulci on the left hemisphere and on the right hemisphere, respectively. The SPAM model is represented in blue and the UNET (+ reg. + cut.) is represented in pink. The significant differences ($pvalue < 0.05$) are marked with a star. The star is black when the difference is still significant after controlling the false discovery rate through the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995). Sulci are sorted from top to bottom, from the smallest to the largest. The average sulci sizes, ranging from about 15 mm^3 to more than 2000 mm^3 on average per subject, are represented on the black graph.

5.2. HPMAS

With regard to HPMAS, the choice to use sulci pairs to form patches was questionable, since there was no evidence suggesting that two sulci are sufficient to prevent spurious hits, especially when two small sulci are associated. In order to create distinguishable local shapes, patches containing three or more sulci should also be considered. However, it would be too expensive to take into account all combinations of three neighboring sulci, as it is done for pairs of sulci. To remedy this, criteria for selecting relevant patch types should be determined, but none of the criteria we tested improves the results sufficiently to be considered here.

5.3. PCNN and UNET

Compared to the approach proposed by Ciresan et al. (2012), the PCNN approach has a major difference. In (Ciresan et al., 2012), several patch sizes, processed by several neural networks in parallel, were used to label each pixel, yet our PCNN approach is based on only one patch size. Moreover, the neural network

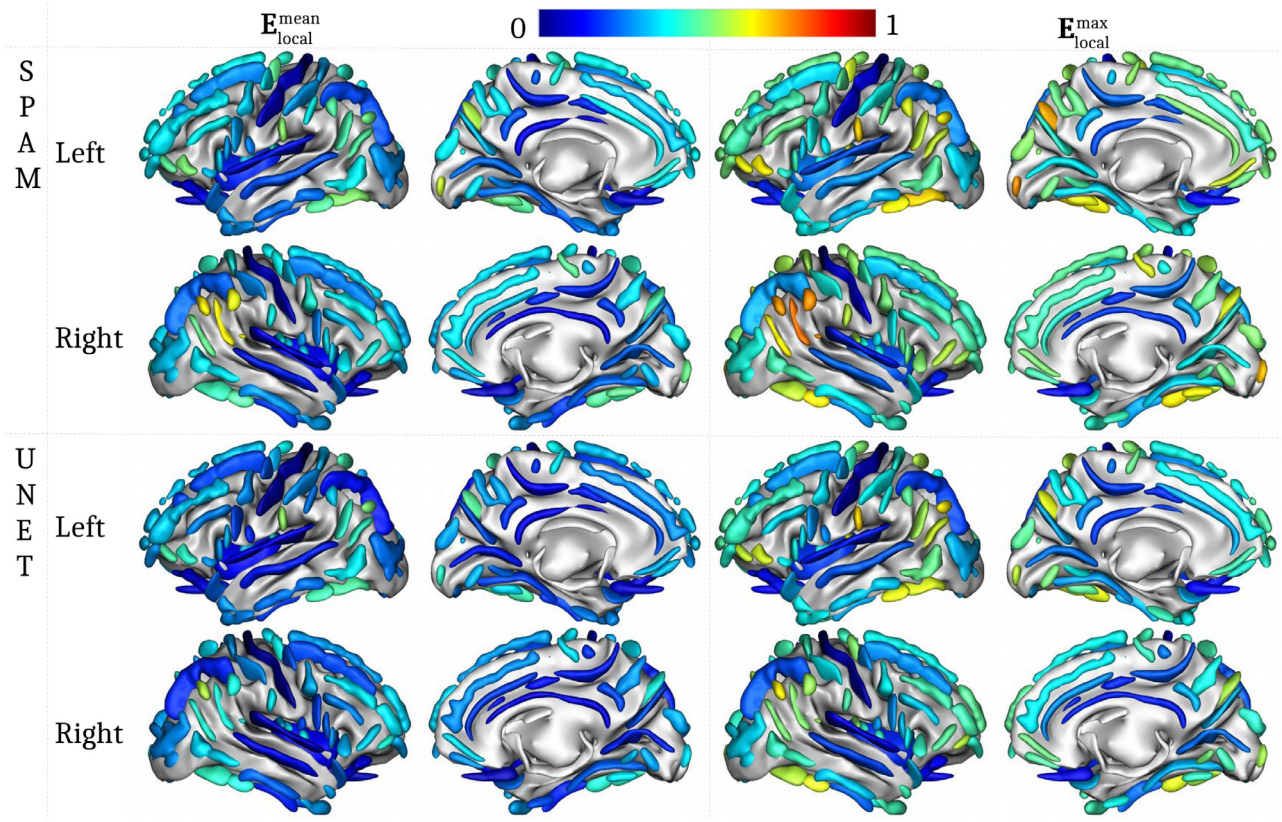

Fig. 15. E_{local} error rate per sulcus for SPAM and UNET models. The UNET model corresponds to the one after re-division of the elementary folds. Once the 10 segmentations have been labeled by hemisphere, we consider the average error per sulcus in the left column and the maximum error in the right column. The external and internal sides are represented for each of the right and left hemispheres.

