Neuroinformatics (2025) 23:0
https://doi.org/10.1007/s12021-024-09700-7

RESEARCH

A Self-supervised Deep Learning Model for Diagonal Sulcus Detection with Limited Labeled Data

Delfina Braggio[1,2] · Hernán C. Külsgaard[1,2] · Mariana Vallejo-Azar[1,3] · Mariana Bendersky[4] · Paula González[1,3] · Lucía Alba-Ferrara[3,5] · José Ignacio Orlando[1,2] · Ignacio Larrabide[1,2]

Accepted: 24 October 2024

Abstract
Sulci are a fundamental part of brain morphology, closely linked to brain function, cognition, and behavior. Tertiary sulci, characterized as the shallowest and smallest subtype, pose a challenging task for detection. The diagonal sulcus (*ds*), located in a crucial area in language processing, has a prevalence between 50% and 60%. Automatic detection of the *ds* is an unexplored field: while some sulci segmenters include the *ds*, their accuracy is usually low. In this work, we present a deep learning based model for *ds* detection using a fine-tuning approach with limited training labeled data. A convolutional autoencoder was employed to learn specific features related to brain morphology with unlabeled data through self-supervised learning. Subsequently, the pre-trained network was fine-tuned to detect the *ds* using a less extensive labeled dataset. We achieved a mean F1-score of 0.7176 (SD=0.0736) for the test set and a F1-score of 0.72 for a second held-out set, surpassing the results of a standard software and other alternative deep learning models. We conducted an interpretability analysis of the results using occlusion maps and observed that the models focused on adjacent sulci to the *ds* for prediction, consistent with the approach taken by experts in manual annotation. We also analyzed the challenges of manual labeling by conducting a thorough examination of interrater agreement on a small dataset and its relationship with our model's performance. Finally, we applied our method on a population analysis and reported the prevalence of *ds* in a case study.

Keywords Tertiary sulci · Diagonal sulcus · Machine learning · Automatic classification · Fine-tuning

Introduction

Cortical folding serves as a distinctive feature of brain topography, and its variability is closely linked to differences in brain function, cognition and behavior at both interindividual and intergroup levels (Akula et al., 2023). Over the last years, numerous works have studied whether sulci and gyrus could be functionally and cognitively differentiated. For example, Amiez and Petrides (2018) related different foci of functional activity and specific sulci of the lateral frontal cortex; Yang et al. (2019), on the other hand, studied the temporal variability of cortical gyral-sulcal resting state functional activity and its association with fluid intelligence measures. Troiani et al. (2020) studied the relationship between the sulcogyral anatomy of the H-sulcus and the orbitofrontal cortex in response to certain stimulus categories.

Tertiary sulci (Ono et al., 1990), being the last ones to develop in utero and continuing postnatally (Welker, 1990; Fernández & Borrell, 2023; Williams et al., 2023), are typically the shallowest and smallest among cortical folds. It was previously reported that their identification presents methodological challenges, being frequently overlooked (Vallejo-Azar et al., 2023). Consequently, despite being evolutionary new structures, the role of tertiary sulci in human cognition remains poorly understood due to limited research in this area. Clinical diagnosis or surgical planification based on magnetic resonance images (MRI) analysis could be influenced by the local analysis of sulcal patterns. The correct identification of non-pathological anatomical variants among individuals improves the analysis, avoiding misdiagnosis linked to a poor knowledge of the normal variability.

Even though the role of tertiary sulci in neuroanatomy was not studied as deep as primary sulci, some works analyzed

Delfina Braggio and Hernán C. Külsgaard contributed equally to this work.

Extended author information available on the last page of the article

Published online: 26 December 2024

their characteristics and contribution in human cognition. Weiner et al. (2014), for example, demonstrated that the mid-fusiform sulcus is a key landmark characterizing both the cytoarchitectonic and functional partitions of the ventral temporal cortex in humans. Garrison et al. (2015) found that patients with hallucinations had shorter paracingulate sulcus than healthy controls and showed that this association is specific to patients with a psychotic disorder. Voorhies et al. (2021) showed that the depth of some tertiary sulci located in the lateral prefrontal cortex is associated with individual differences in reasoning scores beyond age. These findings underscore the importance of examining the tertiary sulci in the human brain.

Diagonal sulcus (*ds*) (Fig. 1) is a tertiary sulcus located in the frontal operculum (*FO*), an area that covers parts of the ventral frontal cortex adjacent to the insula. The *FO* is an essential cortical region for expressive speech, and includes the Broca's area, which has diverse roles in language processing. This region is structurally and functionally heterogeneous, and includes both a language-selective region, which is part of the frontotemporal core language network, and a domain-general region, which is a component of the frontoparietal multiple-demand network (Fedorenko & Blank, 2020). A correct identification of Broca's area and neighboring structures is mandatory to avoid misdiagnosis in MRI studies and to ensure their protection during surgery. Gaining a thorough understanding of the sulcal morphology in this region is also important for the success of functional and structural neuroimaging studies exploring language.

There is no agreement about a reliable structural asymmetry between Broca's area in the left and right hemispheres. Some investigators found a leftward asymmetry while others a rightward one or no asymmetry at all (Sprung-Much et al., 2022). One of the reasons for this discrepancy could be the presence of a *ds*, that modifies volume measurements in this region (Vallejo-Azar et al., 2023). As studied in Sprung-Much and Petrides (2018) and Vallejo-Azar et al. (2023), *ds* is highly prevalent, being present in 60% of the studied population in Vallejo-Azar et al. (2023), and in 51.25% of the hemispheres studied by Sprung-Much and Petrides (2018).

Identifying tertiary sulci in MRI, specifically *ds*, is challenging since they are frequently excluded from neuroanatomical atlases (Desikan et al., 2006; Destrieux et al., 2010). The *ds* may occasionally be confused with the inferior precentral sulcus in cases where the *ds* is long and connects with the inferior precentral sulcus. There may be an indirect connection between the inferior precentral sulcus and the lateral (Sylvian) fissure when the former connects with the *ds*, which in turn connects with the lateral fissure (Keller et al., 2009; Sprung-Much et al., 2022). Also, anatomists faced difficulties in accurately identifying them on the brain surface in postmortem tissue (Miller et al., 2021). As a result, most neuroimaging software packages are not reliable in recognizing them (Willbrand et al., 2022). They are capable of locating the region in which the *ds* could be located, but fail to distinguish it in cases where it is not similar to the atlas reference (Cointepas et al., 2001).

In the state of art, there is a small number of sulci classification methods, and the majority of them do not focus on a specific sulcus but rather on sulci in general (Yang & Kruggel, 2009; Perrot et al., 2011; Borne et al., 2020). Therefore, tertiary sulci are classified with low accuracy or simply excluded. To the best of our knowledge, there is no automatic classification methods trained to detect the presence of the *ds* exclusively with high accuracy.

The aim of this work is to develop a novel deep learning-based tool to accurately identify the *ds* in MRI. The primary objective is to detect the presence of the *ds* and analyze its prevalence across various populations. Combining the classification task with saliency maps, enables the identification of the brain morphological patterns that contribute to the detection of this sulcus, thereby facilitating the identification of anatomical associations in the examined area. As mentioned earlier, manual labeling is challenging and requires specific anatomical knowledge, especially with respect to tertiary sulci. As a result, the amount of labeled data is

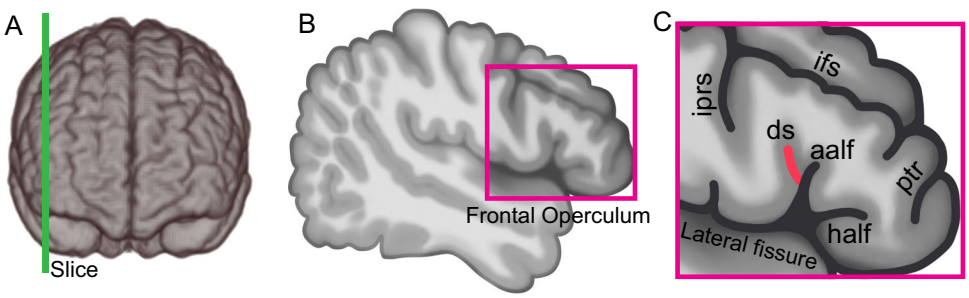

Fig. 1 Sulci reference in the frontal operculum. (A) Slice located in the *ds* approximately. (B) Sagittal plane at the location of the *ds* in the frontal operculum (pink color rectangle).(C) Sulcal map in the frontal operculum.*aalf*: ascending ramus of the lateral fissure; *ds*: diagonal sulcus (draw in red); *half*: horizontal ramus of the lateral fissure; *ifs*: inferior frontal sulcus; *iprs*: inferior precentral sulcus

limited. To overcome this limitation, we propose the use of a convolutional neural network (CNN) to automatically classify the presence of *ds* by combining a self-supervised pre-training approach and a supervised fine-tuning step. In the first stage, we train a convolutional autoencoder using an extensive and heterogeneous unlabeled dataset to learn decisive features from the images. In the second stage, we propose to reshape the network architecture as a classification CNN by fine-tuning the pre-trained model with a limited amount of manually labeled data. We evaluated the performance of our model on two held-out test sets, comparing the results with manual labels. We observed that for both datasets, our model achieved better results in detecting *ds* than a standard software and other deep learning models. Additionally, we analyzed the difficulties of manual labeling on a limited dataset and its impact on the performance of our model. We observed that patches presenting challenges in classification by our model are also more visually difficult to identify. We used occlusion maps to interpret the model and analyzed crucial parts for prediction. We observed that these areas are consistent with the guidelines used for manual labeling. Finally, we applied our method to analyze the prevalence of *ds* on a population study and contrasted the results with the existing literature.

