Learning to combine complementary segmentation methods for fetal and 6-month infant brain MRI segmentation.

Gerard Sanroma[a,*], Oualid M. Benkarim[a], Gemma Piella[a], Karim Lekadir[a], Nadine Hahner[b], Elisenda Eixarch[b], Miguel A. González Ballester[a,c]

[a] *Universitat Pompeu Fabra, Dept. of Information and Communication Technologies, Tànger 122-140, 08018 Barcelona, Spain.*
[b] *Fetal i+D Fetal Medicine Research Center, BCNatal - Barcelona Center for Maternal-Fetal and Neonatal Medicine (Hospital Clínic and Hospital Sant Joan de Déu), IDIBAPS, University of Barcelona, Spain*
[c] *ICREA, Pg. Lluis Companys 23, 08010 Barcelona, Spain*

Abstract

Segmentation of brain structures during the pre-natal and early post-natal periods is the first step for subsequent analysis of brain development. Segmentation techniques can be roughly divided into two families. The first, which we denote as registration-based techniques, rely on initial estimates derived by registration to one (or several) templates. The second family, denoted as learning-based techniques, relate imaging (and spatial) features to their corresponding anatomical labels. Each approach has its own qualities and both are complementary to each other. In this paper, we explore two ensembling strategies, namely, *stacking* and *cascading* to combine the strengths of both families. We present experiments on segmentation of 6-month infant brains and a cohort of fetuses with isolated non-severe ventriculomegaly (INSVM). INSVM is diagnosed when ventricles are midly enlarged and no other anomalies are apparent. Prognosis is difficult based solely on the degree of ventricular enlargement. In order to find markers for a more reliable prognosis, we use the resulting segmentations to find abnor-

*Corresponding author
Email address: gerard.sanroma@upf.edu (Gerard Sanroma)

malities in the cortical folding of INSVM fetuses. Segmentation results show that either combination strategy outperform all of the individual methods, thus demonstrating the capability of learning systematic combinations that lead to an overall improvement. In particular, the cascading strategy outperforms the ensembling one, the former one obtaining top 5, 7 and 13 results (out of 21 teams) in the segmentation of white matter, gray matter and cerebro-spinal fluid in the iSeg2017 MICCAI Segmentation Challenge. The resulting segmentations reveal that INSVM fetuses have a less convoluted cortex. This points to cortical folding abnormalities as potential markers of later neurodevelopmental outcomes.

Keywords: fetal brain MRI segmentation, multi-atlas label fusion, stacking, cascading, isolated non-severe ventriculomegaly

1. Introduction

Studying the brain in the pre-natal and early post-natal stages allows understanding the mechanisms of both normal and abnormal brain development. With the recent advances in brain magnetic resonance imaging (MRI), high-quality images with excellent contrast among several anatomical structures can be obtained. The morphological analyisis of such structures promises to discover disease biomarkers with the subsequent identification of individuals at risk for possible early intervention (Benkarim et al., 2017). In the case of fetuses, ventriculomegaly (VM) is the most frequent brain abnormality in prenatal ultrasound examination (Huisman et al., 2012). It consists of an enlargement of the ventricles, as measured by an atrial diameter $\geq$ 10mm. When the enlargement is between 10mm and 15mm, and there are no other anomalies (e.g., infections, malformations, ...), it is called isolated non-severe VM (INSVM). Prognosis in INSVM fetuses cannot be predicted solely from the degree of ventricular enlargement (Beeghly et al., 2010). To find more reliable prognostic

markers, recent works have searched for abnormalities in the cortex of INSVM fetuses, both in volume (Kyriakopoulou et al., 2014) and folding (Scott et al., 2013).

Segmentation of brain structures is the first step required for such analyses, which is usually done with T1 and/or T2 MRI modalities since they offer good anatomical contrast. Compared to the adult brain, segmentation of the developing brain poses several challenges. Due to fetal motion, 3D brain volumes have to be reconstructed from motion corrupted stacks (Murgasova et al., 2012), which may compromise image quality. Furthermore, rapid development causes dramatic changes in shape and image intensities in short periods of time. These challenges motivate the use of specific templates for different age ranges (Gholipour et al., 2017; Shi et al., 2011). Commonly used techniques for segmenting both the adult and infant brain can be roughly divided into registration- and learning-based.

Registration-based techniques first obtain a rough estimate of the location of the anatomical structures by registering the target image to one or several templates. Then, these estimates are refined to better fit the target anatomy based on either 1) parametric image intensity models (Makropoulos et al., 2014; Leemput et al., 1999; Avants et al., 2011), 2) non-parametric weighted voting techniques (Coupé et al., 2011; Wang et al., 2013; Koch et al., 2014) or 3) a combination of both (Ledig et al., 2015; Sanroma et al., 2014). Multi-atlas label fusion falls in the second kind of approaches, where the label on each target point is obtained as a consensus among the local atlas labels.

Another family of methods, which we refer to as learning-based techniques, aim at computing a mapping from image features to anatomical labels, typically using some machine learning algorithm. Features are usually derived from intensity information but may also incorporate spatial information. Different

machine learning techniques have been used including k-nearest-neighbors (Anbeek et al., 2013), support vector machines (SVM) (Moeskops et al., 2015), random forest (Wang et al., 2015) and more recently, deep learning (Moeskops et al., 2016; Kamnistas et al., 2016).

One of the differences between both families is the way in which spatial and intensity information are treated. Learning-based techniques integrate both spatial and intensity information as features into some machine learning algorithm, whereas registration-based techniques adopt a more sequential approach where spatial information is used as prior or initialization to the subsequent intensity-based modelling.

Both approaches have their drawbacks and advantages. The spatial constraints used in registration-based techniques tend to produce specific models for each region and therefore can better discriminate between the anatomical subtleties of adjacent similar structures. For example, in multi-atlas label fusion, the decision on each point is done taking into consideration only the neighboring atlas locations. As a downside, they tend to be sensitive to registration errors, which cause the model to use the wrong information. Learning-based approaches, in contrast, are more robust to registration errors since they use a global model and thus, not specific for a given region. On the other hand, because of this reason, they might fail in distinguishing between adjacent structures with similar intensity patterns. Similar concerns have previously been raised in the context of fetal brain segmentation by Wright et al. (2012).

Motivated by the complementarity of registration- and learning-based approaches, in this paper we propose to combine them based on two different ensembling strategies, namely, *stacking* and *cascading*. Stacking aims at learning which regions each method works best to combine them accordingly. Cascading, on the other hand, is incremental, and results from one method are fed onto the

other aiming at their improvement.

