



HAL
open science

Multilabel SegSRGAN - A framework for parcellation and morphometry of preterm brain in MRI

Guillaume Dollé, G. Loron, Margaux Alloux, Vivien Kraus, Quentin Delannoy, Jonathan Beck, Nathalie Bednarek, François Rousseau, Nicolas Passat

► To cite this version:

Guillaume Dollé, G. Loron, Margaux Alloux, Vivien Kraus, Quentin Delannoy, et al.. Multilabel SegSRGAN - A framework for parcellation and morphometry of preterm brain in MRI. PLoS ONE, In press. hal-04576760v1

HAL Id: hal-04576760



<https://hal.science/hal-04576760v1>

Submitted on 16 May 2024 (v1), last revised 21 Oct 2024 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Multilabel SegSRGAN — A framework for parcellation and morphometry of preterm brain in MRI

Guillaume Dollé¹^{*}, Gauthier Loron^{2,3}, Margaux Alloux^{3,4}, Vivien Kraus², Quentin Delannoy², Jonathan Beck³, Nathalie Bednarek^{2,3}, François Rousseau⁵, Nicolas Passat²


1 Université de Reims Champagne Ardenne, CNRS, LMR, UMR 9008, Reims, France

2 Université de Reims Champagne Ardenne, CRESTIC, Reims, France

3 Service de médecine néonatale et réanimation pédiatrique, CHU de Reims, France

4 Unité d'aide méthodologique - Pôle Recherche, CHU de Reims, France

5 IMT Atlantique, LaTIM INSERM U1101, 29238 Brest, France

 These authors contributed equally to this work.

* guillaume.dolle@univ-reims.fr

Abstract

Magnetic Resonance Imaging (MRI) is a powerful tool for observing and assessing the properties of cerebral tissues and structures. In particular, in the context of neonatal care, MR images can be used to analyze neurodevelopment issues that may occur in the preterm newborn. However, the intrinsic properties of the newborn MR images, combined with the high variability of MR acquisition in a clinical context, result in complex and heterogeneous images. It is then challenging to accurately compute and analyze morphometric biomarkers inferred from MRI of newborn. This task is often carried out in 2-dimensions with interactive tools. In recent works, we proposed a new method, namely SegSRGAN, designed for various image processing and analysis tasks, including the segmentation of specific brain structures. In this article, we first propose an extension of SegSRGAN from binary segmentation to multilabel segmentation, then leading to a parcellation of an MR image into several labels, each corresponding to a specific brain tissue / area. Second, we propose a quality control protocol dedicated to assess the performance of our proposed method with respect to this specific parcellation task in neonatal MR imaging. In particular, we combine scores derived from experts' analysis, from morphometric measures and from topological properties of the investigated structures. Based on this protocol, we study the strengths and weaknesses of SegSRGAN and its potential ability to be used for clinical research in the context of morphometric analysis of preterm brain structure, and to possibly design new biomarkers of neurodevelopment. The proposed study involves MR images from the EPIRMEX dataset, collected in the context of a national cohort study.

1 Introduction 1

1.1 Context and objectives 2

Prematurity is still associated with a high risk of neurodevelopmental impairment [1, 2]. 3
The neurodevelopmental outcome of this population is a significant concern in terms of 4
public health, due to the increased survival of extremely preterm infants [2]. Identifying 5
risk factors of impaired neurodevelopment and high-risk infants stays a prioritized need 6

for optimizing brain development, neuroprotection and providing adapted care. White matter injuries are the most frequently diagnosed lesions in preterm infants in association with neurodevelopmental impairment [3]. These injuries are associated with axonal and neuronal abnormalities, involving structures such as basal ganglia, brainstem, cerebellum and cortex. These lesions are consequences of direct (inflammation direct impact on white matter) and impaired maturative processes (altered neurogenesis and synaptogenesis), known as the encephalopathy of prematurity [4].

Brain Magnetic Resonance Imaging (MRI) at term equivalent age is recommended for identifying such structural lesions [5]. However, the neurodevelopmental trajectory is not exclusively correlated to these brain injuries visible on qualitative MRI. Of note, preterm birth is associated with growth impaired brain volume, even in the absence of brain injury [6, 7]. It is also demonstrated that cerebrum and cerebellar growth is associated with neurocognitive outcomes [8, 9]. For all these reasons, the analysis of brain MRI at term corrected requires a structural and volumetric approach to attempt of prediction the long-term neurological outcome in preterm infants. MRI brain volumes of gray matter and white matter as well as regional volumes may identify biomarkers for the evaluation of the impact of preterm birth and its adverse effects.

Brain MRI segmentation is a sophisticated method explored and developed for the last two decades [10]. At present time and despite large literature, neonatal MRI segmentation [11] is still a research tool and has no application in routine. In [12], we recently proposed a new segmentation method, namely SegSRGAN, which was specifically dedicated to neonatal brain MRI segmentation. SegSRGAN relies on the paradigm of Generative Adversarial Networks (GAN) and aims to provide both a super-resolution (SR) reconstruction of the neonatal MR images (often acquired at a low resolution) and a segmentation of cerebral structures at the super-resolution level. In [12], the relevance of SegSRGAN was already proved by comparison with various state-of-the-art methods, especially regarding the challenging problem of cortex segmentation.

In this article, we propose a methodological and experimental framework, built upon SegSRGAN, which is dedicated to parcellation and morphometric analysis of brain structures from preterm MR images. In particular, our contributions are threefold.

First, we propose a multilabel version of SegSRGAN. The initial version of the method, proposed in [12], could perform the binary segmentation, i.e. the extraction of one specific kind of tissue. The new multilabel SegSRGAN, proposed in this article, is now able to perform multilabel segmentation, i.e. the extraction of an arbitrary number of specific kinds of tissues, thus leading to the possibility to propose a parcellation of the whole brain into regions of interest chosen by the user.

Second, we propose a quality control (QC) strategy for brain MR image parcellation, dedicated to preterm issues. This QC strategy, inspired by recent efforts of the community towards these issues, relies on three main categories of evaluations: (1) qualitative assessment by clinical experts, that aims to link the visual quality of the parcellation with standard quality scores usually considered for segmentation; (2) quantitative assessment of the segmentation by comparison between the morphometric measures carried out manually by clinical experts, and automatically from the segmentation; and (3) quantitative assessment of the topological correctness of the segmentation by correlation of connectivity and adjacency measures between the segmented regions and the ground truth regions used for training the method.

Third and last, we experimentally assess the quality of the multilabel SegSRGAN. To this end, we consider MR images acquired in a clinical context. These images are part of a national cohort, namely EPIRMEX. The purpose of this QC of SegSRGAN on “real” data is to validate the approach and determine the strengths, limits and biases as a prerequisite to its involvement into the processing of the whole cohort for further

clinical studies.

The remainder of this article is organized as follows. In Section 2, we briefly describe recent works in the different domains connected to the topics of this article, namely clinical aspects of brain MRI analysis, neonatal brain segmentation and QC of brain MRI segmentation. In Section 3.1, we present SegSRGAN. We first recall the initial, binary version of the method. Then, we present its extension in order to handle the case of multilabel segmentation (i.e. parcellation) of the brain from MR images. We also describe a post-processing step for cleaning the results, especially with respect to extracranial artifacts. In Section 3.2, we describe our QC protocol. We detail its three modules, which are dedicated to qualitative, morphometric and topological assessment, respectively. In Section 3.3, we apply this QC protocol to the multilabel version of SegSRGAN on a dataset built from the EPIRMEX cohort. We provide the complete numerical results of this analysis and discuss on the strengths, biases and limits of SegSRGAN regarding its ability to explore a whole MR image cohort.

2 Related works

In this section, we describe some recent contributions related to the three main issues dealt with in this article: the clinical interest of preterm brain MRI analysis (Section 2.1); the recent methods for neonate brain segmentation / parcellation (Section 2.2); and the development of QC for brain MRI segmentation (Section 2.3).

2.1 Preterm brain MRI analysis: clinical aspects

Over the past three decades, MRI of the neonatal brain has shown that large, overt lesions are associated with severe neurological outcome [13]. High-grade haemorrhage and parenchymal infarcts are associated with cerebral palsy, low IQ and death [14]. Clinical consequences of venous infarcts (i.e. Volpe’s infarcts) vary with location and size [15]. Cerebellar infarcts have a significant impact on neurodevelopment outcome, especially when vermis, or both hemispheres, are involved [16]. Cystic white matter lesions are highly associated with cerebral palsy, but currently represent only 1% of white matter lesions. Overall, moderate and severe overt brain lesions on MRI are quite good predictors for cerebral palsy and severe neurodevelopmental delay [17–19]. However, these overt injuries are not the only potential consequences of premature birth on the developing brain. Many former preterm babies have mild to moderate neurodevelopmental disorders, including mild cognitive impairment, social cognition, neurodevelopmental disorders and learning disabilities, behavioral disorders [2, 20]. Brain MRI does a poor job of predicting these mild to moderate cognitive dysfunctions by analyzing overt lesions only.

Indeed, preterm birth induces diffuse alterations in brain developmental trajectories, including structural changes of the subplate, of neuro-axonal organization and cortical lamination [21]. In infants born preterm, advanced analysis of brain MRI has highlighted these structural and functional changes: gyration, structural and functional connectivity, regional volumes are altered in infants born preterm, with or without associated overt lesions. It is beyond the scope of this paper to describe all those alterations; the reader may find more information in dedicated reviews [22–24].

Finally, children born preterm exhibit alteration of regional brain volumes that persist in childhood [25, 26], and even in adulthood by drawing a morphological pattern of the “brain of infant born preterm” [27]. These alterations seem to correlate with neurodevelopmental prognosis [8, 28, 29]. The respective contribution of: (1) regional brain volumes [30], (2) their growth kinetics [30, 31] and (3) their asymmetry [24] for prognosis of neurodevelopment is still controversial and under research. In our opinion,

biases related to the methods of image processing and their validation must be systematically considered in these analyses, as the performance and validation of the tool may greatly contribute to the relevance of the biomarker considered.

2.2 Neonatal brain segmentation

Studying developing brain involves several major image analysis challenges which concern the development of appropriate approaches that can cope with low contrast-to-noise ratio, rapid change of size of brain structures, complex brightness changes in structural MRI reflecting rapid white matter structuring through myelination, rapid change and large variability of anatomical shapes. To address these challenges, many methods have been proposed in the literature [11, 32].

In image segmentation tasks, deep learning-based algorithms have been at the leading edge of development in recent years, including in neonatal brain imaging. The U-Net architecture [33], which provides a multiscale representation of the data, is probably the most widely used model in segmentation, especially for neonatal data [34, 35]. One can also mention the use of other architectures such that hyperdense-net [36], transformer weighted network [37] or attention-based networks [38].

In the context of neonatal brain imaging, deep learning segmentation algorithms are trained on large image databases, such as the dHCP project data [39], for which the ground truth has been estimated with the DrawEM method [40].

Deep learning methods have shown high quality segmentation results on these research databases. However, their application on clinical data remains a challenge because of the motion artifacts present in the images, the appearance variabilities of multisite data, and the anisotropic resolution of clinical data. To this end, Khalili et al. [41] proposed a method based on Generative Adversarial Networks (GANs) to reduce artifacts related to subject motion during acquisition. Grigorescu et al. [42] studied two unsupervised data adaptation methods to transfer learning from one database to another. Chen et al. [43] investigated the use of GAN methods for segmentation harmonization. Finally, Delannoy et al. [12] proposed a GAN-based method to reconstruct the data in highly isotropic resolution and jointly estimate a segmentation of the cortex.

In this work, we focus on the SegSRGAN method [12] to analyze anisotropic clinical data from the EPIRMEX cohort [28] associated with EPIPAGE 2 study [44].

