Fetal brain tissue annotation and segmentation challenge results

Payette, Kelly ; Li, Hongwei Bran ; de Dumast, Priscille ; Licandro, Roxane ; Ji, Hui ; Siddiquee, Md Mahfuzur Rahman ; Xu, Daguang ; Myronenko, Andriy ; Liu, Hao ; Pei, Yuchen ; Wang, Lisheng ; Peng, Ying ; Xie, Juanying ; Zhang, Huiquan ; Dong, Guiming ; Fu, Hao ; Wang, Guotai ; Rieu, ZunHyan ; Kim, Donghyeon ; Kim, Hyun Gi ; Karimi, Davood ; Gholipour, Ali ; Torres, Helena R ; Oliveira, Bruno ; Vilaça, João L ; Lin, Yang ; Avisdris, Netanell ; Ben-Zvi, Ori ; Bashat, Dafna Ben ; Fidon, Lucas ; Menze, Bjoern ; Jakab, András ; et al

Abstract: In-utero fetal MRI is emerging as an important tool in the diagnosis and analysis of the developing human brain. Automatic segmentation of the developing fetal brain is a vital step in the quantitative analysis of prenatal neurodevelopment both in the research and clinical context. However, manual segmentation of cerebral structures is time-consuming and prone to error and inter-observer variability. Therefore, we organized the Fetal Tissue Annotation (FeTA) Challenge in 2021 in order to encourage the development of automatic segmentation algorithms on an international level. The challenge utilized FeTA Dataset, an open dataset of fetal brain MRI reconstructions segmented into seven different tissues (external cerebrospinal fluid, gray matter, white matter, ventricles, cerebellum, brainstem, deep gray matter). 20 international teams participated in this challenge, submitting a total of 21 algorithms for evaluation. In this paper, we provide a detailed analysis of the results from both a technical and clinical perspective. All participants relied on deep learning methods, mainly U-Nets, with some variability present in the network architecture, optimization, and image pre- and post-processing. The majority of teams used existing medical imaging deep learning frameworks. The main differences between the submissions were the fine tuning done during training, and the specific pre- and post-processing steps performed. The challenge results showed that almost all submissions performed similarly. Four of the top five teams used ensemble learning methods. However, one team's algorithm performed significantly superior to the other submissions, and consisted of an asymmetrical U-Net network architecture. This paper provides a first of its kind benchmark for future automatic multi-tissue segmentation algorithms for the developing human brain in utero.

DOI: https://doi.org/10.1016/j.media.2023.102833

Karimi, Davood; Gholipour, Ali; Torres, Helena R; Oliveira, Bruno; Vilaça, João L; Lin, Yang; Avisdris, Netanell; Ben-Zvi, Ori; Bashat, Dafna Ben; Fidon, Lucas; Menze, Bjoern; Jakab, András; et al (2023). Fetal brain tissue annotation and segmentation challenge results. Medical Image Analysis, 88:102833.
DOI: https://doi.org/10.1016/j.media.2023.102833

Medical Image Analysis 88 (2023) 102833

Contents lists available at ScienceDirect

Medical Image Analysis

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

Fetal brain tissue annotation and segmentation challenge results

Kelly Payette [a,b,*], Hongwei Bran Li [c,d], Priscille de Dumast [e,f], Roxane Licandro [g,h], Hui Ji [a,b], Md Mahfuzur Rahman Siddiquee [i,j], Daguang Xu [j], Andriy Myronenko [j], Hao Liu [k], Yuchen Pei [k], Lisheng Wang [k], Ying Peng [l], Juanying Xie [l], Huiquan Zhang [l], Guiming Dong [m], Hao Fu [m], Guotai Wang [m], ZunHyan Rieu [n], Donghyeon Kim [n], Hyun Gi Kim [o], Davood Karimi [p], Ali Gholipour [p], Helena R. Torres [q,r,s,t], Bruno Oliveira [q,r,s,t], João L. Vilaça [q], Yang Lin [u], Netanell Avisdris [v,w], Ori Ben-Zvi [w,x], Dafna Ben Bashat [w,x,y], Lucas Fidon [z], Michael Aertsen [aa], Tom Vercauteren [z], Daniel Sobotka [ab], Georg Langs [ab], Mireia Alenyà [ac], Maria Inmaculada Villanueva [ad,ae], Oscar Camara [ac], Bella Specktor Fadida [v], Leo Joskowicz [v], Liao Weibin [af], Lv Yi [af], Li Xuesong [af], Moona Mazher [ag], Abdul Qayyum [ah], Domenec Puig [ag], Hamza Kebiri [e,f], Zelin Zhang [ai], Xinyi Xu [ai], Dan Wu [ai], Kuanlun Liao [aj], Yixuan Wu [aj], Jintai Chen [aj], Yunzhi Xu [ai], Li Zhao [ai], Lana Vasung [ak,al], Bjoern Menze [c], Meritxell Bach Cuadra [e,f], Andras Jakab [a,b,am]

[a] *Center for MR Research, University Children's Hospital Zurich, University of Zurich, Zurich, Switzerland*
[b] *Neuroscience Center Zurich, University of Zurich, Zurich, Switzerland*
[c] *Department of Quantitative Biomedicine, University of Zurich, Zurich, Switzerland*
[d] *Department of Informatics, Technical University of Munich, Munich, Germany*
[e] *Department of Radiology, Lausanne University Hospital and University of Lausanne, Lausanne, Switzerland*
[f] *CIBM, Center for Biomedical Imaging, Lausanne, Switzerland*
[g] *Laboratory for Computational Neuroimaging, Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital/Harvard Medical School, Charlestown, MA, United States*
[h] *Department of Biomedical Imaging and Image-guided Therapy, Computational Imaging Research Lab (CIR), Medical University of Vienna, Vienna, Austria*
[i] *Arizona State University, United States*
[j] *NVIDIA, United States*
[k] *Shanghai Jiaotong University, China*
[l] *School of Computer Science, Shaanxi Normal University, Xi'an 710119, China*
[m] *School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, Chengdu, China*
[n] *Research Institute, NEUROPHET Inc., Seoul 06247, South Korea*
[o] *Department of Radiology, The Catholic University of Korea, Eunpyeong St. Mary's Hospital, Seoul 06247, South Korea*
[p] *Boston Children's Hospital and Harvard Medical School, Boston, MA, United States*
[q] *2Ai – School of Technology, IPCA, Barcelos, Portugal*
[r] *Algoritmi Center, School of Engineering, University of Minho, Guimarães, Portugal*
[s] *Life and Health Sciences Research Institute (ICVS), School of Medicine, University of Minho, Braga, Portugal*
[t] *ICVS/3B's - PT Government Associate Laboratory, Braga Guimarães, Portugal*
[u] *Department of Computer Science, Hong Kong University of Science and Technology, China*
[v] *School of Computer Science and Engineering, The Hebrew University of Jerusalem, Israel*
[w] *Sagol Brain Institute, Tel Aviv Sourasky Medical Center, Israel*
[x] *Sagol School of Neuroscience, Tel Aviv University, Israel*
[y] *Sackler Faculty of Medicine, Tel Aviv University, Israel*
[z] *School of Biomedical Engineering & Imaging Sciences, King's College London, London SE1 7EU, United Kingdom*
[aa] *Department of Radiology, University Hospitals Leuven, Leuven 3000, Belgium*
[ab] *Computational Imaging Research Lab, Department of Biomedical Imaging and Image-guided Therapy, Medical University of Vienna, Vienna, Austria*
[ac] *BCN-MedTech, Department of Information and Communications Technologies, Universitat Pompeu Fabra, Barcelona, Spain*
[ad] *Department of Information and Communications Technologies, Universitat Pompeu Fabra, Barcelona, Spain*
[ae] *Institut d'Investigacions Biomèdiques August Pi i Sunyer, Barcelona, Spain*
[af] *School of Computer Science, Beijing Institute of Technology, China*
[ag] *Department of Computer Engineering and Mathematics, University Rovira i Virgili,Spain*

* Corresponding author.
E-mail address: kelly.payette@kispi.uzh.ch (K. Payette).

https://doi.org/10.1016/j.media.2023.102833
Received 20 April 2022; Received in revised form 16 March 2023; Accepted 20 April 2023
Available online 22 April 2023

[ah] Université de Bourgogne, France
[ai] Key Laboratory for Biomedical Engineering of Ministry of Education, Department of Biomedical Engineering, College of Biomedical Engineering & Instrument Science, Zhejiang University, Yuquan Campus, Hangzhou, China
[aj] Zhejiang University, Hangzhou, China
[ak] Division of Newborn Medicine, Department of Pediatrics, Boston Children's Hospital, United States
[al] Department of Pediatrics, Harvard Medical School, United States
[am] University Research Priority Project Adaptive Brain Circuits in Development and Learning (AdaBD), University of Zürich, Zurich, Switzerland

ARTICLE INFO

Keywords:
Multi-class image segmentation
Fetal brain MRI
Congenital disorders
Super-resolution reconstructions

ABSTRACT

In-utero fetal MRI is emerging as an important tool in the diagnosis and analysis of the developing human brain. Automatic segmentation of the developing fetal brain is a vital step in the quantitative analysis of prenatal neurodevelopment both in the research and clinical context. However, manual segmentation of cerebral structures is time-consuming and prone to error and inter-observer variability. Therefore, we organized the Fetal Tissue Annotation (FeTA) Challenge in 2021 in order to encourage the development of automatic segmentation algorithms on an international level. The challenge utilized FeTA Dataset, an open dataset of fetal brain MRI reconstructions segmented into seven different tissues (external cerebrospinal fluid, gray matter, white matter, ventricles, cerebellum, brainstem, deep gray matter). 20 international teams participated in this challenge, submitting a total of 21 algorithms for evaluation. In this paper, we provide a detailed analysis of the results from both a technical and clinical perspective. All participants relied on deep learning methods, mainly U-Nets, with some variability present in the network architecture, optimization, and image pre- and post-processing. The majority of teams used existing medical imaging deep learning frameworks. The main differences between the submissions were the fine tuning done during training, and the specific pre- and post-processing steps performed. The challenge results showed that almost all submissions performed similarly. Four of the top five teams used ensemble learning methods. However, one team's algorithm performed significantly superior to the other submissions, and consisted of an asymmetrical U-Net network architecture. This paper provides a first of its kind benchmark for future automatic multi-tissue segmentation algorithms for the developing human brain in utero.

1. Introduction

Fetal *in-utero* magnetic resonance imaging (MRI) is a powerful tool to investigate the developing human brain in fetuses with and without pathological features (De Asis-Cruz et al., 2021; Hosny and Elghawabi, 2010). It can be used to portray the complex neurodevelopmental events during human gestation, which remain to be completely characterized (Vasung et al., 2019). Clinically, it is becoming an important adjunct to ultrasound in the detection and diagnosis of congenital disorders (Hart et al., 2020), and can be used to aid during prenatal care (Gholipour et al., 2014).

Automated segmentation and quantification of the highly complex and rapidly changing brain morphology using MRI prior to birth has great potential to improve the diagnostic process, as manual segmentation is both time consuming and subject to human error and inter-rater variability. It is clinically relevant to analyze the morphometry of the developing brain, where measures such as the volume or the shape can be objectively compared with population-based references of normative development. Many congenital and acquired disorders manifest in reduced brain volume or altered anatomical structure of cerebral tissue compartments, for example, slower cortical growth (Clouchoux et al., 2013; Egaña-Ugrinovic et al., 2013) or reduced white matter volume (Rollins et al., 2021). Existing MRI based data of brain growth is mainly based on normally developing brains (Jarvis et al., 2019; Kyriakopoulou et al., 2017; Prayer et al., 2006), leaving brain growth in various numerous pathologies and congenital disorders largely unexplored.