Other works have adopted a cascading strategy for brain MRI segmentation (Wang and Yushkevich, 2013; Tu and Bai, 2009; Kim et al., 2013; Sanroma et al., 2015). Compared to these works, the main methodological contribution of this paper consists in the observation that learning- and registration-based label fusion methods are complementary and how to best combine them to their advantage. The cascading strategy is therefore one of the two proposed strategies to address this question.

A preliminary version of this work was presented in Sanroma et al. (2016), where we proposed the stacking strategy and obtained excellent results in the NeoBrainS12 Neonatal Brain Segmentation Challenge (`http://neobrains12.isi.uu.nl/`). In the current paper, we 1) include the cascading scheme, 2) compare both of them, 3) present further brain MRI segmentation experiments on 6-month infants and a cohort of fetuses with isolated non-severe ventriculomegaly (INSVM) and 4) analyze cortical folding abnormalities in INSVM fetuses at later gestational ages than in Scott et al. (2013), when there is more gyrification.

2. Method

In the following we describe our proposed combination strategies. We use as baseline methods 1) multi-atlas joint label fusion (JLF) (Wang et al., 2013) as representative of registration-based methods and 2) SVM (Cortes and Vapnik, 1995) classifiers as representative of learning-based methods. In multi-atlas joint label fusion, each target label is computed as a weighted combination of local labels from a set of registered atlases, based on the local similarity between atlas and target image patches. In SVM-based segmentation, a classifier is learnt to discriminate the anatomical label on each point based on a set of ex-

tracted features, similarly as done in (Moeskops et al., 2015). Both registration- and learning-based approaches require a set of multiple annotated images (i.e., atlases).

2.1. Cascading

We propose to combine complementary segmentation methods in a cascading approach so that the results of one of them are used for guiding the segmentation with the other. Fig. 1 shows the pipeline for the testing phase of our proposed cascading approach.

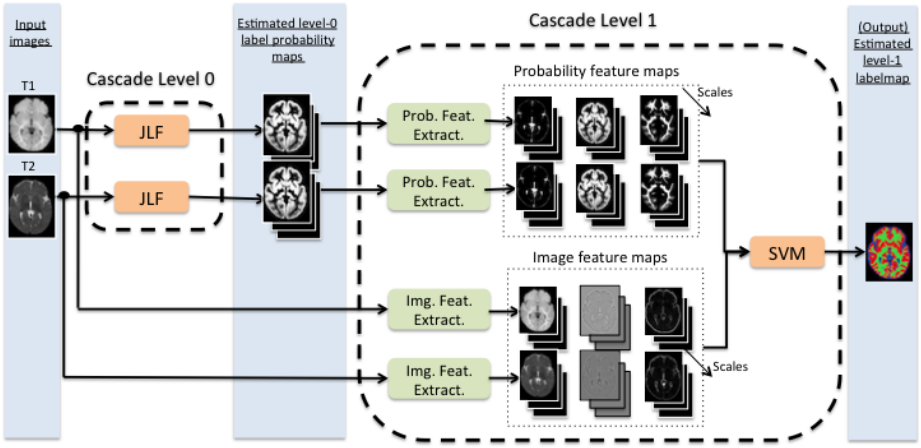


Figure 1: Pipeline for the testing phase of cascading. Input images, intermediate output and final results are shown in blue rectangles on the left, middle and right sides of the figure, respectively. Registration-based approach joint label fusion (JLF) with different modalities is represented in orange boxes. Learning-based approach is composed of feature extraction and learning, represented in green and orange boxes, respectively. Note that the learning based approach is equipped with a dual pathway of feature extraction, for probabilistic estimates and images, respectively. Dashed boxes contain details of each cascade level. The output of the level-0 is used as input for level-1.

The level-0 of the cascade segments (possibly multi-modal) images (we use T1 and T2 as illustrative example), with multi-atlas joint label fusion (Wang et al., 2013) applied independently to each modality. The estimated probability maps along with the original images are fed onto the level-1 of the cascade.

In level-1, first, a dual pathway is implemented for extracting multi-scale features from both input images and level-0 probability maps, respectively. Image features are extracted using 1) Gaussian, 2) Laplacian-of-Gaussian, and 3) gradient magnitude images convolved with Gaussians at multiple scales for each modality. Probability features are obtained by convolving the level-0 probability maps with Gaussians at multiple scales. We use the following scales $\sigma = [1.0, 5.0, 10.0, 20.0]$ mm, which correspond to the standard deviation of the Gaussian kernel. The multi-scale image and probability-map features are fed to a SVM classifier (Cortes and Vapnik, 1995) that outputs the final estimated labelmap. Adding features at different scales allows the classifier to incorporate local appearance information.

The training phase consists in standard SVM training where each sample is built with the (image and spatial probability) features extracted from each voxel from the training set.

2.2. Stacking

We propose to learn an optimal spatial combination of the probabilistic estimates of the baseline segmentation methods. Using probabilistic estimates has better generalization abilities than using discretized segmentations (Li et al., 2014), as it allows for a finer quantification of the performance of each baseline method during training. Fig. 2, shows the pipeline for the testing phase of our proposed stacking approach.

As learning-based method, we use a similar SVM classifier as in the cascading approach, but instead of a dual feature extraction pathway for probabilistic estimates and image features, respectively, it implements a single image feature-extraction pathway. This is because the stacking approach fuses the probabilistic estimates at a later stage. Since our stacking approach draws upon probabilistic segmentations, we use the approach by Wu et al. (2004) to obtain probabilistic

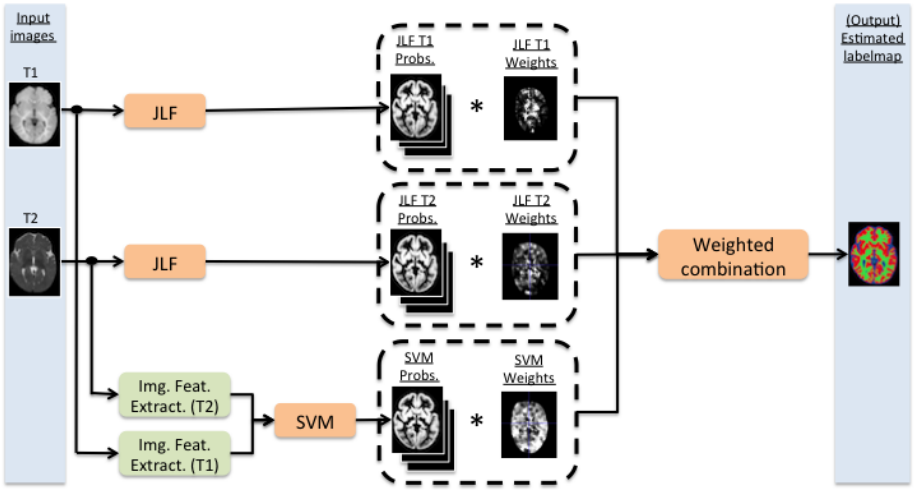


Figure 2: Pipeline for the testing phase of stacking. Input images and final results are shown in blue rectangles. Registration-based approach joint label fusion (JLF) with different modalities is represented in orange boxes. Learning-based approach is composed of feature extraction and learning, represented in green and orange boxes, respectively. Note that a single image feature extraction pathway (for both modalities) is used for the learning-based method. Dashed boxes shows the results of each method along with the optimal combination weights, computed during training as described below. Final results are obtained as a weighted combination of the results from the different methods.