Materials and Methods

Subject Data

HEC-HR

One hundred healthy volunteers (43 males and 57 females) with no history of neurological or psychiatric diseases participated in this study. The participants had a mean age of 31.4 years, ranging from 18 to 57 years. Image acquisition followed the protocol detailed in Vallejo-Azar et al. (2023). The scans were conducted using two different MRI scanners: a 3T Philips Achieva (n = 55) located at Hospital El Cruce in Florencio Varela (Argentina) and a 3T Siemens Trio (n = 45) located at Instituto de Oncología Ángel Roffo in Buenos Aires (Argentina).

Volumetric T1 images were acquired using a 3D FFE sequence with the following parameters for the 3T Philips Achieva: TE = 3.3 ms, TR = 2,300 ms, TI = 900 ms, flip angle = 9°, FOV = 240 × 240 × 180, voxel size = 1 × 1 × 1 mm^3, and 239 slices. For the 3T Siemens Trio, the MP-RAGE sequence was used with the following parameters: TE = 2.27 ms, TR = 2,000 ms, TI = 900 ms, inverted angle = 80°, FOV = 250 × 250, voxel size = 1 × 1 × 1 mm^3, and 204 slices.

All acquired images were converted from DICOM to NIFTI format using Dcm2Nii (Li et al., 2016).

Preprocessing

Images were preprocessed through several steps. Initially, they were visually examined to identify artifacts that could affect subsequent preprocessing. Each of the images was reoriented, with reference to the RAS orientation. Skull stripping, denoising and bias correction were carried out. Images were then spatially normalized without modulation, using the MNI152 template as a reference, with a spacing of 1.5 x 1.5 x 1.5 mm per voxel, 121 x 145 x 121 voxels in total. Since the images were obtained from different scanners, segmentation is an important step to minimize differences arising from image acquisition. Hence, normalized images were segmented into tissues, and gray matter segments were utilized for subsequent analysis. The segmentation output was a probability map in which the value of each voxel represented the concentration of gray matter at that specific brain coordinate.

We conducted a thorough visual inspection of each image in the dataset at every stage of preprocessing. This ensured that the normalized images were properly aligned and free of artifacts or distortion. Images that did not meet our quality standards were discarded. We also overlaid the gray matter segments with the T1 images to check for any segmentation issues. This step allowed us to identify and address potential problems arising during the segmentation process. After preprocessing, one individual was excluded as its segmentation did not meet the desired image quality.

Finally, for each image, two areas - one for each hemisphere - were extracted, centered in the region where the *ds* is located (if applicable). Coordinates were consistent for all images and were visually selected based on the MNI152 template. Since the images had been previously spatially normalized, each area corresponded approximately to the same anatomical area, centered on the *FO*. During the training process, we extracted patches from each area of interest with voxel dimensions of 32x64x64 (see Fig. 2). By incorporating these diverse patches, we aimed to enhance the model's ability to learn from a wide range of anatomical variations, ultimately improving its generalization capabilities in subsequent analyses.

All preprocessing was conducted using SPM12 (Penny et al., 2011), specifically with the CAT12 plugin developed by Gaser et al. (2022), along with custom Python routines. All software tools are publicly available and widely applied in the neuroimaging field.

Labeling

Each patch was visually examined and categorized based on the presence of the *ds*. Images were classified into two classes: *Class 0* (no *ds* present) and *Class 1* (with *ds*). The analysis was conducted collaboratively and iteratively by a group of experts (n=3). In cases where a consensus was

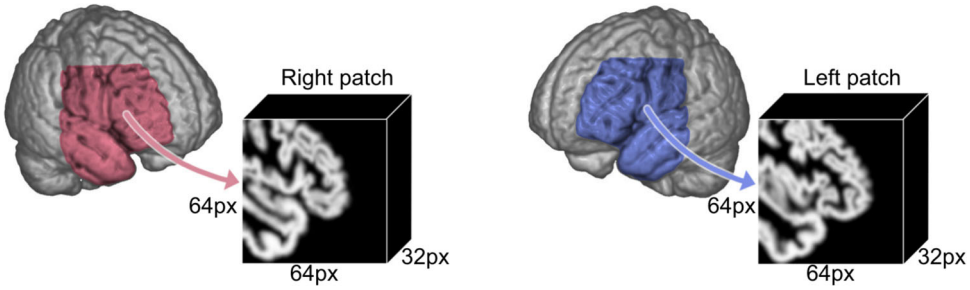

Fig. 2 Representation of the patches extracted for each hemisphere from a gray matter segmentation image. The size of those patches is 32x64x64 pixels

not reached, a voting strategy was employed, and the final label was determined based on the majority opinion. Consequently, 51% of the patches (n=101) exhibited the presence of the *ds*.

The anatomical criteria (Sprung-Much & Petrides, 2018) followed by the experts for identifying the *ds* included:

- It is located within the inferior frontal gyrus, relatively vertical, within the pars opercularis.
- It is always behind the ascending ramus of the lateral fissure (*aalf*), and it can assume various patterns: *type II*, *type III a*, *b* and *c* (Fig. 3).
- It differs from the ascending ramus of the lateral sulcus (*aalf*) in that both the ascending (*aalf*) and horizontal ramus (*half*) deeply reach the circular sulcus of the insula.
- Based on the aforementioned, the *ds* is solely a superficial sulcus.
- The ascending ramus (*aalf*) forms an angle of approximately 90° with the horizontal ramus (*half*). It is worth noting that in hemispheres with a *type III a* sulcus (Fig. 3), the ascending ramus of the lateral fissure (*aalf*) takes a superiorly anterior direction.

Data Splitting

To train the model, a 5-fold cross-validation strategy was employed, resulting in the data being split into five different non-overlapping sets, each containing between 38 and 42 images. This selection was performed to ensure data balance, with both patches from each individual belonging to the same subset. For example, the left patch of subject 1 cannot be in the training set, while the right one for that subject is in the validation set.

In each of the five iterations, a model k_i was trained using four folds for training and validation, while the remaining one was used for testing. This resulted in training with 156 to 160 images and testing with 38 to 42 images. We extracted patches of 32x64x64 dimension from the images as described in Section "Data Splitting", and used complete 3D patches for training. Each model's test set was distinct and non-overlapping, ensuring that the model was trained and

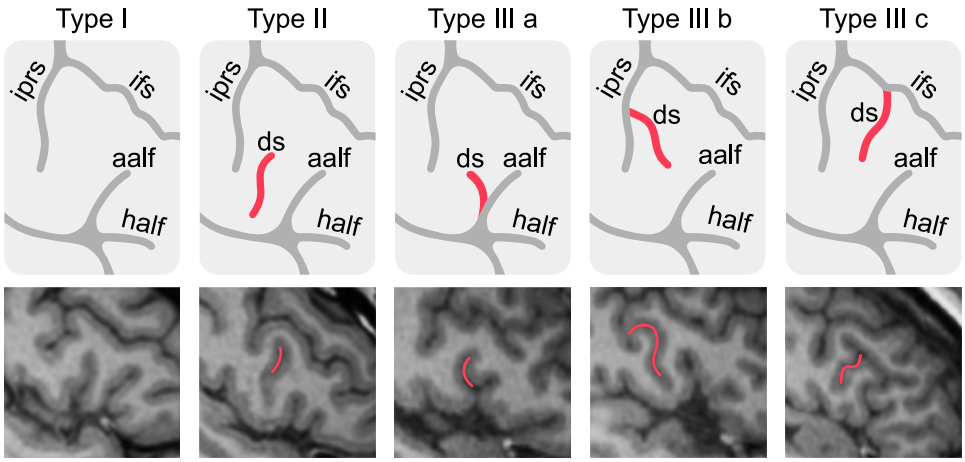

Fig. 3 Morphological patterns of the different types of *ds* and its neighboring sulci based on Sprung-Much and Petrides 2018, illustrated in sagittal planes. *aalf:* ascending ramus of the lateral fissure; *half:* horizontal ramus of the lateral fissure; *ifs:* inferior frontal sulcus; *iprs:* inferior precentral sulcus. Top row: reference illustrations of the types of *ds* (drawn in red). Lower row: examples of each *ds* type, extracted from our dataset. *Type I*: absence of the *ds*. *Type II*: the *ds* lies between the *aalf* and the *iprs*. *Type III a,b,c*: the *ds* is attached to one of the neighboring sulcus, the *aalf*, the *iprs* or the *ifs*, respectively

tested on different datasets during each cycle. This approach enhanced the robustness of the models, resulting in five different k_i models being trained. Data division for each of the k_i models is detailed in Table 1.

Data Augmentation

We employed data augmentation to mitigate overfitting, increasing the amount of training data by artificially generating new images. These new samples were created from the originals by applying various transformations to them. In our study, we utilized the TorchIO library (Pérez-García et al., 2021) to generate these augmented images. Specifically, for each scan in the training set, we applied a random transformation with a probability of 0.6, including random Gaussian noise and random affine transformations with probabilities of 0.7 and 0.3, respectively. Affine transformations involved random rotations along three independent axes of up to 5 degrees, random resizing and random translations of up to 2 mm along each dimension. Gaussian noise was randomly applied with a mean value of 0 and a standard deviation of 0.1.

Additional Data

OASIS-ADNI

For pre-training, we used an additional set of 5388 brain images extracted from two public datasets, namely ADNI[1] (Mueller et al., 2005) and OASIS[2] (LaMontagne et al., 2019). To be more specific, we selected 2005 images from the ADNI 1.5T 1 Year Complete Collection dataset, comprising longitudinal images from 639 different subjects. These scans included healthy controls (n=195), subjects with Mild Cognitive Impairment (n=311), and patients with Alzheimer's disease (n=133). ADNI images were acquired using various 1.5 T scanners and are T1 weighted. Additionally, we used 3383 longitudinal images from 1097 subjects from the OASIS dataset taken with different scanners. This dataset included healthy controls (n=703), patients with Alzheimer's disease (n=286), and subjects with other types of dementia (n=108). Images were preprocessed as described in Section "Preprocessing". A multicenter dataset was considered to enhance variability and ensure robustness across different image types and patient cohorts. The images were acquired using various scanners and included a range of clinical conditions, with both healthy subjects and those with pathologies. This variability in the training dataset enables our model to effectively recognize characteristics across different patient cohorts, enhancing its applicability to a broader range of clinical scenarios.