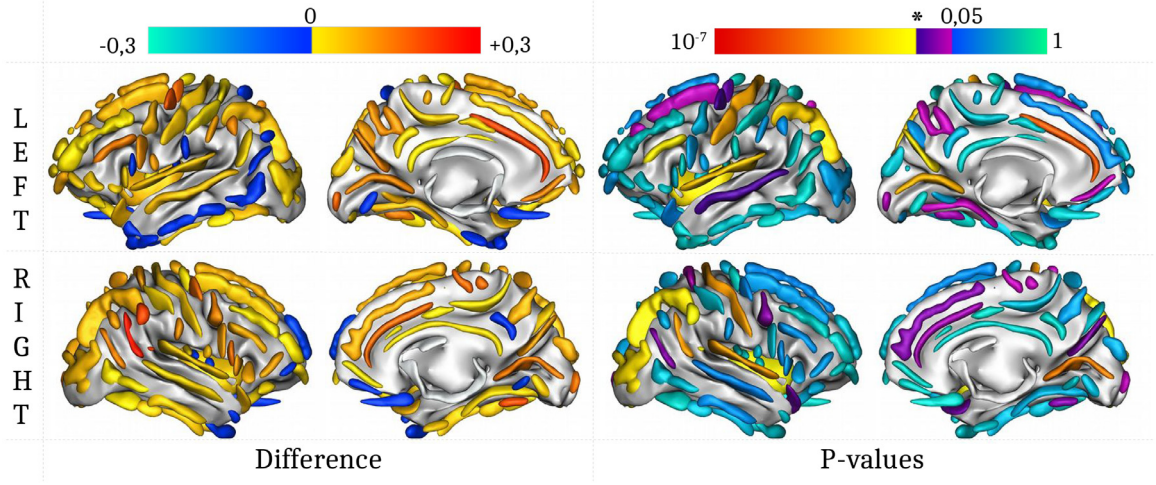

Fig. 16. Comparison of E^{max}_{local} error rates between the SPAM model and the UNET model. The left column represents the difference between the E^{max}_{local} of the SPAM model and of the UNET model. The right column shows the p-value of the Wilcoxon test between each model. Note that the scale of the color palette used to represent p-values is logarithmic. In order to visualize the sulci significantly better recognized, the threshold 0.05 is indicated and the threshold at the star corresponds to the first sulci considered significantly better by controlling the false discovery rate through the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995).

used for PCNN is not deep (only one hidden layer) compared to (Ciresan et al., 2012). However, after trying to make the network architecture more complex by increasing the number of hidden layers or using multiple patch sizes, we did not observe significant improvements in the results. It is imperative to note that the PCNN model achieves performances comparable to the UNET model while the U-Net architecture is much deeper and previous studies show that it is supposed to achieve better results (Ronneberger et al., 2015).

5.4. Unknown label

In this paper, except for the HPMAS model, the “unknown” label in the manually labeled database is treated like the other sulci labels. However, although the “unknown” label represents about 0.5% of the skeleton voxels of manual labeling, this proportion is null if we consider the labels of the HPMAS model. Moreover, the PMAS and PCNN models label around 0.02% of voxels as “unknown” and the SPAM and UNET models 0.04%. These figures show that