135 predictions from SVM.

The output segmentation is obtained as the weighted combination of the base probabilistic segmentations, as follows:

$$\mathbf{F}_i = \sum_k \omega_i^k \mathbf{P}_i^k \quad \text{s.t.} \sum_k \omega_i^k = 1 \ , \tag{1}$$

where $\mathbf{P}_i^k$ is the label probability-vector assigned by method k to voxel i and ω_i^k is the weight denoting its contribution, whose computation is the goal of the training phase and is described in the following.

2.2.1. Training

Instead of simply using the discrete ground-truth labels for training, we use probabilistic estimates, which allows us to account for the confidence in the prediction of each method when learning the optimal combination weights. We

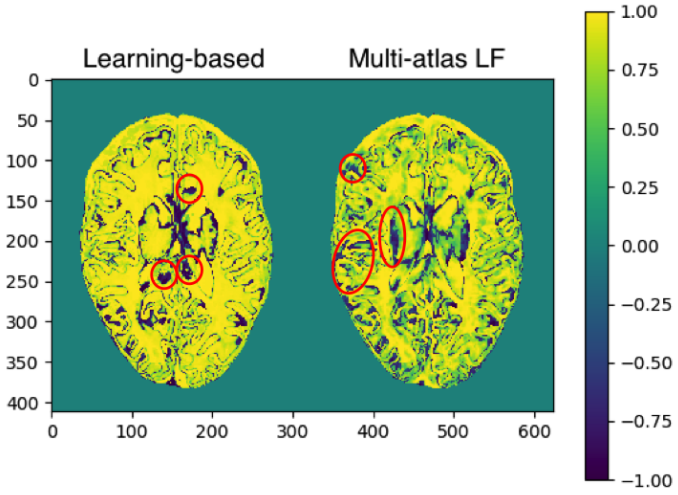


Figure 3: Margins for the learning-based and multi-atlas label fusion methods in a particular example. Red circles indicate parts where the methods are complementary.

define the margin as performance measure for each baseline method, which is positive (negative) in case it predicts the correct (wrong) label, with a magnitude proportional to its confidence. That is,

$$m_i^k = \Lambda_i^k P_{ic}^k \tag{2}$$

where P_{ic}^k is the probability assigned to the predicted label (i.e., c) by method k at voxel i, and $\Lambda_i^k \in \{1, -1\}$ indicates whether the predicted label c is correct (1) or not (-1). By using probabilistic estimates, we are not limited to just determining whether the prediction is correct or not but we can furthermore quantify the confidence of such prediction. Fig. 3 shows the margins for two segmentation methods on a sample image, which visually captures the notion of complementarity.

Substituting the probability of the estimated label in Eq. (2) by the one corresponding to the ensemble segmentation of Eq. (1), the margin of the ensemble

for point i is defined as:

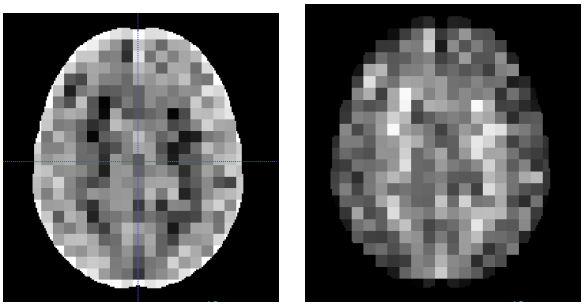
$$m_i\left(\mathbf{w}_i\right) = \sum_k \omega_i^k \Lambda_i^k P_{ic}^k \quad , \tag{3}$$

where the weights vector $\mathbf{w}_i = \left[\omega_i^1, \ldots, \omega_i^k\right]$ is the parameter of the ensemble to be estimated.

Finally, we seek the optimal weights for each point that maximize the margin. Instead of computing the combination weights for each point, we aggregate the points in spatial neighborhoods $\mathcal{N}$. In this way, we increase the number of samples in the optimization to improve the stability of the results. For the points in the neighborhood $i \in \mathcal{N}$, we compute the weights that minimize the following quadratic loss:

$$\min_{\mathbf{w}} \sum_{\forall i \in \mathcal{N}} \left(1 - m_i\left(\mathbf{w}\right)\right)^2 + \lambda \|\mathbf{w}\|^2 = \min_{\mathbf{w}} \|\mathbf{u} - M\mathbf{w}\|_2^2 + \lambda \|\mathbf{w}\|_2^2 \quad , \tag{4}$$

where $\mathbf{u}$ is a vector of ones, M is a matrix with each column M_k containing the margins of method k for all the neighborhood (i.e., $M_k = \left[m_1^k, \ldots, m_i^k, \ldots, m_{|\mathcal{N}|}^k\right]^\top$) and λ is a regularization parameter. This minimization can be solved with standard convex optimization packages such as CVX (Grant and Boyd, 2014, 2008). Fig. 4 shows the weights obtained by two segmentation methods for a given training set of segmented images. Qualitative inspection of the weights reveals that learning-based methods perform better in the cortex whereas multi-atlas label fusion performs better in the sub-cortical structures. These results are consistent with the fact that learning-based approaches work best in the cortex, where registration is difficult because it is a highly convoluted structure but has a distinct intensity profile compared to the surrounding tissues. On the other hand, registration-based approaches work best in the subcortical structures which have less distinguishable intensity profiles but a better defined shape



(a) Learning-based (b) Multi-atlas LF

Figure 4: Weights for the learning-based and multi-atlas label fusion methods. Bright denotes higher weight. Points are aggregated in spatial neighborhoods to compute the weights.

and hence, more easily registrable.

3. Experiments and results

We evaluate the proposed combination strategies in 1) 6-month infant brain segmentation using the iSeg2017 dataset (`http://iseg2017.web.unc.edu/`) and 2) fetal segmentation using an in-house dataset (Benkarim et al., 2018).