From a technical standpoint, there are many challenges that an automatic segmentation method of the fetal brain would need to overcome. The cerebral structures are constantly growing and developing in complexity throughout gestation, which results in a gradually changing appearance in shape, size, and image intensity on MRI. In addition, the quality of the images can be poor due to fetal and maternal movement and imaging artefacts (Glenn, 2010). The boundary between tissues is often unclear on MR images due to partial volume effects (Bach Cuadra et al., 2009). Furthermore, fetal brains with abnormal features can have radically different morphology than those in a non-pathological brain. This can make it challenging for an automatic method to correctly identify these structures.

Fetal MRI requires no special MRI equipment, is noninvasive, safe (Gowland, 2011; Zvi et al., 2020), and its value in the diagnosis of certain central nervous system or somatic disorders is being increasingly recognized (Griffiths et al., 2019; Nagaraj et al., 2022). The development of ultra-fast MRI sequences such as the single shot T2-weighted sequence have also led to the increasing popularity of fetal MRI as these images have excellent soft tissue contrast and reduced motion artefact (Gholipour et al., 2014). As a result, fetal MRI is more frequently performed at diagnostic and surgical centers worldwide. There is also an increase in the number of studies focused on developing computational tools to quantitatively analyze the fetal brain. Some studies have focused on segmenting a specific tissue for analysis, such as the cortical plate (Benkarim et al., 2018; de Dumast et al., 2020; Fetit et al., 2020; Hong et al., 2020). Other studies have developed multi-tissue segmentation algorithms using a limited in-house dataset (e.g., clinically acquired anisotropic coronal images of normal fetuses (Khalili et al., 2019), or images of a specific pathology (Sanroma et al., 2018), or with atlas based frameworks (Dittrich et al., 2011; Gholipour et al., 2017; Licandro et al., 2016; Wu et al., 2021b; Xu et al., 2022)). Recently, research groups have developed more extensive in-house datasets with which to train automatic segmentation networks, but these datasets remain private (Karimi et al., 2023; Zhao et al., 2022). The field of developing automated tools for fetal MRI has been understudied due to both challenges in imaging and the lack of public, curated, and annotated ground truth data. Such shared datasets are currently the backbone for developing computer-aided diagnostic support systems.

In this paper we describe the Fetal Brain Tissue Annotation and Segmentation Challenge (FeTA) and outline the challenge organization, the submitted segmentation frameworks, and a detailed evaluation of the challenge results, with reporting based on the BIAS method (Maier-Hein et al., 2020). The aim of the FeTA Challenge was to develop reliable, valid, and reproducible methods of analyzing high resolution reconstructed MR images of the developing fetal brain from gestational week 20–35. The FeTA Challenge used an expanded version of the original FeTA Dataset to develop automatic fetal brain tissue segmentation methods (Payette et al., 2021a). Our evaluation compares and

analyzes the algorithms on a test dataset hidden to the participants. The submitted algorithms are also tested on various subsets of the testing dataset in order to determine whether they perform better or worse under various circumstances such as image quality or reconstruction method. Finally, we investigated two real life applications outside the scope of the FeTA Challenge evaluation: First, the performance of the submitted algorithms to estimate intracranial volume was evaluated, an application relevant to the characterization of developmental delay in many conditions, such as intrauterine growth restriction or congenital heart defects (Polat et al., 2017; Sadhwani et al., 2022; Skotting et al., 2021). Second, we looked at the ability of the algorithms to segment younger (<29 weeks) versus older (≥29 weeks) fetal brains). The algorithms developed as part of the FeTA Challenge will have the potential to help better understand the underlying causes of congenital disorders and ultimately to guide the development of perinatal guidelines and clinical risk stratification tools for early interventions, treatments, and care management decisions.

2. Materials and methods

2.1. Challenge organization

The FeTA Challenge was held as part of the international Medical Image Computing and Computer

Assisted Intervention (MICCAI) 2021 Conference (https://feta.grand-challenge.org/). Participants were to create a fully automatic multi-class segmentation algorithm of the fetal brain (with optional inputs of gestational age and whether the brain was pathological or not). The training dataset was made available to the participants on May 3rd, 2021 on Synapse (https://www.synapse.org/#!Synapse:syn25649159/wiki/610007), (Payette and Jakab, 2021) to train their own methods. Participants were able to use other publicly available datasets for training if they wished to, as long as it was documented in their algorithm description. Participants created a Docker container which stored the algorithm, and submitted this container to the organizers by July 30, 2021. Organizers were allowed to submit containers, but were not eligible for prizes. This container was run by the challenge organizers locally on the hidden testing dataset in order to compare the algorithms. *Re*-submission of the Docker container was only allowed in cases of technical difficulties or bugs identified during evaluation. The top teams received their results on September 1, 2021 in order to prepare presentations. The complete results and awards to the top three teams were presented on Oct 1, 2021 at the MICCAI Conference FeTA Challenge Session. Dockers of teams who provided permission are available on Dockerhub (https://hub.docker.com/u/fetachallenge). For the complete overview of the challenge, see the final challenge proposal (Payette et al., 2021b).

2.2. Mission of the challenge

The mission of the FeTA Challenge is to boost the development of accurate and automatic multi-class segmentation algorithms for the developing human brain with fetal MRI, and to create a benchmark for future algorithms. There were a total of eight classes: external cerebrospinal fluid (eCSF), gray matter (GM), white matter (WM), ventricles (including cavum), cerebellum, deep gray matter (deep GM), brainstem, and background. The target cohort for the FeTA Challenge were pregnant mothers who, after an initial ultrasound examination, were clinically referred for a fetal MRI. The acquired fetal MRI images were then reconstructed into a 3-dimensional volume using a super-resolution method (for details see Section 2.3). The task of the challenge was to segment these super-resolution volumes into different brain tissues. The challenge cohort was made up of two subgroups: fetuses with normal and abnormal development of the nervous system, and covers a gestational age (GA) range of 20–35 weeks. The accuracy of the automatically generated fetal brain segmentations was evaluated in the challenge cohort in order to determine the optimal segmentation method for fetal brain MRI.

2.3. Challenge dataset

For the challenge, a clinically acquired dataset from a single institution was used for both the training and testing data. 120 fetal MRI brain scans were acquired. Recorded gestational age was modified by a random value within the range of ±3 days to further anonymize the data. Several T2-weighted single shot Fast Spin Echo (ssFSE) images were acquired for each subject in all three planes with a reconstructed resolution of 0.5mm x 0.5mm x 3 to 5mm. The images were acquired on either a 1.5T or 3T clinical GE whole-body MRI scanners (Signa Discovery MR450 and MR750) using an 8-channel cardiac or body coil with the following sequence parameters: TR: 2000–3500ms, TE: 120ms (minimum), flip angle: 90°, sampling percentage 55%. Field of view (200–240mm) and image matrix (1.5T: 256×224; 3T: 320×224) were adjusted depending on the gestational age and size of the fetus. The data was acquired at the University Children's Hospital Zurich in Zurich, Switzerland by trained radiographers using clinically defined protocols.

For each subject, the acquired images were reviewed, and images of good quality, at least one image in each of the axial, sagittal, coronal planes with respect to the fetal brain, were chosen. A high-resolution fetal brain reconstruction was performed with the chosen scans using a super-resolution (SR) method (60 cases reconstructed with the mialSR method (Pierre Deman et al., 2020; Tourbier et al., 2019; S. 2015) and 60 cases reconstructed with the Simple IRTK method (Kuklisova--Murgasova et al., 2012)). Fetal brain masks were created where necessitated by the SR algorithm, either manually or with a custom MeVisLab module (Pierre Deman et al., 2020; Tourbier et al., 2015). Cases reconstructed with mialSR were reoriented prior to reconstruction through the MeVisLab module. Cases reconstructed with the Simple IRTK method were registered to an atlas after reconstruction (Serag et al., 2012). After reconstruction, each fetal brain volume had an isotropic resolution of approximately 0.5mmx0.5mmx0.5mm, with some deviation in exact dimensions between the SR methods. Each reconstructed image was then histogram-matched using Slicer (Kikinis et al., 2014), and zero-padded to be 256×256×256 voxels. For each reconstruction method, 40 cases were included in the training dataset available to the challenge participants (for a total of 80 cases), and 20 cases were included in testing dataset not available to the participants (for a total of 40 cases). Note that maternal tissue was excluded from the super-resolution reconstruction, only the fetal brain was reconstructed. Examples of non-pathological fetal brains across the range of gestational ages included in the dataset and their corresponding label maps can be seen in Fig. 1.

The training and testing datasets consisted of fetuses with both typical and atypical features. In the group with atypical features, a variety of cerebral pathologies of varying severities were included (such as Chiari-II malformation or ventricular dysmorphology seen in ventriculomegaly). There were slightly more pathological than neurotypical cases, as in the clinic where the scans were performed it is more common to see pathologic brains (Fig. 2). Fetuses with a gestational age range of 20 to 35 gestational weeks were included (mean gestational age: 27.0 ±3.60 weeks), with the distribution of ages and pathologies equal between the training and testing datasets (see Fig. 3). The gestational age and the label of "neurotypical/pathological" was made available to the participants. Each case's label map was manually segmented by individuals with experience in segmenting medical images using the method described in (Payette et al., 2021a). Each case consists of a 3D super-resolution reconstruction of a fetal brain (256×256×256 voxels) and the associated manually segmented label map. There is no overlap of subjects between the training and testing dataset, each dataset is unique. The dataset and affiliated custom license is publicly available on Synapse (Payette and Jakab, 2021).

Mothers of the healthy fetuses participating in the BrainDNIU study

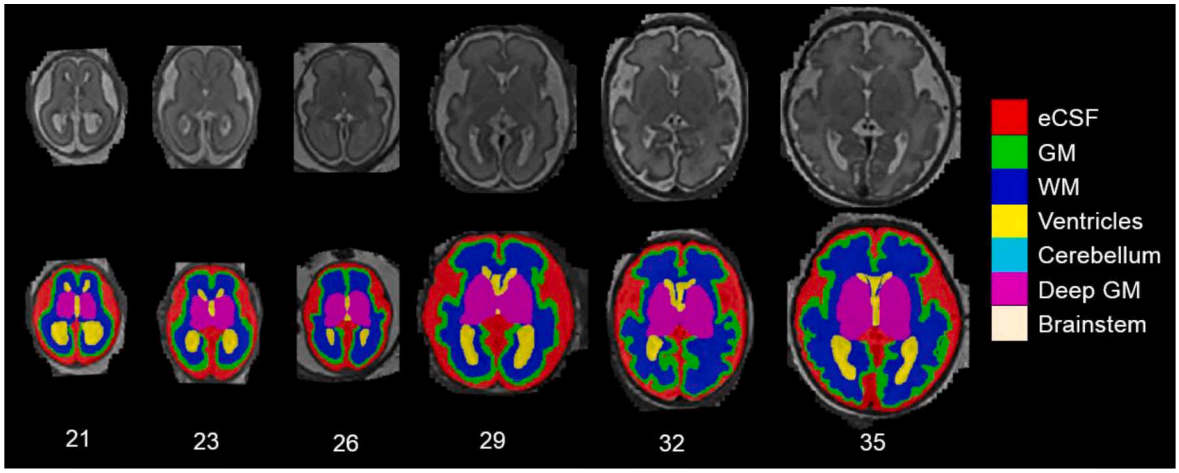

Fig. 1. Fetal Brain Segmentations by gestational age.

were prospectively informed about the inclusion in the FeTA Dataset by members of the research team and gave written consent for their participation. Mothers of all other fetuses included in the current work were scanned as part of their routine clinical care and gave informed written consent for the re-use of their data for research purposes. The ethical committee of the Canton of Zurich, Switzerland approved the prospective and retrospective studies that collected and analyzed the MRI data (Decision numbers: 2017-00885, 2016-01019, 2017-00167), and a waiver for an ethical approval was acquired for the release of a fully anonymous dataset for research purposes.