IXI

We collected 20 additional patches from 10 unseen T1 scans from IXI dataset (2012) [3]. Images were acquired in Hammersmith Hospital using a Philips 3T scanner with an acquisition matrix of 208 x 208 voxels and the following parameters: TE = 4.6 ms, TR = 9.6 ms, flip angle = 8.0º. Preprocessing is described in Section "Preprocessing". These images were not used to train our model and were employed to assess its performance with a new set of data featuring acquisition parameters distinct from those used during training. By analyzing this, we can analyze how robust our model is to generalize to new, unseen images.

Each of these images were analyzed and labeled by a group of experts, composed with three anatomy experts and two technical experts (n=5). The labeling process was made individually and blindly by each of the experts and categorized in *Class 1* (with *ds*) and *Class 0* (no *ds* present). The final label for each image was decided using a majority voting strategy and resulted in 55% of *Class 1* patches.

Classification

To facilitate automatic classification, we propose *Ft-Encoder*, a convolutional neural network (CNN) trained by combining a self-supervised pre-training and a fine-tuning step. The primary objective was for the network to acquire the ability to identify the *ds* in new images not included in the training dataset. To evaluate its results, we classified the same dataset using various baseline methods. The subsequent sections will provide descriptions of our model *Ft-Encoder* and the different baseline variants.

Ft-Encoder

Motivated by the limited availability of labeled data, we employed a combination of self-supervised learning and a fine-tuning step.

Initially, a convolutional autoencoder was trained with a pretext task of reconstructing the original input patches, thus learning relevant features from them. The task in this step involved minimizing the difference between the original image and its reconstruction, and we utilized the publicly available images described in Section "OASIS-ADNI" as a training set. We intentionally chose these heterogeneous datasets, acquired with different scanners and encompassing a variety of clinical conditions, with the aim of training a more adaptable model. The network was designed to be

[1] https://adni.loni.usc.edu

[2] https://www.oasis-brains.org/

[3] https://brain-development.org/ixi-dataset/

Table 1 Data distribution in splits for cross-validation

Setting	Training			Validation			Test		
	class 0	class 1	class 1%	class 0	class 1	class 1%	class 0	class 1	class 1%
1	68	58	0.539	14	18	0.437	15	25	0.375
2	68	60	0.531	13	17	0.433	16	24	0.4
3	65	65	0.5	13	17	0.433	19	19	0.5
4	56	68	0.451	16	16	0.5	25	17	0.595
5	61	65	0.484	14	20	0.411	22	16	0.578

Each row represents a different setting. We present the distribution for each set, including training, validation, and test. For each set, we categorize the data according to its class: class 0 (no *ds*) and class 1 (with *ds*). Additionally, we provide the class 1 ratio, which represents the percentage of images in the split with a *ds*

capable of identifying the *ds* in diverse types of brain images, not limited to a specific data acquisition protocol or healthy controls alone. This approach allowed us to fine-tune the pre-trained network with a different subset of labeled images without requiring retraining. The convolutional autoencoder was trained with the following hyperparameters: binary cross entropy loss, ADAM optimizer, a reduce-on-plateau scheduler with a reducing factor of 0.2, a batch size of 8 images and an initial learning rate of 0.001.

After the self-supervised learning step, we employed a fine-tuning step, and used the encoder layers as an initialization point to train a new network with the label data (Section "Preprocessing"). The architecture of the encoder is described in Fig. 4, and the choice of hyperparameters was carefully considered. For the loss function, we opted for cross-entropy, which is a commonly used metric in classification tasks. To optimize our model, we employed the Adam optimizer, and our learning rate was set to a small value of 0.0001. Figure 5 illustrates the general pipeline used during fine-tuning. We added an additional fully-connected layer for image classification to the encoder, and the output provided a binary label indicating the presence or absence of the *ds*.

Baseline Models

To evaluate the performance of our approach, several baseline models were considered. We describe them on the following sections.

ResNet10

In this baseline approach, a 3D adaptation of the *ResNet10* CNN architecture (He et al., 2016) was trained from scratch for automatic classification. Its architecture is described in Fig. 4 and as hyperparameters we used a cross-entropy loss, an Adam optimizer, and an initial learning rate of 0.0001. The model was trained with the labeled patches from the dataset described in Section "HEC-HR" and evaluated on its classification performance. This baseline approach utilized the same architecture as in the fine-tuning step of the Ft-Encoder. The main difference between the two approaches lies in the pre-training step. Unlike the Ft-Encoder, that used a self-supervised approach, the ResNet10 model involved training the network from scratch without any unsupervised phase. Model parameters were adjusted solely during supervised learning, focusing on classifying the labeled patches from the HEC-HR dataset.

Med3D

With the same architecture and hyperparameters as *ResNet10*, this strategy used the pre-trained *Med3D* network (Chen et al., 2019) as an initial point for a fine-tuning step. Unlike ResNet10, this approach included a pre-training phase. The *Med3D* network was trained on numerous small-scale datasets from various medical domains, covering different imaging modalities, target organs, and pathological manifestations. This diverse pre-training, which involved multiple types of 3D medical images, positions the network as a viable solution for *ds* classification.

BrainVisa

We used the sulci recognition method provided by BrainVisa

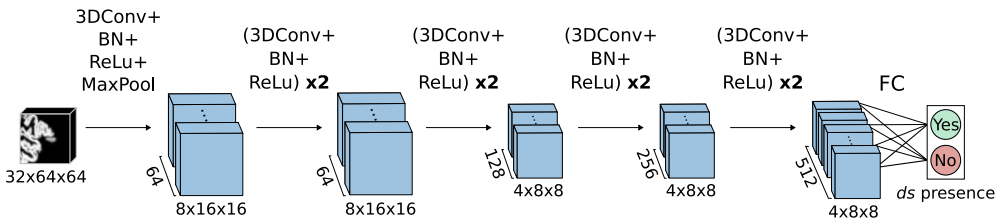

Fig. 4 ResNet10 architecture (He et al., 2016). *3DConv*: 3D convolutional layer, *BN*: Batch Normalization layer, *ReLu*: Rectified linear unit, *MaxPool*: Max-pooling layer, *FC*: Fully connected layer

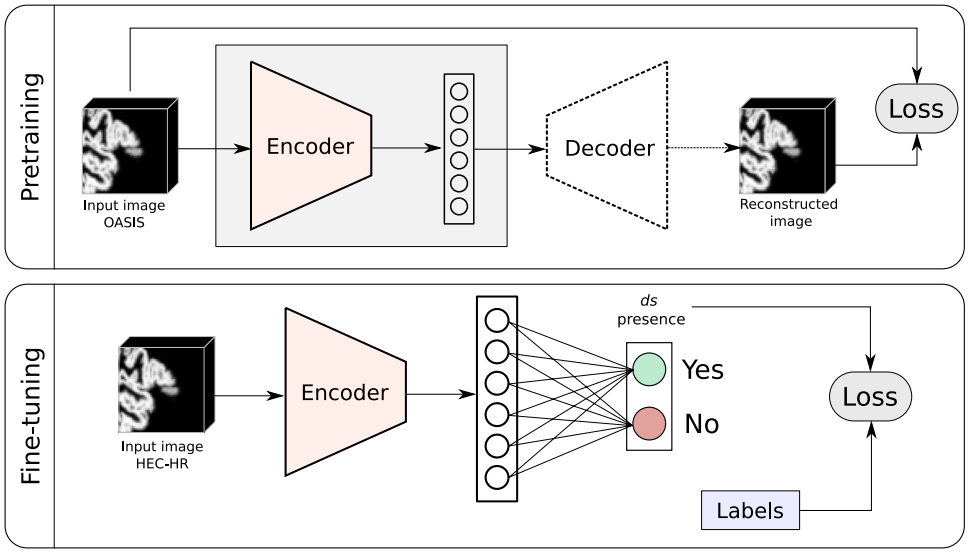

Fig. 5 Structure of the Ft-Encoder model. Top subfigure: Autoencoder architecture, trained with unlabeled data. Bottom subfigure: Supervised CNN architecture. This network uses the pre-trained encoder from the previous step as a starting point to train a classification model with labeled data. A fully connected layer is added to the encoder indicating if the *ds* is present

(Cointepas et al., 2001) to identify the *ds*. This is a software platform that contains several tools for brain morphology analysis. The Morphologist toolbox includes a recent update of its sulci recognition method based on CNN designed by Borne et al. (2020). This methodology is able to precisely identify up to 63 sulci in the right hemisphere and 64 in the left hemisphere, including the *ds* for each one. All the subjects from the HEC-HR (Section "HEC-HR") and IXI dataset (Section "IXI") were processed using the pipeline of the Morphologist toolbox for sulci recognition.

Model Interpretability

We used a model interpretability approach based on occlusion maps (Zeiler & Fergus, 2014) to understand the results from each network. With this approach, we translated which parts of the images were more important for the classifier to make a decision. This was done by systematically occluding different parts of the input image using a sliding window of 3x3x3 voxel dimension and analyzing its effect on the output label. If a portion is covered and the probability of the correct class drops, it can be assumed that that specific portion of the image is relevant for the model to make a decision. Otherwise, if the output probability is not affected by the occlusion, it can be assumed that the specific covered part is not relevant to the decision making of the model.