When both T1 and T2 images are available, we include the modality in the acronym of the registration-based method, i.e., JLF_T1 and JLF_T2 (we use both modalities simultaneously in the learning-based method, when available). To evaluate the importance of including spatial features in the learning-based approaches, we test two versions of the SVM-based method: one including rough atlas spatial priors as features (SVM) and the other without (SVM_nospat).

As pre-processing, we first matched the intensity histograms of the images of both datasets to a reference template. We used the templates by Shi et al. (2011) and Gholipour et al. (2017) for the infant and fetal datasets, respectively. Next, we non-rigidly registered all the images to the above templates using ANTs (Avants et al., 2008). Non-rigid registrations were used by 1) JLF to obtain pair-wise registrations by concatenating registrations through the template

and 2) SVM-based segmentation to obtain rough spatial priors to be included as features. No post-processing steps were applied after the proposed combination methods.

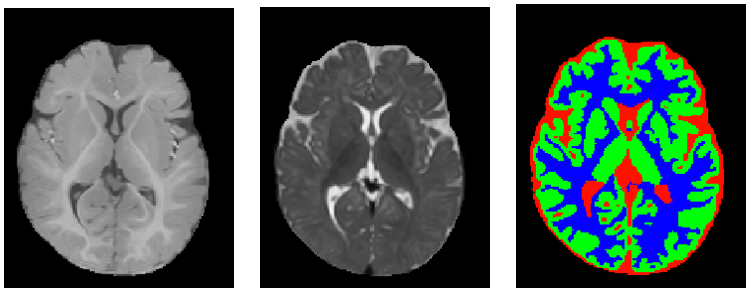
The computational time for segmenting each subject was ~ 30 min., which was mostly spent by the baseline methods in equal proportion (AMD Opteron Abu Dhabi 6378 processor).

3.1. 6-month infants

Images for the iSeg 2017 segmentation challenge were obtained from the Baby Connectome Project, an initiative jointly held by the University of North Carolina (UNC) at Chapel Hill and the University of Minnesota (UMN) with the aim of understanding how the human brain develops from birth through early childhood.

Results were evaluated in 3-fold cross-validation experiments on the 10 annotated images provided for training purposes. Annotated tissues included white matter (WM), gray matter (GM) and cerebro-spinal fluid (CSF). Datasets contained both T1- and T2-weighted scans at 1 mm isotropic resolution. Skull and cerebellum were extracted by the organizers with the iBEAT tool (`http://www.nitrc.org/projects/ibeat/`). Fig. 5 shows the T1 and T2 images along with the ground-truth tissue annotations for an example subject.

We select the best parameters according to 3-fold cross-validation experiments. We selected the set of parameters that, in average, performed best across the 3 folds. Then, we fixed these parameters for the segmentation experiments in all the folds. Specifically, for JLF, we set patch radius of 2 for both modalities and search window of 7 and 5 for T1 and T2, respectively. For SVM, we set the regularization constant to $C = 5$, we use an RBF kernel and we normalize the feature vectors to zero-mean and unit standard deviation. For stacking, we set the regularization parameter to $\lambda = 10^{-3}$ (although we found



(a) T1 (b) T2 (c) Labemap

Figure 5: Images and labelmap for an example subject in iSeg 2017 database.

little performance differences for a range of values).

Table 1 shows the mean and standard deviation Dice coefficients obtained by each method in each tissue. To get a better understanding of the performance of the methods, Fig. 6 shows a boxplot with the distribution of average Dice coefficients (across tissues) for all the methods.

Method	CSF	GM	WM
JLF_T1	88.18 ± 0.98	86.84 ± 1.12	84.35 ± 1.72
JLF_T2	84.42 ± 1.91	83.18 ± 1.64	79.02 ± 1.70
SVM	$90.96 \pm 1.16\star$	86.27 ± 1.27	83.22 ± 1.68
SVM_nospat	$90.69 \pm 1.24\star$	85.89 ± 1.40	82.81 ± 1.91
Stacking	$90.83 \pm 1.15\star$	87.81 ± 1.21	$85.32 \pm 1.57\dagger$
Cascading	$\mathbf{91.48 \pm 1.07}\star$	$\mathbf{88.44 \pm 1.14}\star\dagger$	$\mathbf{86.86 \pm 1.80}\star\dagger$

Table 1: Mean ($\pm$ st.d.) Dice coefficient of each method and each tissue in the iSeg 2017 database. Star ($\star$) and dagger ($\dagger$) denote significantly better than the base methods JLF_T1 and SVM, respectively, according to Wilcoxon signed-rank test ($p < 0.05$).

As we can see from Table 1 and Fig. 6, the proposed combination methods perform better than any of the individual baseline methods. We argue that this is because the proposed combination methods succeed in exploiting the complementarity of the baseline methods. In particular, the cascading approach performed better than stacking. We did not find any improvements by adding further levels to the cascading approach based on SVM classifiers, thus suggesting that results from previous layers were not useful for further improving the results. Including spatial features to SVM increases its performance by ~ 0.4

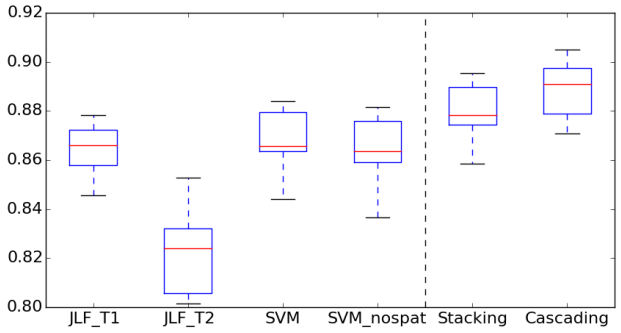


Figure 6: Boxplot with the distribution of the average Dice coefficients across tissues for each method.

Dice points, as we can see from the difference between SVM and SVM_nospat. Finally, the T2 modality might not be playing an important role in this particular dataset, based on the results of JLF_T2.

Out of the 21 participating teams in the challenge, our Cascading approach ranked 5, 7 and 13 in the segmentation of WM, GM and CSF, respectively, according to the Dice scores in the testing set. The deep learning methods based on the methodologies by Moeskops et al. (2016) and Kamnistas et al. (2016) were among the best performing ones in the challenge. It is difficult to point to a single factor explaining the performance of deep learning methods since there is already a great variability in performance among them in the i-Seg 2017 challenge. One common feature of deep learning methods compared to the proposed SVM-based approach is that feature extraction is performed as part of the classification problem instead of being done in two separate stages, and therefore the extracted features are optimized for the classification task. An interesting avenue of future research would be to substitute the SVM approach by a deep learning method in the proposed ensembles. In the cascading approach, this would imply to design a new deep learning method that uses the spatial probability maps derived from multi-atlas segmentation as additional

input channels.