Participants were free to choose if they wanted to work with the data in a 2D or 3D format. A validation dataset was not provided to the participants, it was up to the team's discretion to decide how to train their data and what to use for validation. The following section outlines the evaluation metrics used to determine the ranking of participants in the challenge.

2.4. Assessment method

2.4.1. Evaluation metrics

Three different metrics were chosen to compute the rankings of the FeTA Challenge: the Dice Similarity Coefficient (DSC), The Volume Similarity (VS), and the Hausdorff distance (HD). The DSC was chosen, as it is the most popular segmentation overlap metric for segmentation evaluation (Dice, 1945). However, we were also interested in assessing volume, as relevant biomarker for fetal development, and surface-based error. Therefore the HD (surface, (Hausdorff, 1991)), and VS (volume) metrics were chosen as well (Taha and Hanbury, 2015), and the final ranking will take all three metrics into account.

The DSC measures the amount of overlap between the manual segmentations (MS) label and the new segmentation (NS) generated by the participant's algorithm, and is defined as

$$DSC = \frac{2\,|MS\ \cap\ NS|}{|MS| + |NS|}$$

The VS is a volumetric metric that measures the similarity between the volume of the GT and NS label map and is defined as

$$VS\ = 1 - \frac{|MS_{vol} - NS_{vol}|}{MS_{vol} + NS_{vol}}$$

The HD is a distance metric that evaluates the distance between two

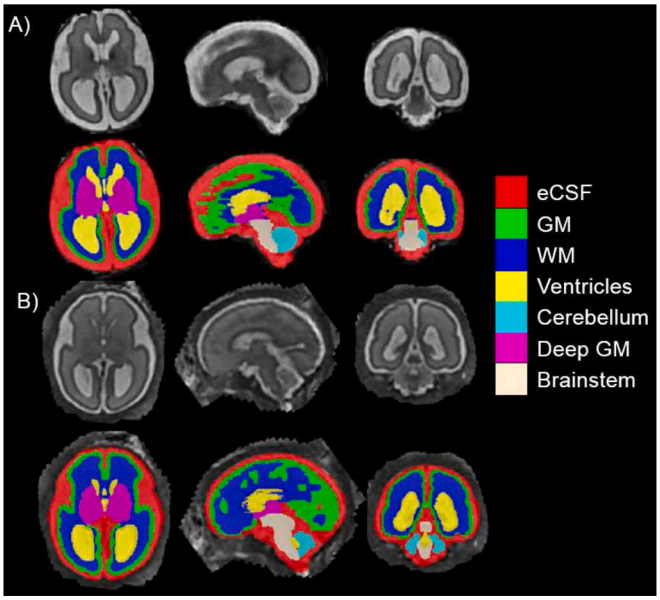

Fig. 2. Pathological fetal brain viewed in axial, sagittal, and coronal directions A): mialSR reconstruction, 27.3 GA; B) Simple IRTK SR Reconstruction, 26.9 GA.

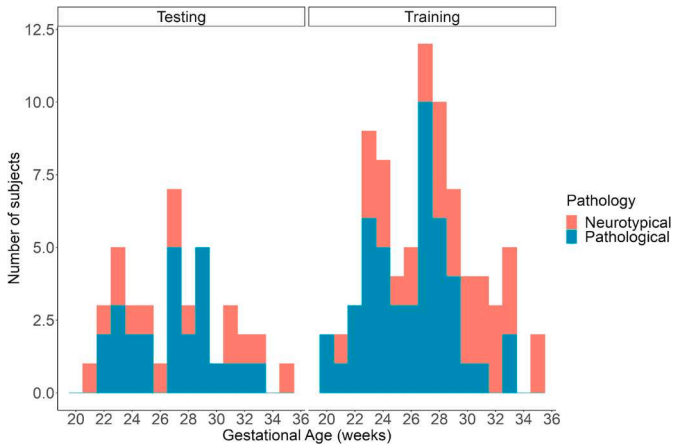

Fig. 3. Dataset Age Range: Histogram of the gestational age range of neurotypical and pathological cases within the testing and training dataset for the FeTA Challenge.

finite point sets A and B.

The Hausdorff distance (HD) is a spatial metric helpful in evaluating the contours of segmentations as well as the spatial positions of the voxels. The HD between two finite point sets A and B is defined as

$$HD(A,B) = max(h(A,B),\ h(B,A))$$

$$h(A,B) = ||a-b||$$

Note: The original challenge design had stated that the 95th percentile HD of maximum distances would be used to exclude possible outliers. However, after the challenge it was discovered that there was an error in the implementation of the 95th percentile, and the values reported were close to the maximum HD, and therefore these are the scores reported in this paper. This makes the HD values reported within this report slightly more susceptible to outliers. However, as we take three different metrics into account for the final ranking, the overall impact of outliers is reduced. For each metric, the implementation described in (Taha and Hanbury, 2015) was used (EvaluateSegmentation Tool, v2017.04.25).

2.4.2. Ranking

Each of the participating teams was ranked based on each evaluation metric, and then the final rankings combined the rankings from all of the metrics (DSC, HD95, VS). The DSC, HD95, and VS were calculated for each label within each of the corresponding predicted label maps of the fetal brain volumes in the testing set. The mean and standard deviation of each label for all test cases was calculated, and the participating algorithms were ranked from low to high (HD95), where the lowest score received the highest scoring rank (best), and from high to low (DSC, VS), where the highest value received highest scoring rank (best) based on the calculated mean across all labels and test cases. If there were missing results, the worst possible value is used. For example, if a label does not exist in the NS label map but is present in the GT label map, it will receive a DSC and VS score of 0, and the HD95 score will be double the max value of the other algorithms submitted. This ranking procedure was developed in order to take three different metric types equally into account.

2.4.3. Further analysis

In addition to the ranking above, several other analyses were performed on the submitted algorithms. Per-label rankings of the entire dataset were analyzed. In addition, the algorithms were evaluated in the categories 'Non Pathological cases' and 'Pathological cases', SR reconstruction method (mialSR and Simple IRTK) as well as 'Excellent Quality', 'Good Quality' and 'Poor Quality', with the identical ranking methodology for each category. The pathology of each fetal brain was determined by an experienced radiologist. The quality of the fetal brain SR reconstructions were determined based on ratings (Excellent, Good, Poor) from three independent raters, and the correlation of the reviewers was calculated using the Gwet AC coefficient using R (v4.0.2, (Gwet, 2019)). As the ratings are ordinal data, the median of the ratings were considered to be the final rating of the SR volume. The participating algorithms were also evaluated on the different SR reconstruction methods.

Intracranial volume was calculated and compared to the manual segmentation's intracranial volume as well, but not used in the rankings. Intracranial volume was calculated by adding all labels except the background together.

An analysis of the performance of the algorithms based on gestational age was also performed, as the structure of the fetal brain changes greatly throughout development, especially in the cortex where there is increased cortical complexity, disappearance of transient subplate zone related to cortical maturation (blurring of white matter and gray matter border) and partial volumes (blurring of white matter/gray matter border in gyral crest, blurring of CSF/gray matter border because of narrow sulci). Because of this, the random error in segmentation of the gray matter between 29 and 35 GW might be increased. Therefore, in order to determine if gestational age impacts the success of a segmentation algorithm, our testing dataset was split into two age groups (21–28 weeks, and 29–35 weeks), and the differences between these two groups was analyzed by looking for differences in the evaluation metrics between each label for each of the submitted algorithms.

3. Results

3.1. Training and testing data

A Kolmogorov–Smirnov test was performed in R v4.0.0 (R Core Team, 2020) in order to compare the distribution of GA and non-pathological/pathological fetal brains between the training and testing data. No significant differences were found between the training and testing datasets (GA: p=0.88; pathology: p=1).

3.2. Challenge submission

In total, 21 teams submitted algorithms to the FeTA Challenge. One team's Docker container was not able to be fixed prior to the deadline and was thereby excluded. One team (Ichi-love) submitted two algorithms, meaning the total number of participating teams was 20, and the total number of valid submissions was 21. Each team submitted a written description of their algorithm, which can be found in the Appendix. Each algorithm is summarized in Table 1, and the pre-processing

Table 1
Overview of algorithms submitted to the FeTA Challenge ordered from best to worst.

Team Name	Network	Loss Function	2D/ 3D	Patch Size	Post-Processing	Convolution Kernel Size	Optimizer	Initialization	Learning Rate	Cross-Validation	Epochs	GPU Used	# of Layers	# of Trainable Parameters
NVAUTO	MONAI (SegResNet), OCR modules	Dice	3D	224×224×144	Ensemble learning	3 × 3 × 3	AdamW	Random	0.0002, decrease to 0 at final epoch with cosine annealing scheduler	5-fold	300	4 x Nvidia V100 32GB	5 desc / 5 asc	75 819 624
SJTU_EIEE_2–426Lab	two steps: coarse to fine, 1. nnU-net and 3D UNet with residual architecture; 2. 5 3D Res-Unets	1. Cross-entropy and Dice; 2. Haus-dorff and Dice	3D	128×128×128 - nn-UNet. only	Ensemble learning	3 × 3 × 3	Adam	Random	1. 1E-3; 2. 1E-4	No	1. 500; 2. 1000	Nvidia RTX 3090	6 desc / 6 asc	1. 2 235 680 (UNet); 31 199 584 (nnUNet) 2. 214 58 929 (first UNet); 85 823 969 (other 4 UNets)
pengyy	nnU-Net	Cross-entropy and Dice	3D	128×128×128	Ensemble learning	3 × 3 × 3	Stochastic Gradient Descent	Random	0.01 with reduction	10-fold	1000	Nvidia GeFor-ce RTX 3090	6 desc / 6 asc	72 142 688
Hilab	nnU-Net	Cross-entropy and Dice	3D	128×128×128	Ensemble learning	3 × 3 × 3	Stochastic Gradient Descent	Random	0.01 with decay	5-fold	400	Nvidia GeFor-ce RTX 2080 Ti	6 desc / 6 asc	30 847 564
Neurophet	U-Net	sum of Cross-entropy and Dice	3D	64×64×64	Isolated segmented voxels removed	3 × 3 × 3	AdamW	Random	1.00E-05	No	500	3 x Tesla V100	5 desc / 5 asc	314 999 688
davoodkari-mi	U-Net with additional short and long skip connections	Novel loss function derived from mean absolute error	3D	128×128×128	Label Fusion	3 × 3 × 3	Adam	He	1E-4 with reduction	No	400	Nvidia GeFor-ce GTX 1080	5 desc / 5 asc	18 500 000
2Ai	U-Net/nnU-Net	Dice	3D	128×128×128	Isolated segmented voxels removed	3 × 3 × 3	Adam	Xavier	1E-3 with decay	No	800	1 x GTX1070	6 desc / 6 asc	29 971 032
xlab	U-Net/nnUnet	Cross-entropy and Dice	2D	No	None	3 × 3	Adam	Random	3.00E-04	5-fold	1000	Nvidia RTX 3090	5 desc / 5 asc	–
Ichilove-axe	Two step networks. Dynamic U-Net with pre-trained ResNET34 network blocks (desc) and pixelShuffle ICNR blocks (asc)	Lovasz-Softmax loss	2D	No	None	3 × 3	OneCycle	ResNet34 - encoder, ICNR - Decoder	1.00E-03	No	60	1 x GTX1080Ti	4 desc / 4 asc	41 221 768
TRABIT	DynU-Net from MONAI; 10 networks	Label-set Loss function: Leaf-Dice and marginalized cross entropy	3D	128×160×128	Ensemble learning	3 × 3 × 3	Stochastic Gradient Descent	He	0.01 with decay	No	2200	1 x Tesla V100-SXM2–32GB	6 desc / 6 asc	31 195 784
Ichilove-Combi	Two step networks. One for ROI, 3 for each axis (Coronal, Axial, Sagittal). Dynamic U-Net with pre-trained	Lovasz-Softmax loss	2D	No	Label Fusion	3 × 3	OneCycle	ResNet34 - encoder, ICNR - Decoder	1.00E-03	No	60	1 x GTX1080Ti	4 desc / 4 asc	103 054 420