2.3 Quality control for brain MRI segmentation

Quality control (QC) of brain segmentation is a key step in ensuring the results of a morphometric study. Manual QC strategies are currently the gold standard, although not being feasible for large neuroimaging samples. Automated QC options have been proposed, offering potential reproducible and time-efficient alternatives. For instance, we can mention Qoala-T [45] which is a supervised tool for QC of FreeSurfer segmentation maps, or MRIQC [46] that uses T1w or T2w images as input. Monereo et al. [47] recently investigated the impact of these two tools for QC and concluded that global morphological estimates should be avoided to detect outliers. This study has also shown that features like the Euler number could be useful to detect inaccurate segmentation maps. To the best of our knowledge, there is no QC study dedicated to neonatal brain MRI. In this work, we propose qualitative and quantitative scores to characterize the segmentation maps from neonatal brain MR images.

3 Materials and methods

3.1 Super-resolution reconstruction and segmentation – SegSRGAN

In this work, we aim to investigate the relevance of SegSRGAN for the analysis of neonatal brain MR images. SegSRGAN is a hybrid method based on Generative Adversarial Networks (GANs) [48], that aims to carry out simultaneously super-resolution (SR) reconstruction and segmentation of low-resolution images. Initially, the segmentation module of SegSRGAN was designed for binary segmentation. We first recall (Section 3.1.1) this initial method, that was published and validated by comparison with state of the art approaches in [12]. Then, we propose an extended version of SegSRGAN which is able to carry out multilabel segmentation, i.e. to provide a parcellation of the intracranial volume into different regions. We present the modifications of this new multilabel SegSRGAN vs. the binary SegSRGAN (Section 3.1.2). Since SegSRGAN is a pixel-based segmentation / parcellation approach, we also propose a post-processing procedure that aims to regularize the segmentation results in a region-based paradigm, in order to remove semantic noise (Section 3.1.3).

3.1.1 SegSRGAN: Reminder of the initial (binary) version

SegSRGAN is both a SR reconstruction and a segmentation method. We first discuss on its SR reconstruction side. A SR method aims at estimating a high resolution (HR) image $\mathbf{X} \in \mathbb{R}^m$ from a low resolution (LR) image $\mathbf{Y} \in \mathbb{R}^n$ ($m > n$). Such a problem can be formulated by a linear observation model:

$$\mathbf{Y} = H_{\downarrow} \mathbf{B} \mathbf{X} + N = \Theta \mathbf{X} + N \quad (1)$$

where $N \in \mathbb{R}^n$ is an additive noise, $B \in \mathbb{R}^{m \times m}$ is a scattering matrix, $H_{\downarrow} \in \mathbb{R}^{n \times m}$ is a decimation matrix, and $\Theta = H_{\downarrow} B \in \mathbb{R}^{n \times m}$.

A common way of tackling this SR problem is to define the matrix Θ^{-1} as the combination of a restoration operator $F \in \mathbb{R}^{m \times m}$ and an interpolation operator $S^{\uparrow} \in \mathbb{R}^{m \times n}$ that computes the interpolated LR image $\mathbf{Z} \in \mathbb{R}^m$ associated to \mathbf{Y} (i.e. $\mathbf{Z} = S^{\uparrow} \mathbf{Y}$). In the context of supervised learning, given a set of HR images \mathbf{X}_i and their corresponding LR images \mathbf{Y}_i , this restoration operator F can be estimated such that:

$$\hat{F} = \arg \min_F \sum_i d(\mathbf{X}_i - F(\mathbf{Z}_i)) \quad (2)$$

where d can be e.g. a ℓ_2 norm, a ℓ_1 norm or a differentiable variant of ℓ_1 such as defined in [49].

We now focus on the segmentation side of SegSRGAN. In order to handle the trade-off between the contributions of the SR image and the segmentation in the cost function, the image segmentation problem is seen as a supervised regression problem:

$$\mathbf{S}_{\mathbf{X}} = R(\hat{\mathbf{X}}) \quad (3)$$

where R is a non-linear function from the interpolated image $\hat{\mathbf{X}}$ to the segmentation map $\mathbf{S}_{\mathbf{X}}$. As for the SR problem, we assume that we have a set of interpolated images $\hat{\mathbf{X}}_i$ associated to the images \mathbf{X}_i together with their corresponding segmentation maps $\mathbf{S}_{\mathbf{X}_i}$. A general approach for solving this segmentation problem is to find the correspondence R such that:

$$\hat{R} = \arg \min_R \sum_i d(\mathbf{S}_{\mathbf{X}_i} - R(\hat{\mathbf{X}}_i)) \quad (4)$$

The GAN approaches rely on two networks. The first network, called generator, aims to estimate, for a given interpolated input image, the corresponding HR image and segmentation map. The second network, called discriminator, aims to differentiate the “real” couples of HR images and segmentation ones from the “generated” ones.

Cost function In order to avoid possible issues related to the gradient saturation that may occur with cost function, so-called “minimax”, usually considered in GANs, the alternative cost function WGAN-GP [50] is used. In this context, the purpose is to minimize the Wasserstein distance between two distributions \mathbb{P}_r and \mathbb{P}_g (corresponding here to the real and generated data):

$$W(\mathbb{P}_r, \mathbb{P}_g) = \inf_{\gamma \in \Pi(\mathbb{P}_r, \mathbb{P}_g)} \mathbb{E}_{(x,y) \sim \gamma} [\|x - y\|] \quad (5)$$

$$= \sup_{\|f\|_L \leq 1} \mathbb{E}_{x \sim \mathbb{P}_r} [f(x)] - \mathbb{E}_{x \sim \mathbb{P}_g} [f(x)] \quad (6)$$

where $\Pi(\mathbb{P}_r, \mathbb{P}_g)$ is the set of all the distributions which margins are \mathbb{P}_r and \mathbb{P}_g , respectively, and the supremum is computed within the 1-Lipschitz functions f .

Here, the discriminator learns the parametered function f while the generator aims to minimize the distance. Then, the antagonistic part of the cost function is:

$$\mathcal{L}_{adv} = \mathbb{E}_{\mathbf{X} \sim \mathbb{P}_{\mathbf{X}}, \mathbf{S}_{\mathbf{X}} \sim \mathbb{P}_{\mathbf{S}_{\mathbf{X}}}} [D((\mathbf{X}, \mathbf{S}_{\mathbf{X}}))] - \mathbb{E}_{\mathbf{Z} \sim \mathbb{P}_{\mathbf{Z}}} [D(G(\mathbf{Z}))] \quad (7)$$

where \mathbf{X} and $\mathbf{S}_{\mathbf{X}}$ are the true HR image and segmentation map, respectively, D is the discriminator, G is the generator and \mathbf{Z} is the interpolated image.

Finally, the cost function to minimize is:

$$\mathcal{L}_{dis} = \lambda_{gp} \mathbb{E}_{\widehat{\mathbf{X}\mathbf{S}}} [(\|\nabla_{\widehat{\mathbf{X}\mathbf{S}}} D(\widehat{\mathbf{X}\mathbf{S}})\|_2 - 1)^2] - \mathcal{L}_{adv} \quad (8)$$

with:

$$\widehat{\mathbf{X}\mathbf{S}} = (1 - \varepsilon)(\mathbf{X}, \mathbf{S}_{\mathbf{X}}) + \varepsilon G(\mathbf{Z}) \quad (9)$$

and $\varepsilon \sim U[0, 1]$, where ∇ and $\lambda_{gp} > 0$ are the gradient operator and its penalization coefficient, respectively.

The cost function of the generator is built by adding a pointwise comparison term ρ [49] between the target and the estimated images:

$$\mathcal{L}_{gen} = \lambda_{adv} \mathcal{L}_{adv} + \mathbb{E}_{\mathbf{X} \sim \mathbb{P}_{\mathbf{X}}, \mathbf{S}_{\mathbf{X}} \sim \mathbb{P}_{\mathbf{S}_{\mathbf{X}}}} [\rho((\mathbf{X}, \mathbf{S}_{\mathbf{X}}) - G(\mathbf{Z}))] \quad (10)$$

with:

$$\rho((x_1, \dots, x_{2m})) = \frac{1}{2m} \sum_{i=1}^{2m} \sqrt{(x_i^2 + \nu^2)} \quad (11)$$

and $\nu = 10^{-3}$.

Network architecture The generator network (Figure 1(a)) is a convolution-based network with residual blocks. It takes as input the interpolated LR image. It is composed of 18 convolutional layers: 3 for the encoding part, 12 for the residual part and 3 for the decoding part. Let $C_j^i-S^k$ be a block consisting of the following layers: a convolution layer of j filters of size i^3 with stride of k , an instance normalization layer (InsNorm) [51] and a rectified linear unit (ReLU). R_k denotes a residual block as Conv-InsNorm-ReLU-Conv-InsNorm that contains 3^3 convolution layers with k filters. U_k denotes layers as Upsampling-Conv-InsNorm-ReLU layers with k filters of 3^3 and stride of 1. The generator architecture is then: $C_{16}^7-S^1$, $C_{32}^3-S^2$, $C_{64}^3-S^2$, R_{64} , R_{64} , R_{64} , R_{64} , R_{64} , R_{64} , U_{32} , U_{16} , $C_2^7-S^1$. During the encoding, the number of kernels is

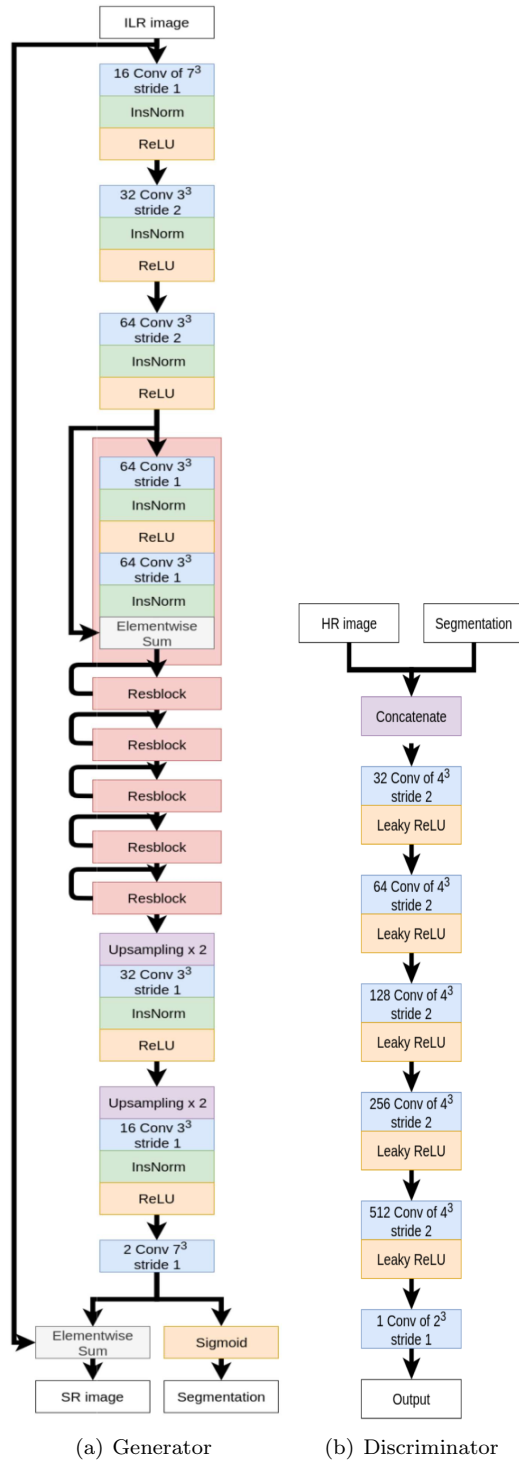


Fig 1. Initial SegSRGAN architecture. (a) Generator architecture. (b) Discriminator architecture.

multiplied by 2 at each convolution, from 16 to 64. The last convolutional layer produces two 3D images: the first will be turned into a class probability map (using a

218
219

sigmoid activation); the second will be summed with the original interpolated image. In order to improve the training procedure performance, instance normalization layers are used on the result of each convolution (before application of activation function).