Evaluation Metrics

To assess the models' performance in detecting the *ds*, we employed various metrics, providing insights into the models' ability to detect the sulcus. We calculated the number of true positives (*TP*), true negatives (*TN*), false positives (*FP*), and false negatives (*FN*) across all samples. In this context, *TP* represents the patches containing the sulcus that the model correctly detected. *FP* signifies cases where the model erroneously identified a sulcus. *TN* indicates instances where the model accurately recognized the absence of the sulcus, and *FN* denotes cases where a *ds* was present in the patch, but the model failed to detect it. The following metrics were derived from these:

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Specificity} = \frac{TN}{FP + TN}$$

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

$$\text{Balanced accuracy} = \frac{\text{Specificity} + \text{Sensitivity}}{2}$$

Interrater Agreement

As described in Section "IXI", a held-out dataset was individually and blindly labeled by a group of five experts. Similarly, each of these images received five different labels from the distinct k_i models of the *Ft-Encoder*. As detailed in Section "Data Splitting", we adopted a cross-validation approach, training five different k_i models with the same architecture and training configuration. Consequently, for every patch, each of these separate models produced a different result, and the final label was determined using a voting strategy.

It is important to note that the agreement among experts (or models) is not uniform across all patches. Some images, due to their anatomical characteristics, are more challenging to label than others. Taking this into account, for each new

patch, we calculated the *Certainty* of the experts and the *Certainty* of *Ft-Encoder* as follows:

$$c = |\frac{p-n}{N}|$$

Here, p represents the number of experts (or k_i models) that labeled the patch as *Class 1*, n stands for the number of experts (or k_i models) that labeled the patch as *Class 0*, and N is the total number of experts (or k_i models). Finally, we defined three different categories of *Certainty*: *Slight*, *Moderate* and *Perfect* as follows:

$$\text{Certainty} = \begin{cases} \text{Slight if } c < 0.5 \\ \text{Moderate if } 0.5 \leq c < 1 \\ \text{Perfect if } c = 1 \end{cases} \quad (1)$$

This measure provides insight into the certainty of the observers and the *Ft-Encoder* when classifying each of the new patches. When the value of Certainty is *Perfect*, it indicates unanimous agreement, eliminating any uncertainty regarding the label.

Results

Model Evaluation

Figure 6 summarizes the metrics of the various models. Every group of bars represents one of the estimated metrics, and each bar within the group corresponds to a different model. Each of the subfigures corresponds to a different dataset.

In Fig. 6a, we show the results of the evaluation on the HEC-HR dataset (described in Section "HEC-HR") for different models, including the *ResNet10*, *Med3D*, *Ft-Encoder*, and *BrainVisa*. As outlined in Section "Data Splitting", for the first three variants, we employed a cross-validation approach, training five different models. Since each of these models yields a different score, we presented the information using the mean value and standard error, which were estimated based on the five values obtained.

We observe that *Ft-Encoder* outperforms other architectures on all metrics, except for *Specificity*, where *BrainVisa* had a better value on the HEC-HR dataset. Analyzing its *Sensitivity* value, we note that *BrainVisa* over classified the *Class 0* cases. *ResNet10* and *Med3D* showed similar results with high variability in each model's outcomes, while our model exhibited more stability.

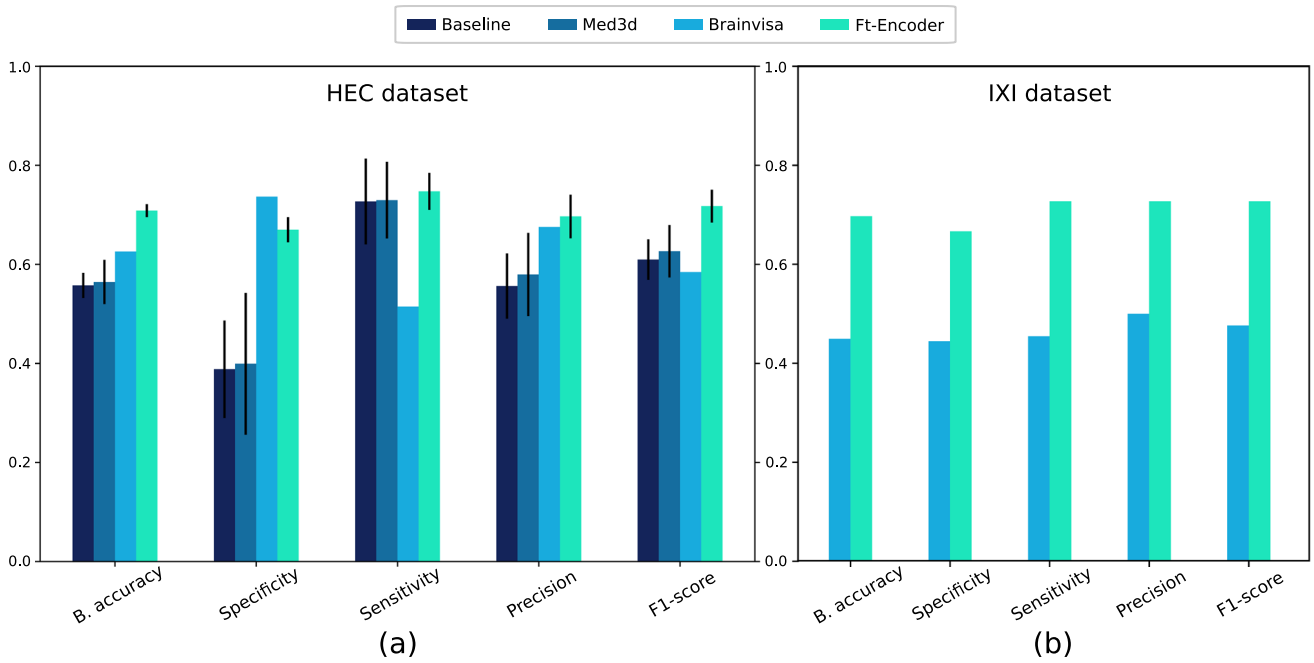

Fig. 6 Summary of the performance of each model for the different datasets. Each group of bars represents a different metric, including Balanced Accuracy, Specificity, Sensitivity, Precision, and F1-score. Each color corresponds to a different model: *ResNet10*, *Med3D*, *BrainVisa* and *Ft-Encoder*. For *ResNet10*, *Med3D* and *Ft-Encoder* each metric is calculated as the mean of each cross-validated model along with its respective standard error. (a) Results from the HEC-HR dataset. (b) Results from the IXI dataset

Study Case using a Held-out Test Set

We analyzed the performance of *Ft-Encoder* on a new set of images. Using our model, we classified each of the new patches described in Section "IXI", computed metrics to evaluate its performance, and the results can be seen in Fig. 6b. Overall, we observe that the metrics achieved with our classifier on the new data are similar to those obtained with the test data. Conversely, each of *BrainVisa*'s metrics dropped when tested on the IXI dataset, performing worse than the expected 0.5 value if the model assigned labels randomly. We can also observe that *Ft-Encoder* outperformed *BrainVisa* in every measure evaluated.

As described in Section "Interrater Agreement", *Certainty* values of the experts and the *Ft-Encoder* were calculated. We illustrate the relationship between these values and the *Ft-Encoder* prediction in Fig. 7. As observed in Fig. 7a, among all cases, 45% of them showed a *Slight Certainty* among experts. Additionally, we observe that as the experts' *Certainty* increases, the number of misclassified cases decreases. There is a 33% misclassification rate when the *Certainty* is *Slight*, 25% when the *Certainty* is *Moderate*, and no misclassified cases when the agreement among experts is *Perfect*. This trend is consistent for both FP and FN cases. In contrast, we can observe on Fig. 7b that on 45% of all cases the model's Certainty is Perfect. In Fig. 7c, we illustrate the relationship between the *Ft-Encoder*'s *Certainty*, experts' *Certainty*, and the classification results. We can observe that when the experts' *Certainty* is *Perfect*, the *Ft-Encoder*'s *Certainty* is also *Perfect*, and the classification results are correct. There is a slight tendency for the model's *Certainty* to decrease when the experts' *Certainty* decreases.

Model Interpretability

As explained in Section "Model Interpretability", we employed a model interpretability approach to visualize the most relevant sections of the image for classification. Figure 8 provides visual representations of occlusion maps generated using randomly selected images inputs from our dataset. These occlusion maps offer valuable insights into how each model identifies and focuses on key features within the input images.

Figure 8 shows two patches of the same individual. Figure 8a shows an image of the right hemisphere, where the *ds* is present and it was correctly identified by the *Ft-Encoder* classifier. We can observe that the model focused mainly on the image features related to the ascending ramus of the lateral fissure and not in the *ds* itself. Figure 8b shows another

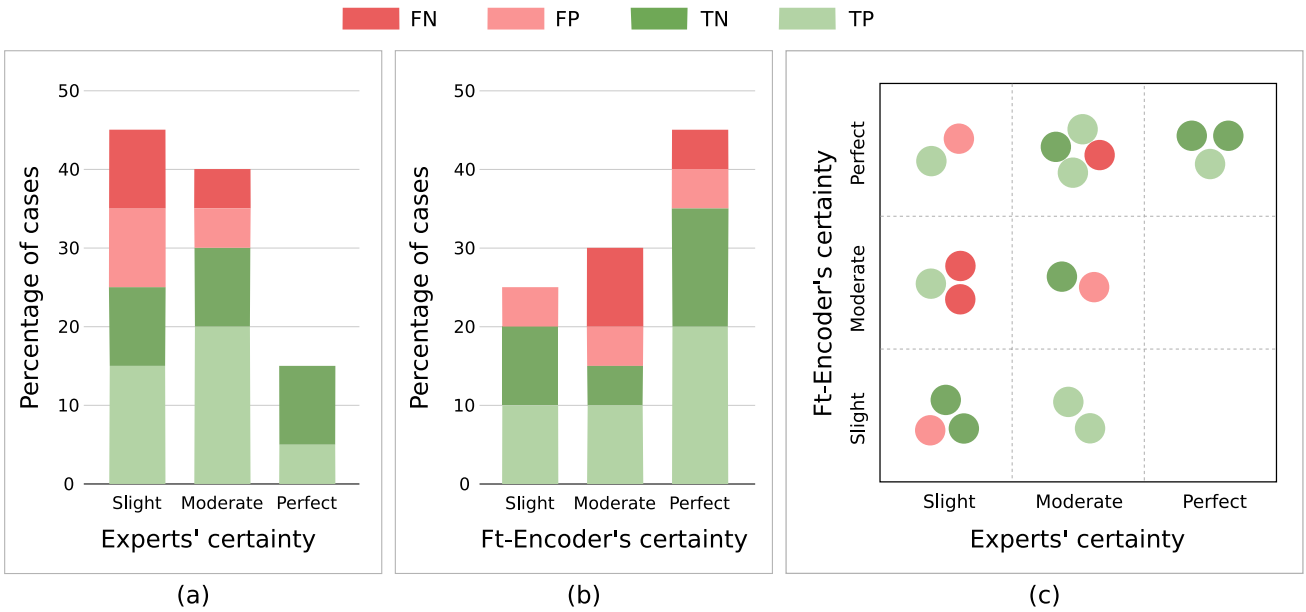

Fig. 7 *Ft-Encoder*'s predictions related with experts' and *Ft-Encoder*'s *Certainty*. (a) Correlation between the expert's *Certainty* and the classification results of the *Ft-Encoder*. (b) Correlation between the *Ft-Encoder*'s *Certainty* and the classification results. (c) Relationship among experts' *Certainty*, *Ft-Encoder*'s *Certainty* and classification results