3.2. Fetuses

3.2.1. Subjects

We selected 32 subjects from a cohort within a research project on congenital isolated ventriculomegaly, containing 19 controls and 13 cases of INSVM [1]. INSVM was defined as unilateral or bilateral ventricular width between 10-14.9 mm. Out of the 13 INSVM cases, 2 were left, 7 right and 4 bilateral. All fetuses were from singleton pregnancies without other malformations or risk of abnormal neurodevelopment. Ages of the included subjects range between 26 and 29.3 gestational age in weeks (GA).

3.2.2. MRI Acquisition

T2-weighted MR imaging was performed on a 1.5-T scanner (SIEMENS 105 MAGNETOM Aera syngo MR D13; Munich, Germany) with a 8-channel body coil. All images were acquired without sedation and following the American college of radiology guidelines for pregnancy and lactation. Half Fourier acquisition single shot turbo spin echo (HASTE) sequences were used with the following parameters: echo time of 82 ms, repetition time of 1500 ms, number of averaging $= 1, 2.5$ mm of slice thickness, 280×280 mm field of view and voxel size of $0.5 \times 0.5 \times 2.5$ mm. For each subject, multiple orthogonal acquisitions were performed: 4 axial, 2 coronal and 2 sagittal stacks. Brain location and extraction from 2D slices was carried out in an automatic manner using the approach by Keraudren et al. (2014), followed by high-resolution 3D volume reconstruction using the method by Murgasova et al. (2012).

[1] Approval was obtained for the study protocol from the Ethics Committee of the Hospital Clínic in Barcelona - Spain (HCB/2014/0484) and all patients gave written informed consent

3.2.3. Segmentation

Ground-truth segmentations were obtained for the following tissues and structures: extracerebellar cerebro-spinal fluid (CSF), cortical gray matter (CoGM), white matter (WM), lateral ventricles (LV), cerebellum (CB) and brain stem (BS). To obtain the ground-truth structures, first, 4 subjects were manually segmented by two expert raters. Then, the remaining subjects were segmented using the automatic method by Sanroma et al. (2016) and the automatic segmentations were manually corrected by the same expert raters. Fig. 7 shows an example of raw acquisitions, the final reconstructed volume and the ground-truth segmentations.

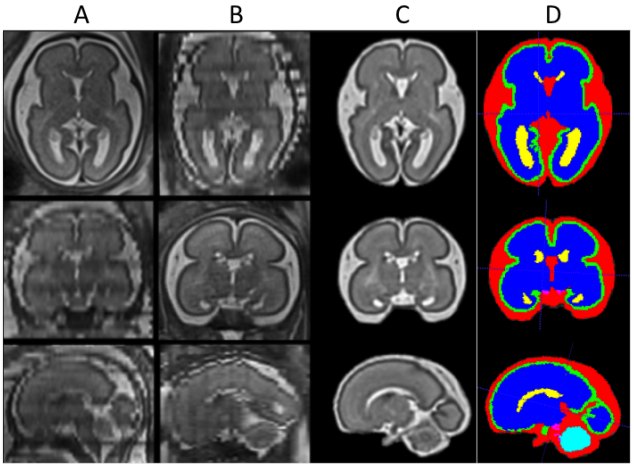


Figure 7: Brain MRI of a 26 weeks old (GA) healthy control reconstructed from 8 stacks of 2.5 mm slice thickness. From top to bottom: axial, coronal and sagittal views of axial (A) and coronal (B) raw stacks, final reconstruction (C) and ground-truth segmentations (D).

For JLF, we set patch radius of 2 and search window of 3. For SVM, we set the regularization constant to $C = 1$, we use an RBF kernel and we normalize the feature vectors to zero-mean and unit standard deviation. Likewise as in the iSeg 2017 experiments, we set the stacking regularization parameter to $\lambda = 10^{-3}$.

We select 3 subjects as atlases and segment the remaining 29 ones. Table 2 shows the average (and st.d.) in Dice coefficients across the 29 testing subjects.

To get a better understanding, Fig. 8 shows a boxplot with the distribution of average Dice coefficients (across tissues) for all the methods except SVM_nospat, which obtained considerably lower values.

Method	CSF	CoGM	WM
JLF	94.93 ± 0.99	89.71 ± 1.12	97.62 ± 0.58
SVM	95.20 ± 1.26	89.37 ± 1.47	97.42 ± 0.41
SVM_nospat	89.02 ± 4.61	75.66 ± 12.02	85.53 ± 18.14
Stacking	$\mathbf{95.56 \pm 1.14}\star$	$90.37 \pm 1.14 \star \dagger$	$97.71 \pm 0.36\dagger$
Cascading	95.46 ± 1.14	$\mathbf{90.67 \pm 1.06} \star \dagger$	$\mathbf{97.88 \pm 0.24} \star \dagger$
Method	**LV**	**CB**	**BS**
JLF	93.67 ± 2.51	96.15 ± 0.46	94.07 ± 0.88
SVM	93.10 ± 2.48	95.86 ± 0.53	94.48 ± 0.43
SVM_nospat	49.95 ± 21.67	43.56 ± 15.67	31.16 ± 13.38
Stacking	93.49 ± 2.47	96.02 ± 0.51	$\mathbf{94.59 \pm 0.43}\star$
Cascading	$\mathbf{94.22 \pm 2.22}$	$\mathbf{96.23 \pm 0.50}\dagger$	93.93 ± 0.92

Table 2: Mean ($\pm$ st.d.) Dice coefficient of each method and each tissue in the fetal brain database. Star ($\star$) and dagger ($\dagger$) denote significantly better than the base methods JLF and SVM, respectively, according to Wilcoxon signed-rank test ($p < 0.05$).

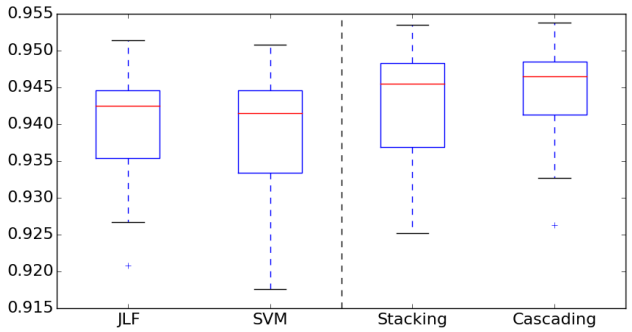


Figure 8: Boxplot with the distribution of average Dice coefficients across tissues for each method.