The discriminator network (Figure 1(b)) is fully convolutional. It takes as input a HR image and a segmentation map. It contains 5 convolutional layers with an increasing number of filter kernels, increasing by a factor of 2 from 32 to 512 kernels. Let C_k be a block consisting of the following layers: a convolution layer of k filters of size 4^3 with stride of 2 and a Leaky ReLU with a negative slope of 0.01. The last layer C_1^2 is a 2^3 convolution filter with stride of 1. No activation layer is used after the last layer. The discriminator then consists of $C_{32}, C_{64}, C_{128}, C_{256}, C_{512}, C_1^2$.

For the generator as for the discriminator, the number of output channels for each convolutional layer is multiplied by 2 at each layer.

3.1.2 Multilabel SegSRGAN

The initial SegSRGAN method described in Section 3.1.1 has been extended in order to segment the intracranial volume into k labels ($k > 2$), with the hypothesis that each point x_i of the image \mathbf{X} be assigned a unique label. This multilabel extension mainly requires two modifications compared to the initial binary version.

First, the final part of the generator network dedicated to the segmentation now relies on k convolution modules (instead of one for the binary part). Each one of these convolution modules is dedicated to a specific label, and the output of the k convolutions is then merged to produce the final segmentation map.

Second, the error measure ρ which uniquely relied on the Charbonnier metric, defined in Eq. (11) now relies on two distinct metrics: Charbonnier for the SR reconstruction part and multilabel Dice for the segmentation segmentation part. The new measure ρ_{multi} is then defined as:

$$\begin{aligned} \rho_{\text{multi}}((\mathbf{X}, \mathbf{S}_{\mathbf{X}}), G(\mathbf{Z})) &= \rho_{\text{multi}}((\mathbf{X}, \mathbf{S}_{\mathbf{X}}), (\mathbf{X}^G, \mathbf{S}_{\mathbf{X}}^G)) \\ &= \rho_{\text{Charbonnier}}(\mathbf{X} - \mathbf{X}^G) + \rho_{\text{Dice}}(\mathbf{S}_{\mathbf{X}}, \mathbf{S}_{\mathbf{X}}^G) \end{aligned} \quad (12)$$

where $\rho_{\text{Charbonnier}}$ is defined as in Eq. (11) (by modifying $2m$ into m) and ρ_{Dice} is the multilabel version of the Dice measure [52]:

$$\begin{aligned} \rho_{\text{Dice}}(\mathbf{S}_{\mathbf{X}}, \mathbf{S}_{\mathbf{X}}^G) &= \frac{2 \cdot TP(\mathbf{S}_{\mathbf{X}}, \mathbf{S}_{\mathbf{X}}^G)}{2 \cdot TP(\mathbf{S}_{\mathbf{X}}, \mathbf{S}_{\mathbf{X}}^G) + FP(\mathbf{S}_{\mathbf{X}}, \mathbf{S}_{\mathbf{X}}^G) + FN(\mathbf{S}_{\mathbf{X}}, \mathbf{S}_{\mathbf{X}}^G)} \\ &= 2 \cdot (1 + m/TP(\mathbf{S}_{\mathbf{X}}, \mathbf{S}_{\mathbf{X}}^G))^{-1} \end{aligned} \quad (14)$$

with m the size of the image and TP , FP and FN the true positives, false positives, and false negatives, respectively.

3.1.3 Post-processing

The output of the segmentation process designed in the multilabel extension of SegSRGAN is a mapping $S : \Omega \rightarrow L$ where

$\Omega = \llbracket 0, \dim_x - 1 \rrbracket \times \llbracket 0, \dim_y - 1 \rrbracket \times \llbracket 0, \dim_z - 1 \rrbracket \subset \mathbb{Z}^3$ is the support of the MR image and $L = \{\ell_i\}_{i=0}^k$ is the set of labels, with ℓ_0 corresponding to the background (“no anatomical label”) and the k other ℓ_i corresponding each to a specific anatomical region.

The following two post-processing steps, mainly based on mathematical morphology and digital topology, aim to improve the quality of the result by removing artifacts and noise.

Extracranial artifact removal The proposed segmentation pipeline does not include a skull stripping preprocessing. Indeed, such approaches are sometimes not sufficiently robust and may induce in particular some false negative results in the intracranial region in case of failure. By contrast, we chose to process the whole MR image, which may lead to false positives in the extracranial regions, and to post-process the results to remove these artifacts afterwards, thus securing the results inside the intracranial region.

The most frequent artifacts are caused by an overestimation of the external cerebrospinal fluid (CSF), that may lead to leakage of the segmentation with the subsequent segmentation of specific extracranial structures, e.g. the eyes. Based on these assumptions, the proposed post-processing is as follows.

1. We build a first volume which is the principal connected component (noted $\mathcal{CC}(\cdot)$) of the part of Ω composed by the labels which are neither the background (BG) nor the CSF. This first (connected) volume is noted T . In particular, by noting X_\star the region of a given label \star , we have:

$$T = \mathcal{CC}(\Omega \setminus (X_{\text{BG}} \cup X_{\text{CSF}})) \quad (16)$$

We define a second volume V as the union of T and X_{CSF} . We then have $V = T \cup X_{\text{CSF}}$ with $T \cap X_{\text{CSF}} = \emptyset$.

2. Given a spherical structuring element B_ρ of radius ρ , we first apply an erosion of V by B_ρ . Then we preserve only the largest connected component of the result. We dilate this connected component by B_ρ and we finally recover the part T of V (which must not be discarded from the result). The overall process can be seen as a connectivity-based morphological opening [53] topologically constrained by the non-CSF brain tissues. It leads to the construction of a final volume V_ρ parametered by ρ , defined as:

$$V_\rho = ((\mathcal{CC}(V \ominus B_\rho)) \oplus B_\rho) \cup T \quad (17)$$

In particular, for any $\rho \in \mathbb{R}_+$, we have:

$$T \subseteq V_\rho \subseteq V \quad (18)$$

and for any two $\rho_1, \rho_2 \in \mathbb{R}_+$, we have:

$$\rho_1 \geq \rho_2 \implies V_{\rho_1} \subseteq V_{\rho_2} \quad (19)$$

3. The definition of V_ρ depends on ρ and the optimal result may not be the same for the various processed images. This optimal value $\hat{\rho}$ is determined for each image by an elbow-curve analysis of the size of the volumes V_ρ .

The optimal volume $V_{\hat{\rho}}$ allows to discard the extracranial artifact regions by assigning the BG label (non-cerebral tissue) to all the points, i.e.:

$$x \in \Omega \setminus V_{\hat{\rho}} \implies x \in X_{\text{BG}} \quad (20)$$

Topological noise removal The multilabel SegSRGAN method, similarly to most multilabel segmentation methods does not natively provide guarantees with respect to the topological correctness of the results. In particular, it may happen that the segmentation result be corrupted by “label” noise, i.e. that isolated voxels (or very small regions) may be erroneously assigned a given label, leading to a multilabel analogue of the binary salt-and-pepper noise.

In order to tackle this denoising issue whereas avoiding as much as possible to modify the segmentation result provided by SegSRGAN, we propose the following

post-processing, that can be seen as a multilabel version of morphological area opening [54].

Let Π be the partition of Ω induced by the segmentation S and composed by the connected components of Ω for each label. Given a limit size $s \in \mathbb{N}$ (which can be defined as a parameter or computed by an Otsu thresholding of the histogram of the size of the connected components of the label image), our aim is to modify Π to remove all the connected components $X \in \Pi$ of size $|X| < s$. This post-processing is composed by the following steps:

1. Computation of a partially labeled image $S_0 : \Omega \rightarrow L \cup \{\perp\}$ from S as follows:

- $\forall j, |X_j| < s \Rightarrow \forall x \in X_j, S_0(x) = \perp$
- $\forall j, |X_j| \geq s \Rightarrow \forall x \in X_j, S_0(x) = S(x)$

We note $\Omega_{\perp} = \{x \in \Omega \mid S_0(x) = \perp\}$.

2. Computation of a totally labeled image $S_1 : \Omega \rightarrow L$ from S_0 as follows:

- $\forall x \in \Omega \setminus \Omega_{\perp}, S_1(x) = S_0(x)$
- $\forall x \in \Omega_{\perp}, S_1(x) = S_0(y)$ with $y = \arg_{\tilde{y} \in \Omega \setminus \Omega_{\perp}} \min d(x, \tilde{y})$ where d is the geodesic distance inside Ω_{\perp} .

Step 1 is a simple operation, similar to a thresholding. Step 2 can be easily implemented by an iterative process of geodesic dilations on a label image, in a framework similar to the one defined in [55]. Here, the topological modeling of the image relies on the standard framework of digital topology [56], and the connectedness is derived from the strong adjacency (a.k.a. 6-adjacency) in \mathbb{Z}^3 .

3.2 Quality control protocol

Assessing the quality of an image processing / analysis method, especially in the context of medical image segmentation [57, 58], generally relies on the computation of usual error metrics (e.g. Dice, Hausdorff distance) which evaluate the similarity between the obtained results and handcrafted annotations provided on a test dataset. In the context of neonatal, and a fortiori premature newborn, MR image segmentation, annotations are generally not available. It is then reasonable to design alternative protocols for evaluating the quality of segmentation. In this section, we propose such a quality control (QC) protocol. It is composed of three parts, which are motivated as follows.

The first part of the protocol builds upon the idea that a segmentation result is good if it is considered as so by experts. This part of the QC protocol is then an expert-based analysis that consists to assign scores related to specific qualitative properties that should be fulfilled by a correct segmentation result. This first part, that requires the direct involvement of human experts, is described in Section 3.2.1. The second part of the protocol builds upon the idea that a segmentation result is good if it allows to successfully carry out a subsequent analysis on the processed data. In the context of neonatal MRI, such analysis often relies on morphometric measures (e.g. length, area) on slices [59, 60]. This part of the QC protocol, described in Section 3.2.2, requires an indirect involvement of human experts, since it consists of comparing the morphometric measures made by medical practitioners directly from the images, to morphometric measures derived from the segmentation results. The third part of the protocol builds upon the idea that a segmentation result is good if it has correct intrinsic properties. Such properties are notably related to the structure, i.e. the topology, of the segmented objects, independently of their spatial embedding. This part of the QC protocol, described in Section 3.2.3, does not require any involvement of human experts. Indeed,

it consists of comparing the topological properties of the segmented structures with the topological properties of the actual structures (which, in particular, do not depend on MR images but on anatomy).

In a previous work [12], we already evaluated the relevance of SegSRGAN compared to other state-of-the-art methods. Here, our purpose is different: we aim to assess the ability of SegSRGAN to segment some clinical data provided by clinical cohorts. We initially thought and designed the proposed QC protocol with this objective in mind. In particular, in this section, we describe this QC protocol with some given parameters (e.g. number of regions) and hyperparameters (e.g. morphometric measures, topological features. . .) which are geared towards our own experimental study, proposed in Section 3.3. Of course, these elements may be tuned for dealing with other kinds of images / applications that would be of interest for the reader. Keeping this in mind, this QC protocol must be considered as a generic and adaptable framework that proposes general guidelines but no hard rules.

3.2.1 Qualitative analysis

The segmentation results provided by SegSRGAN provide a parcellation of the brain into k regions. In our case, we set $k = 14$ (the corresponding cerebral regions are fully discussed in Section 3.3). Our goal in this first part of the QC protocol is to propose a simple reading form to validate manually the segmentation results.

Here, the segmentation quality is defined by the so-called FCOOT score, which is a vectorial score composed by five criteria: (F)rontier, (C)onnectedness, (O)verlap, (O)verflow and (T)rust. These criteria are detailed in Table 1. The FCOOT score provides an evaluation of the region morphology. The first four criteria (F,C,O,O) are complementary and determine a local anatomical score for each of the k specific regions. The last criterion (T) is a more global quality score for each of the k regions. Although not being equivalent, it may be noticed that these five scores are somehow related to usual quality metrics, namely:

- (F)rontier: with the Hausdorff distance;
- (C)onnectedness: with the first Betti number;
- (O)verlap: with sensitivity;
- (O)verflow: with precision;
- (T)rust: with Dice or Jaccard scores.

A FCOOT score has to be provided for each labeled region of the segmentation result. This motivates the fact that these scores are mainly binary (0: incorrect; 1: correct) except the (T)rust which is ternary (0: unsatisfactory; 1: medium; 2: satisfactory) for a better precision.