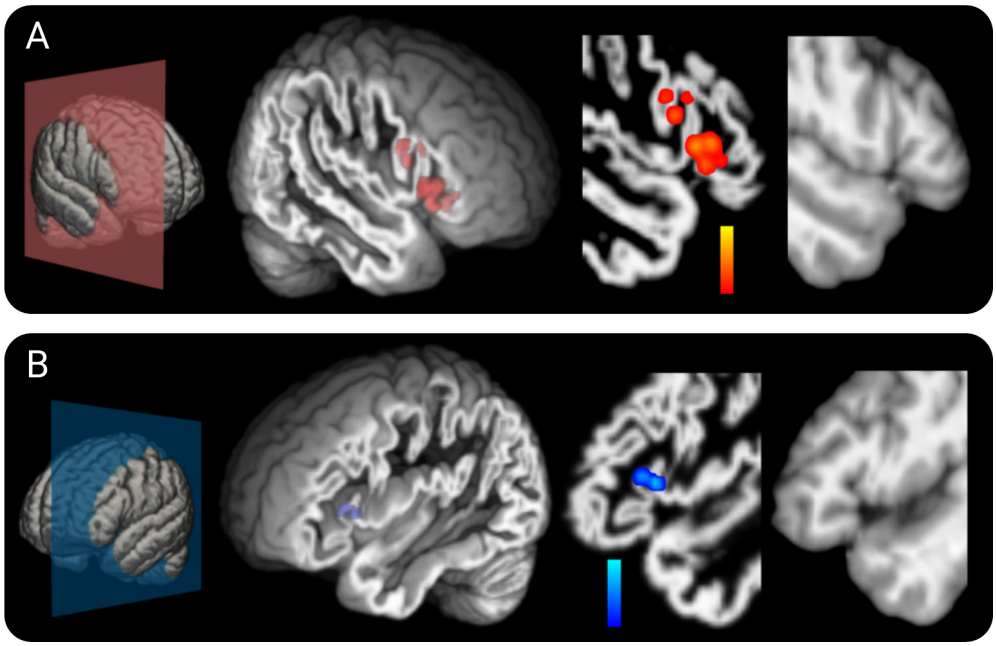

Fig. 8 Model interpretability. Classification examples and its correspondent occlusion maps. From left to right: 3D representation of the brain slice, 3D view of the gray matter with the occlusion region highlighted (red and blue color), slice from the patch on the sagittal plane of the segmented gray matter and the same slice for combination of white and gray matter. (a) TP example of subject with a *ds* in the right hemisphere classified as *Class 1*. (b) FP example of subject without a *ds* in the left hemisphere misclassified as *Class 1*

image, which does not contain the *ds* but the model erroneously identified it. As we can observe on the occlusion map the model focused only on the region of the ascending ramus of the lateral fissure, but not on the misclassified *ds*.

To analyze a general pattern across all *Ft-Encoder* outputs, for each of the k_i models trained using a cross-validation approach (Section "Data Splitting"), we computed a frequency map. These maps were generated by combining the results from the occlusion maps of every subject. Each voxel of the maps represents the number of images for which that particular location was considered relevant for classification by the corresponding k_i model.

In Fig. 9 the results of the right hemisphere of HEC-HR dataset for all the *Ft-Encoder* models are shown. As we can see, the focus was mainly located on the image features related to the ascending ramus of the lateral fissure and the *ds*, being also the most relevant guidelines used during manual classification. The lower part of the inferior precentral sulcus was also highlighted, but to a lesser extent, specially in the model k_4 and k_5. Similar results were observed for the horizontal ramus of the lateral sulcus in the k_1 model, and for the inferior frontal sulcus in the k_1, k_4, and k_5 models, respectively. There were also some outliers located in regions distant from the *ds*.

In the case of the left hemisphere (Fig. 10), the results were similar. The ascending ramus of the lateral fissure and the *ds* were part of the main focus of the models, however, the highlighted zones were more spread around the *ds*. For the k_1 model, the inferior precentral sulcus and the inferior frontal sulcus were also discerning features for the classification. Similar results were observed on the other models but to a lesser extent.

The same occlusion and frequency maps were computed for the outputs from the IXI dataset, and the results were consistent with the HEC-HR dataset for all the models. The corresponding figures are presented in the supplementary material.

Finally, to qualitative analyze the performance of *Ft-Encoder*, we individually examined each of the misclassified patches and analyzed their characteristics (Fig. 11). FN cases mainly occurred in patches where the *ds* was of an infrequent type like *type III c* (cases *a* and *b*) or *II* (cases *c* and *d*). Additionally, we noticed that among the FN cases, there were instances where the *ds* was short or resembled the curvature of the ascending ramus of the lateral fissure.

We also examined the FP cases and correlated these results with their occlusion maps. We observed several cases where the ascending ramus of the lateral fissure was curved, and

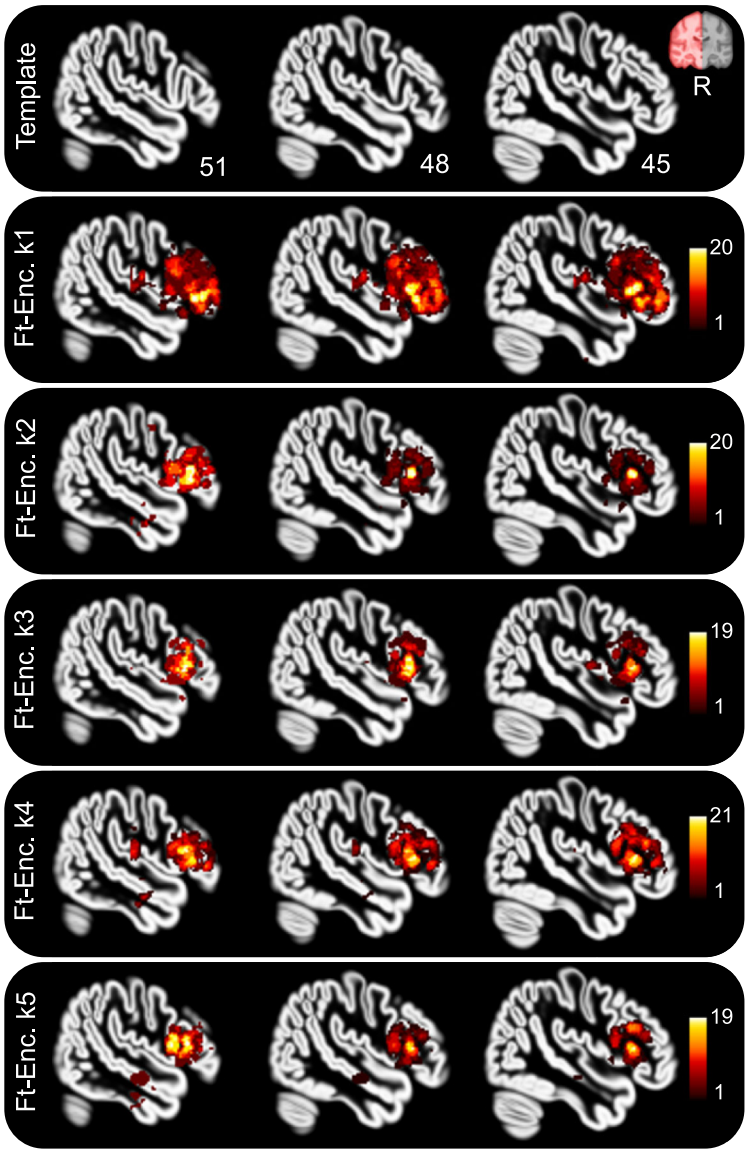

Fig. 9 Occlusion frequency maps for the right hemisphere. This maps were calculated combining the results from the occlusion maps of each subject for the *Ft-Encoder* models from k_1 to k_5. From left to right: slices from the sagittal plane. Top row: gray matter MNI152 template image used as reference

the model incorrectly interpreted it as a small *type III a ds* (Fig. 11, cases *e* and *f*). In some other FP cases, the triangular sulcus was present within the patch (cases *g* and *h*) and its presence misled the correct identification of the *ds* by our model.

Case Study

Finally, we applied our model in a population study. We utilized *Ft-Encoder* to classify the prevalence of *ds* within a specific study group. We employed a subset of 1715 individuals from the dataset described in Section "OASIS-ADNI" to apply the *Ft-Encoder* model. Since the original dataset contained multiple images of the same individual, we selected only one image per subject. To analyze the prevalence of *ds* among different study groups, we included normal controls, individuals with Alzheimer's disease, and those with mild cognitive impairment. Our model was then applied, and the presence of *ds* was reported. Additionally, we analyzed the population by gender and by condition, reporting the prevalence of *ds* in each of the analyzed subgroups. The results are detailed in Table 2.