(*continued on next page*)

Table 1 (*continued*)

Team Name	Network	Loss Function	2D/3D	Patch Size	Post-Processing	Convolution Kernel Size	Optimizer	Initialization	Learning Rate	Cross-Validation	Epochs	GPU Used	# of Layers	# of Trainable Parameters
	ResNET34 network blocks (desc) and pixelShuffle ICNR blocks (asc)													
muw_dsobotka	multi-task U-Net with two decoders (segmentation and reconstruction)	Homoscedastic uncertainty, cross-entropy, mean squared error	3D	128×96×96	None	3 × 3 × 3	Adam	Random	0.001	No	100	Nvidia GeFor-ce RTX 2080 Ti	3 desc / 3 asc	6 491 385
Physense-UPF Team	nnU-Net	Cross-entropy and Generali-zed Dice	3D	128×128×128	None	2 × 2 × 2	Stochastic Gradient Descent	Random	0.01	5-fold	100	1x Nvidia GEFORCE GTX 1080 Ti	6 desc / 6 asc	31 199 584
SingleNets	U-Net	Soft Dice and Contour Dice	3D	96×96×96	Majority Voting, clipping using "skull" (background-foreground) network response with threshold 0.5	3 × 3 × 3	Adam	Fine tuning from previously trained networks on smaller training set	0.005 with reduction	No	100	Tesla M60	5 desc / 5 asc	4 727 841
BIT_LILAB	CNN-Transformer Hybrid (Trans-U-Net)	Cross-entropy and Dice	2D	16×16	None	3 × 3	Stochastic Gradient Descent	Pre-trained ResNet-50 and ViT	1E-2 with decay	No	150	4 x Nvidia GTX 1080Ti GPU	5 desc / 5 asc	54 000 000
Moona Mazher	DenseNet	Binary Cross-entropy	2D	No	Label Fusion	3 × 3	Adam	Random	0.0003	5-fold	1000	4 x Nvidia V100	5 desc / 5 asc	49 510 728
MIAL	U-Net	hybrid loss (Dice and Cross-entropy)	2D	64×64	Majority Voting	3 × 3	Adam	Random	0.001 with decay	5-fold	100	NVIDIA RTX 2070	5 desc / 5 asc	–
ZJUWULAB	U-Net with conv downsampling instead of max pooling downsampling	L1 Regularization and feature-matching with a pre-trained VGG19 Network	2D	No	None	3 × 3	Adam	Random	0.002	No	100	4 x RTX 3080Ti	5 desc / 5 asc	7 765 442
FeVer	Res-Unet	Dice	3D	48×224×224	Ensemble learning	3 × 3 × 3	QHAdam	Random	0.005–0.0005	No	300	1 x RTX 3090	5 desc / 5 asc	2 369 496
Anonymous	U-Net	Focal Loss	2D	No	None	3 × 3	Adam	Random	0.002	No	30, backbone frozen for 15	-	–	-
A3	V-Net with PReLU activation	Binary Cross-entropy	3D	Crop-ped & pad-ded to 192×192×192; down-sampled to 128×128×128	None	3 × 3 × 3	Adam	Random	1E-4 with reduction	No	200	2 x NVIDIA P100,	3 desc / 3 asc	283 886 304

Table 2
Overview of the data augmentation, and pre-processing used in each submission.

Team Name	Data Augmentation	External Dataset used	Pre-processing
NVAUTO	Rotation, Flipping, Zoom, contrast adjustment, Gaussian noise, Gaussian smoothing	No	Normalize images to zero mean
SJTU_EIEE_2–426Lab	Rotation, Scaling, Flipping	No	Normalize images to zero mean, cropping in 2nd stage
pengyy	Rotation, scaling, elastic deformation, mirroring, Gaussian noise, Gamma Correction	No	resample dimensions to .5x.5x.5 mm; z-score normalization
Hilab	Pathological Cases copied 3 times in training data, rotation, scaling, Gaussian noise, Gaussian blur, brightness, contrast, simulation of low resolution, gamma augmentation, mirroring	No	Cropping and normalization
Neurophet	Affine, Blur	No	Intensity Normalization, classification of images into poor and good quality
davoodkarimi	Flipping, rotation, elastic deformation, label perturbation and smoothing	No	Intensity Normalization
2Ai	Flipping, rotation, scaling, grid distortion, optical distortion, elastic transformations, noise, brightness, contrast, gamma transformations	No	Image normalization (mean value zero)
xlab	Mirroring, rotation, scaling, gamma correction, random elastic transformation	No	nnUNet standard preprocessing
Ichilove-axe	Intensity, contrast, scaling, normalization, rotation, intensity inhomogeneity	No	No
TRABIT	Flipping, zooming, rotation, Gaussian noise, spatial smoothing, gamma augmentation	Uses external fetal brain atlases and neonatal MRI's segmented with dHCP	generated brain mask (from atlas + niftyReg), registered brain to atlas, resampled to 0.8 mm isotropic; skull stripping; thresholding of intensity percentiles; z-score normalization
Ichilove-Combi	Intensity, contrast, scaling, normalization, rotation, intensity inhomogeneity	No	No
muw_dsobotka	Elastic deformation, flipping, rotation, contrast, Gaussian noise, Poisson noise	No	z-score normalized patches
Physense-UPF Team	Rotation, elastic deformation, scaling, Gaussian noise, Gaussian blur, gamma transform, mirror, brightness, contrast, low resolution simulation, zoom	No	Cropping, resampling, normalization, classification of quality of images, registration to Gholipour atlas
SingleNets	Flipping, rotation, translation, scaling, Poisson noise, contrast, intensity	No	Thresholding, cropping, windowing, normalization, downscaling by 0.5 in all axes
BIT_LILAB	Rotation and flipping	Yes, Synapse multi-organ segmentation dataset (for pre-training)	None
Moona Mazher	Cropping, Flipping, Brightness and Contrast, Random Gamma	No	None
MIAL	Flipping, Rotation	No	Intensity standardization
ZJUWULAB	No	yes, pre-trained VGG19network	Normalize, color map labels
FeVer	Flipping, Mixup	No	Intensity-based image filtering, resampling voxels to have equal spacing, remove slices with only background label
Anonymous	No	ResNet backbone pre-trained on Kinect400	Intensity re-scaling; ResNet backbone pre-trained on ImageNet
A3	Shifting, rotation, flipping	No	Image normalization (mean value zero)

and data augmentation used by each team is outlined in Table 2. All teams submitted a deep learning-based method, most of which were variants based on the U-Net architecture (Çiçek et al., 2016; Ronneberger et al., 2015). The top five teams used similar loss functions (mainly the combination of Dice loss and cross-entropy loss), and four of the five (excluding Neurophet) used an ensemble learning method. Every method used a 3×3 (or 3×3×3) convolutional kernel except for one team (Physense-UPF) who used a 2×2×2 kernel. Most submissions (14 of 21) used a random initialization of network parameters. The networks were of varying depths, between 3 and 6 layers on each of the ascending and descending layers of the networks. Thirteen of the submitted networks were 3D networks, the remainder were 2D or 2.5D. Seven teams used cross-validation. A variety of different data augmentation strategies were used, and only two team did not employ data augmentation at all. Only four teams used external datasets, either during the training step or used pre-trained network backbones trained on publicly available datasets.

3.3. Metric values and rankings

Statistical analysis of the metrics of the challenge and images displayed in this section were created using the ChallengeR tool (Wiesenfarth et al., 2021). The individual metrics for each team (all labels combined) can be found in Figs. 4-6. The final ranking of all teams and their average evaluation metrics can be found in Table 3. The full reports (DSC, HD95, VS of all labels combined) created by the ChallengeR Tool, including details on the statistical tests performed can be found in Sections 2-4 of the Appendix. In the significance maps displayed, the testing was done using a one-sided Wilcoxon signed rank test at a 5% significance level, with adjustments for multiple comparisons. In all cases, the x-axis in the boxplots are ranked according to the mean values of the respective evaluation metric, and the black bar indicates the median value.

The top three teams according to the DSC were NVAUTO, SJTU_EIEE_2–426Lab, and Neurophet. The top three teams according to the HD95 were NVAUTO, Hilab, and 2Ai. The top three teams according to the VS were ichilove-axe, NVAUTO, and SJTU_EIEE_2–426Lab. With a few exceptions, there was no statistically significant differences between the top 10–12 teams in all three metrics, suggesting that a plateau has been reached. The highest and lowest average DSC were: 0.786 (team NVAUTO) and 0.534 (team A3). The lowest and highest average HD95 were: 14.012 (team NVAUTO) and 39.608 (team A3). The highest and lowest average VS were: 0.888 (team ichilove-axe) and 0.791 (team A3). However, when the bootstrapping and significance maps are investigated, it is clear that NVAUTO is the top team for the DSC metric, placing first in 100% of the bootstrap sampling, and is statistically significant to

all but one of the algorithms (Team pengyy). There is no difference between teams in places 2 to 4 for the DSC metric in both the bootstrapping and statistical significance testing. The same trend appears when looking at the HD95 metric, with NVAUTO being the clear winner, and the teams in places 2 to 4 performing equivalently. Some differences exist in the VS metrics, with ichilove-axe as the first place, but with not as clear of a lead with no statistical difference from any of the top teams, and a less clear winner when looking at the bootstrapping.

3.4. Further analysis

A variety of subsets of the data were created in order to determine if the algorithms perform better or worse based on various criteria such as image quality, SR method used, and normal vs pathological brains. The rankings of the teams based on the different subsets can be seen in Fig. 7. A large amount of variability in the rankings is present depending on the subset of data being investigated. However, NVAUTO remains in the number 1 ranking spot in all subsets except two (Excellent Quality and IRTK_SR).

3.4.1. Per-label metric values and ranking

Each team's algorithm was analyzed separately per tissue label. The average DSC, HD95, and VS scores for each team and label can be found in Figs. 8- 10. The order of the teams on the x-axis in each graph is ordered from best to worst, left to right. When looking at the DSC, team NVAUTO placed first in all labels except eCSF (MIAL), deepGM (TRABIT), and brainstem (SJTU_EIEE_2–426Lab). When looking at the HD95, team NVAUTO placed first in all labels except eCSF (ichilove-combi), Ventricles (Hilab), and deepGM (SJTU_EIEE_2–426Lab). In the VS metric, almost every label had a different top team: eCSF (MIAL), GM (2Ai), WM (NVAUTO), Ventricles (NVAUTO), Cerebellum (ichilove-axe), deepGM (A3), and brainstem (SJTU_EIEE_2–426Lab).

