Distributionally Robust Segmentation of Abnormal Fetal Brain 3D MRI

Lucas Fidon[1](✉), Michael Aertsen[2], Nada Mufti[1,3,4], Thomas Deprest[2], Doaa Emam[4,6], Frédéric Guffens[2], Ernst Schwartz[5], Michael Ebner[1], Daniela Prayer[5], Gregor Kasprian[5], Anna L. David[3,4], Andrew Melbourne[1], Sébastien Ourselin[1], Jan Deprest[2,3,4], Georg Langs[5], and Tom Vercauteren[1]

[1] School of Biomedical Engineering and Imaging Sciences, King's College London, London, UK
lucas.fidon@kcl.ac.uk
[2] Department of Radiology, University Hospitals Leuven, Leuven, Belgium
[3] Institute for Women's Health, University College London, London, UK
[4] Department of Obstetrics and Gynaecology, University Hospitals Leuven, Leuven, Belgium
[5] Department of Biomedical Imaging and Image-guided Therapy Medical University of Vienna, Vienna, Austria
[6] Department of Gynecology and Obstetrics, University Hospitals Tanta, Tanta, Egypt

Abstract. The performance of deep neural networks typically increases with the number of training images. However, not all images have the same importance towards improved performance and robustness. In fetal brain MRI, abnormalities exacerbate the variability of the developing brain anatomy compared to non-pathological cases. A small number of abnormal cases, as is typically available in clinical datasets used for training, are unlikely to fairly represent the rich variability of abnormal developing brains. This leads machine learning systems trained by maximizing the average performance to be biased toward non-pathological cases. This problem was recently referred to as hidden stratification. To be suited for clinical use, automatic segmentation methods need to reliably achieve high-quality segmentation outcomes also for pathological cases. In this paper, we show that the state-of-the-art deep learning pipeline nnU-Net has difficulties to generalize to unseen abnormal cases. To mitigate this problem, we propose to train a deep neural network to minimize a percentile of the distribution of per-volume loss over the dataset. We show that this can be achieved by using Distributionally Robust Optimization (DRO). DRO automatically reweights the training samples with lower performance, encouraging nnU-Net to perform more consistently on all cases. We validated our approach using a dataset of 368 fetal brain T2w MRIs, including 124 MRIs of open spina bifida cases and 51 MRIs of cases with other severe abnormalities of brain development.

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-87735-4_25) contains supplementary material, which is available to authorized users.

C. H. Sudre et al. (Eds.): UNSURE 2021/PIPPI 2021, LNCS 12959, pp. 263–273, 2021.
https://doi.org/10.1007/978-3-030-87735-4_25

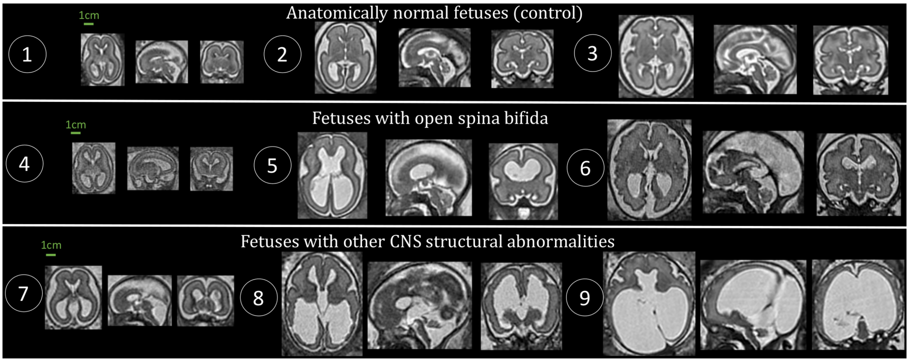
Anatomically normal fetuses (control)
1cm
1
2
3
Fetuses with open spina bifida
1cm
4
5
6
Fetuses with other CNS structural abnormalities
1cm
7
8
9

Fig. 1. Illustration of the anatomical variability in fetal brain across gestational ages and diagnostics. 1: Control (22 weeks); 2: Control (26 weeks); 3: Control (29 weeks); 4: Spina bifida (19 weeks); 5: Spina bifida (26 weeks); 6: Spina bifida (32 weeks); 7: Dandy-walker malformation with corpus callosum abnormality (23 weeks); 8: Dandy-walker malformation with ventriculomegaly and periventricular nodular heterotopia (27 weeks); 9: Aqueductal stenosis (34 weeks).