To validate these results, we conducted a visual confirmation on a subset of images. An expert individually analyzed 60 randomly selected images from the dataset, approximately 5% of the total. We compared the manual labeling with our model's results, achieving an F1-score of 0.7663. We observed that, according to our model's results, *ds* was present in 73.46% of the individuals, occurring in 50.20% of the studied hemispheres. It was found more frequently in the left hemisphere. This was consistent for both female and male patients, as well as for normal controls,

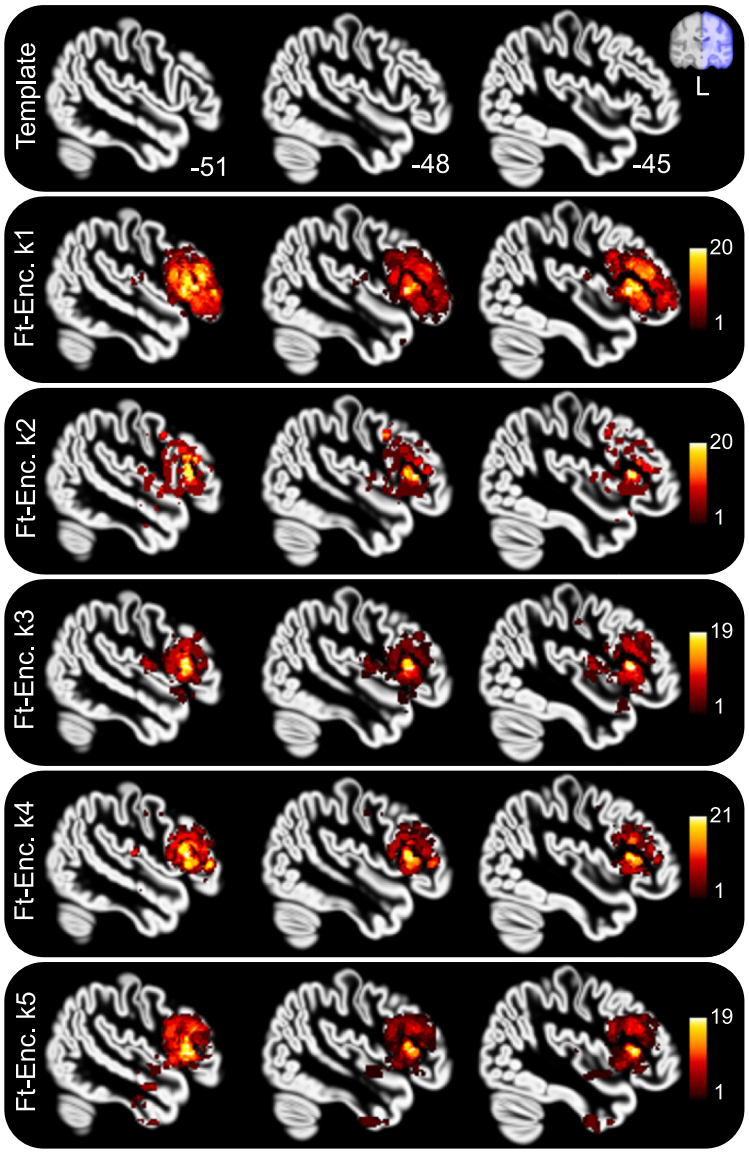

Fig. 10 Occlusion frequency maps for the left hemisphere. This maps were calculated combining the results from the occlusion maps of each subject for the *Ft-Encoder* models from k_1 to k_5. From left to right: slices from the sagittal plane. Top row: gray matter MNI152 template image used as reference

patients with Alzheimer's disease and mild cognitive impairment. Additionally, we observed a higher prevalence of *ds* in both hemispheres among the male population compared to females. This trend was also observed in patients with Alzheimer's disease compared to those with mild cognitive impairment and normal controls.

Discussion

The identification of tertiary sulci on brain images is challenging because they are shallower and not always present in all individuals. Several studies (Weiner et al., 2014; Garrison et al., 2015; Voorhies et al., 2021; Yao et al., 2023) have investigated the importance of these sulci in brain functionality in general, while other works have specifically focused on the *FO* (Sprung-Much & Petrides, 2018; Vallejo-Azar et al., 2023). The *FO* is an essential cortical region for expressive speech. Accurate identification of Broca's area and its adjacent structures is essential for avoid misdiagnosis and to protect them during surgical interventions. Moreover, a comprehensive understanding of the sulcal morphology of this area is also important for the neuroimaging studies exploring language. To our knowledge, this is the first work that automatically detects the *ds* using a deep learning approach.

Employing a self-supervised approach with brain patches from the area of interest, combined with supervised fine-tuning using labeled brain data, outperformed a standard software and two simpler models: one trained from scratch with a supervised deep learning approach and another pre-trained with 3D unlabeled medical data, then fine-tuned with labeled data. We quantified this improvement, as shown

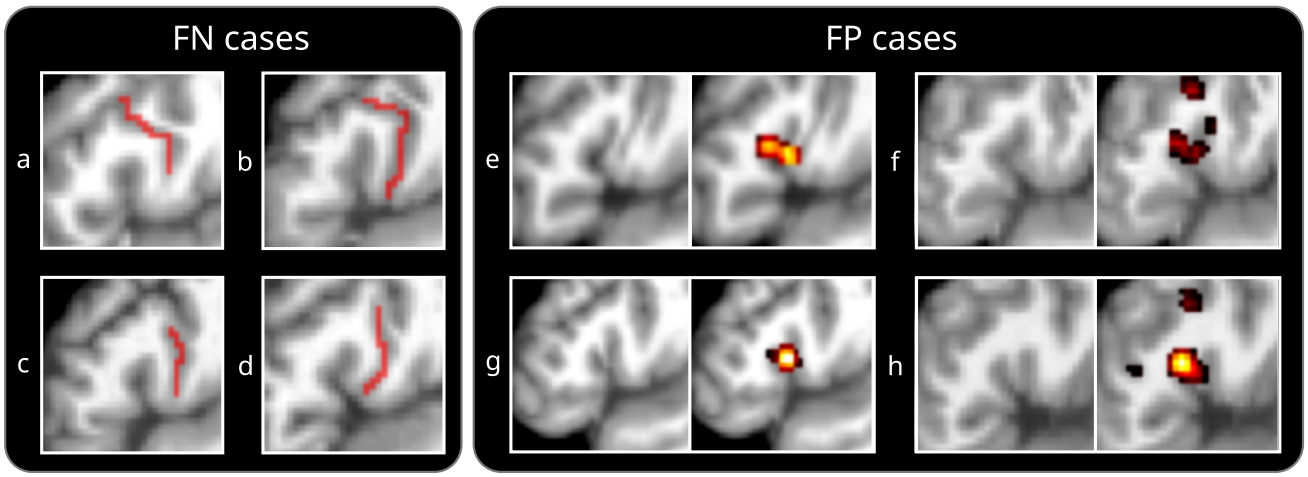

Fig. 11 Misclassified examples. The left subfigure displays instances where the model failed to identify the *ds* present in the patches, with its corresponding segmentation marked in red. Cases a and b correspond to a *type III c ds*. Cases c and d correspond to a *type II ds*. The right subfigure exhibits FP examples with their corresponding occlusion maps. On the first row, cases e and f correspond to two instances where the ascending ramus of the lateral fissure sulcus was curved, and the model interpreted it as a *type III a ds*. On the bottom row, cases g and h correspond to two different instances where the triangular sulcus was present, and the model misclassified it as the *ds*

in Fig. 6, where all the metrics considered were enhanced by applying our model. This combined approach of self-supervision and fine-tuning with brain data demonstrates strong performance on the classification task without requiring large volumes of labeled data.

We found that the *Ft-Encoder* is robust and can be applied to datasets with characteristics different from the training data. Conducting a quantitative analysis of *Ft-Encoder*'s performance on a new, held-out dataset (Fig. 6b), we confirmed its ability to achieve similar metrics in detecting the *ds* in patches taken from this new dataset. This demonstrates the utility of our model for application to new images without the need for retraining. Additionally, comparing its results with those obtained using *BrainVisa*, we observed that *Ft-Encoder* outperformed it. While our method demonstrated strong performance across different datasets, we did not assess its effectiveness on various image modalities. We encourage using the Ft-Encoder with T1-weighted images, as they are considered the optimal choice for segmenting brain tissues. T1-weighted images provide better contrast between different brain structures, facilitating the identification of brain tissues.

Table 2 Prevalence of *ds* in a case study

	Left (%)	Right (%)	Both (%)	Either (%)
All	51.45	48.86	26.93	73.46
Male	55.12	51.73	30.75	76.11
Female	48.19	46.16	23.36	70.99
NC	49.21	47.86	25.39	71.69
AD	56.28	52.09	31.67	76.70
MCI	52.68	48.11	26.07	74.73

NC: Normal Control, AD: Alzheimer's Disease, MCI: Mild Cognitive Impairment. Values correspond to the percentage of subjects presenting *ds* in the left hemisphere, right hemisphere, both hemispheres, and either hemisphere. The population was analyzed by gender and divided based on the presence of Alzheimer's disease or mild cognitive impairment

We qualitatively analyzed the cases when *Ft-Encoder* mislabeled the data. We can observe that most FN cases correspond to small *ds* or less frequent types of *ds* as *type II* or *type III c* (Fig. 11). This distribution aligned with our expectations, as the training data had more samples of *type III a* sulcus, making it more challenging for the model to generalize and learn features from the less common sulcus types. We can presume that this could be minimized by using a larger dataset during training, containing more samples of each type of sulcus, or by providing morphological information to the model, such as the type of the sulcus or a segmentation mask of the *ds*. Additionally, we observed that most FP cases corresponded to patches with a curvature in the ascending ramus of the lateral fissure or with the triangular sulcus present. This could also be improved by adding more labeled data to the training set, including label information about the triangular sulcus, which is another tertiary and inconstant sulcus of the *FO*.