3.2.2 Morphometric analysis

We also aim to go beyond qualitative analysis of the segmented data. In order to obtain quantitative information, we rely on morphometric measures generally recognized as relevant in the literature. In particular, we focus on 1-dimensional (length) and 2-dimensional (area) measures. Basically, our purpose is to quantify in which extent such measures carried out “manually” by a human expert on a native image are similar to the same measures obtained from the binary objects given by the segmentation results.

In our study, we considered some of the measures proposed in [59] and [60]. In these pioneering works, the measures were carried out by human experts, from their visual

Table 1. FCOOT score definition (the higher the score value, the better each criterion). See Section 3.2.1.

Criteria	Score	Evaluated features
(F)rontier	{0, 1}	Boundary of the region.
(C)onnectedness	{0, 1}	Expected number of connected components.
(O)verlap	{0, 1}	No false negatives.
(O)verflow	{0, 1}	No false positives.
(T)rust	{0, 1, 2}	Overall correctness with respect to true positives, false positives and false negatives, i.e. “good overall area”, which is quantified by the following scores: $\sim 100\% \rightarrow 2$; $\sim 75\% \rightarrow 1$; $< 75\% \rightarrow 0$.

analysis of the data in slices of the images in the principal orientations (sagittal, coronal, axial).

Based on these previous works, we chose to consider three specific metrics:

- biparietal diameter (BPD);
- transcerebellar diameter (TCD);
- deep grey matter area (DGA).

The first two ones (BPD, TCD) are length metrics; the third (DGA) is an area metric. In particular, the paradigm considered here is that a good segmentation is a segmentation that allows to obtain accurate morphological measures, thus saving time and efforts for medical practitioners.

We define hereafter the protocol used by clinicians for providing manually the metrics, considered as “ground truth” (Section 3.2.2) and the protocol designed to reproduce the same metrics from the segmented images (Section 3.2.2).

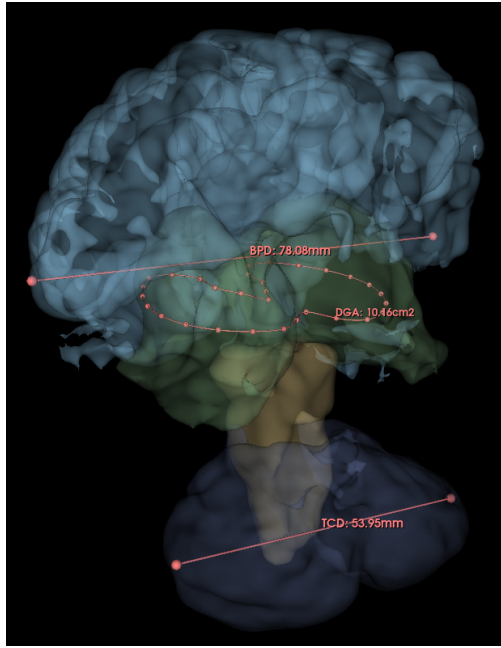
Manual measurements Each MR image is analysed by an experimented clinician. (In our case, one expert analyzed 30 images, while a second expert analysed 10 of these 30 images, in order to assess the inter-expert agreement; the processing was carried out with 3D Slicer¹.)

The two length metrics (BPD, TCD) are obtained by computing the Euclidean distance between two landmark points positioned in specific coronal slices (see Figure 2). The surface metric (DGA) is obtained by computing the area of a surface defined by a spline contour generated from control points positioned in a specific axial slice.

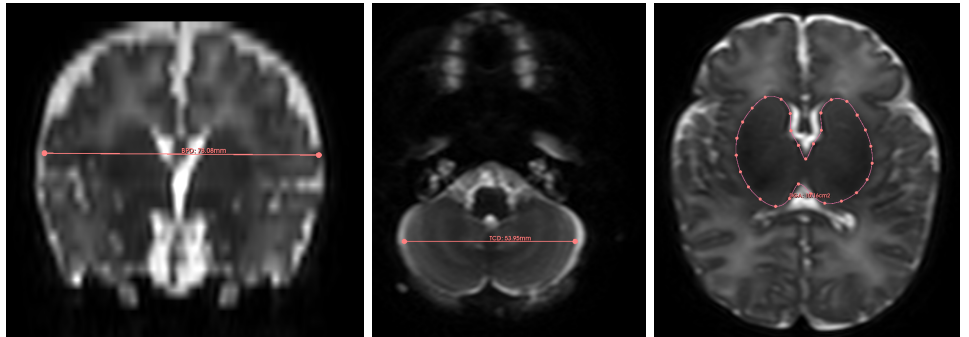
Biparietal diameter (BPD) The coronal slice is chosen as the first one located in front of the brainstem (visualized in the median sagittal slice). The start of the cochlea should be visible. Two points p_{BPD} and q_{BPD} are defined by the clinician. The biparietal diameter is then defined as $\text{BPD}_{\text{man}} = \|q_{\text{BPD}} - p_{\text{BPD}}\|_2$.

Transcerebellar diameter (TCD) The coronal slice is chosen as the one where the diameter of the cerebellum is visually assessed as maximal. The plexus can be visible and may be a reference to locate the slice. Two extremal points p_{TCD} and q_{TCD} are defined by the clinician. The transcerebellar diameter is then defined as $\text{TCD}_{\text{man}} = \|q_{\text{TCD}} - p_{\text{TCD}}\|_2$.

¹<https://www.slicer.org/>



(a) Global view



(b) Biparietal diameter

(c) Transcerebellar diameter

(d) Deep grey matter area

Fig 2. Illustration of the manual computation of the metrics. (a) 3-dimensional view of the three (length and area) measures. (b–d) 2-dimensional view of the three measures. (a) Biparietal diameter (BPD): the length is computed in the coronal slice. (b) Transcerebellar diameter (TCD): the length is computed in the axial slice. (c) Deep grey matter area (DGA): the area is computed in the axial slice. See Section 3.2.2.

Deep grey matter area (DGA) The axial slice is chosen as the one where the DGA region is visually assessed as maximal. A series of points p_{DGA}^i are set by the clinician, thus defining the contour C_{DGA} of a closed surface $S_{\text{DGA}} \subset \mathbb{R}^2$. The deep grey matter area is then defined as $\text{DGA}_{\text{man}} = \iint S_{\text{DGA}}$.

Segmentation-based measurements In order to evaluate the quality of the proposed segmentation, we compared these manual measures with measures induced by the labeled regions.

Biparietal diameter (BPD) The points p_{BPD} and q_{BPD} define a line \mathcal{L}_{BPD} . This line is intersected with the region R obtained from the label corresponding to the region

“Frontal Nocingulate”, thus providing a segment $\mathcal{S}_{\text{BPD}} = R \cap \mathcal{L}_{\text{BPD}}$. The biparietal diameter estimated from the segmentation is then defined as $\text{BPD}_{\text{seg}} = \|\mathcal{S}_{\text{BPD}}\|_2$.

Transcerebellar diameter (TCD) The points p_{TCD} and q_{TCD} define a line \mathcal{L}_{TCD} . This line is intersected with the region R_{Cer} corresponding to the “Cerebellum” label, thus providing a segment $\mathcal{S}_{\text{TCD}} = R_{\text{Cer}} \cap \mathcal{L}_{\text{TCD}}$. The transcerebellar diameter estimated from the segmentation is then defined as $\text{TCD}_{\text{seg}} = \|\mathcal{S}_{\text{TCD}}\|_2$.

Deep grey matter area (DGA) In the axial slice S chosen by the clinician, the region R_{DGA} corresponding to the “Deep grey matter” label provides a surface $\widehat{S}_{\text{DGA}} = S \cap R_{\text{DGA}}$ which is the segmentation analogue of the surface S_{DGA} defined by the clinician. The deep grey matter area estimated from the segmentation is then defined as $\text{DGA}_{\text{seg}} = \iint \widehat{S}_{\text{DGA}}$.

Comparison of manual and segmentation-based measurements At this stage, for each of the three metrics, we have two, manual and segmentation-based measurements. The error of the segmentation-based measurement with respect to the manual measurement can be computed in absolute and relative ways as:

$$\rho_{\text{M}}^{\text{abs}} = M_{\text{seg}} - M_{\text{man}} \quad (21)$$

and

$$\rho_{\text{M}}^{\text{rel}} = \frac{M_{\text{seg}} - M_{\text{man}}}{M_{\text{man}}} \quad (22)$$

with $M = \text{BPD}, \text{TCD}$ and DGA .

3.2.3 Topological analysis

Discrete topology provides efficient tools for digital image analysis, especially in the context of medical imaging [61]. In addition to the previous quality scores, that derive from ground-truth and/or clinical expert analysis, i.e. from extrinsic information, it is possible to design topological metrics which assess the intrinsic quality of the segmentation. More precisely, such topological metrics aim to quantify the degree of correctness of the segmentation maps from a structural point of view with respect to the relational properties of the training label maps, that model the topological properties of the brain structures.

In our study, we consider a first topological metric that assesses the connectedness of the k labels. To this end, we define two connectedness vectors:

$$C = [C_{\ell}]_{\ell=1}^k \quad (23)$$

and

$$C(S) = [C_{\ell}(S)]_{\ell=1}^k \quad (24)$$

In the first one, each value C_{ℓ} indicates that the region of label ℓ is anatomically composed of C_{ℓ} connected components. In the second, each value $C_{\ell}(S)$ indicates that the segmented region related to the label ℓ is composed of $C_{\ell}(S)$ connected components. For each label ℓ , the mean error over a population of n patients associated to n segmentations S_i ($1 \leq i \leq n$) is given by:

$$\mathcal{E}_C^{\ell} = \frac{1}{n} \sum_{i=1}^n |C_{\ell}(S_i) - C_{\ell}| \quad (25)$$

For the whole set of labels $\ell \in \llbracket 1, k \rrbracket$, the mean error over a population of n patients associated to n segmentations S_i ($1 \leq i \leq n$) is given by:

$$\mathcal{E}_C = \frac{1}{k} \left\| (\mathcal{E}_C^\ell)_{\ell=1}^k \right\|_1 = \frac{1}{k} \sum_{\ell=1}^k |\mathcal{E}_C^\ell| = \frac{1}{k} \sum_{\ell=1}^k \mathcal{E}_C^\ell \quad (26)$$

In particular, we have $\mathcal{E}_C^\ell, \mathcal{E}_C(S) \in \mathbb{R}_+$ and the lower the error, the better the segmentation quality with regard to connectedness (with the best score being 0).

We consider a second topological measure, related to the adjacency relation between the different label regions. Anatomically, each labeled region is adjacent to p other labeled regions ($1 \leq p \leq k$) and non-adjacent to the other $k - p$ regions. It is then possible to design an adjacency matrix, namely a square $k \times k$ Boolean, symmetric matrix:

$$A = (a_{i,j})_{1 \leq i,j \leq k} = \begin{pmatrix} a_{1,1} & \dots & a_{1,j} & \dots & a_{1,k} \\ \vdots & & \vdots & & \vdots \\ a_{i,1} & \dots & a_{i,j} & \dots & a_{i,k} \\ \vdots & & \vdots & & \vdots \\ a_{k,1} & \dots & a_{k,j} & \dots & a_{k,k} \end{pmatrix} \quad (27)$$

where $a_{i,i} = 1$ for all labels i and $a_{i,j} = 1$ (resp. 0) if the regions of distinct labels i and j are adjacent (resp. non-adjacent). A segmentation map S , endowed with an adjacency matrix $A(S) = (a_{i,j}(S))_{1 \leq i,j \leq k}$ is defined the same way. In this matrix, the elements $a_{i,i}(S)$ of the diagonal are set to 1 if the label i is present in the final segmentation, and 0 otherwise. (In particular, the trace of this matrix then assesses the ability of the method to consider all the labels in the segmentation.) This matrix $A(S)$ should satisfy $A = A(S)$ if it is fully correct with regard to the adjacency between the labeled regions.