Similarly as with the previous database, the proposed combination strategies outperform the baseline methods, however in this case the difference is smaller. Among the combination strategies, cascading slightly outperforms stacking. The spatial information (derived through registration) plays a critical role in this dataset, as can be seen by the poor results of SVM_nospat. With T2 contrast alone and without spatial information, it is indeed difficult to discriminate some

deep brain structures such as the BS and LV.

3.2.4. Cortical Folding Analysis

We extracted the inner cortical surface from the resulting segmentations to assess the folding. We separated the two hemispheres by registration to the developing brain atlas by Makropoulos et al. (2014). We smoothed the WM binary masks using a 2 mm full width at half-maximum Gaussian kernel and we reconstructed cortical surface meshes for each hemisphere with the marching cubes algorithm (Lorensen and Cline, 1987).

Cortical folding alterations due to ventricular enlargement were investigated in each hemisphere independently using a curvature-based approach. A similar approach has been used previously to study cortical folding in fetuses (Wright et al., 2014; Wu et al., 2015). For each vertex on the cortical surface, principal curvatures were obtained, denoted as k_1 and k_2, and the following four representative folding measures were computed:

- Curvedness index: $\mathrm{CI} = \sqrt{\frac{k_1+k_2}{2}}$
- Positive mean curvature: $\mathrm{PMC} = \left(\frac{k_1+k_2}{2}\right)^+$
- Squared mean curvature: $\mathrm{SMC} = \left(\frac{k_1+k_2}{2}\right)^2$
- Positive Gaussian curvature: $\mathrm{PGC} = (k_1 \cdot k_2)^+$

The overall folding for each hemisphere was determined by a weighted average of these measures across all vertices, with weights corresponding to the mean area of the cells incident at each vertex. For all these measures, the more folding, the higher the measure. Unilateral INSVM cases (i.e., left or right) were considered as controls in the analysis of the opposite hemisphere.

Figures 9 and 10 show plots of the average folding measures across GA in weeks for the left and right hemispheres, respectively. At the top of each plot,

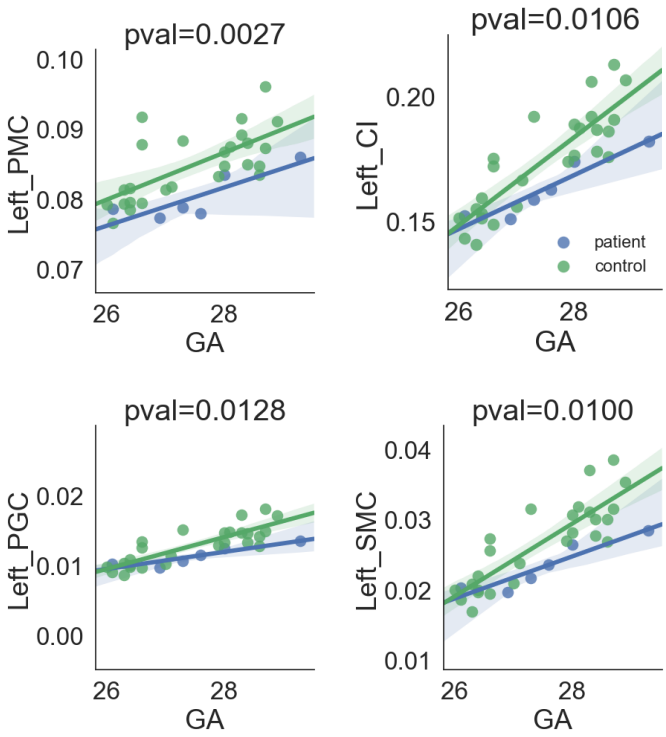


Figure 9: Different cortical folding measures in the left hemisphere w.r.t. gestational age in weeks for the patient (blue) and control (green) groups, respectively. We overlay linear fits for each diagnostic group respectively, along with their confidence intervals. P-values of the association between cortical folding and diagnostic group are displayed in the top of each plot.

we display the p-values of the association between diagnostic group (INSVM / control) and cortical folding, corrected by age, estimated with a generalized linear model according to the following expression: folding $\sim$ INSVM + age.

Results show that there is indeed an association between ventricular enlargement in INSVM and cortical folding, especially in the left hemisphere. Cortical folding alterations have been linked to neuro-developmental problems (Wolosin et al., 2009; Batty et al., 2015). These findings point to cortical folding abnormalities as potential makers for INSVM prognosis in addition to ventricular enlargement (Beeghly et al., 2010). To test the efficacy of cortical folding abnormalities as markers for INSVM prognosis, post-natal follow up cognitive test

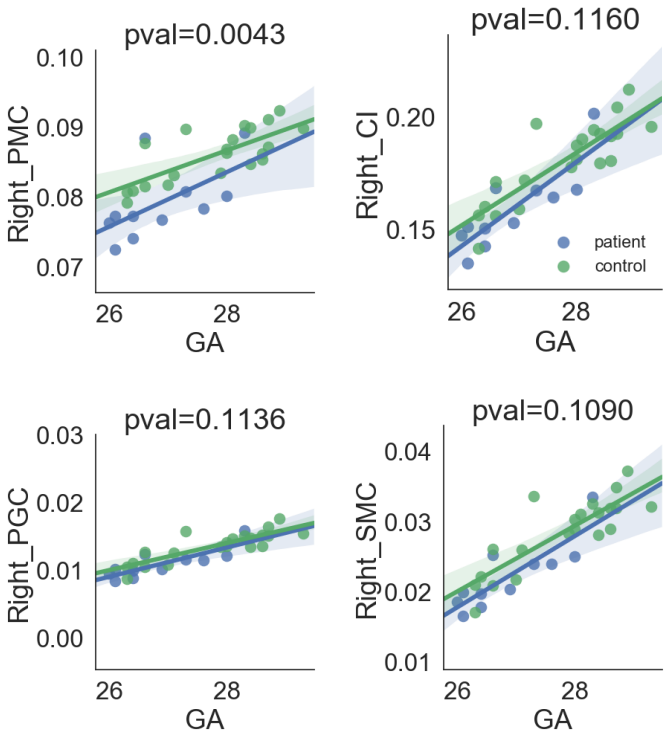


Figure 10: Different cortical folding measures in the right hemisphere w.r.t. gestational age in weeks for the patient (blue) and control (green) groups, respectively. We overlay linear fits for each diagnostic group respectively, along with their confidence intervals. P-values of the association between cortical folding and diagnostic group are displayed in the top of each plot.

data would be necessary.