One of the main constraints in understanding deep learning approaches is the difficulty in interpreting results and visualizing the features learned by the classifier. In this work, we visualized the areas considered by the *Ft-Encoder* to make decisions in the classification. We used a model interpretability approach and created occlusion maps highlighting voxels used for determining the presence of the *ds*. We

observed that the *Ft-Encoder* does not always focus solely on the *ds* but also examine the surrounding area, primarily in the ascending ramus of the lateral fissure sulcus. These results are aligned with the main areas used as guidelines during manual classification. Occlusion maps could also serve as an additional output of the classifier, highlighting important areas and guiding the experts in the correct identification of the sulcus. They can also give an insight of the local morphology when the sulcus is present.

We analyzed the challenges of manually labeling the data. It was evident that during the image labeling process, the level of interobserver agreement varied depending on the anatomical characteristics of each image. Not only did we identify labeling difficulties, but we also classify the extent of interobserver reliability (Fig. 7). We concluded that the labeling process is a challenging task, even for experts in the field, and it is not always possible to achieve complete agreement. Additionally, we found that our method exhibited a solid performance in cases where observers reached a perfect agreement and this performance was reduced as the certainty in the labeling process decreased. This analysis underscores the need for a robust automatic labeling method that can assist experts in arriving at a consensus.

Examining the variability in the labeling process, we can conclude that labeling images manually would require a group of well-trained experts to carefully discuss and analyze each image to reach an agreement. This process presents various challenges, including being time-consuming and necessitating specific training and expertise in the field, which is not easily attainable. Our method can mitigate this issue by automatically detecting the *ds* with no need of human intervention and minimizing inter and intra-individual differences in the labeling process.

Given the reported difficulty in identifying the *ds*, the amount of labeled data was limited. In this work, we focused on developing a tool capable of detecting *ds* and analyzing its prevalence across different populations. By combining the classification task with saliency maps, we could also analyze the morphological characteristics of Broca's area and identify the main patterns contributing to *ds* detection.

Our initial approach to the *ds* identification problem concentrated on accurately detecting the presence of the *ds* and optimizing the model for this task. While this approach lacks sulcus localization, we supplement the output with saliency maps as an additional result. Another valuable output of the model would be to provide users with *ds* subtype information. Based on these initial findings, it would be interesting to explore training a different network capable of generating a voxel mask as output, offering more detailed subtype information in subsequent stages of this research.

To achieve such precision, expanding the training dataset by labeling more images would be beneficial. Our approach could contribute to a future tool that not only identifies the *ds* but also facilitates the delineation of 3D masks of the sulcus. Furthermore, the weights of our model could be utilized as a pre-training stage for the encoder of a U-Net-like architecture, potentially reducing the need for a large number of samples for training.

Finally, we applied *Ft-Encoder* to analyze the prevalence of the *ds* in a studied population and found a prevalence of the *ds* in a 73.46% of the individuals. Moreover, we detected the presence of the *ds* in 50.20% of the studied hemispheres, which aligns with the results reported by Sprung-Much and Petrides (2018). Additionally, we observed that the *ds* is more frequently observed in the left hemisphere than in the right hemisphere, this is consistent across all subgroups analyzed and with the results reported by Vallejo-Azar et al. (2023).

In this work, we achieved an F1-score of 0.7176 evaluating on the test data and a F1-score of 0.7272 on a held-out dataset. As we mentioned, the performance of our model could likely be enhanced by incorporating a larger dataset, specifically one labeled by a more extensive group of well-trained experts in the field. Engaging a group of specialists for the labeling process could significantly improve the level of inter-rater agreement, thereby providing more consistent and reliable annotations. This consistency is crucial for training robust models, as it allows the model to better learn and understand the anatomical characteristics of the images. Furthermore, a larger and more diverse dataset would expose the model to a wider range of variations in anatomical features and pathologies, which could improve its generalization capabilities. This would not only facilitate more accurate decision-making but also enhance the model's ability to identify subtle differences in imaging that are critical for classification. By training on a larger dataset with well-defined labels, we could mitigate the risk of overfitting and improve the overall performance of our method. While leveraging a larger dataset would likely further reduce the risk of overfitting, we achieved promising results even with a smaller dataset.

Structural MRI is commonly used to explore potential variations in brain morphology among both clinical and healthy control groups (Lee et al., 2020; McCarthy et al., 2018; Vijayakumari et al., 2023). A key aspect of this distinction is the emphasis on reproducibility in MRI findings, which aims for the standardization and optimization of imaging methods across different settings and populations. Ensuring consistency in MRI protocols is essential for reliable comparisons, especially when investigating subtle anatomical differences that may have significant clinical implications. Our main objective in this work is to help avoid misdiagnosis, as the presence of the *ds* is a frequent anatomical finding that can potentially influence the results of automatic gray matter quantification performed by popular software tools (Vallejo-Azar et al., 2023). Misinterpretations related to the *ds* can lead to erroneous assessments

of brain morphology, which may impact clinical decision-making and research outcomes. Our work contributes to future exploratory studies that could investigate the correlation between *ds* and biomarkers associated with neurological conditions. Such studies could enhance our understanding of the implications of *ds* in both healthy and clinical populations, ultimately improving diagnostic accuracy and patient care.

Conclusions

We developed an automatic and robust tool to assist in identifying the *ds* in brain MRI by using a limited number of labeled data. To the best of our knowledge, this is the first article focused on locating the *ds*. We employed existing machine learning techniques, combining them to effectively detect the *ds* despite the limited amount of labeled data. As a result, we present a useful tool to researchers interested in studying Broca's area efficiently.

We conducted a thorough analysis of its performance, providing various metrics and a detailed examination of cases where the model disagreed with manual labeling. Additionally, we presented a visual analysis of how the model determines the presence of the sulcus. We investigated which regions the model considered for detecting the *ds* and performed a separate analysis for both hemispheres. We also applied our model to a population study, analyzing *ds* prevalence on a study group. The importance of a tool that assists in identification and labeling becomes evident due to the challenges associated with manually labeling the corresponding sulcus, as we explored in this study. The significance of identifying tertiary sulci is justified and supported by numerous studies emphasizing their importance in various contexts (Weiner et al., 2014; Garrison et al., 2015; Voorhies et al., 2021; Miller & Weiner, 2022; Willbrand et al., 2022; Yao et al., 2023) specially in the opercular area, as cited in the literature (Keller et al., 2007; Knaus et al., 2007; Sprung-Much & Petrides, 2018; Vallejo-Azar et al., 2023).

In this initial approach, the focus was on accurately identifying the sulcus and adjusting the *Ft-Encoder* model to achieve this. Results providing an accurate segmentation mask of the *ds* and information about its subtype would be an interesting next step in this research. Based on these initial findings, our future work will focus on providing valuable *ds* information such as its localization, segmentation, and subtype.

Information Sharing Statement

The data used in this article includes MRI images from various public datasets and an in-house dataset. The IXI dataset is available at http://braindevelopment.org/ixidataset, the ADNI dataset is available at http://adni.loni.usc.edu/, and the OASIS dataset is available at https://www.oasisbrains.org/. The private data (HEC-HR dataset) used in this study is not publicly available to protect patient privacy. The code used in this study is available on GitHub at https://github.com/hkulsgaard/sulcus.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12021-024-09700-7.

Acknowledgements We thank the technician Sergio Morganti, who acquired the images; the volunteers at III Normal Anatomy Department, Facultad de Medicina, Universidad de Buenos Aires (Argentina): Santiago Lasalle, César Gómez, Melanie Catena Baudo, Martina Arfili Perez y Lucía Canestrari who participated in our study in the process of image labeling; and the volunteers of Hospital El Cruce and Hospital Angel Roffo for their participation in this study.

Author Contributions Conception and study design: D.B., H.C.K., J.I.O., I.L. Data processing and analysis: D.B., H.C.K. Data Acquisition: M.V., M.B., P.G., L.A. Software development: D.B, H.C.K. Interpretation of results: D.B., H.C.K., J.I.O., I.L. Drafting the manuscript work and revising it: All authors Approval of final version to be published: All authors

Funding This work was partially funded by PIP GI 2021-2023 - 11220200102472CO (CONICET) and PICT 2016-0116 (ANPCyT).

Data Availability Data used in this article comprises MRI images taken from various public datasets and an in-house dataset. The IXI dataset is available at http://braindevelopment.org/ixi-dataset, the ADNI dataset is available at http://adni.loni.usc.edu/, and the OASIS dataset is available at https://www.oasisbrains.org/.
The private data (HEC-HR dataset) used in this study is not publicly available to protect patient privacy.

Code Availability Code is available on GitHub at https://github.com/hkulsgaard/sulcus.

Declarations

Conflicts of Interest The authors declare that there are no conflicts of interest in this work.

Consent Written informed consent was obtained from all patients/participants prior to their involvement in the study.

Ethics Approval The ethical committees of Hospital El Cruce and Hospital Angel Roffo in Buenos Aires, Argentina, reviewed and approved the studies involving human participants.

Competing Interests The authors declare no competing interests.

References

Akula, S. K., Exposito-Alonso, D., & Walsh, C. A. (2023). Shaping the brain: The emergence of cortical structure andfolding. *Developmental Cell, 58*(24), 2836–2849.

Amiez, C., & Petrides, M. (2018). Functional rostro-caudal gradient in the human posterior lateral frontal cortex. *Brain Structure and Function, 223*(3), 1487–1499.