Fig. 11 shows example error maps of the gray matter level for two test cases. The label maps of the top 5 teams were analyzed to show voxels where many teams mis-identified the cortical gray matter, and where all top 5 teams were able to correctly identify it.

3.4.2. Image quality

The dataset was split into three subsets based on the quality of the SR reconstructions as determined by experienced raters (excellent quality SR: n=11 (mialSR/IRTK: 1/10); good quality SR: n=25 (mialSR/IRTK: 15/10); poor quality SR: n=4 (mialSR/IRTK: 4/0)). Each team's algorithm was analyzed with the average metrics across all labels. The average DSC, HD95, and VS scores for each team and label can be found in Fig. 12. The order of the teams on the x-axis in each graph is ordered from best to worst, left to right. Team pengyy performed the best (according to the DSC) when the fetal brain reconstructions were of excellent quality, while NVAUTO performed the best for good and poor quality reconstructions. Complete ranking information taking all three metrics into account based on SR reconstruction quality can be found in Fig. 7.

3.4.3. SR reconstruction

The dataset was split into two subsets based on SR reconstruction method used. Each team's algorithm was analyzed with the average metrics across all labels. The average DSC, HD95, and VS scores for each team and label can be found in Fig. 13. The order of the teams on the x-axis in each graph is ordered from best to worst, left to right. Team NVAUTO performed the best (according to the DSC) with the mialSR reconstruction, and Team SJTU_EIEE_2–426Lab performed the best (according to the DSC) with the IRTK SR reconstruction. Complete ranking information taking all three metrics into account for each SR reconstruction can be found in Fig. 7.

3.4.4. Pathology

The dataset was split into two subsets based on whether the fetal brain contained a pathology (n=25 (mialSR/IRTK: 14/11)) or not (neurologically normal, n=15 (mialSR/IRTK: 6/9)). Each team's algorithm was analyzed with the average metrics across all labels. The average DSC, HD95, and VS scores for each team and label can be found

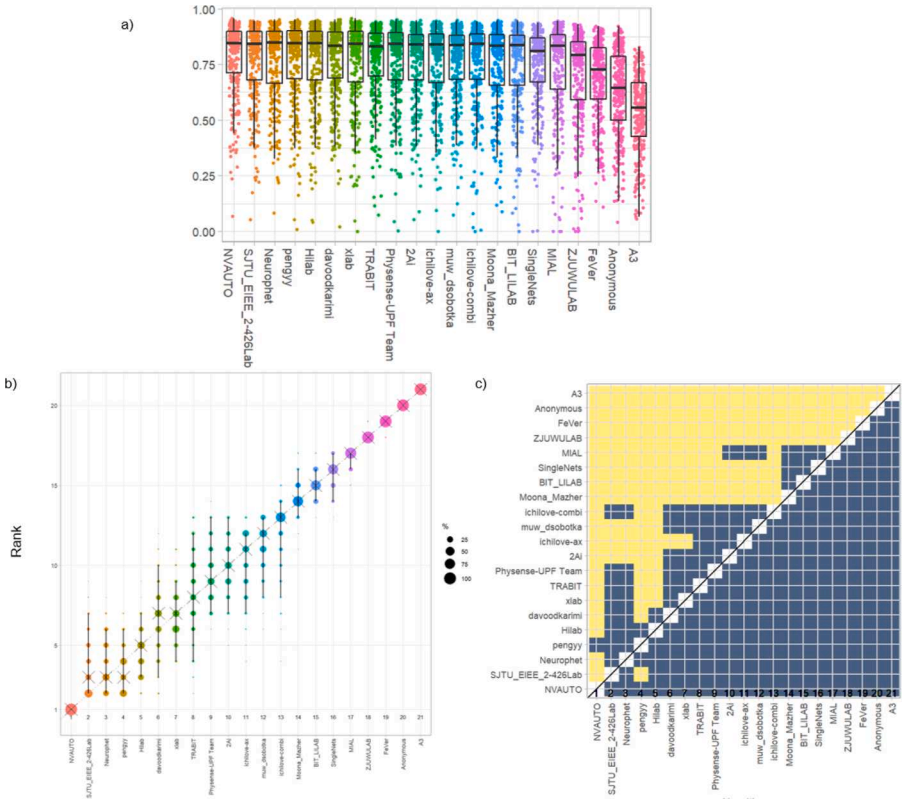

Fig. 4. DSC values of FeTA Challenge participants a) Dot and box plot; b) Blob plot for visualizing ranking stability based on bootstrap sampling, black cross indicated the median rank for each algorithm and 95% bootstrap intervals across samples are indicated by black lines; c) Significance maps for visualizing ranking stability based on statistical significance (Yellow: metrics from the algorithm on the x-axis were significantly superior to the algorithm on the y-axis, blue color indicates no significant difference). Figures were created using the ChallengeR Tool (Wiesenfarth et al., 2021).

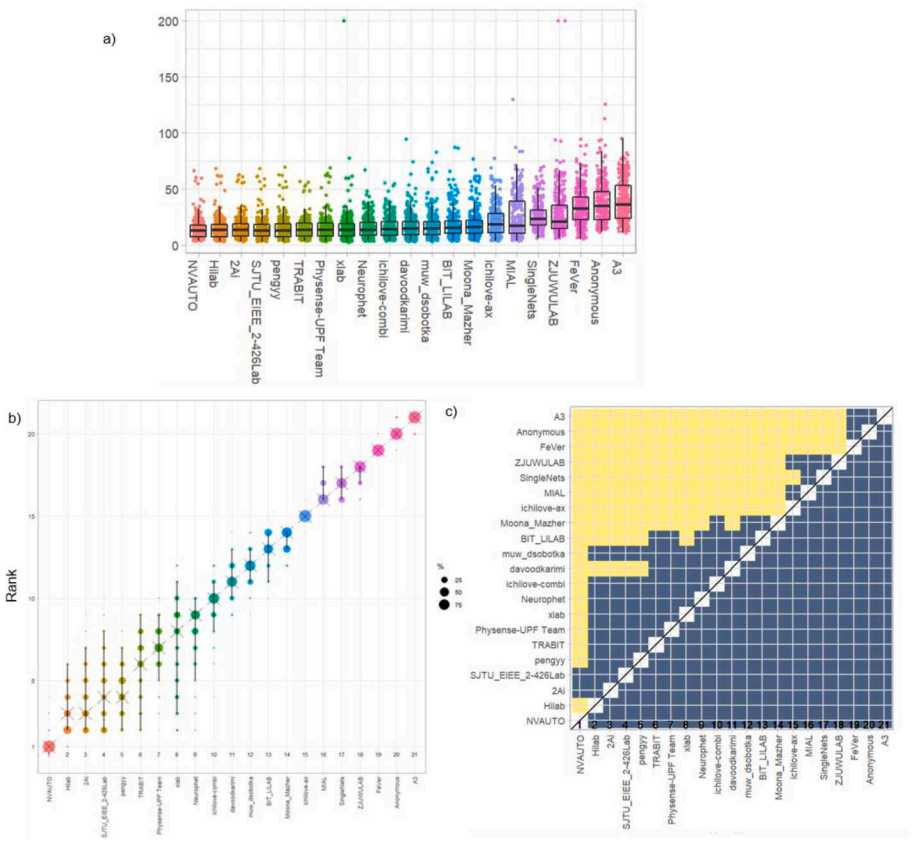

Fig. 5. HD95 values of FeTA Challenge participants a) Dot and box plot; b) Blob plot for visualizing ranking stability based on bootstrap sampling, black cross indicated the median rank for each algorithm and 95% bootstrap intervals across samples are indicated by black lines; c) Significance maps for visualizing ranking stability based on statistical significance (Yellow: metrics from the algorithm on the x-axis were significantly superior to the algorithm on the y-axis, blue color indicates no significant difference.).

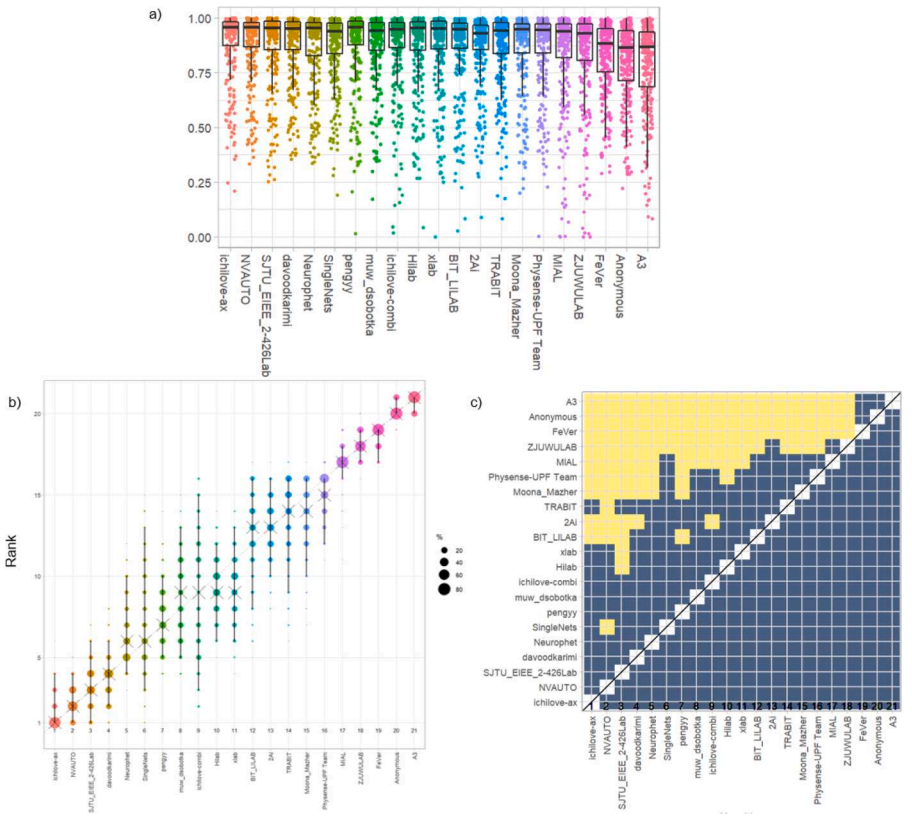

Fig. 6. VS values of FeTA Challenge participants a) Dot and box plot; b) Blob plot for visualizing ranking stability based on bootstrap sampling, black cross indicated the median rank for each algorithm and 95% bootstrap intervals across samples are indicated by black lines; c) Significance maps for visualizing ranking stability based on statistical significance (Yellow: metrics from the algorithm on the x-axis were significantly superior to the algorithm on the y-axis, blue color indicates no significant difference.).

Table 3
Final FeTA Ranking; * indicates a tie.