1 Introduction

The segmentation of fetal brain tissues in MRI is essential for the study of abnormal fetal brain developments [2]. Fetal brain structures segmentation could also support the evaluation and prediction of surgery outcome for open spina bifida [1,4,16,21,22]. Accurate and automatic methods for fetal brain segmentation are necessary as manual segmentation is very time-consuming and suffers from high inter- and intra-rater variability. Recently, deep neural network-based methods for fetal brain T2w MRI segmentation have been proposed [7,8,15,18,19]. On average, deep learning currently achieves state-of-the-art segmentation performance. However, those studies do not evaluate specifically the generalization and robustness properties when applied to fetuses with a pathological central nervous system.

Datasets used to train deep neural networks typically contain some under-represented subsets of cases. These cases are not specifically dealt with by the training algorithms currently used for deep neural networks. This problem has been referred to as hidden stratification [17]. Hidden stratification has been shown to lead to deep learning models with good average performance but poor performance on some clinically relevant subsets of the population [17]. While uncovering the issue, the study of [17], which is limited to classification, does not study the cause or propose a method to mitigate this problem. Cases with abnormal fetal brain development are likely to suffer from hidden stratification effects for two reasons: 1) The presence of abnormalities exacerbates the anatomical variability of the fetal brain between 18 weeks and 38 weeks of gestation, as illustrated in Fig. 1; and 2) The prevalence of those diseases is typically below 1/1000 births [1].

In this work, we study the problem of hidden stratification in fetal brain MRI segmentation using deep learning. We claim that the methodology currently used to train deep neural networks, that is maximizing the average performance across the training volumes, is at the root of the hidden stratification problem. Instead of the average empirical risk, training safe and robust deep learning models requires an asymmetric measure of risk that gives higher weights to the cases for which the algorithm fails (hard examples). Percentiles, also known as value-at-risk, is such a measure of risk that has even been adopted in industry regulations [13]. Given a per-volume fetal brain MRI segmentation metric such as the Dice score and an algorithm, the percentile at 5% is the value of the score below which 5% of the cases fall, i.e. perform worse than the percentile. The percentile relates to hidden stratification effects as it informs us of how badly worst-case examples are performing. Our contributions are four-fold. 1) We empirically show that the state-of-the-art deep learning pipeline nnU-Net [14] trained by maximizing the average segmentation performance leads to clinically significant failures for fetal brain MRI segmentation. 2) We propose to use percentiles of the Dice score on clinically relevant subpopulations as a measure of hidden stratification effects. 3) We propose to train a deep learning network to minimize a percentile of the per-volume loss function. 4) We propose a relaxation of this optimization problem based on distributionally robust optimization that can be solved efficiently in practice. We evaluate the proposed methodology for the automatic segmentation of white matter, ventricles, and cerebellum based on fetal brain 3D T2w MRI. We used a total of 368 fetal brain 3D MRIs including anatomically normal fetuses, fetuses with open spina bifida, and fetuses with other central nervous system pathologies for gestational ages ranging from 19 weeks to 39 weeks. Our empirical results suggests that the proposed training method based on distributionally robust optimization leads to better percentiles values for abnormal fetuses. In addition, qualitative results shows that distributionally robust optimization allows to reduce the number of clinically relevant failures of nnU-Net.

2 Minimization of a Percentile Loss Using Distributionally Robust Optimization

In this section, we study how a deep neural network can be trained to minimize percentiles of the loss function using a distributionally robust optimization (DRO) approach [10].