For each couple of labels (i, j) , the mean error over a population of n patients associated to n segmentations S_i ($1 \leq i \leq n$) is given by:

$$\mathcal{E}_A^{(i,j)} = \frac{1}{n} \sum_{i=1}^n a_{i,j}(S) \oplus a_{i,j} \quad (28)$$

where \oplus is the ‘‘xor’’ operator (defined by $x \oplus y = (1 - x) \cdot y + (1 - y) \cdot x$ where *true* is associated to 1 and *false* to 0). For the whole set of couples of labels $(i, j) \in \llbracket 1, k \rrbracket^2$, the mean error over a population of n patients associated to n segmentations S_i ($1 \leq i \leq n$) is given by:

$$\mathcal{E}_A(S) = \frac{1}{k^2} \left\| (\mathcal{E}_A^{(i,j)})_{1 \leq i,j \leq k} \right\|_1 = \frac{1}{k^2} \sum_{i=1}^k \sum_{j=1}^k |\mathcal{E}_A^{(i,j)}| = \frac{1}{k^2} \sum_{i=1}^k \sum_{j=1}^k \mathcal{E}_A^{(i,j)} \quad (29)$$

In particular, we have $\mathcal{E}_A^{(i,j)}, \mathcal{E}_A(S) \in [0, 1]$ and the lower the error, the better the segmentation quality with regard to adjacency (with the best score being 0).

3.3 Experiments

We initially designed the multilabel version of SegSRGAN (Section 3.1.2) and the QC protocol (Section 3.2) with the purpose to carry out the segmentation of a whole clinical MRI cohort. In particular, our first purpose was to assess the strengths and weaknesses of SegSRGAN with respect to that goal.

3.3.1 Training

Training dataset The images considered for training SegSRGAN are part of the dHCP² project [39]. The first release of the database was used. It includes infants from 37 to 44 weeks of gestational age. T2w and inversion recovery T1w multi-slice fast spin echo anatomical images, were acquired on a 3T Philips Achieva. Infants were sleeping during the acquisition. Only axial T2w images were used for the training set with the following characteristics: $0.8 \times 0.8 \text{ mm}^2$ resolution in axial planes and 1.6 mm slices overlapped.

dHCP provides a parcellation of the brain into 87 labels/classes³. We chose to reduce the number of classes from 87 to 14 in order to train SegSRGAN. This choice was motivated by the following reasons:

- in clinical practice, an excessive precision may be counterproductive and make the tool less clinician-friendly;
- the “manual” part of the QC protocol, which is somehow central to our methodology, becomes more tedious with a large number of labels;
- with constant error, volume determination is more affected for small volumes than for larger ones.

In practice, the labels of the basal ganglia were grouped together, as well as the labels of the ventricular system. Gray and white matter labels of the same lobe were grouped together because we observed a volume interdependence between these two areas depending on imaging quality and degree of myelination. Moreover, from a physiological perspective, the cortex is connected to the underlying white matter, which contains axons from cell bodies located in the cortex. Finally, in the premature brain, the subcortical white matter is occupied by the subplate, which is intimately connected to the cortex. Next, assuming that a median slice plane would allow later individualization of the right and left portions of each volume, we grouped the right and left sides of each volume. Finally, we retained a higher level of segmentation of the temporal lobe to distinguish the auditory and language centers, whose functional maturation is central in premature infants and the subject of much research.

We considered it important to be able to measure different parts of the temporal lobes as accurately as possible. This finally led us to define the 14 macroscopic regions of interest detailed in Table 2.

A visual representation of the induced label map is illustrated in Figure 3. One may note that the cerebrospinal fluid is one of these 14 regions. In practice, the segmentation of this region, which plays in a certain extent the role of the “background” in the intracranial volume, was not assessed in our QC protocol.

Training SegSRGAN For the current study, various sets of parameters were tested to train the GAN architecture (see Figure 1). Based on this analysis, we selected a batch size of 27 and 300 epoch iterations. Regarding images, the training relied on a stride of 20, a 128 patch size and a step 20 between patches. Regarding the discriminator loss \mathcal{L}_{dis} (see Eq. (8)), we chose $\lambda_{gp} = 1e^2$. Regarding the generator loss \mathcal{L}_{gen} (see Eq. (10)), we set $\lambda_{adv} = 1e^{-3}$. We set a learning rate of $1e^{-4}$ for both networks. Testing was performed on a set of 8 images of the dHCP dataset.

²<http://www.developingconnectome.org>

³<https://gin.g-node.org/BioMedIA/dhcp-volumetric-atlas-groupwise/raw/master/config/structures.txt>

Table 2. The 14 labels corresponding to the considered anatomical regions (and their correspondence with the 87 dHCP label identifiers). See Figure 3.

Id	Label	Anatomical region	dHCP Identifiers
1	A	Occipital	22–23, 65–66
2	B	Parietal	38–39, 81–82
3	C	Cerebellum	17–18
4	D	Corpus callosum	48
5	E	Brainstem	19
6	F	Deep grey matter	40–47, 85–87
7	G	Frontal ncingulate	36–37, 79–80
8	H	Frontal cingulate	32–35, 75–78
9	I	Temporal auditory	11–12, 30–31, 57–58, 73–74
10	J	Temporal insula	20–21, 63–64
11	K	Temporal internal	1–6, 9–10, 15–16, 24–27, 51–52, 55–56, 61–62, 67–70
12	L	Temporal lateral	7–8, 13–14, 28–29, 53–54, 59–60, 71–72
13	M	Ventricle lateral	49–50
14	N	Cerebral spinal fluid	83

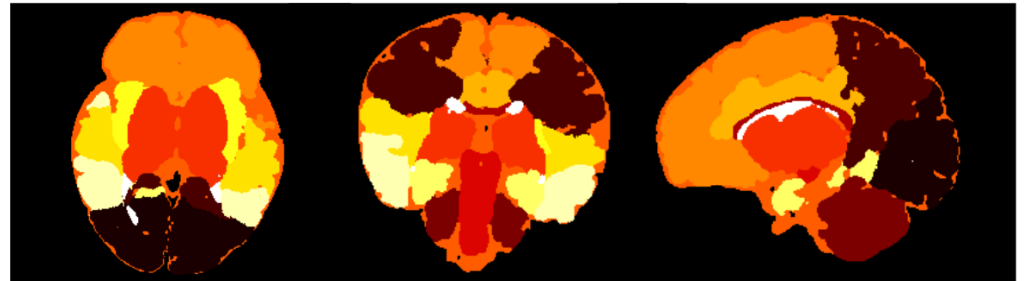


Fig 3. Example of the 14-label map obtained from the the 87-label map of dHCP image. Each colour corresponds to a distinct label. Axial, coronal and sagittal cross-section views.

3.3.2 Data

Epirmex cohort The images considered in this study are part of the EPIRMEX dataset. EPIRMEX is a French research project aimed at correlating brain MRI at birth with the cognitive outcome of extremely preterm infants. It is an ancillary study to the EPIPAGE-2 project⁴ [62], which enrolled 5170 children born before 32 weeks of gestation between 28 March 2011 and 31 December 2011 and collected demographic and clinical data as well as follow-up data up to 12 years. In the EPIRMEX subset, 581 children from 12 hospitals underwent brain MRI at term equivalent age (TEA-MRI). Neonatologists with expertise in interpretation of brain MRI of the newborn were involved in the centralized, expert review of these data. Moreover, DICOM files of the images were collected for image processing purposes.

Choice of a subset of data To avoid introducing uncontrolled bias into the validation process, we performed this validation on a subset of the data. First, we only worked on images acquired in a single hospital center, since the characteristics and settings of the MRI in each center might affect the segmentation. Secondly, we only analysed images acquired with a TE of 280 ms. Indeed, preliminary results showed that TE inversely influenced cortical thickness and white matter volume. The most visually

⁴<https://epipage2.inserm.fr>

satisfying results were obtained around 280 ms, which we kept for future use. The subset of data from centre A, which contained the largest number of MR images at 280 ms, has therefore been retained.

544
545
546

4 Results

547

The subset of EPIRMEX composed of the 70 images described in Section 3.3.2 was processed by the trained occurrence of SegSRGAN. A segmentation result for one of these images is given in Figure 4, for illustration purposes. These segmentation maps were used as input for the QC protocol described in Section 3.2.

548
549
550
551

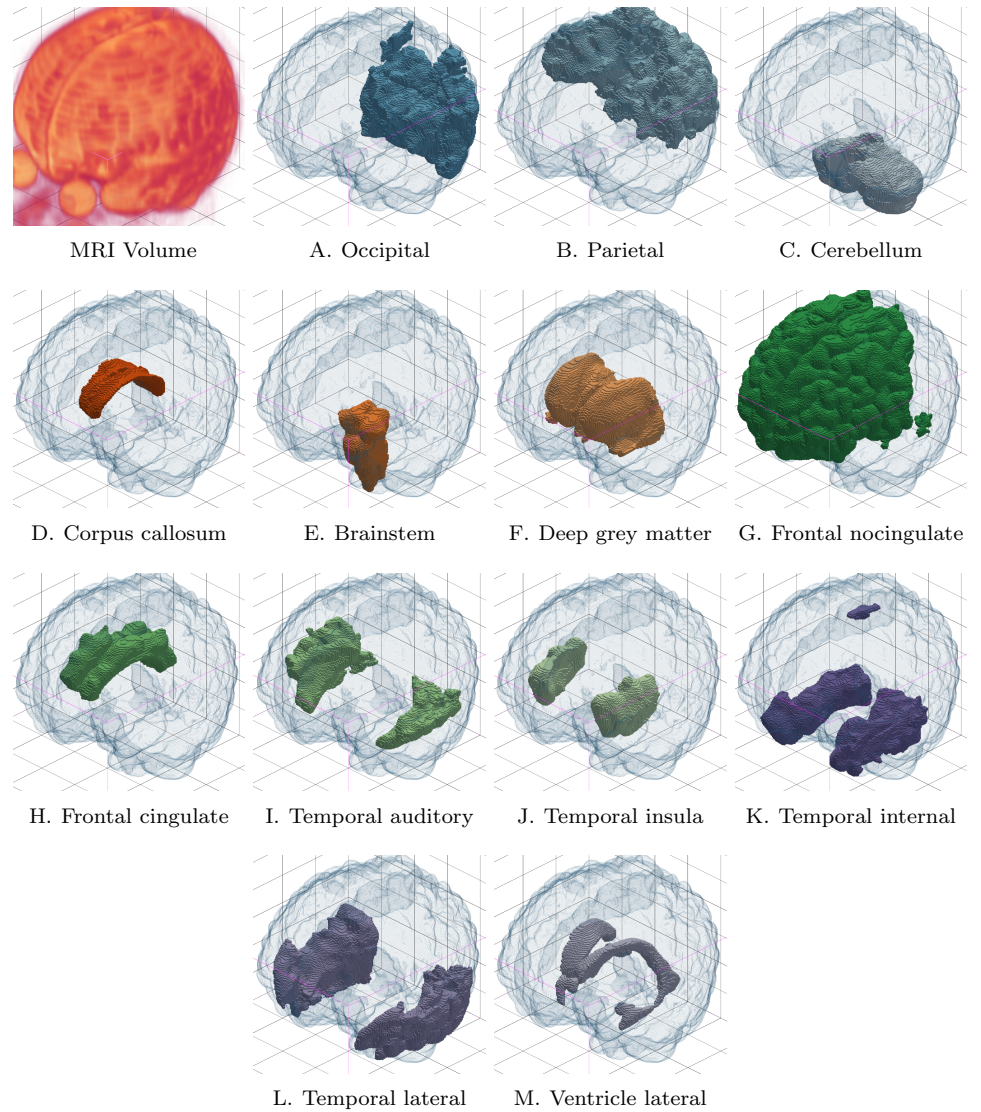


Fig 4. Segmentation result (labels A–M, see Table 2) on one MR image of the dataset. For the sake of visualization, each of the labels is represented standalone, as a binary segmentation map.