Borne, L., Rivière, D., Mancip, M., & Mangin, J.-F. (2020). Automatic labeling of cortical sulci using patch-or cnn-based segmentation techniques combined with bottom-up geometric constraints. *Medical Image Analysis, 62*, 101651.

Chen, S., Ma, K., & Zheng, Y. (2019). Med3d: Transfer learning for 3d medical image analysis. arXiv preprint arXiv:1904.00625

Cointepas, Y., Mangin, J.-F., Garnero, L., Poline, J.-B., & Benali, H. (2001). Brainvisa: Software platform for visualization and analysis of multi-modality brain data. *Neuroimage, 13*(6), 98.

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P., Hyman, B. T., Albert, M. S., & Killiany, R. J. (2006). An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. *Neuroimage, 31*(3), 968–980.

Destrieux, C., Fischl, B., Dale, A., & Halgren, E. (2010). Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *Neuroimage, 53*(1), 1–15.

Fedorenko, E., & Blank, I. A. (2020). Broca's area is not a natural kind. *Trends in Cognitive Sciences, 24*(4), 270–284.

Fernández, V., & Borrell, V. (2023). Developmental mechanisms of gyrification. *Current Opinion in Neurobiology, 80*, 102711.

Garrison, J. R., Fernyhough, C., McCarthy-Jones, S., Haggard, M., 7, A. S. R. B. C. V...S. U.. S. R.. J. A.. M. B.. M. P.. C. S.. H. F.. P. C...L. C., & Simons, J. S. (2015). Paracingulate sulcus morphology is associated with hallucinations in the human brain. *Nature communications, 6*(1), 8956.

Gaser, C., Dahnke, R., Thompson, P. M., Kurth, F., Luders, E., & Initiative, A. D. N. (2022). Cat–a computational anatomy toolbox for the analysis of structural mri data. *biorxiv,* 2022–06.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770–778).

IXI Dataset (2012). https://brain-development.org/ixi-dataset/

Keller, S. S., Crow, T., Foundas, A., Amunts, K., & Roberts, N. (2009). Broca's area: Nomenclature, anatomy, typology and asymmetry. *Brain and Language, 109*(1), 29–48.

Keller, S. S., Highley, J. R., Garcia-Finana, M., Sluming, V., Rezaie, R., & Roberts, N. (2007). Sulcal variability, stereological measurement and asymmetry of broca's area on mr images. *Journal of Anatomy, 211*(4), 534–555.

Knaus, T. A., Corey, D. M., Bollich, A. M., Lemen, L. C., & Foundas, A. L. (2007). Anatomical asymmetries of anterior perisylvian speech-language regions. *Cortex, 43*(4), 499–510.

LaMontagne, P. J., Benzinger, T. L., Morris, J. C., Keefe, S., Hornbeck, R., Xiong, C., Grant, E., Hassenstab, J., Moulder, K., Vlassenko, A. G., Raichle, M. E., Cruchaga, C., & Marcus, D. (2019). Oasis-3: Longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and alzheimer disease. *MedRxiv,* 2019–12.

Lee, P., Kim, H.-R., Jeong, Y., & Initiative, A. D. N. (2020). Detection of gray matter microstructural changes in alzheimer's disease continuum using fiber orientation. *BMC Neurology, 20*, 1–10.

Li, X., Morgan, P. S., Ashburner, J., Smith, J., & Rorden, C. (2016). The first step for neuroimaging data analysis: Dicom to nifti conversion. *Journal of neuroscience methods, 264*, 47–56.

McCarthy, J., Collins, D. L., & Ducharme, S. (2018). Morphometric mri as a diagnostic biomarker of frontotemporal dementia: A systematic review to determine clinical applicability. *NeuroImage: Clinical, 20*, 685–696.

Miller, J. A., Voorhies, W. I., Lurie, D. J., D'Esposito, M., & Weiner, K. S. (2021). Overlooked tertiary sulci serve as a meso-scale link between microstructural and functional properties of human lateral prefrontal cortex. *Journal of Neuroscience, 41*(10), 2229–2244.

Miller, J. A., & Weiner, K. S. (2022). Unfolding the evolution of human cognition. *Trends in Cognitive Sciences, 26*(9), 735–737.

Mueller, S. G., Weiner, M. W., Thal, L. J., Petersen, R. C., Jack, C., Jagust, W., Trojanowski, J. Q., Toga, A. W., & Beckett, L. (2005). The alzheimer's disease neuroimaging initiative. *Neuroimaging Clinics, 15*(4), 869–877.

Ono, M., Kubik, S., & Abernathey, C. D. (1990). *Atlas of the Cerebral Sulci*. New York: Thieme.

Penny, W. D., Friston, K. J., Ashburner, J. T., Kiebel, S. J., & Nichols, T. E. (2011). *Statistical Parametric Mapping: The Analysis of Functional Brain Images*. London: Elsevier.

Pérez-García, F., Sparks, R., & Ourselin, S. (2021). Torchio: A python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. *Computer Methods and Programs in Biomedicine, 208*, 106236.

Perrot, M., Rivière, D., & Mangin, J.-F. (2011). Cortical sulci recognition and spatial normalization. *Medical image analysis, 15*(4), 529–550.

Sprung-Much, T., Eichert, N., Nolan, E., & Petrides, M. (2022). Broca's area and the search for anatomical asymmetry: Commentary and perspectives. *Brain Structure and Function, 227*(2), 441–449.

Sprung-Much, T., & Petrides, M. (2018). Morphological patterns and spatial probability maps of two defining sulci of the posterior ventrolateral frontal cortex of the human brain: The sulcus diagonalis and the anterior ascending ramus of the lateral fissure. *Brain Structure and Function, 223*, 4125–4152.

Troiani, V., Patti, M. A., & Adamson, K. (2020). The use of the orbitofrontal h-sulcus as a reference frame for value signals. *European Journal of Neuroscience, 51*(9), 1928–1943.

Vallejo-Azar, M. N., Alba-Ferrara, L., Bouzigues, A., Princich, J. P., Markov, M., Bendersky, M., & Gonzalez, P. N. (2023). Influence of accessory sulci of the frontoparietal operculum on gray matter quantification. *Frontiers in Neuroanatomy, 16*, 134.

Vijayakumari, A. A., Fernandez, H. H., & Walter, B. L. (2023). Mri-based multivariate gray matter volumetric distance for predicting motor symptom progression in parkinson's disease. *Scientific Reports, 13*(1), 17704.

Voorhies, W. I., Miller, J. A., Yao, J. K., Bunge, S. A., & Weiner, K. S. (2021). Cognitive insights from tertiary sulci in prefrontal cortex. *Nature Communications, 12*(1), 5122.

Weiner, K. S., Golarai, G., Caspers, J., Chuapoco, M. R., Mohlberg, H., Zilles, K., Amunts, K., & Grill-Spector, K. (2014). The mid-fusiform sulcus: A landmark identifying both cytoarchitectonic and functional divisions of human ventral temporal cortex. *Neuroimage, 84*, 453–465.

Welker, W. (1990). Why does cerebral cortex fissure and fold? a review of determinants of gyri and sulci. *Cerebral Cortex: comparative structure and evolution of Cerebral Cortex, Part, II,* 3–136.

Willbrand, E., Parker, B., Voorhies, W., Miller, J., Lyu, I., Hallock, T., Aponik-Gremillion, L., Koslov, S., Null, N., Bunge, S., Foster, B. L., & Weiner, K. S. (2022). Uncovering a tripartite landmark in posterior cingulate cortex. *Science Adventure, 8*, eabn9516.

Williams, L. Z., Fitzgibbon, S. P., Bozek, J., Winkler, A. M., Dimitrova, R., Poppe, T., Schuh, A., Makropoulos, A., Cupitt, J., O'Muircheartaigh, J., Duffc, E. P., Cordero-Grande, L., Price, A. N., Hajnal, J. V., Rueckert, D., Smith, S. M., Edwards, A. D., & Robinson, E. C. (2023). Structural and functional asymmetry of the neonatal cerebral cortex. *Nature Human Behaviour,* 1–14.

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

Yang, S., Zhao, Z., Cui, H., Zhang, T., Zhao, L., He, Z., Liu, H., Guo, L., Liu, T., Becker, B., Kendrick, K. M., & Jiang, X. (2019). Temporal variability of cortical gyral-sulcal resting state functional activity correlates with fluid intelligence. *Frontiers in Neural Circuits, 13*, 36.

Yao, J. K., Voorhies, W. I., Miller, J. A., Bunge, S. A., & Weiner, K. S. (2023). Sulcal depth in prefrontal cortex: A novel predictor of working memory performance. *Cerebral Cortex, 33*(5), 1799–1813.

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13* (pp. 818–833). Springer.

Authors and Affiliations

Delfina Braggio[1,2] · Hernán C. Külsgaard[1,2] · Mariana Vallejo-Azar[1,3] · Mariana Bendersky[4] · Paula González[1,3] · Lucía Alba-Ferrara[3,5] · José Ignacio Orlando[1,2] · Ignacio Larrabide[1,2]

✉ Delfina Braggio
delfinabraggio@pladema.exa.unicen.edu.ar

Hernán C. Külsgaard
hkulsgaard@pladema.exa.unicen.edu.ar

1 Consejo Nacional de Investigaciones Científicas y Técnicas, CONICET, Buenos Aires, Argentina

2 Yatiris, PLADEMA, Facultad de Ciencias Exactas, UNICEN, Tandil, Buenos Aires, Argentina

3 Unidad Ejecutora de Estudios en Neurociencias y Sistemas Complejos, CONICET, UNAJ, Hospital El Cruce, Florencio Varela, Argentina

4 Laboratorio de Anatomía Viviente, Facultad de Medicina, Universidad de Buenos Aires, Ciudad Autónoma de Buenos Aires, Buenos Aires, Argentina

5 Facultad de Ciencias Biomédicas, Universidad Austral, Ciudad Autónoma de Buenos Aires, Buenos Aires, Argentina