Standard deep learning training consists in optimizing the parameters $\boldsymbol{\theta}$ of a deep neural network $f(\cdot;\boldsymbol{\theta})$ by minimizing the average per-example loss $\mathcal{L}$

$$\min_{\boldsymbol{\theta}} \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}\left(f(\boldsymbol{x}_i;\boldsymbol{\theta}),\boldsymbol{y}_i\right) \tag{1}$$

Within this empirical risk minimization framework, $f(\cdot;\boldsymbol{\theta})$ is typically a Convolutional Neural Network (CNN), $\mathcal{L}$ is a smooth per-volume loss function, and $\{(\boldsymbol{x}_i,\boldsymbol{y}_i)\}_{i=1}^{n}$ is the training dataset.

In our case, $\boldsymbol{x}_i$ are the input 3D fetal brain T2w MRI volumes and $\boldsymbol{y}_i$ are the ground-truth manual segmentations. This approach is the one used to train state-of-the-art deep learning methods for segmentation using stochastic gradient descent [14]. Due to the scarcity and the higher anatomical variability of abnormal cases illustrated in Fig. 1, we cannot assume that the set of all possible fetal brain anatomies is sampled uniformly in the training dataset. However, in (1), all brain volumes are given the same weight equal to $\frac{1}{n}$.

Instead of the average per-volume loss, for robust and safe segmentation, we argue that it might be more interesting to minimize the percentile l_α at α (e.g. 5%) of the per-volume loss function. Formally, this corresponds to the minimization problem

$$\min_{\boldsymbol{\theta},\, l_\alpha} \quad l_\alpha \qquad \text{such that} \qquad \mathbb{P}\left(\mathcal{L}\left(f(\mathrm{x};\boldsymbol{\theta}), \mathrm{y}\right) \geq l_\alpha\right) \leq \alpha \tag{2}$$

where $\mathbb{P}$ is the empirical distribution defined by the training dataset. In other words, if $\alpha = 0.05$, the optimal $l_\alpha^*(\boldsymbol{\theta})$ of (2) for a given value set of parameters $\boldsymbol{\theta}$ is the value of the loss such that the per-volume loss function is worse than $l_\alpha^*(\boldsymbol{\theta})$ 5% of the time. As a result, training the deep neural network using (2) corresponds to minimizing the percentile of the per-volume loss function $l_\alpha^*(\boldsymbol{\theta})$.

Unfortunately, the minimization problem (2) cannot be solved directly using stochastic gradient descent to train a deep neural network. We now propose a tractable upper bound for $l_\alpha^*(\boldsymbol{\theta})$ and show that it can be solved in practice using distributionally robust optimization [10].

The Chernoff bound [3] applied to the per-volume loss function and the empirical training data distribution states that for all l_α and $\beta > 0$

$$\mathbb{P}\left(\mathcal{L}\left(f(\mathrm{x};\boldsymbol{\theta}), \mathrm{y}\right) \geq l_\alpha\right) \leq \frac{\exp\left(-\beta l_\alpha\right)}{n} \sum_{i=1}^{n} \exp\left(\beta\, \mathcal{L}\left(f(\boldsymbol{x}_i;\boldsymbol{\theta}), \boldsymbol{y}_i\right)\right) \tag{3}$$

To link this inequality to the minimization problem (2), we set β such that

$$\alpha = \frac{\exp\left(-\beta \hat{l}_\alpha(\boldsymbol{\theta})\right)}{n} \sum_{i=1}^{n} \exp\left(\beta\, \mathcal{L}\left(f(\boldsymbol{x}_i;\boldsymbol{\theta}), \boldsymbol{y}_i\right)\right) \tag{4}$$

$$\iff \hat{l}_\alpha(\boldsymbol{\theta}) = \frac{1}{\beta} \log\left(\frac{1}{\alpha n} \sum_{i=1}^{n} \exp\left(\beta\, \mathcal{L}\left(f(\boldsymbol{x}_i;\boldsymbol{\theta}), \boldsymbol{y}_i\right)\right)\right) \tag{5}$$