4.1 Quality control – Part 1: qualitative analysis

As stated in Section 3.2.1, the first part of the QC protocol relies on a qualitative analysis formalized by FCOOT scores for each of the 13 labeled regions. For the first four scores, namely (F)rontier, (C)onnectedness, (O)verlap, (O)verflow, and for each label, the mean value over the set of 70 patients was computed. The results are gathered in the four Kiviat diagrams depicted in Figure 5 (one diagram per score). These diagrams are oriented from 0 (center of the diagram) to 1 (border of the diagram). The closer to this border / the closer to 1, the better the value of the mean score for a given score and a given label.

The correlation between the five FCOOT scores is given in Figure 6. Correlation expresses a notion of link between variables. We aim in particular to observe the dependency that exists between the criteria taken in pairs, which can be useful in the cases where the medical experts do not have enough time for fully / accurately carrying out the extensive assessments of all the scores. The formula used is:

$$\frac{Cov(X, Y)}{\sigma_X \sigma_Y} = \frac{E[XY] - E[X]E[Y]}{\sigma_X \sigma_Y} \quad (30)$$

where $E[\cdot]$, $Cov(\cdot, \cdot)$ and σ are expected value, covariance and standard deviation, respectively, and X, Y are the investigated two criteria.

4.2 Quality control – Part 2: morphometric analysis

As stated in Section 3.2.2, the morphometric analysis part of the proposed QC protocol can be carried out by computing the error between the hand-made measures (length, area) of some structures of interest obtained from the native images, and the same measures obtained from the segmentation of these structures. Here, we focus on three measures: the biparietal diameter (BPD), the transcerebellar diameter (TCD) and the deep grey matter area (DGA). For each of them, 30 patients of the dataset were involved. The absolute and relative errors obtained from these experiments are summarized by the histograms in Figure 7.

4.3 Quality control – Part 3: topological analysis

In order to assess the quality of the segmentation results with respect to connectedness and to adjacency relation, it is mandatory to determine the ground-truth for these two features, i.e. to define the connectedness vector C (Eq. (23)) and the adjacency matrix A (Eq. (27)). In particular we set the connectedness vector as:

$$C = [C_\ell]_{\ell=1}^{13} = [1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 1] \quad (31)$$

Anatomically, each labeled region is connected, i.e. composed of one connected component, except the regions with symmetric (left and right) parts, which are composed of two connected components. The labeled regions in a segmented image should satisfy the same connectedness properties.

Given a label ℓ , we note $C_\ell(S)$ the number of connected components of the region of label ℓ in the segmentation map S . A segmentation map S which is correct with regard to connectedness should then present a vector $C(S) = [C_\ell(S)]_{\ell=1}^k$ equal to the vector C (see Eq. (24)).

The overall quality of the segmentation S with respect to the connectedness feature is then given by the label-wise and global error measures \mathcal{E}_C^ℓ and \mathcal{E}_C defined in Eqs. (25) and (26), respectively. Here, the global error is $\mathcal{E}_C = 0.9593$. The 13 label-wise error measures \mathcal{E}_C^ℓ are depicted in Figure 8.

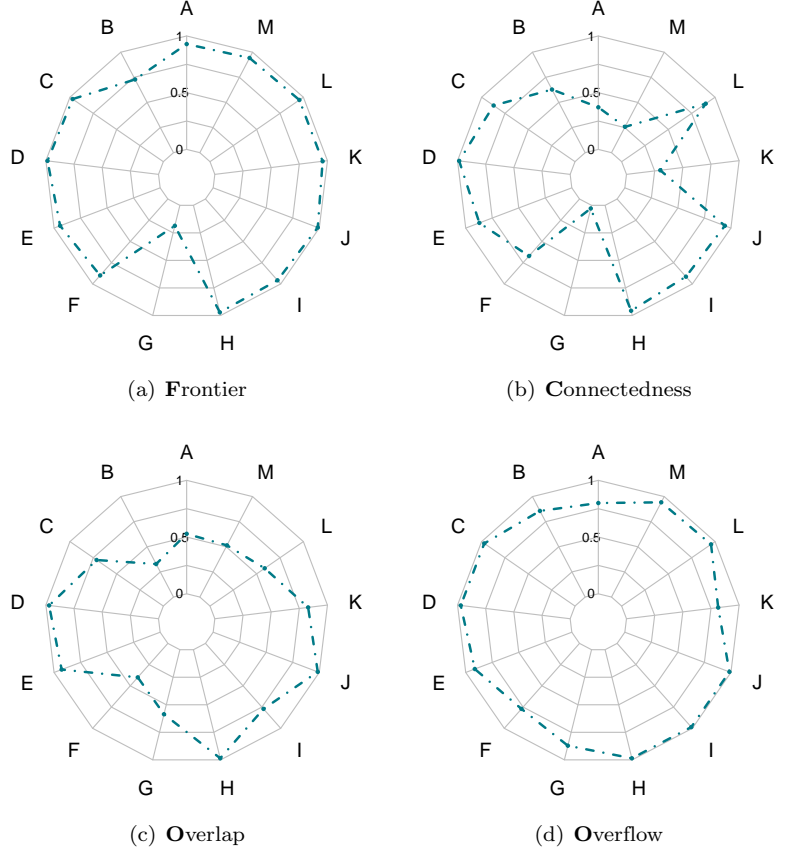


Fig 5. Kiviat diagrams for the qualitative analysis of the QC protocol: (a) Frontier; (b) Connectedness; (c) Overlap; (d) Overflow. Each point of a diagram corresponds to a mean score in $[0, 1]$ obtained as the mean value over the tested segmentations (See Table 1 and Table 2).

Regarding the adjacency error measure, we set the adjacency matrix as induced by the dHCP ground truth:

$$A = (a_{i,j})_{1 \leq i,j \leq 13} = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix} \quad (32)$$

The overall quality of the segmentation S with respect to the adjacency feature is then given by the couple-wise and global error measures $\mathcal{E}_A^{i,j}$ and \mathcal{E}_A defined in Eqs. (28) and (29), respectively. Here, the global error is $\mathcal{E}_A = 0.1534$. The 91 label-wise error

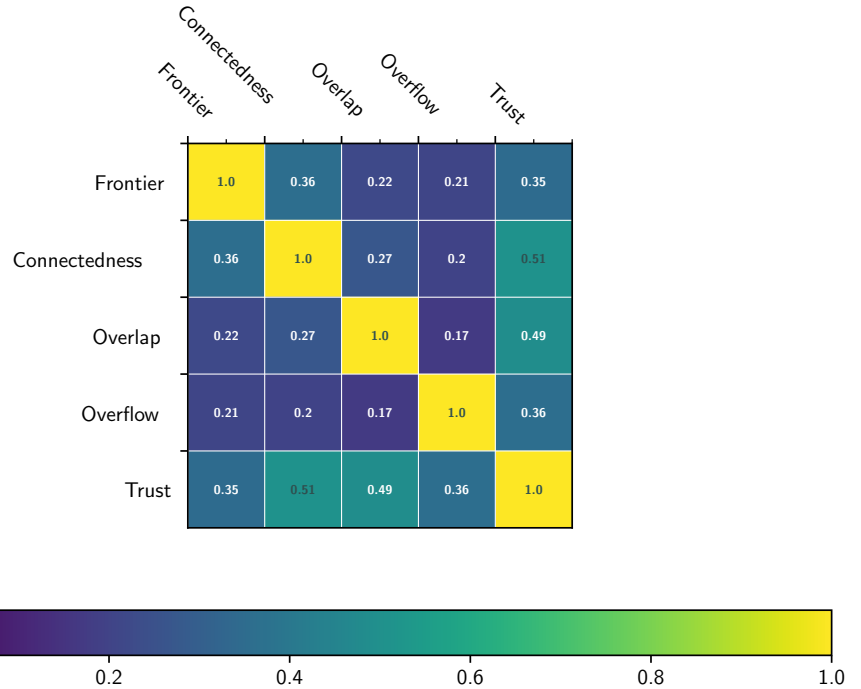


Fig 6. Correlation (symmetric) matrix between the five FCOOT scores.

measures $\mathcal{E}_A^{i,j}$ are depicted in the (symmetric) matrix of Figure 9.

5 Discussion

In this section, we discuss on the results stated in Section 4, both from methodological and clinical points of view.

First, the qualitative results exemplified in Figure 4 emphasize the ability to correctly segment the structures and tissues with salient contours. The Kiviat diagrams given in Figure 5, which summarize the expert-based scores confirm the robustness of the method in terms of frontier accuracy of the segmented regions. Indeed, for 12 of the 13 regions, the associated scores are very good. Still based on the Kiviat diagrams, the overflow quality also appears as very good. By contrast, the connectedness and overlap seem less constant, with regions exhibiting excellent results, while others are more mitigated. Regarding the correlation between these scores, summarized in Figure 6, we observe a low pairwise correlation for the four FCOO scores (0.17 to 0.36). This tends to confirm the relevance of considering these 4, complementary scores. In the meantime, we observe a greater correlation of each of these 4 FCOO scores with the (T)rust score (0.35 to 0.51). This correlation may allow the medical expert to reduce his/her analysis to this unique score when a trade-off has to be found between the time cost of the analysis of the segmented images and the expected quality of this analysis.

Regarding the morphological scores, we observe a low dispersion of the error between the segmented-based and the expert-based measures. This error varies from -5% to $+5\%$ for the biparietal diameter and -3% to $+3\%$ for the transcerebellar diameter with respect to the maximum of the histogram. It varies from -10% to $+10\%$ for the deep grey matter area. This confirms the ability of a segmented-based morphometric measure to remain compliant with a human based morphometric measure. We observe, however, a shift of the maxima of the histograms. Both for the biparietal diameter and the

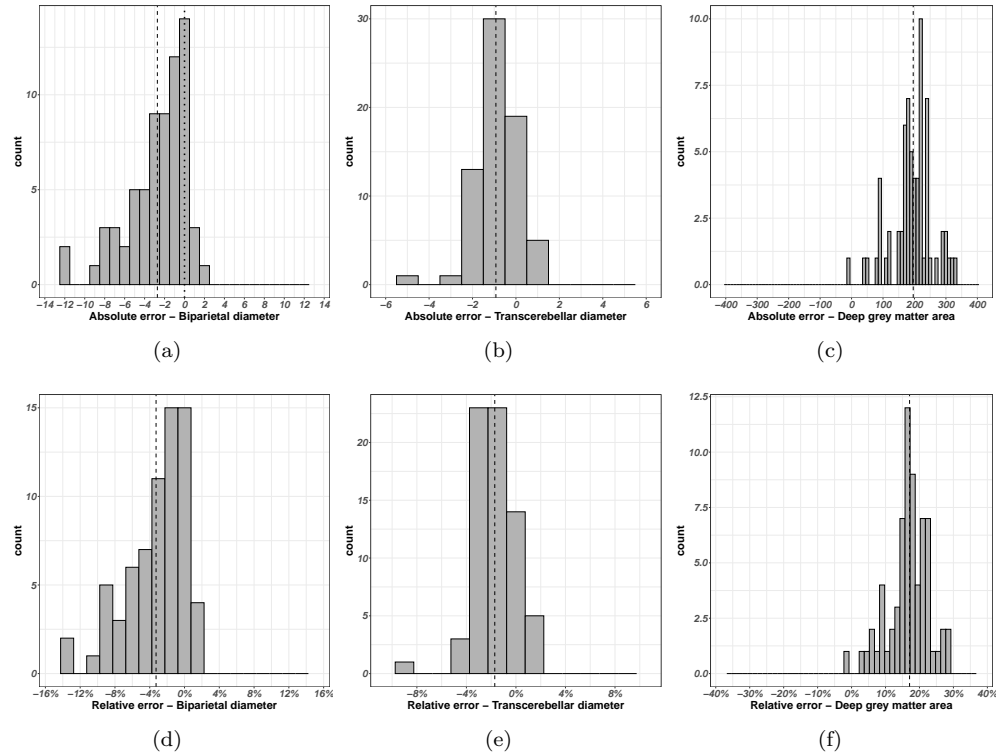


Fig 7. Histograms of the errors between hand-made morphometric measures and segmentation-guided morphometric measures (see Sections 3.2.2 and 4.2). (a-c) Absolute errors. (d-f) Relative errors. (a,d) Biparietal diameter (BPD). (b,e) Transcerebellar diameter (TCD). (c,f) Deep grey matter area (DGA). For the sake of visualization, the number of bins has been optimized with respect to the distributions. The vertical dashed line corresponds to the average error.