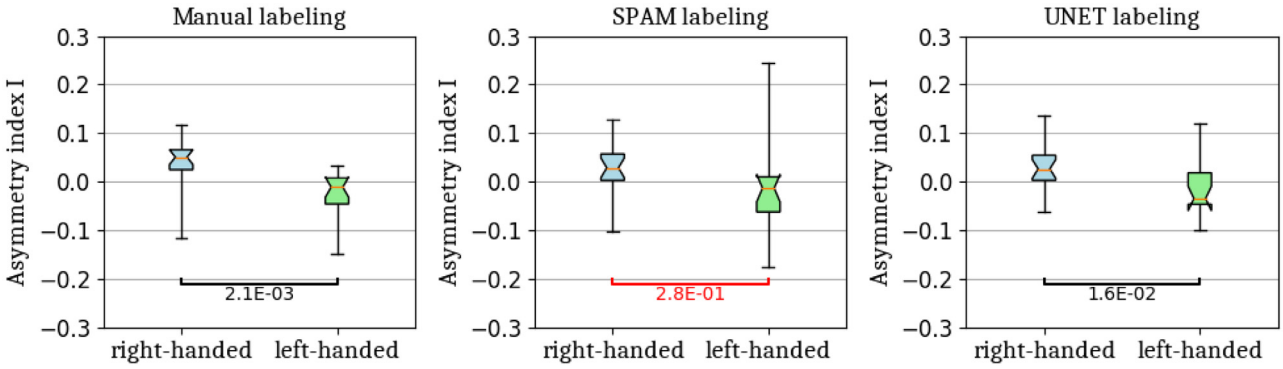

Fig. 17. Comparison of the asymmetry index I between right-handed and left-handed people. The left/middle/right graphs respectively show the results obtained with manual/SPAM/UNET labeling. The index for right-handed people is represented in blue and the one for left-handed people in green. The p-values of the T-test for the means of these two independent samples of scores are indicated on the graphs. With manual labeling, there is a significant difference: in left-handed people, the central sulcus is longer in the right hemisphere than in the left, while this is the opposite in right-handed people. The same significant difference is observed with the UNET model labeling but not with the SPAM model labeling. The box extends from the lower to upper quartile index I values, with a line at the median. The whiskers extend from the box to show the minimum and maximum values.

treating the "unknown" label as other labels is insufficient. Models should also assign the "unknown" label to structures where it is not sufficiently confident.

However, since all the new methods were compared to the SPAM model, which treated the unknown label as sulci labels, we chose not to address this point in this paper.

6. Conclusion

To summarize, the new methods presented in this paper outperform the current SPAM model provided by the Morphologist toolbox of BrainVISA. Compared to the SPAM model, the best models have a 4% higher recognition rate and 15% of sulci are significantly better recognized. By automatically re-dividing the elementary folds, the new models are considerably more robust to under-segmentation errors. In practice, these improvements make it possible to reproduce findings that were previously only possible with manual labeling. The UNET model will soon be available in the BrainVISA/Morphologist toolbox.

In this paper, the application of methods based on MAS or CNNs give approximately the same results for the automatic recognition of cortical sulci. However, although CNN-based methods have a particularly long training process compared to MAS-based methods, which are significantly faster. Therefore, CNNs-based methods are far more productive in practice. The UNET method labels a brain in only twenty seconds, whereas the SPAM method takes about ten minutes. It is interesting to note that patch MAS approaches are also beginning to integrate deep learning techniques (Manjón et al., 2018), probably due to their ability to effectively summarize the data and for their rapidity of execution.

Furthermore, the top-down refinement of bottom-up regularization significantly improves the results. Indeed, voxel-wise labeling is used to give a top-down perspective to a traditional bottom-up pattern recognition process that agglomerates the voxels into elementary folds: these folds can therefore be automatically re-divided when necessary. Thus, the labeling is robust to under-segmentation errors, unlike the SPAM method, which does not provide voxel-wise labeling. Note that despite the definition of elementary folds specific to the problem posed here, defining a coherent geometric entity is a legitimate concern addressed in many segmentation problems, for example by using super-pixels (Giraud et al., 2017; Soltaninejad et al., 2017) that group the most similar connected pixels together so that they have the same label.

In order to improve the current performance of the model, several options remain to be considered. Second, the inputs currently contain the fold skeleton in order to normalize the data for acquisition and age biases. However, the input can be enriched by taking into account grey/white matter segmentation or directly the normalized MRI. For instance, we could consider integrating this data into new input channels for CNN-based approaches. Finally, in order to take advantage of the large unlabeled databases currently available, a semi-supervised strategy would be particularly attractive to better represent the variability of the cortical folds.

In the near future, considering that the labeling model seems sufficiently reliable to us, we would like to reconsider the number of sulci used in the nomenclature on the basis of the sulci most often confused by the model. Indeed, the error rates of some small sulci are still too high to be used in morphological studies. By allowing the user to choose the level of granularity of the nomenclature, he will be able to use sufficiently stable labeling of the structures of interest to him.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 785907 (HBP SGA2), No. 720270 (HBP SGA1) and No. 604102 (HBP's ramp-up phase), and from the FR-MDIC20161236445.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.media.2020.101651.

References

Auzias, G., Colliot, O., Glaunes, J.A., Perrot, M., Mangin, J.-F., Trouve, A., Baillet, S., 2011. Diffeomorphic brain registration under exhaustive sulcal constraints. IEEE Trans. Med. Imaging 30 (6), 1214–1227.