$\hat{l}_\alpha(\boldsymbol{\theta})$ is therefore an upper bound for $l_\alpha^*(\boldsymbol{\theta})$, independently to the value of $\boldsymbol{\theta}$. We propose to relax the minimization problem (2) by

$$\min_{\boldsymbol{\theta}} \frac{1}{\beta} \log\left(\sum_{i=1}^{n} \exp\left(\beta\, \mathcal{L}\left(f(\boldsymbol{x}_i;\boldsymbol{\theta}), \boldsymbol{y}_i\right)\right)\right) \tag{6}$$

where $\beta > 0$ is a hyperparameter, and where the term $\frac{1}{\beta}\log\left(\frac{1}{\alpha n}\right)$ was dropped as being independent of $\boldsymbol{\theta}$. While in (6), α does not appear in the optimization

Table 1. Training and testing dataset details. Other Abn: other brain structural abnormalities. There is no overlap of subjects between training and testing.

Train/Test	Origin	Condition	Volumes	Gestational age (in weeks)
Training	Atlas [12]	Control	18	$[21, 38]$
Training	FeTA [18]	Control	5	$[22, 28]$
Training	UHL and MUV	Control	116	$[20, 35]$
Training	UHL and MUV	Spina Bifida	28	$[22, 34]$
Training	UHL and MUV	Other Abn	10	$[23, 35]$
Testing	FeTA [18]	Control	28	$[20, 34]$
Testing	FeTA [18]	Spina Bifida	31	$[22, 31]$
Testing	FeTA [18]	Other Abn	16	$[20, 34]$
Testing	UHL and MUV	Control	26	$[26, 37]$
Testing	UHL and MUV	Spina Bifida	65	$[19, 33]$
Testing	UHL and MUV	Other Abn	25	[21, 40]

problem directly anymore, β essentially acts as a substitute for α. The higher the value of β, the higher weights the per-volume losses with a high value will have in (6).

We give a proof in the supplementary material[1] that (6) is equivalent to solving the distributionally robust optimization problem

$$\min_{\boldsymbol{\theta}} \max_{\boldsymbol{q} \in \Delta_n} \left(\sum_{i=1}^{n} q_i \, \mathcal{L}\left(f(\boldsymbol{x}_i; \boldsymbol{\theta}), \boldsymbol{y}_i\right) - \frac{1}{\beta} D_{KL}\left(\boldsymbol{q} \middle\| \frac{1}{n}\mathbf{1} \right) \right) \tag{7}$$

where a new unknown probabilities vector parameter $\boldsymbol{q}$ is introduced, $\frac{1}{n}\mathbf{1}$ denotes the uniform probability vector $\left(\frac{1}{n}, \ldots, \frac{1}{n}\right)$, D_{KL} is the Kullback-Leibler divergence, Δ_n is the unit n-simplex, and $\beta > 0$ is a hyperparameter. D_{KL} measures the dissimilarity between $\boldsymbol{q}$ and the uniform probability vector $\frac{1}{n}\mathbf{1}$ that corresponds to assign the same weight $\frac{1}{n}$ to each sample. Therefore, β controls how much the samples with a relatively high loss value (hard examples) are weighted.

Recently, hardness weighted sampling [10] was introduced as a principled hard example mining method to solve (7). Here, we proved that it can be used to minimize the proposed relaxed minimization (6) of the percentile loss problem.

3 Anatomically Abnormal Fetal Brain T2w MRI Dataset

In this section, we give details about the fetal brain 3D MRI data, the labelling protocol, and the pre-processing used in our experiments.

[1] Please see the arxiv version for the supplementary material http://arxiv.org/abs/2108.04175.

Public Fetal Brain Datasets. We used the 18 control fetal brain 3D MRI volumes of the spatio-temporal fetal brain atlas[2] [12] for gestational ages ranging from 21 weeks to 38 weeks. We also used 80 volumes from the publicly available FeTA MICCAI challenge dataset[3] [18]. For the 40 MIAL 3D MRIs, corrections of the segmentations were performed by authors MA, LF, and PD to reduce the variability against the published segmentation guidelines that was released with the FeTA dataset [18]. Those corrections were performed as part of our previous work [8] and are publicly available[4]. Brain masks for the FeTA data were obtained via affine registration using two fetal brain atlases[5] [11,12].