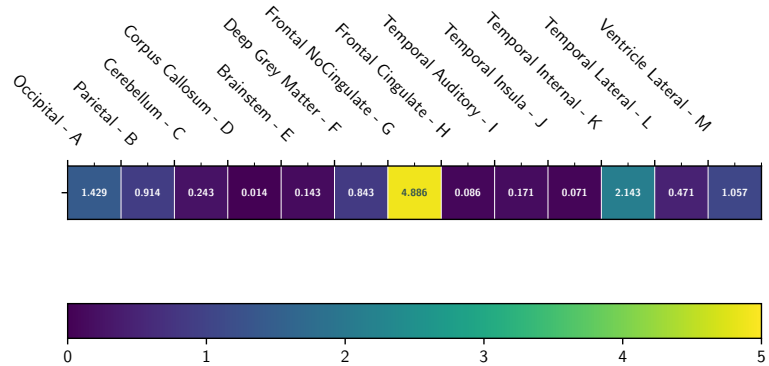


Fig 8. Mean connectedness error $\mathcal{E}_C^\ell(S)$ for each of the 13 labels ℓ , computed over 70 images, with a heatmap coloration.

transcerebellar diameter, this systematic bias is of +2%. For the deep grey matter area, it is around +15%. This may be caused by two (non-mutually exclusive) reasons: (1) the behaviour of the human expert, who may under / over-estimate the position of the landmarks in the MR images, and (2) the position of the borders of the segmentation, which may be influenced by the properties of the images. Such biases may be corrected,

624
625
626
627
628

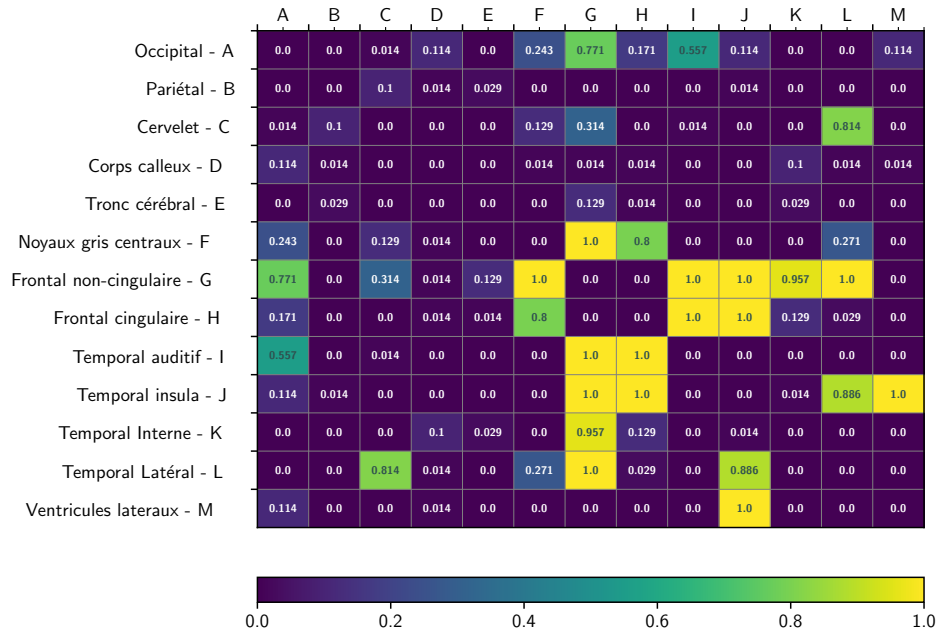


Fig 9. Mean adjacency error $\mathcal{E}_A^\ell(S)$ for each of the couples of labels, computed over 70 images, with a heatmap coloration.

for instance by benchmarking the results of the human experts and of the segmentation on a small sample of data, to identify and correct this bias before the application of the segmentation-based morphometric methods on a larger cohort. This would open the way to the development of automated, segmentation-based morphometric analysis, which could save a precious time for medical practitioners.

Regarding the topological analysis of the segmentation results, the connectedness score of the method is good, with a mean error lower than 1 (i.e. no more that one erroneous connected component per labeled region). In particular, the connectedness scores depicted in Figure 8 are satisfactory for 11 over the 13 regions, with two exceptions, namely the Frontal ncingulate and the Temporal internal. In particular, the region with the worst connectedness score (Frontal ncingulate) was also the region with the worst connectedness score in the Kiviat diagram.

This tends to prove that such topological measures can be assessed automatically, thus saving time and effort for medical practitioners. Regarding the adjacency analysis, the mean error is low, around 0.15. More precisely, when observing the pairwise region adjacencies given in Figure 9, this error is most often equal or very close to 0. In certain cases, this error is very high and in particular often equal to 1. This is explained by two facts. First, the reference adjacency map was created from only one label image of dHCP. It appears, however that for certain frontiers, the adjacency is induced by very small contacts between regions, leading to varying results both for the ground truth data and the segmented ones. Second, the current adjacency matrix provides only a binary characterization, which may not model accurately the “degree of neighbouring” between structures. These flaws may be further improved by (1) defining the adjacency matrix by a metric characterization instead of a symbolic one, and (2) building the ground truth adjacency matrix by agglomeration of the information of several label images. This will constitute some of our further works.

When reading the segmentation, the expert clinicians noted an excellent segmentation of many volumes: the cerebellum, the brainstem, the corpus callosum, the

cingulum, the temporal lobes taken as a whole. However, the experts noted variability in the demarcation line between the temporal, parietal and occipital lobe as segmented by SegSRGAN. Admittedly, these lobes are not anatomically separated by an easily discernible structure. As previously discussed, there were a few connectivity abnormalities in the frontal lobes, but given the overall volume of the frontal lobes, the impact on the final volume estimate is limited. There was an effect of head orientation in the orthogonal plane on the effectiveness of the SegSRGAN. Segmentation performance was significantly reduced when the axis of the head was very far from the orthogonal plane. The FCOOT score is easy for clinicians to use. From the clinician’s perspective, the Trust score provides a quantitative assessment of segmentation quality and degree of accuracy, and FCOO scores provide information about the nature of the error. The clinician’s delineation choices on the low resolution image are partly responsible for the reported error in the basal ganglia surface. In particular, the area behind the posterior limb of internal capsule was delimited in a more restrictive way by the expert than it was by SegSRGAN. In SegSRGAN, the area of the tail of the caudate nucleus was appropriately included in the deep gray matter label, which was often difficult to see in the low-resolution image. The entire validation procedure described in this paper makes it possible to select well-segmented MR images, or some of their labels, that can be used in clinical studies.

Conclusion

In this article, we have presented new contributions related to preterm brain analysis from MR images. In particular, we proposed an extended version of SegSRGAN [12], a super-resolution reconstruction and segmentation approach, that is now able to handle multilabel instead of binary segmentation. We also proposed a quality control protocol dedicated to the multicriteria assessment of multilabel segmentation results, based on expert-based, morphometric and topological features. Both SegSRGAN and the quality control protocol were designed with the purpose of being involved in the analysis of preterm brain MRI analysis. Nonetheless, this whole framework remains essentially generic. In particular, it could be adapted, modified and used for any other data and clinical purposes.

We used this framework for a preliminary analysis of a subset of a large clinical cohort, namely EPIRMEX, composed of multicentric MR images. Here, our purpose was to assess the ability of SegSRGAN to be further applied on the whole cohort, and to identify its strengths, weaknesses and biases. It appears from this study that SegSRGAN seems to be sufficiently robust for such purpose. Based on this conclusion, our next work will consist of applying it more systematically on the whole EPIRMEX dataset, in order to allow for further clinical research studies.

From a methodological point of view, we will also aim to improve / extend the proposed quality control protocol. Regarding the topological part, based on the above discussion, we will investigate the coupling of topological and geometric information in the adjacency matrix, by turning it from a binary to a metric mapping. We will also aim to embed a new module in the quality control by also investigating the computation of uncertainty maps in order to discriminate the regions of the segmentation that are assumed reliable versus those that may be altered by errors.

Supporting information

701

Acknowledgments

702

This work was supported by the French *Agence Nationale de la Recherche* (grants ANR-15-CE23-0009, ANR-19-CHIA-0015, ANR-22-CE45-0034), by the PHRC EPIRMEX, ancillary cohort EPIPAGE 2 and by the American Memorial Hospital Foundation (AMHF).

703

704

705

706

References

1. Pierrat V, Marchand-Martin L, Arnaud C, Kaminski M, Resche-Rigon M, Lebeaux C, et al. Neurodevelopmental outcome at 2 years for preterm children born at 22 to 34 weeks' gestation in France in 2011: EPIPAGE-2 cohort study. *BMJ*. 2017;358:j3448. doi:10.1136/bmj.j3448.
2. Pierrat V, Marchand-Martin L, Marret S, Arnaud C, Benhammou V, Cambonie G, et al. Neurodevelopmental outcomes at age 5 among children born preterm: EPIPAGE-2 cohort study. *BMJ*. 2021;373:n741. doi:10.1136/bmj.n741.
3. Woodward LJ, Anderson PJ, Austin NC, Howard K, Inder TE. Neonatal MRI to predict neurodevelopmental outcomes in preterm infants. *N Engl J Med*. 2006;355:685–694. doi:10.1056/NEJMoa053792.
4. Volpe JJ. Brain injury in premature infants: a complex amalgam of destructive and developmental disturbances. *Lancet Neurol*. 2009;8:110–124. doi:10.1016/S1474-4422(08)70294-1.
5. Inder TE, Wells SJ, Mogridge NB, Spencer C, Volpe JJ. Defining the nature of the cerebral abnormalities in the premature infant: a qualitative magnetic resonance imaging study. *J Pediatr*. 2003;143:171–179. doi:10.1067/S0022-3476(03)00357-3.
6. Padilla N, Alexandrou G, Blennow M, Lagercrantz H, Ådén U. Brain Growth Gains and Losses in Extremely Preterm Infants at Term. *Cereb Cortex*. 2015;25:1897–1905. doi:10.1093/cercor/bht431.
7. Bouyssi-Kobar M, du Plessis AJ, McCarter R, Brossard-Racine M, Murnick J, Tinkleman L, et al. Third Trimester Brain Growth in Preterm Infants Compared With In Utero Healthy Fetuses. *Pediatrics*. 2016;138:e20161640. doi:10.1542/peds.2016-1640.
8. Lind A, Parkkola R, Lehtonen L, Munck P, Maunu J, Lapinleimu H, et al. Associations between regional brain volumes at term-equivalent age and development at 2 years of age in preterm children. *Pediatr Radiol*. 2011;41:953–961. doi:10.1007/s00247-011-2071-x.
9. Rathbone R, Counsell SJ, Kapellou O, Dyet L, Kennea N, Hajnal J, et al. Perinatal cortical growth and childhood neurocognitive abilities. *Neurology*. 2011;77:1510–1517. doi:10.1212/WNL.0b013e318233b215.
10. Despotović I, Goossens B, Philips W. MRI segmentation of the human brain: challenges, methods, and applications. *Comput Math Methods Med*. 2015;2015:450341. doi:10.1155/2015/450341.