Ranking	Team Name	Average DSC	Average HD95 (voxels)	Average VS
1	NVAUTO	0.786 ± 0.161	14.012 ± 9.285	0.885 ± 0.156
2	SJTU_EIEE_2–426Lab	0.775 ± 0.173	14.671 ± 9.917	0.883 ± 0.166
3	Pengyy	0.774 ± 0.182	14.699 ± 10.049	0.875 ± 0.182
4	Hilab*	0.774 ± 0.181	14.569 ± 9.954	0.873 ± 0.180
4	Neurophet*	0.775 ± 0.171	15.375 ± 9.277	0.877 ± 0.165
6	davoodkarimi	0.771 ± 0.171	16.755 ± 11.443	0.882 ± 0.156
7	2Ai*	0.767 ± 0.170	14.625 ± 9.892	0.867 ± 0.166
7	xlab*	0.771 ± 0.183	15.262 ± 14.769	0.873 ± 0.182
9	ichilove-axe	0.766 ± 0.176	21.329 ± 13.241	0.888 ± 0.158
10	TRABIT	0.769 ± 0.174	14.901 ± 9.049	0.866 ± 0.173
11	ichilove-combi*	0.762 ± 0.188	16.039 ± 9.395	0.873 ± 0.183
11	muw_dsobotka*	0.765 ± 0.171	17.159 ± 11.905	0.874 ± 0.168
11	Physense-UPF Team*	0.767 ± 0.182	15.018 ± 10.145	0.863 ± 0.180
14	SingleNets	0.748 ± 0.172	26.121 ± 12.072	0.876 ± 0.154
15	BIT_LILAB	0.752 ± 0.190	18.162 ± 12.644	0.868 ± 0.183
16	Moona Mazher	0.755 ± 0.183	18.548 ± 12.739	0.866 ± 0.179
17	MIAL	0.740 ± 0.211	25.107 ± 19.425	0.845 ± 0.213
18	ZJUWULAB	0.703 ± 0.217	27.948 ± 22.400	0.835 ± 0.218
19	FeVer	0.683 ± 0.180	34.419 ± 15.990	0.828 ± 0.164
20	Anonymous	0.621 ± 0.192	37.385 ± 18.249	0.801 ± 0.181
21	A3	0.534 ± 0.178	39.608 ± 18.249	0.791 ± 0.199

in Fig. 14. The order of the teams on the x-axis in each graph is ordered from best to worst, left to right. Team NVAUTO performed best for both pathological and non-pathological brains. No details of the specific pathologies were available to the challenge participants. Complete ranking information taking all three metrics into account for the pathological and non-pathological datasets can be found in Fig. 7.

3.4.5. *Intracranial volume*

The intracranial volume of each case in the test set was calculated using all labels (excluding the background) and compared to the intracranial volumes determined by each participant in the challenge. While most methods had some outliers, all teams except for five had a median percent difference from GT within ±1% (Fig. 15).

3.4.6. *Gestational age comparison in the GM*

The evaluation metrics for the cortex (gray matter label) were calculated based on age of the fetus. The top-scoring teams for the younger fetuses (GA 21–28; n=28) were pengyy, NVAUTO, and Hilab.

The top scoring teams for the older fetuses (GA 29–35; n=12) were xlab, pengyy, and Hilab. When all ages were combined together, the top teams for the GM label were pengyy, Hilab, and NVAUTO (see Table 4), showing that gestational age does play a small role in the success of the algorithms in segmenting the cortex. There are fewer cases included in the older group mainly due to the smaller gestational age range, and the fact that the majority of fetal scans at the center used to collect the data happen by the 32nd gestational week. The evaluation metrics of the GM from both the older and younger fetuses can be seen in Fig. 16 and 17.

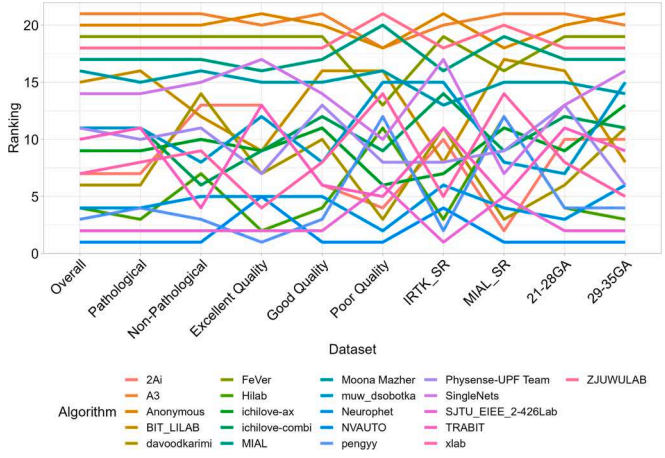

Fig. 7. Ranking of each algorithm for each subset of data in Section 3.3.

4. Discussion and conclusion

In this paper we present the results of the first FeTA Challenge held at the MICCAI 2021 conference. All submissions to the FeTA Challenge were deep-learning based submissions. Other machine-learning methods or purely atlas-based approaches were not submitted. This demonstrates that deep learning is currently the leading method for fetal brain medical image segmentation, and confirms its dominance in medical image segmentation more broadly. Indeed, the top three teams all had very similar network architectures. The majority of participating teams obtained very similar evaluation metrics; however, one team performed significantly better than all other teams on the complete testing dataset.

4.1. *Top methods*

When all labels are combined together and the entire testing dataset is used, Team NVAUTO submitted the top algorithm of the challenge. They ranked first in two out of the three evaluation metrics (DCS and HD95), and came second in the third (VS). In addition, the bootstrapping and significance testing showed that NVAUTO was the clear winner in the DSC and HD95 coefficients. The VS metric was more ambiguous across all participants, with no statistically significant difference among the first 9 teams. This suggests that while volumetry is a valuable biomarker when performing imaging studies, it is potentially not a sensitive evaluation metric for this challenge as it is unable to show differences in the performance. However, these results could also point to the fact that volumetry is indeed a stable measurement for fetal MRI studies, as the challenge results demonstrate that this measurement was stable across different segmentation methods.

There were many methodological similarities among the top five ranking teams. All were 3D U-Nets, all used either a Dice or cross-entropy/Dice combination loss function, none used an external dataset, and all used standard data augmentation techniques such as rotation, flipping, scaling, addition of Gaussian noise, Gaussian smoothing, gamma correction, affine transformations, and contrast adjustment. Four of the teams used Pytorch, while the fifth used MONAI, which is Pytorch-based (MONAI Consortium, 2020). All used the same convolution kernel size (3×3×3) with random initialization. Four out of five

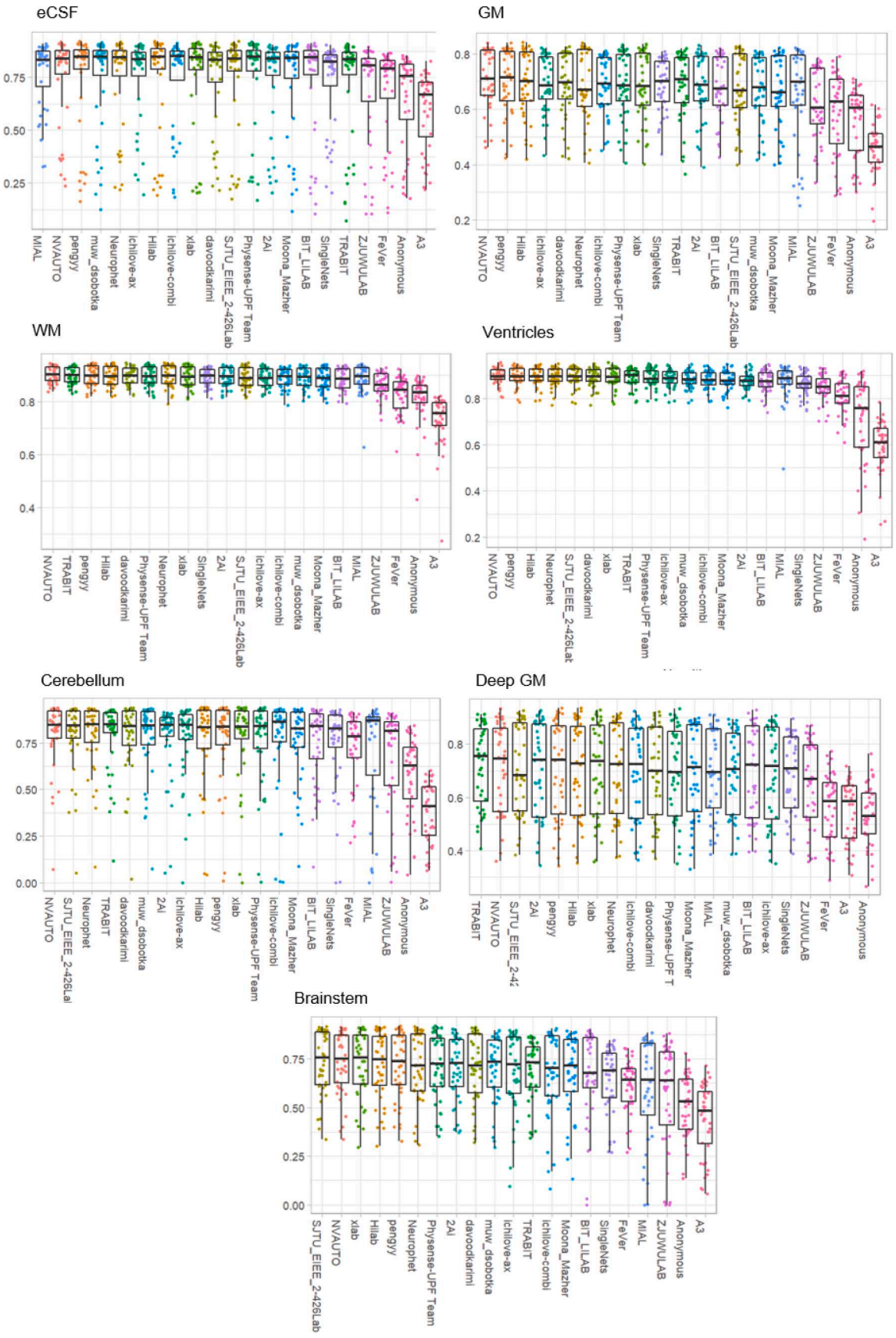

Fig. 8. DSC per label for each team (teams ranked from best to worst are visualized from left to right on the x-axis of each graph).