Image Acquisition and Preprocessing for the Private Dataset. All images in the private dataset were part of routine clinical care and were acquired at UHL and MUV due to congenital malformations seen on ultrasound.

In total, 93 cases with open spina bifida, 35 cases with other central nervous system pathologies, and 142 cases with other malformations, though with normal brain, and referred as controls, were included. The gestational age at MRI ranged from 19 weeks to 40 weeks. We have started to make fetal brain T2w 3D MRIs publicly available[6]. For each study, at least three orthogonal T2-weighted HASTE series of the fetal brain were collected on a 1.5T scanner using an echo time of 133 ms, a repetition time of 1000 ms, with no slice overlap nor gap, pixel size 0.39 mm to 1.48 mm, and slice thickness 2.50 mm to 4.40 mm. A radiologist attended all the acquisitions for quality control.

The reconstructed fetal brain 3D MRIs were obtained using `NiftyMIC` [6] a state-of-the-art super resolution and reconstruction algorithm. The volumes were all reconstructed to a resolution of 0.8 mm isotropic and registered to a fetal brain atlas [12]. Our pre-processing improves the resolution, and removes motion between neighboring slices and motion artefacts present in the original 2D slices [6]. We used volumetric brain masks to mask the tissues outside the fetal brain. Those brain masks were obtained using the automatic segmentation method described in [6,20].

Labelling Protocol. The labelling protocol used for white matter, ventricles and cerebellum is the same as in [18]. The three tissue types were segmented for our private dataset by a trained obstetrician and medical students under the supervision of a paediatric radiologist specialized in fetal brain anatomy, who quality controlled and corrected all manual segmentations.

Separation of the Data into Training and Testing. A summary of the number of fetal brain 3D MRIs used at training and testing for each central

[2] http://crl.med.harvard.edu/research/fetal_brain_atlas/.

[3] DOI: 10.7303/syn25649159.

[4] DOI: 10.5281/zenodo.5148611.

[5] DOI: 10.7303/syn25887675.

[6] https://www.cir.meduniwien.ac.at/research/fetal/.

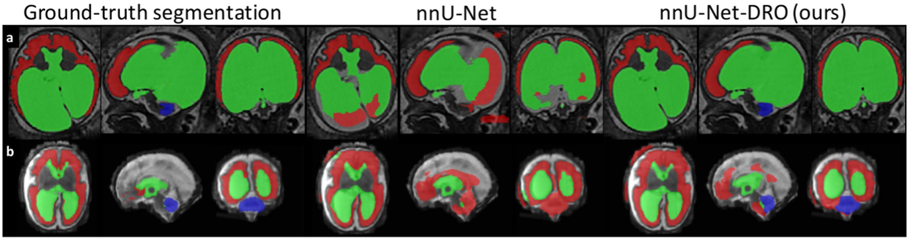

Fig. 2. Qualitative results. a) Fetus with aqueductal stenosis (34 weeks). b) Fetus with open spina bifida (27 weeks). For those two cases, nnU-Net [14] misses completly the cerebellum and achieves poor segmentation for the white matter and the ventricles. Our nnU-Net-DRO achieves satisfactory segmentation for the cerebellum for the two cases, and for all tissue types for the aqueductal stenosis case.

nervous system condition can be found in Table 1. The training dataset contains a total of 177 cases with a majority of 139 controls and only 38 abnormal cases which is typical in clinical datasets. Five controls from the FeTA dataset were added in the training dataset because we found in preliminary experiments that nnU-Net [14] fails on most of the FeTA data at testing when it is trained using only data from UHL and MUV and the fetal brain atlas [12]. The testing dataset contains 193 volumes with a majority of abnormal cases which is necessary to cover the anatomical variability of abnormal cases in our evaluation.