11. Makropoulos A, Counsell SJ, Rueckert D. A review on automatic fetal and neonatal brain MRI segmentation. *NeuroImage*. 2017;170:231–248. doi:10.1016/j.neuroimage.2017.06.074.
12. Delannoy Q, Pham CH, Cazorla C, Tor-Díez C, Dollé G, Meunier H, et al. SegSRGAN: Super-resolution and segmentation using generative adversarial networks—Application to neonatal brain MRI. *Comput Biol Med*. 2020;120:103755. doi:10.1016/j.compbiomed.2020.103755.
13. Banihani R, Seesahai J, Asztalos E, Terrien Church P. Neuroimaging at Term Equivalent Age: Is There Value for the Preterm Infant? A Narrative Summary. *Children*. 2021;8:227.
14. Keunen K, Išgum I, van Kooij BJM, Anbeek P, van Haastert IC, Koopman-Esseboom C, et al. Brain Volumes at Term-Equivalent Age in Preterm Infants: Imaging Biomarkers for Neurodevelopmental Outcome through Early School Age. *J Pediatr*. 2016;172:88–95.
15. Soltirovska Salamon A, Groenendaal F, van Haastert IC, Rademaker KJ, Benders MJNL, Koopman C, et al. Neuroimaging and neurodevelopmental outcome of preterm infants with a periventricular haemorrhagic infarction located in the temporal or frontal lobe. *Dev Med Child Neurol*. 2014;56:547–555.
16. Brossard-Racine M, Limperopoulos C. Cerebellar injury in premature neonates: Imaging findings and relationship with outcome. *Semin Perinatol*. 2021;45:151470.
17. van't Hooft J, van der Lee JH, Opmeer BC, Aarnoudse-Moens CSH, Leenders AGE, Mol BWJ, et al. Predicting developmental outcomes in premature infants by term equivalent MRI: Systematic review and meta-analysis. *Syst Rev*. 2015;4:71. doi:10.1186/s13643-015-0058-7.
18. Kline JE, Illapani VSP, He L, Parikh NA. Automated brain morphometric biomarkers from MRI at term predict motor development in very preterm infants. *NeuroImage: Clinical*. 2020;28:102475. doi:10.1016/j.nicl.2020.102475.
19. Rees P, Callan C, Chadda KR, Vaal M, Diviney J, Sabti S, et al. Preterm Brain Injury and Neurodevelopmental Outcomes: A Meta-analysis. *Pediatrics*. 2022;150:e2022057442.
20. Linsell L, Johnson S, Wolke D, O'Reilly H, Morris JK, Kurinczuk JJ, et al. Cognitive trajectories from infancy to early adulthood following birth before 26 weeks of gestation: A prospective, population-based cohort study. *Arch Dis Child*. 2018;103:363–370.
21. Volpe JJ. Dysmaturation of Premature Brain: Importance, Cellular Mechanisms, and Potential Interventions. *Pediatr Neurol*. 2019;95:42–66.
22. Brenner RG, Wheelock MD, Neil JJ, Smyser CD. Structural and functional connectivity in premature neonates. *Semin Perinatol*. 2021;45:151473.
23. Vo Van P, Alison M, Morel B, Beck J, Bednarek N, Hertz-Pannier L, et al. Advanced Brain Imaging in Preterm Infants: A Narrative Review of Microstructural and Connectomic Disruption. *Children*. 2022;9:356.
24. Bisiacchi P, Cainelli E. Structural and functional brain asymmetries in the early phases of life: A scoping review. *Brain Struct Funct*. 2022;227:479–496.

25. Monson BB, Anderson PJ, Matthews LG, Neil JJ, Kapur K, Cheong JL, et al. Examination of the Pattern of Growth of Cerebral Tissue Volumes From Hospital Discharge to Early Childhood in Very Preterm Infants. *JAMA Pediatr.* 2016;170:772–779.
26. Haebich KM, Willmott C, Scratch SE, Pascoe L, Lee KJ, Spencer-Smith MM, et al. Neonatal brain abnormalities and brain volumes associated with goal setting outcomes in very preterm 13-year-olds. *Brain Imaging Behav.* 2020;14:1062–1073. doi:10.1007/s11682-019-00039-1.
27. Kelly CE, Shaul M, Thompson DK, Mainzer RM, Yang JY, Dhollander T, et al. Long-lasting effects of very preterm birth on brain structure in adulthood: A systematic review and meta-analysis. *Neurosci Biobehav Rev.* 2023;147:105082.
28. Morel B, Bertault P, Favrais G, Tavernier E, Tosello B, Bednarek N, et al. Automated brain MRI metrics in the EPIRMEX cohort of preterm newborns: Correlation with the neurodevelopmental outcome at 2 years. *Diagn Interv Imaging.* 2021;102:225–232. doi:10.1016/j.diii.2020.10.009.
29. Pagnozzi AM, van Eijk L, Pannek K, Boyd RN, Saha S, George J, et al. Early brain morphometrics from neonatal MRI predict motor and cognitive outcomes at 2-years corrected age in very preterm infants. *NeuroImage.* 2023;267:119815. doi:10.1016/j.neuroimage.2022.119815.
30. Moeskops P, Išgum I, Keunen K, Claessens NHP, van Haastert IC, Groenendaal F, et al. Prediction of cognitive and motor outcome of preterm infants based on automatic quantitative descriptors from neonatal MR brain images. *Sci Rep.* 2017;7:2163.
31. L G, Loukas S, F L, Hüppi PS, Meskaldji DE, Borradori Tolsa C. Longitudinal study of neonatal brain tissue volumes in preterm infants and their ability to predict neurodevelopmental outcome. *NeuroImage.* 2019;185:728–741.
32. Devi CN, Chandrasekharan A, Sundararaman VK, Alex ZC. Neonatal brain MRI segmentation: A review. *Comput Biol Med.* 2015;64:163–178. doi:10.1016/j.compbimed.2015.06.016.
33. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: *MICCAI, Proceedings*; 2015. p. 234–241.
34. Fetit AE, Cupitt J, Kart T, Rueckert D. Training deep segmentation networks on texture-encoded input: application to neuroimaging of the developing neonatal brain. In: *MIDL, Proceedings*; 2020. p. 230–240.
35. Richter L, Fetit AE. Accurate segmentation of neonatal brain MRI with deep learning. *Front Neuroinform.* 2022;16:1006532. doi:10.3389/fninf.2022.1006532.
36. Ding Y, Acosta R, Enguix V, Suffren S, Ortmann J, Luck D, et al. Using deep convolutional neural networks for neonatal brain image segmentation. *Front Neurosci.* 2020;14:207. doi:10.3389/fnins.2020.00207.
37. Zhang S, Ren B, Yu Z, Yang H, Han X, Chen X, et al. TW-Net: Transformer Weighted Network for Neonatal Brain MRI Segmentation. *IEEE J Biomed Health Inform.* 2022;doi:10.1109/JBHI.2022.3225475.
38. Fan X, Shan S, Li X, Li J, Mi J, Yang J, et al. Attention-modulated multi-branch convolutional neural networks for neonatal brain tissue segmentation. *Comput Biol Med.* 2022;146:105522. doi:10.1016/j.compbimed.2022.105522.

39. Makropoulos A, Robinson EC, Schuh A, Wright R, Fitzgibbon S, Bozek J, et al. The developing human connectome project: A minimal processing pipeline for neonatal cortical surface reconstruction. *NeuroImage*. 2018;173:88–112. doi:10.1016/j.neuroimage.2018.01.054.
40. Makropoulos A, Gousias IS, Ledig C, Aljabar P, Serag A, Hajnal JV, et al. Automatic Whole Brain MRI Segmentation of the Developing Neonatal Brain. *IEEE Trans Med Imaging*. 2014;33:1818–1831. doi:10.1109/TMI.2014.2322280.
41. Khalili N, Turk E, Zreik M, Viergever MA, Benders MJNL, Išgum I. Generative adversarial network for segmentation of motion affected neonatal brain MRI. In: *MICCAI, Proceedings*; 2019. p. 320–328.
42. Grigorescu I, Vanes L, Uus A, Batalle D, Cordero-Grande L, Nosarti C, et al. Harmonized segmentation of neonatal brain MRI. *Front Neurosci*. 2021;15:662005. doi:10.3389/fnins.2021.662005.
43. Chen J, Sun Y, Fang Z, Lin W, Li G, Wang L, et al. Harmonized neonatal brain MR image segmentation model for cross-site datasets. *Biomed Signal Process Control*. 2021;69:102810. doi:10.1016/j.bspc.2021.102810.
44. Ancel PY, Goffinet F. EPIPAGE 2: a preterm birth cohort in France in 2011. *BMC pediatrics*. 2014;14(1):1–8. doi:10.1186/1471-2431-14-97.
45. Klapwijk ET, Van De Kamp F, Van Der Meulen M, Peters S, Wierenga LM. Qoala-T: A supervised-learning tool for quality control of FreeSurfer segmented MRI data. *NeuroImage*. 2019;189:116–129. doi:10.1016/j.neuroimage.2019.01.014.
46. Esteban O, Moodie CA, Triplett W, Poldrack RA, Gorgolewski KJ. MRIQC: Automated assessment and quality reporting of MRI scans. In: *ISMRM, Proceedings*; 2017.
47. Monereo-Sánchez J, de Jong JJA, Drenthen GS, Beran M, Backes WH, Stehouwer CDA, et al. Quality control strategies for brain MRI segmentation and parcellation: Practical approaches and recommendations-insights from the Maastricht study. *NeuroImage*. 2021;237:118174. doi:10.1016/j.neuroimage.2021.118174.
48. Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial networks. *Commun ACM*. 2020;63(11):139–144. doi:10.1145/3422622.
49. Charbonnier P, Blanc-Féraud L, Aubert G, Barlaud M. Deterministic edge-preserving regularization in computed imaging. *IEEE Trans Image Process*. 1997;6(2):298–311. doi:10.1109/83.551699.
50. Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville AC. Improved training of Wasserstein GANs. In: *NIPS, Proceedings*; 2017. p. 5769–5779.
51. Ulyanov D, Vedaldi A, Lempitsky VS. Instance Normalization: The Missing Ingredient for Fast Stylization. *CoRR*. 2016;abs/1607.08022. doi:10.48550/arXiv.1701.02096.
52. Dice LR. Measures of the amount of ecologic association between species. *Ecology*. 1945;26:297–302. doi:10.2307/1932409.

53. Ronse C, Heijmans HJAM. The algebraic basis of mathematical morphology : II. Openings and closings. *CVGIP Image Underst.* 1991;54(1):74–97. doi:10.1016/1049-9660(91)90076-2.
54. Ouzounis GK, Pesaresi M, Soille P. Differential Area Profiles: Decomposition Properties and Efficient Computation. *IEEE Trans Pattern Anal Mach Intell.* 2012;34(8):1533–1548. doi:10.1109/TPAMI.2011.245.
55. Ronse C, Agnus V. Morphology on Label Images: Flat-Type Operators and Connections. *J Math Imaging Vis.* 2005;22(2-3):283–307. doi:10.1007/s10851-005-4895-1.
56. Rosenfeld A. Digital topology. *Am Math Mon.* 1979;86:621–630. doi:10.1080/00029890.1979.11994873.
57. Shattuck DW, Prasad G, Mirza M, Narr KL, Toga AW. Online resource for validation of brain segmentation methods. *NeuroImage.* 2009;45(2):431–439. doi:10.1016/j.neuroimage.2008.10.066.
58. Gerig G, Jomier M, Chakos M. Valmet: A New Validation Tool for Assessing and Improving 3D Object Segmentation. In: *MICCAI, Proceedings; 2001.* p. 516–523.
59. Nguyen The Tich S, Anderson PJ, Shimony JS, Hunt RW, Doyle LW, Inder TE. A novel quantitative simple brain metric using MR imaging for preterm infants. *AJNR Am J Neuroradiol.* 2009;30:125–131. doi:10.3174/ajnr.A1309.
60. Kidokoro H, Neil JJ, Inder TE. New MR imaging assessment tool to define brain abnormalities in very preterm infants at term. *AJNR Am J Neuroradiol.* 2013;34:2208–2214. doi:10.3174/ajnr.A3521.
61. Saha PK, Strand R, Borgfors G. Digital Topology and Geometry in Medical Imaging: A Survey. *IEEE Trans Med Imaging.* 2015;34(9):1940–1964. doi:10.1109/TMI.2015.2417112.
62. Ancel PY, Goffinet F, EPIPAGE 2 Writing Group. EPIPAGE 2: A preterm birth cohort in France in 2011. *BMC Pediatrics.* 2014;14:97. doi:10.1186/1471-2431-14-97.