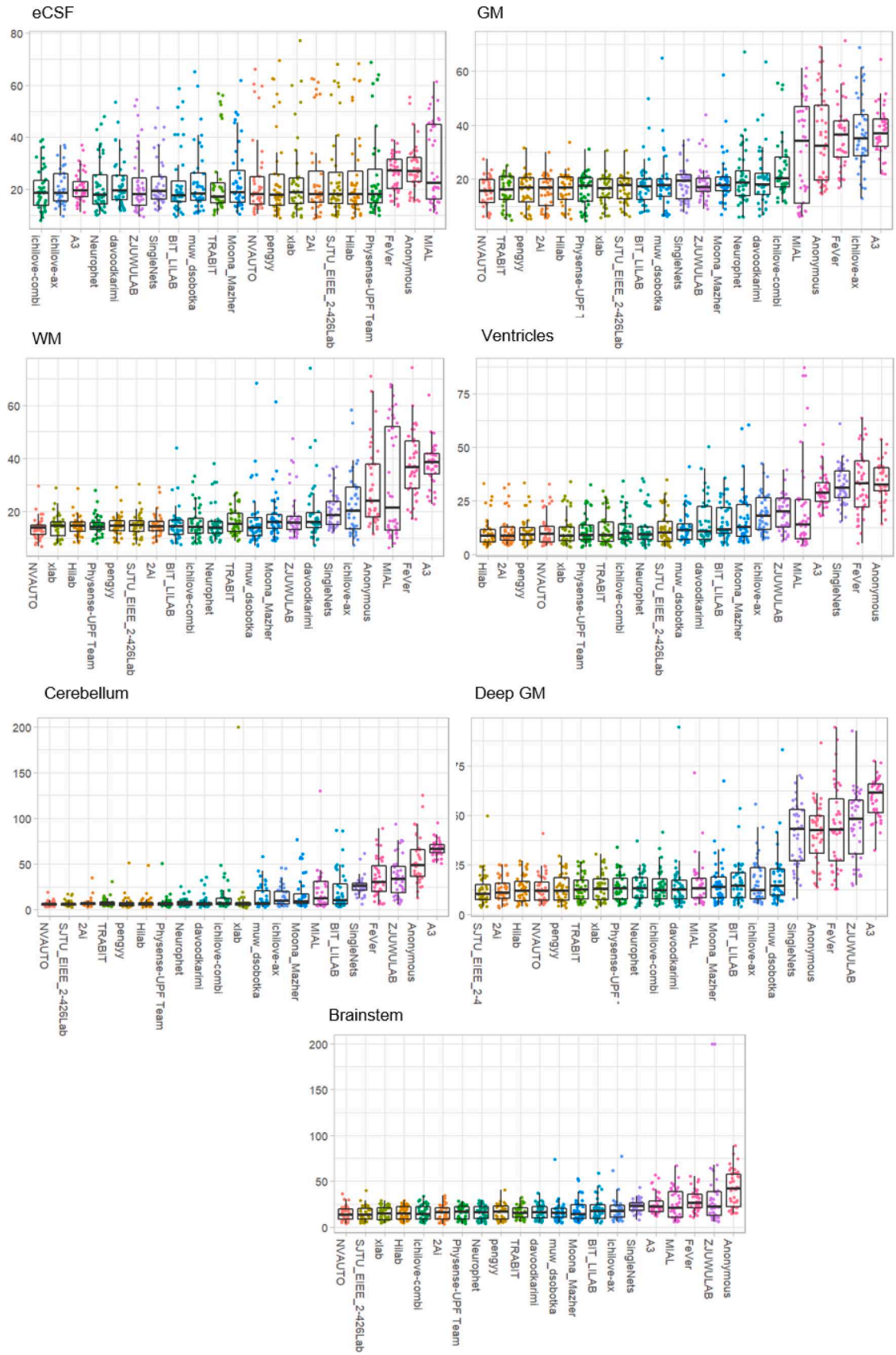

Fig. 9. HD95 per label for each team (teams ranked from best to worst are visualized from left to right on the x-axis of each graph).

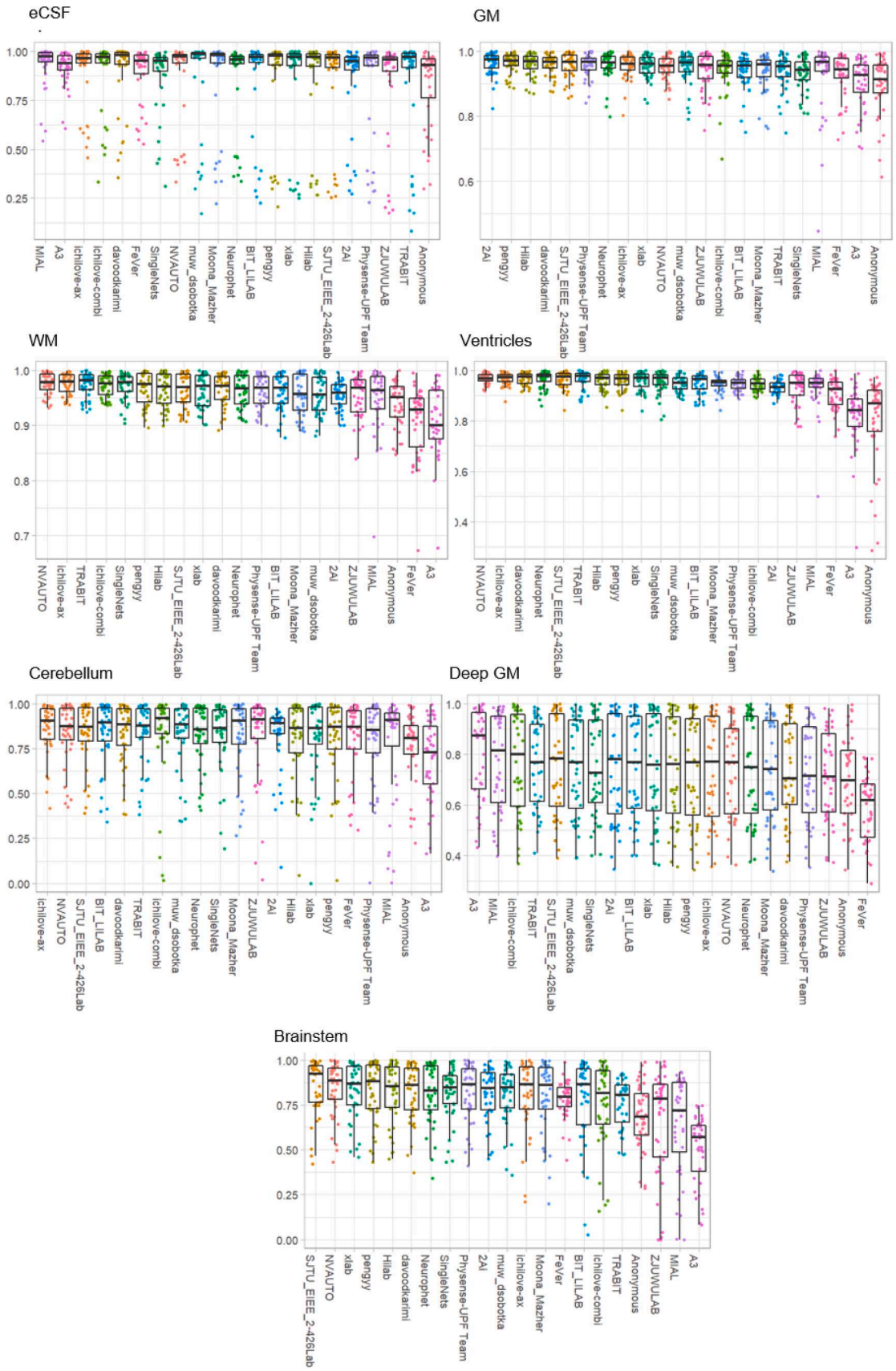

Fig. 10. VS per label for each team (teams ranked from best to worst are visualized from left to right on the x-axis of each graph).

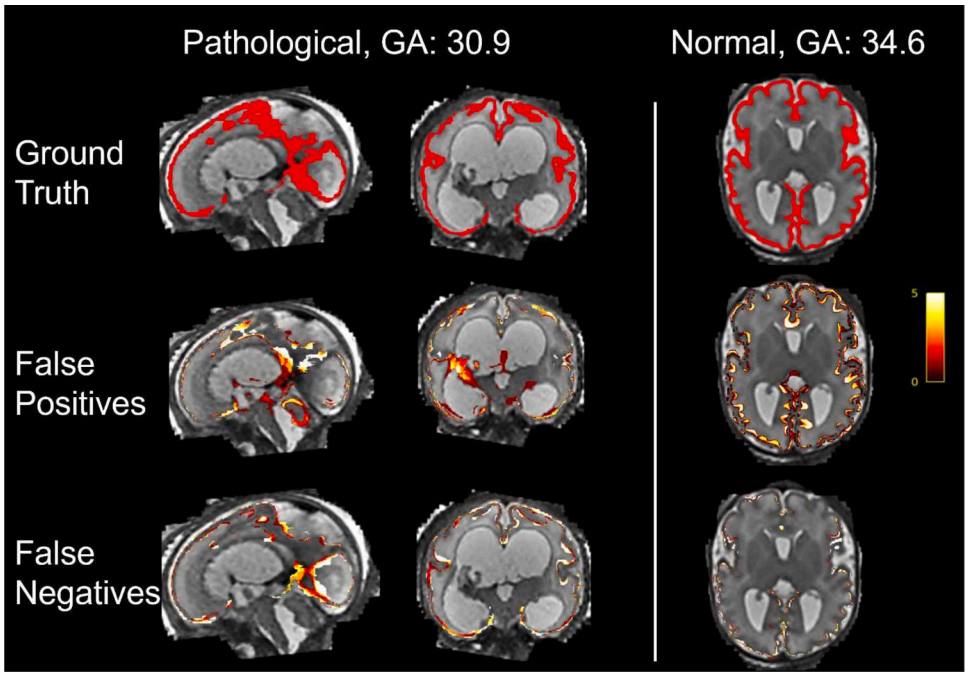

Fig. 11. Error maps for the top 5 teams of the gray matter for two test cases. Top Row: Ground truth gray matter label; Middle Row: Error map of false positives, meaning voxels that teams identified as cortex but weren't; Bottom Row: Error map of false negatives, meaning the voxels which should have been identified as cortex but weren't. Left two columns: Pathological brain (ventriculomegaly and other associated malformations); Right Column: Neurotypical brain. In the pathological brain, many of the top 5 algorithms misidentified some parts of the cerebellar cortex as cerebral cortex, and misidentified septum pellucidum as cortex as well. In the neurotypical brain, errors were mainly located at the interface between the white and gray matter.

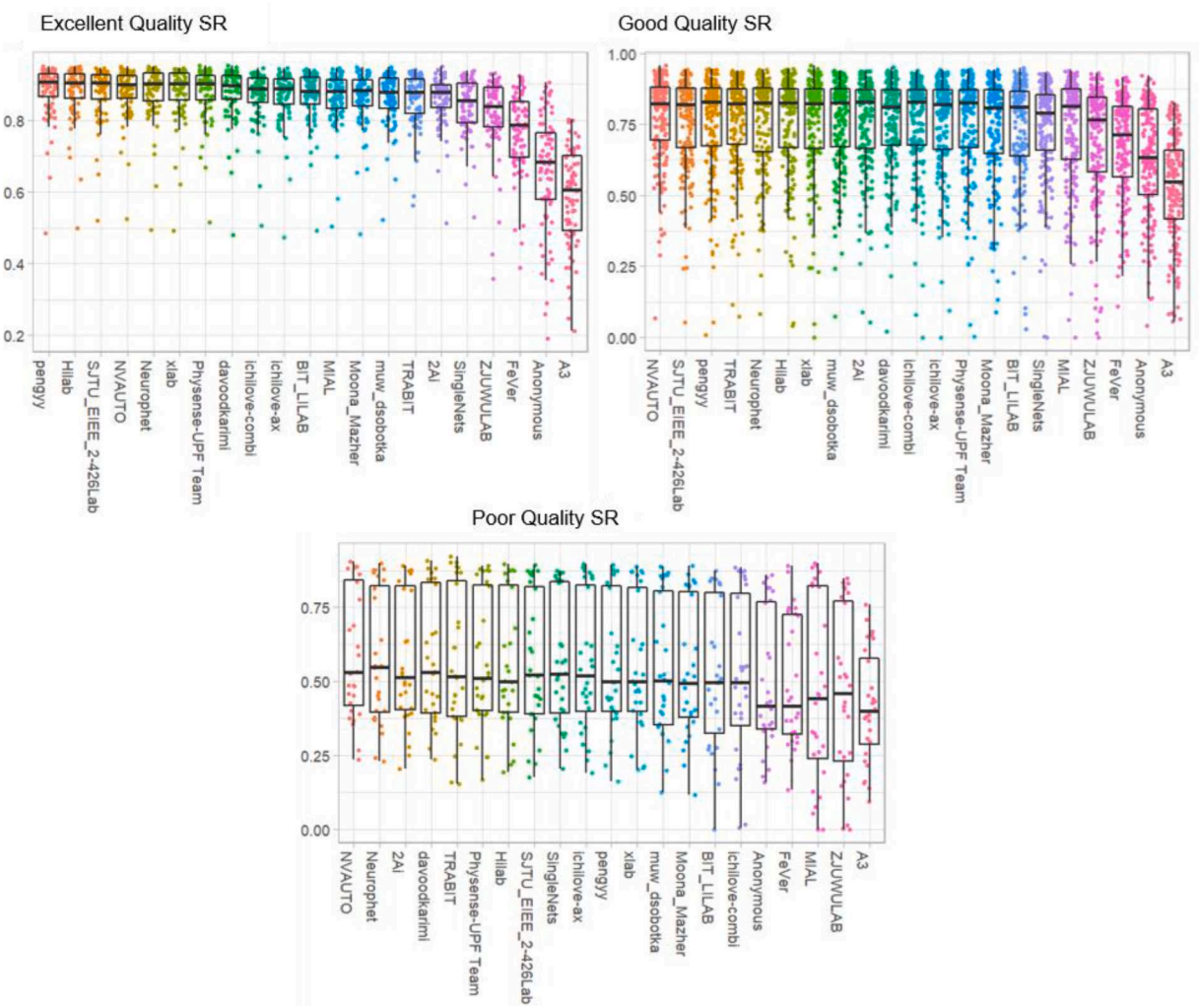

Fig. 12. DSC across all labels for each team, based on the quality of the SR reconstruction (teams ranked from best to worst are visualized from left to right on the x-axis of each graph).