4 Experiments

Common Deep Learning Pipeline. We used nnU-Net [14], a generic deep learning pipeline for medical image segmentation, that has been shown to outperform other deep learning pipelines on 23 public datasets without the need to tune the loss function or the deep neural network architecture. Specifically, we used nnU-Net version 2 in 3D-full-resolution mode which is the recommended mode for isotropic 3D MRI data. nnU-Net automatically splits the training data into 5 folds 80% training/20% validation used to train 5 networks for each method. The predicted class probability maps of the 5 models are averaged at inference to improve robustness [14]. We used NVIDIA Tesla V100 GPUs with 16 GB of memory. Training each network took from 4 to 6 days.

Specificities of Each Method. The baseline consists in using nnU-Net [14] without any modification. Our method, nnU-Net-DRO, also uses nnU-Net. The only difference is that we changed the sampling strategy to use the hardness weighted sampler for DRO [10]. We used the default hyper-parameter values for the hardness weighted sampler, i.e. $\beta = 100$ with importance sampling and clipping values $w_{min} = 0.1$ and $w_{max} = 10$ as described in [10]. No other values were tested. Our implementation of the nnU-Net-DRO training procedure is publicly available at https://github.com/LucasFidon/HardnessWeightedSampler. It provides an implementation of the hardness weighted sampler described in [10].

Table 2. Evaluation of distribution robustness with respect to the pathology (193 3D MRIs). **WM**: White matter, **Vent**: Ventricles, **Cer**: Cerebellum. $\mathbf{p}_X$: X^{th} percentile of the Dice score distribution in percentage. Best values are in bold.

Method	CNS	ROI	Dice Score (%)					
			Mean	Std	$\mathbf{p}_{50}$	$\mathbf{p}_{25}$	$\mathbf{p}_{10}$	$\mathbf{p}_{5}$
(baseline) nnU-Net	**Controls** (54 cases)	**WM**	**93.9**	**2.9**	**94.1**	**91.5**	**90.6**	**89.3**
		Vent	87.8	6.8	89.7	82.1	78.1	**76.8**
		Cer	**94.5**	3.2	**94.6**	92.4	**90.7**	**89.8**
	Spina Bifida (98 cases)	**WM**	89.9	7.9	92.5	89.1	79.9	73.4
		Vent	90.6	10.6	93.0	88.6	84.8	80.7
		Cer	78.2	28.7	**89.8**	**84.2**	13.9	**0.0**
	Other Abn. (41 cases)	**WM**	**90.3**	9.8	**92.7**	**89.7**	**82.7**	70.1
		Vent	87.1	7.3	87.1	82.5	77.7	75.2
		Cer	89.7	14.7	**92.8**	89.4	85.1	81.6
(ours) nnU-Net-DRO	**Controls** (54 cases)	**WM**	93.8	3.0	93.9	91.2	90.1	89.2
		Vent	**87.9**	**6.7**	**89.9**	**82.6**	**78.3**	76.7
		Cer	94.4	**3.1**	**94.6**	**92.6**	**90.7**	89.5
	Spina Bifida (98 cases)	**WM**	**90.3**	**7.5**	**92.9**	**89.2**	**81.5**	**73.7**
		Vent	**90.9**	**10.3**	**93.2**	**89.2**	**85.1**	**81.7**
		Cer	**79.7**	**27.6**	89.7	84.1	**40.4**	**0.0**
	Other Abn. (41 cases)	**WM**	**90.3**	**9.5**	92.5	89.6	82.5	**72.0**
		Vent	**87.5**	**7.1**	**87.5**	**82.7**	**80.4**	**76.7**
		Cer	**90.6**	**10.5**	**92.8**	**89.8**	**85.5**	**82.9**

Evaluation Method. We evaluate the quality of the automatic fetal brain MRI segmentations using the Dice score [5,9]. We are particularly interested in measuring the statistical risk of the results as a way to evaluate the robustness of the different methods. To this end, in addition to the mean and standard deviation, we also report the percentiles of the Dice score at 50%, 25%, 10%, and 5%. In Table 2, we report those quantities for the Dice scores of the three tissue types white matter, ventricular system, and cerebellum.