4. Conclusions

We have presented two strategies for combining the strengths of complementary brain MRI segmentation methods. Stacking learns the optimal spatial combination of baseline segmentation methods and cascading uses the results of one of them as input to the other. As complementary baseline methods, we use one representative of registration-based methods, namely, multi-atlas joint label fusion (Wang et al., 2013) and one representative of learning-based methods, namely, SVM-based segmentation (Cortes and Vapnik, 1995). We compare

the proposed methods in 6-month infant and fetal brain MRI segmentation experiments. Results show that the proposed combination strategies outperform the individual baseline methods, suggesting that systematic combinations can be learnt capable of improving the results. We found that cascading is a more successful combination strategy than stacking in the presented experiments. Adding more levels to the cascade did not further improve the results. One possible reason might be that the additional classifiers were not methodologically different from the ones in previous levels, although further research is needed to determine the exact reason. Analysis of the resulting fetal brain MRI segmentations shows that INSVM fetuses have a less convoluted cortex. This suggests the potential of cortical folding abnormalities as marker of later neurodevelopmental outcomes. However, this remains to be tested when follow-up cognitive test data is available.

Acknowledgements

The first author is co-financed by the Marie Curie FP7-PEOPLE-2012-COFUND Action, Grant agreement no: 600387. This study was partly supported by Instituto de Salud Carlos III (PI16/00861) integrados en el Plan Nacional de I+D+I y cofinanciados por el ISCIII-Subdirección General de Evaluación y el Fondo Europeo de Desarrollo Regional (FEDER) "Una manera de hacer Europa"; additionally the research leading to these results has received funding from "la Caixa" Foundation. This work is partly supported by the Spanish Ministry of Economy and Competitiveness under the Maria de Maeztu Units of Excellence Programme (MDM-2015-0502).

References

Anbeek, P., Isgum, I., van Kooij, B.J.M., Mol, C.P., Kersbergen, K.J., Groenendaal, F., Viergever, M.A., de Vries, L.S., Benders, M.J.N.L., 2013. Automatic Segmentation of Eight Tissue Classes in Neonatal Brain MRI. PLoS ONE 8.

Avants, B.B., Epstein, C.L., Grossman, M., Gee, J.C., 2008. Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. Medical Image Analysis 12, 26–41.

Avants, B.B., Tustison, N.J., Wu, J., Cook, P.A., Gee, J.C., 2011. An open source multivariate framework for n-tissue segmentation with evaluation on public data. Neuroinformatics 9, 381–400.

Batty, M.J., Palaniyappan, L., Sceriff, G., Groom, M.J., Liddle, E.B., Liddle, P.F., Hollis, C., 2015. Morphological abnormalities in prefrontal surface area and thalamic volume in attention deficit/hyperactivity disorder. Psychiatry Research 233, 225–32.

Beeghly, M., Ware, J., Soul, J., du Plessis, A., Khwaja, O., Senapati, G.M., Robson, C.D., Robertson, R.L., Poussaint, T.Y., Barnewolt, C.E., Feldman, H.A., Estroff, J.A., Levine, D., 2010. Neurodevelopmental outcome of fetuses referred for ventriculomegaly. Ultrasound Obstet Gynecol 35, 405–416.

Benkarim, O.M., Hahner, N., Piella, G., Gratacos, E., González Ballester, M.A., Eixarch, E., Sanroma, G., 2018. Cortical folding alterations in fetuses with isolated non-severe ventriculomegaly. NeuroImage Clinical (in press).

Benkarim, O.M., Sanroma, G., Zimmer, V.A., Muñoz-Moreno, E., Hahner, N., Eixarch, E., Camara, O., González Ballester, M.A., Piella, G., 2017. Toward the automatic quantification of in utero brain development in 3D structural

MRI: A review. Human Brain Mapping. Human Brain Mapping 38, 2772–2787.

Clouchoux, C., du Plessis, A.J., Bouyssi-Kobar, M., Tworetzky, W., McElhinney, D.B., Brown, D.W., Gholipour, A., Kudelski, D., Warfield, S.K., McCarter, R.J., Robertson, R.L., Evans, A.C., Newburger, J.W., Limperopoulos, C., 2013. Delayed cortical development in fetuses with complex congenital heart disease. Cerebral Cortex 23, 2932–43.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Machine Learning 20, 273–297.

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

Egaña-Ugrinovic, G., Sanz-Cortes, M., Figueras, F., Bargalló, N., Gratacos, E., 2013. Differences in cortical development assessed by fetal MRI in late-onset intrauterine growth restriction. American Journal of Obstetrics & Gynecology 209, 126.e1–8.

Gholipour, A., Rollins, C.K., Velasco-Annis, C., Ouaalam, A., Akhondi-Asl, A., Afacan, O., Ortinau, C.M., Clancy, S., Limperopoulos, C., Yang, E., Estroff, J.A., Warfield, S.K., 2017. A normative spatiotemporal MRI atlas of the fetal brain for automatic segmentation and analysis of early brain growth. Nature Scientific Reports 7, 476.

Grant, M., Boyd, S., 2008. Graph implementations for nonsmooth convex programs, in: Blondel, V., Boyd, S., Kimura, H. (Eds.), Recent Advances in Learning and Control, Springer-Verlag Limited. pp. 95–110.

Grant, M., Boyd, S., 2014. CVX: Matlab Software for Disciplined Convex Programming, version 2.1. http://cvxr.com/cvx.

Huisman, T., Tekes, A., Poretti, A., 2012. Brain malformations and fetal ventriculomegaly: What to look for? Journal of Pediatric Neuroradiology 1, 185–195.

Kamnistas, K., Ledig, C., Newcombe, V.F., Simpson, J.P., Kane, A.D., Menon, D.K., Rueckert, D., Glocker, B., 2016. Efficient Multi-Scale 3D CNN with Fully Connected CRF for Accurate Brain Lesion Segmentation. Medical Image Analysis 36, 61–78.

Keraudren, K., Kuklisova-Murgasova, M., Kyriakopoulou, V., Malamateniou, C., Rutherford, M., Kainz, B., Hajnal, J.V., Rueckert, D., 2014. Automated fetal brain segmentation from 2d mri slices for motion correction. NeuroImage 101, 633–643.

Kim, M., Wu, G., Li, W., Wang, L., Son, Y.D., Cho, Z.H., Shen, D., 2013. Automatic hippocampus segmentation of 7.0 tesla mr images by combining multiple atlases and auto-context models. NeuroImage 83, 335–345.

Koch, L.M., Wright, R., Vatansever, D., Kyriakopoulou, V., Malamateniou, C., Patkee, P.A., Rutherford, M.A., Hajnal, J.V., Aljabar, P., Rueckert, D., 2014. Graph-based label propagation in fetal brain MR images, in: MICCAI Workshop on Machine Learning in Medical Imaging, Springer International Publishing. pp. 9–16.