Auzias, G., Lefevre, J., Le Troter, A., Fischer, C., Perrot, M., Régis, J., Coulon, O., 2013. Model-driven harmonic parameterization of the cortical surface: hip-hop. IEEE Trans. Med. Imaging 32 (5), 873–887.

Barnes, C., Shechtman, E., Finkelstein, A., Goldman, D.B., 2009. Patchmatch: a randomized correspondence algorithm for structural image editing. In: ACM Transactions on Graphics (ToG), vol. 28. ACM, p. 24.

Behnke, K.J., Rettmann, M.E., Pham, D.L., Shen, D., Resnick, S.M., Davatzikos, C., Prince, J.L., 2003. Automatic classification of sulcal regions of the human brain cortex using pattern recognition. In: Medical Imaging 2003: Image Processing, vol. 5032. International Society for Optics and Photonics, pp. 1499–1511.

Belaggoune, M., Saliha, O., Nadjia, B., 2014. Ontology driven graph matching approach for automatic labeling brain cortical sulci. Int. Conf. Inf. Technol. Organ. Dev. 162.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B 57 (1), 289–300.

Besl, P.J., McKay, N.D., 1992. Method for registration of 3-d shapes. In: Sensor Fusion IV: Control Paradigms and Data Structures, vol. 1611. International Society for Optics and Photonics, pp. 586–607.

Borne, L., Mangin, J.-F., Rivière, D., 2018. A patch-based segmentation approach with high level representation of the data for cortical sulci recognition. In: International Workshop on Patch-based Techniques in Medical Imaging. Springer, pp. 114–121.

Caliński, T., Harabasz, J., 1974. A dendrite method for cluster analysis. Commun. Stat.-Theory Methods 3 (1), 1–27.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O., 2016. 3d u-net: learning dense volumetric segmentation from sparse annotation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 424–432.

Ciresan, D., Giusti, A., Gambardella, L.M., Schmidhuber, J., 2012. Deep neural networks segment neuronal membranes in electron microscopy images. In: Advances in Neural Information Processing Systems, pp. 2843–2851.

Collins, D.L., Neelin, P., Peters, T.M., Evans, A.C., 1994. Automatic 3d intersubject registration of mr volumetric data in standardized talairach space.. J. Comput. Assisted Tomogr. 18 (2), 192–205.

Coupé, P., Manjón, J.V., Fonov, V., Pruessner, J., Robles, M., Collins, D.L., 2011. Patch-based segmentation using expert priors: application to hippocampus and ventricle segmentation. NeuroImage 54 (2), 940–954.

Fischl, B., Van Der Kouwe, A., Destrieux, C., Halgren, E., Ségonne, F., Salat, D.H., Busa, E., Seidman, L.J., Goldstein, J., Kennedy, D., et al., 2004. Automatically parcellating the human cerebral cortex. Cereb. Cortex 14 (1), 11–22.

Giraud, R., Ta, V.-T., Bugeau, A., Coupé, P., Papadakis, N., 2017. Superpatchmatch: an algorithm for robust correspondences using superpixel patches. IEEE Trans. Image Process. 26 (8), 4068–4078.

Giraud, R., Ta, V.-T., Papadakis, N., Manjón, J.V., Collins, D.L., Coupé, P., Initiative, A.D.N., et al., 2016. An optimized patchmatch for multi-scale and multi--feature label fusion. NeuroImage 124, 770–782.

Holz, D., Ichim, A.E., Tombari, F., Rusu, R.B., Behnke, S., 2015. Registration with the point cloud library: a modular framework for aligning in 3-d. IEEE Rob. Autom. Mag. 22 (4), 110–124.

Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv:1502.03167.

Le Guen, Y., Philippe, C., Riviere, D., Lemaitre, H., Grigis, A., Fischer, C., Dehaene-Lambertz, G., Mangin, J.-F., Frouin, V., 2019. eqtl of kcnk2 regionally influences the brain sulcal widening: evidence from 15,597 UK biobank participants with neuroimaging data. Brain Struct. Funct. 224 (2), 847–857.

LeCun, Y.A., Bottou, L., Orr, G.B., Müller, K.-R., 2012. Efficient backprop. In: Neural Networks: Tricks of the Trade. Springer, pp. 9–48.

Lohmann, G., von Cramon, D.Y., 2000. Automatic labelling of the human cortical surface using sulcal basins. Med. Image Anal. 4 (3), 179–188.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440.

Malandain, G., Bertrand, G., Ayache, N., 1993. Topological segmentation of discrete surfaces. Int. J. Comput. Vis. 10 (2), 183–197.