For each method, nnU-Net is trained 5 times using different train/validation splits and different random initializations. The 5 same splits, computed randomly, are used for the two methods. The results in Table 2 are for the ensemble of the 5 3D U-Nets. Ensembling is known to increase the robustness of deep learning methods for segmentation [14]. It also makes the evaluation less sensitive to the random initialization and to the stochastic optimization.

Evaluation of nnU-Net and nnU-Net-DRO. Quantitative evaluation of nnU-Net and nnU-Net-DRO for the three different central nervous system conditions control, spina bifida, and other abnormalities can be found in Table 2.

For spina bifida and other brain abnormalities, the proposed nnU-Net-DRO achieves same or higher mean Dice scores and lower standard deviations than nnU-Net [14] for the three tissue types. For controls, the mean Dice scores and standard deviation of nnU-Net-DRO and nnU-Net differ by less than 0.1 percentage points (pp) for the three tissue types.

The comparison of the percentiles of the Dice score allows us to compare methods at the tail of the Dice scores distribution where segmentation methods reach their worst-case performance. For spina bifida, nnU-Net-DRO achieves higher values of percentiles than nnU-Net for the white matter (+0.6pp for $\mathbf{p}_{10}$), for the ventricular system (+1.0pp for $\mathbf{p}_5$), and for the cerebellum (+26.5pp for $\mathbf{p}_{10}$). And for other brain abnormalities, nnU-Net-DRO achieves higher values of percentiles than nnU-Net for the white matter (+1.9pp for $\mathbf{p}_5$), for the ventricular system (+1.5pp for $\mathbf{p}_5$ and +2.7pp for $\mathbf{p}_{10}$), and for the cerebellum (+1.3pp for $\mathbf{p}_5$). All the other percentile values differ by less than 0.5pp of Dice score between the two methods. This suggests that nnU-Net-DRO achieves better worst case performance than nnU-Net for abnormal cases.

It is worth noting that the Dice scores decrease for the white matter and the cerebellum between controls and spina bifida and abnormal cases. It was expected due to the higher anatomical variability in pathological cases. However, the Dice scores for the ventricular system tend to be higher for abnormal cases than for controls. This can be attributed to the large proportion of pathological cases with enlarged ventricles because the Dice score values tend to be higher for larger region of interests.

As can be seen in the qualitative results of Table 2, there are cases for which nnU-Net predicts an empty cerebellum segmentation while nnU-Net-DRO achieves satisfactory cerebellum segmentation. There were no cases for which the converse was true. Robust segmentation of the cerebellum for spina bifida is particularly relevant for the evaluation of fetal brain surgery for open spina bifida [1,4,21]. Additional qualitative results in the supplementary material[7] illustrates 5 other cases for which nnU-Net-DRO outperforms nnU-Net.

5 Conclusion

The high anatomical variability of the developing fetal brain across gestational ages and pathologies hampers the robustness of deep neural networks trained by maximizing the average per-volume performance. Specifically, it limits the generalization of deep neural networks to abnormal cases for which few cases are available during training. In this paper, we propose to mitigate this problem by training deep neural networks to minimize a percentile of the per-volume performance rather than the average. To allow to do this in practice, we propose to train deep neural networks with Distributionally Robust Optimization (DRO) and we show that the DRO objective is a relaxation of the per-volume loss percentile. We have validated the proposed training method on a multicentric dataset of 368 fetal brain T2w 3D MRIs with various diagnostics. nnU-Net trained with DRO achieved improved segmentation results for pathological

[7] Please see the arxiv version for the supplementary material http://arxiv.org/abs/2108.04175.

cases as compared to the unmodified nnU-Net, while achieving similar segmentation performance for the neurotypical cases. Our results suggest that nnU-Net trained with DRO is more robust to anatomical variabilities than the original nnU-Net.

Acknowledgments. This project has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement TRABIT No 765148. This work was supported by core and project funding from the Wellcome [203148/Z/16/Z; 203145Z/16/Z; WT101957], and EPSRC [NS/A000049/1; NS/A000050/1; NS/A000027/1]. TV is supported by a Medtronic/RAEng Research Chair [RCSRF1819\7\34].