Kyriakopoulou, V., Vatansever, D., Elkommos, S., Dawson, S., McGuiness, A., Allsop, J., Molnár, Z., Hajnal, J., Rutherford, M., 2014. Cortical overgrowth in fetuses with isolated ventriculomegaly. Cerebral Cortex 24, 2141–50.

Ledig, C., Heckemann, R.A., Hammers, A., Lopez, J.C., Newcombe, V.F.J.,

Makropoulos, A., Lötjönen, J., Menon, D.K., Rueckert, D., 2015. Robust whole-brain segmentation: Application to traumatic brain injury. Medical Image Analysis 21, 40–58.

Leemput, K.V., Maes, F., Vandermuelen, D., Suetens, P., 1999. Automated Model-Based Tissue Classification of MR Images of the Brain. IEEE Transactions on Medical Imaging 18, 897–908.

Li, L., Hu, Q., Wu, X., Yu, D., 2014. Exploration of classification confidence in ensemble learning. Pattern Recognition 47, 3120–3131.

Lorensen, W.E., Cline, H.E., 1987. Marching cubes: A high resolution 3D surface construction algorithm, in: SIGGRAPH '87 Proceedings of the 14th annual conference on Computer graphics and interactive techniques, pp. 163–169.

Lyoo, I.K., Noam, G.G., Lee, C.K., Lee, H.K., Kennedy, B.P., Renshaw, P.F., 1996. The corpus callosum and lateral ventricles in children with attention-deficit hyperactivity disorder: a brain magnetic resonance imaging study. Biol Psychiatry 40, 1060–3.

Makropoulos, A., Gousias, I.S., Ledig, C., Aljabar, P., Serag, A., Hajnal, J.H., Edwards, A.D., Counsell, S.J., Rueckert, D., 2014. Automatic Whole Brain MRI Segmentation of the Developing Neonatal Brain. IEEE Transactions on Medical Imaging 33, 1818–1831.

Moeskops, P., Benders, M.J.N.L., Chita, S.M., Kersbergen, K.J., Groenendaal, F., de Vries, L.S., Viergever, M.A., Isgum, I., 2015. Automatic segmentation of MR brain images of preterm infants using supervised classification. NeuroImage 118, 628–641.

Moeskops, P., Viergever, M.A., Mendrik, A.M., de Vries, L.S., Benders,

M.J.N.L., Isgum, I., 2016. Automatic Segmentation of MR Brain Images With a Convolutional Neural Network. IEEE Transactions on Medical Imaging 35, 1252–1261.

Murgasova, M.K., Quaghebeur, G., Rutheford, M.A., Hajnal, J.V., Schnabel, J.A., 2012. Reconstruction of fetal brain MRI with intensity matching and complete outlier removal. Medical Image Analysis 16, 1550–64.

Sanroma, G., Benkarim, O.M., Piella, G., González Ballester, M.A., 2016. Building an ensemble of complementary segmentation methods by exploiting probabilistic estimates, in: Wang, L., Adeli, E., Wang, Q., Shi, Y., Suk, H.I. (Eds.), Machine Learning in Medical Imaging (MLMI), Springer International Publishing. pp. 27–35.

Sanroma, G., Wu, G., Gao, Y., Thung, K.H., Guo, Y., Shen, D., 2015. A transversal approach for patch-based label fusion via matrix completion. Medical Image Analysis 24, 135–148.

Sanroma, G., Wu, G., Thung, K.H., Guo, Y., Shen, D., 2014. Novel multi-atlas segmentation by matrix completion, in: International Workshop on Machine Learning in Medical Imaging, Springer International Publishing. pp. 207–214.

Scott, J.A., Habas, P.A., Rajagopalan, V., Kim, K., Barkovich, A.J., Glenn, O.A., Studholme, C., 2013. Volumetric and surface-based 3D MRI analyses of fetal isolated mild ventriculomegaly: brain morphometry in ventriculomegaly. Brain Structure and Function 218, 645–55.

Shi, F., Yap, P.T., Wu, G., Jia, H., Gilmore, J.H., Lin, W., Shen, D., 2011. Infant brain atlases from neonates to 1- and 2-years old. PLoS ONE 6.

Tu, Z., Bai, X., 2009. Auto-context and its application to high-level vision tasks

and 3d brain image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 32, 1744–57.

Wang, H., Suh, J.W., Das, S.R., Pluta, J.B., Craige, C., Yushkevich, P.A., 2013. Multi-atlas segmentation with joint label fusion. IEEE Transactions on Pattern Analysis and Machine Intelligence 35, 611–623.

Wang, H., Yushkevich, P.A., 2013. Multi-atlas segmentation with joint label fusion and corrective learning-an open source implementation. Frontiers in Neuroinformatics 7.

Wang, L., Gao, Y., Shi, F., Li, G., Gilmore, J.H., Lin, W., Shen, D., 2015. LINKS: Learning-based multi-source IntegratioN frameworK for Segmentation of infant brain images. NeuroImage 108, 160–172.

Wolosin, S.M., Richardson, M.E., Hennessey, J.G., Denckla, M.B., Mostofsky, S.H., 2009. Abnormal cerebral cortex structure in children with ADHD. Human Brain Mapping 30, 175–84.

Wright, I.C., Rabe-Hesketh, S., Woodruff, P.W., David, A.S., Murray, R.M., Bullmore, E.T., 2000. Meta-analysis of regional brain volumes in schizophrenia. American Journal of Psychiatry 157, 16–25.

Wright, R., Kyriakopoulou, V., Ledig, C., Rutherford, M.A., Hajnal, J.V., Rueckert, D., Aljabar, P., 2014. Automatic quantification of normal cortical folding patterns from fetal brain MRI. NeuroImage 91, 21–32.

Wright, R., Vatansever, D., Kyriakopoulou, V., Ledig, C., Wolz, R., Serag, A., Rueckert, D., Rutheford, M.A., Hajnal, J.V., Aljabar, P., 2012. Age dependent fetal MR segmentation using manual and automated approaches, in: Perinatal and Paediatric Imaging, pp. 97–104.

Wu, J., Awate, S., Licht, D., Clouchoux, C., du Plessis, A., Avants, B., Vossough, A., Gee, J., Limperopoulos, C., 2015. Assess- ment of mri-based automated fetal cerebral cortical folding measures in prediction of gestational age in the third trimester. Journal of Neuroradiology 36, 1369–1374.

Wu, T.F., Lin, C.J., Weng, R.C., 2004. Probability estimates for multi-class classification by pairwise coupling. Journal of Machine Learning Research 5, 975–1005.