Mancip, M., Spezia, R., Jeanvoine, Y., Balsier, C., 2018. Tileviz: tile visualization for direct dynamics applied to astrochemical reactions.. Electronic Imaging 2018 (16). 286–1

Mangin, J.-F., Perrot, M., Operto, G., Cachia, A., Fischer, C., Lefèvre, J., Rivière, D., Neurospin, C., 2015. Sulcus identification and labeling. In: Brain Mapping: An Encyclopedic Reference, pp. 365–371.

Manjón, J.V., Coupé, P., Raniga, P., Xia, Y., Desmond, P., Fripp, J., Salvado, O., 2018. Mri white matter lesion segmentation using an ensemble of neural networks and overcomplete patch-based voting. Comput. Med. Imaging Graph. 69, 43–51.

Milletari, F., Navab, N., Ahmadi, S.-A., 2016. V-net: fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV). IEEE, pp. 565–571.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A., 2017. Automatic differentiation in pytorch. Neural Information Processing Systems - Autodiff Workshop.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al., 2011. Scikit-learn: machine learning in python. J. Mach. Learn. Res. 12 (Oct), 2825–2830.

Perrot, M., Riviere, D., Mangin, J.-F., 2011. Cortical sulci recognition and spatial normalization. Med. Image Anal. 15 (4), 529–550.

Riviere, D., Mangin, J.-F., Papadopoulos-Orfanos, D., Martinez, J.-M., Frouin, V., Régis, J., 2002. Automatic recognition of cortical sulci of the human brain using a congregation of neural networks. Med. Image Anal. 6 (2), 77–92.

Rohlfing, T., Brandt, R., Menzel, R., Maurer Jr, C.R., 2004. Evaluation of atlas selection strategies for atlas-based image segmentation with application to confocal microscopy images of bee brains. NeuroImage 21 (4), 1428–1442.

Romero, J.E., Coupé, P., Giraud, R., Ta, V.-T., Fonov, V., Park, M.T.M., Chakravarty, M.M., Voineskos, A.N., Manjón, J.V., 2017. Ceres: a new cerebellum lobule segmentation method. NeuroImage 147, 916–924.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 234–241.

Rousseau, F., Habas, P.A., Studholme, C., 2011. A supervised patch-based approach for human brain labeling. IEEE Trans. Med. Imaging 30 (10), 1852–1862.

Royackkers, N., Desvignes, M., Revenu, M., 1998. Une méthode générale de reconnaissance de courbres 3d: application à l'identification de sillons corticaux en imagerie par résonance magnétique. Traitement du Signal 15 (5), 365–379.

Shi, Y., Tu, Z., Reiss, A.L., Dutton, R.A., Lee, A.D., Galaburda, A.M., Dinov, I., Thompson, P.M., Toga, A.W., 2007. Joint sulci detection using graphical models and boosted priors. In: Biennial International Conference on Information Processing in Medical Imaging. Springer, pp. 98–109.

Snoek, C.G., Worring, M., Smeulders, A.W., 2005. Early versus late fusion in semantic video analysis. In: Proceedings of the 13th annual ACM international conference on Multimedia. ACM, pp. 399–402.

Soltaninejad, M., Yang, G., Lambrou, T., Allinson, N., Jones, T.L., Barrick, T.R., Howe, F.A., Ye, X., 2017. Automated brain tumour detection and segmentation using superpixel-based extremely randomized trees in flair mri. Int. J. Comput. Assisted Radiol. Surg. 12 (2), 183–203.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15 (1), 1929–1958.

Sun, Z.Y., Klöppel, S., Riviere, D., Perrot, M., Frackowiak, R., Siebner, H., Mangin, J.-F., 2012. The effect of handedness on the shape of the central sulcus. Neuroimage 60 (1), 332–339. doi:10.1016/j.neuroimage.2011.12.050.

Ta, V.-T., Giraud, R., Collins, D.L., Coupé, P., 2014. Optimized patchmatch for near real time and accurate label fusion. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 105–112.

Vivodtzev, F., Linsen, L., Hamann, B., Joy, K.I., Olshausen, B.A., 2006. Brain mapping using topology graphs obtained by surface segmentation. In: Scientific Visualization: The Visual Extraction of Knowledge from Data. Springer, pp. 35–48.

Ward Jr, J.H., 1963. Hierarchical grouping to optimize an objective function. J. Am. Stat. Assoc. 58 (301), 236–244.

Wolny, A., enfisan, 2019. wolny/pytorch-3dunet: first stable release. 10.5281/zenodo.2671581

Yang, F., Kruggel, F., 2009. A graph matching approach for labeling brain sulci using location, orientation, and shape. Neurocomputing 73 (1-3), 179–190.