References

1. Aertsen, M., et al.: Reliability of MR imaging-based posterior fossa and brain stem measurements in open spinal dysraphism in the era of fetal surgery. Am. J. Neuroradiol. **40**(1), 191–198 (2019)
2. Benkarim, O.M., et al.: Toward the automatic quantification of in utero brain development in 3D structural MRI: a review. Hum. Brain Mapp. **38**(5), 2772–2787 (2017)
3. Chernoff, H., et al.: A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. Ann. Math. Stat. **23**(4), 493–507 (1952)
4. Danzer, E., Joyeux, L., Flake, A.W., Deprest, J.: Fetal surgical intervention for myelomeningocele: lessons learned, outcomes, and future implications. Dev. Medi. Child Neurol. **62**(4), 417–425 (2020)
5. Dice, L.R.: Measures of the amount of ecologic association between species. Ecology **26**(3), 297–302 (1945)
6. Ebner, M., et al.: An automated framework for localization, segmentation and super-resolution reconstruction of fetal brain MRI. Neuroimage **206**, 116324 (2020)
7. Fetit, A.E., et al.: A deep learning approach to segmentation of the developing cortex in fetal brain MRI with minimal manual labeling. In: Medical Imaging with Deep Learning, pp. 241–261. PMLR (2020)
8. Fidon, L., et al.: Label-set loss functions for partial supervision: application to fetal brain 3D MRI parcellation. arXiv preprint arXiv:2107.03846 (2021)
9. Fidon, L., et al.: Generalised Wasserstein dice score for imbalanced multi-class segmentation using holistic convolutional networks. In: Crimi, A., Bakas, S., Kuijf, H., Menze, B., Reyes, M. (eds.) BrainLes 2017. LNCS, vol. 10670, pp. 64–76. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-75238-9_6
10. Fidon, L., Ourselin, S., Vercauteren, T.: Distributionally robust deep learning using hardness weighted sampling. arXiv preprint arXiv:2001.02658 (2020)
11. Fidon, L., et al.: A spatio-temporal atlas of the developing fetal brain with spina bifida aperta. Open Res. Europe (2021)
12. Gholipour, A., et al.: A normative spatiotemporal MRI atlas of the fetal brain for automatic segmentation and analysis of early brain growth. Sci. Rep. **7**(1), 1–13 (2017)
13. Holton, G.: Value at Risk: Theory and Practice. Academic Press (2003)
14. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods **18**(2), 203–211 (2021)

15. Khalili, N., et al.: Automatic brain tissue segmentation in fetal MRI using convolutional neural networks. Magn. Reson. Imaging **64**, 77–89 (2019)
16. Mufti, N., et al.: Cortical spectral matching and shape and volume analysis of the fetal brain pre-and post-fetal surgery for spina bifida: a retrospective study. Neuroradiology 1–14 (2021)
17. Oakden-Rayner, L., Dunnmon, J., Carneiro, G., Ré, C.: Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In: Proceedings of the ACM Conference on Health, Inference, and Learning, pp. 151–159 (2020)
18. Payette, K., et al.: An automatic multi-tissue human fetal brain segmentation benchmark using the fetal tissue annotation dataset. Sci. Data **8**(1), 1–14 (2021)
19. Payette, K., et al.: Longitudinal analysis of fetal MRI in patients with prenatal spina bifida repair. In: Wang, Q., et al. (eds.) PIPPI/SUSI -2019. LNCS, vol. 11798, pp. 161–170. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32875-7_18
20. Ranzini, M., Fidon, L., Ourselin, S., Modat, M., Vercauteren, T.: MONAIfbs: MONAI-based fetal brain MRI deep learning segmentation. arXiv preprint arXiv:2103.13314 (2021)
21. Sacco, A., et al.: Fetal surgery for open spina bifida. Obstetrician Gynaecol. **21**(4), 271 (2019)
22. Zarutskie, A., et al.: Prenatal brain imaging for predicting need for postnatal hydrocephalus treatment in fetuses that had neural tube defect repair in utero. Ultrasound Obstet. Gynecol. **53**(3), 324–334 (2019)