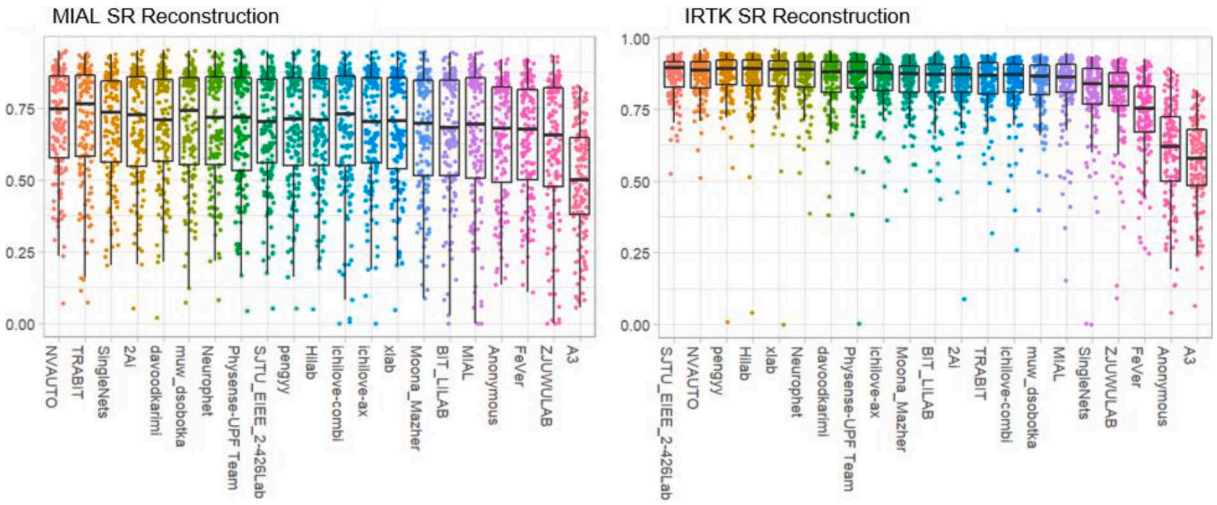

Fig. 13. DSC across all labels for each team, based on the SR reconstruction used (teams ranked from best to worst are visualized from left to right on the x-axis of each graph).

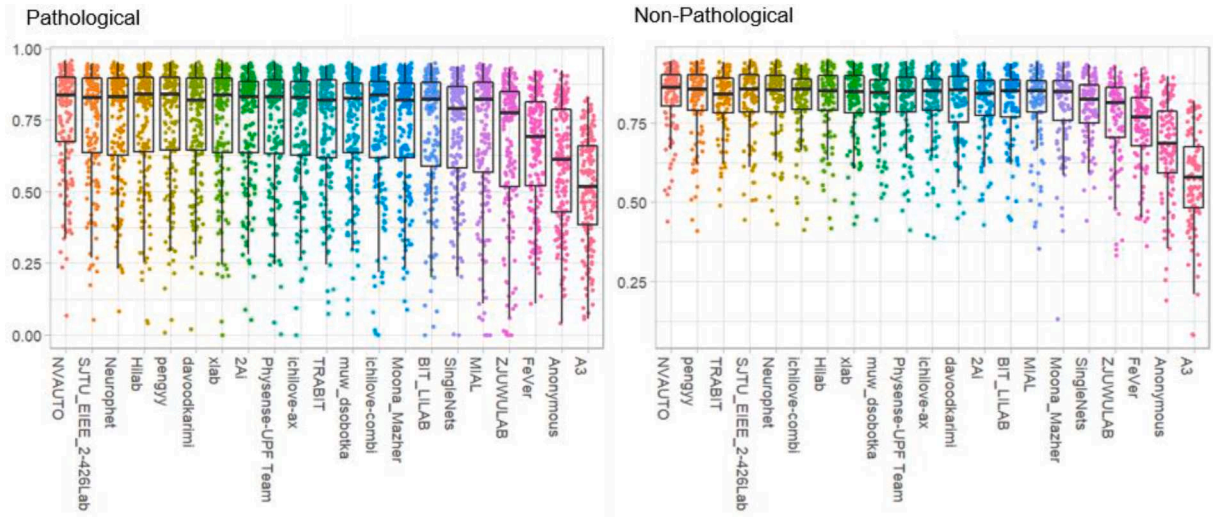

Fig. 14. DSC across all labels for each team, based on whether the fetal brain was pathological or non-pathological (teams ranked from best to worst are visualized from left to right on the x-axis of each graph).

used an ensemble learning strategy, three teams out of the top five used cross-validation. The main differences in networks appeared to be in the training procedures, such as the number of epochs, and in how the learning rate was manipulated throughout training.

When looking at the changes in rankings based on different subsets of the data the interpretation of the results become challenging. As shown in Fig. 7, the rankings change considerably depending on the data subset tested. This is relevant, as different centers have different data, different age ranges at which fetal MRI is acquired and one algorithm may not work well across all sites. As the submitted algorithms were all similar deep learning-based methods, this suggests that fine tuning the networks plays a key role in any potential practical application of these algorithms, depending on the specific clinical or research usage.

4.2. Performance of submitted algorithms

As mentioned already, all submissions were deep learning-based submissions. To go one step further, it was not just that all submissions used deep learning, but 19 out of 21 submissions used some form of U- Net, consisting of a contracting and an expanding path forming a U-shaped network. During the former, higher resolution information is sacrificed for more context. However, U-Net has the capability, using skip connections, of combining this information with the corresponding output from the expanding path. There were many differences within each U-Net, but the overall shape and structure of the network remained consistent, including the depth of the network. Eight teams used the pre-existing medical imaging neural network frameworks nnU-Net (Isensee et al., 2021) or MONAI (MONAI Consortium, 2020). The main differences across the submissions were in how the training was performed (such as the use of cross-validation or changes in the learning rate decay), or in the pre-processing (patch size, how the data was normalized) and post-processing (such as ensemble learning, removal of external label 'blobs'). The plateauing of the top team entries is interesting as well, potentially suggesting that U-Nets have a performance limit in multi-class segmentation tasks with limited data. Interesting to note is that none of the teams took advantage of the option to use the meta data provided as network input (gestational age in weeks, pathological/neurotypical brain classification). This could have potentially allowed some teams to differentiate themselves in the ranking and is still an area for future research.

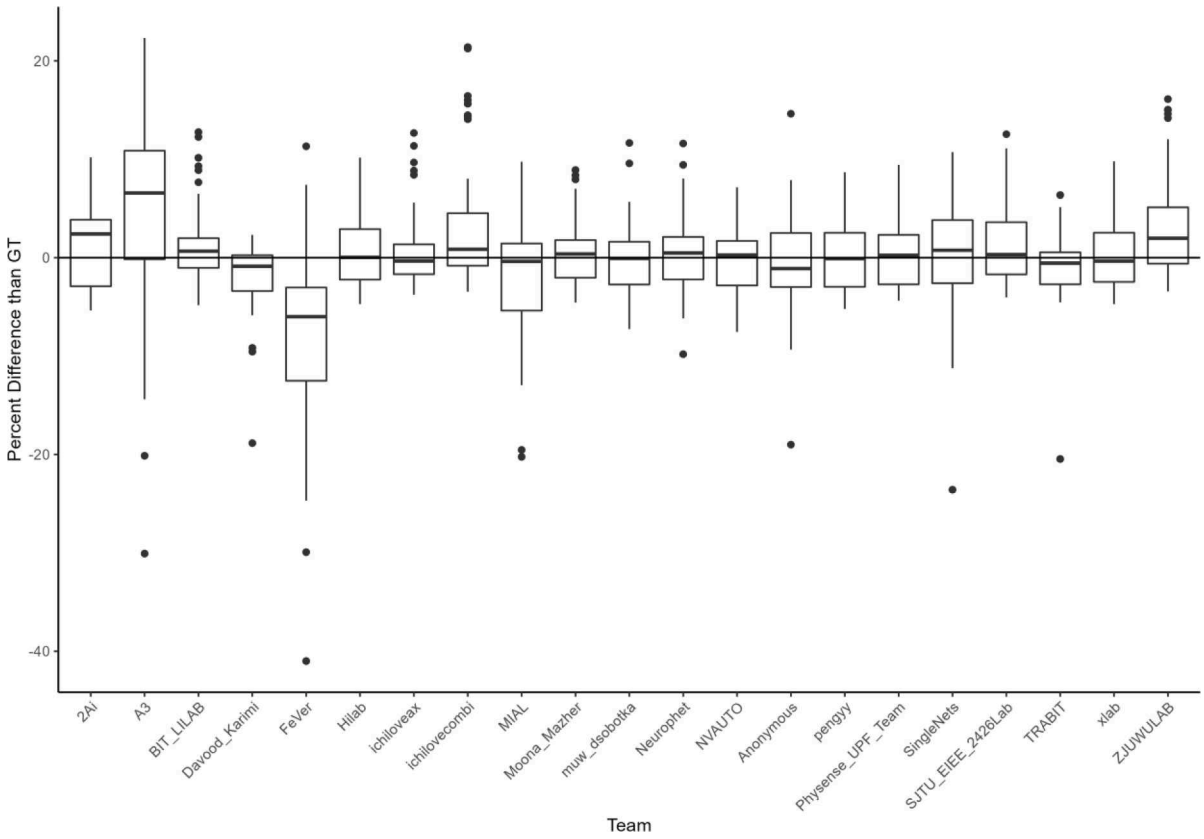

Fig. 15. Percent difference in intracranial volume between the submitted algorithms and the reference label map.

The most likely labels to fail to be segmented (that is, where the algorithm was unable to detect any voxels with the specific tissue) were the brainstem and the cerebellum, in particular in the pathological cases. This could potentially be explained by unclear demarcations of the brainstem and cerebellum in pathological groups which contained some cases of the Chiari-II malformation. Additionally, the overall segmentation accuracy for the cortical gray matter was rather moderate. We noticed that this might be due to the limited image resolution that leads to some degree of blurring of the smaller sulci, as well as due to annotation uncertainty. As a result, the cortical surface is often not topologically correct, contains holes or the thickness of the cortex is not homogeneous. Obtaining topologically correct cortical GM segmentation is an active field of research with some recent publications in the field that may overcome this limitation (de Dumast et al., 2020).The most likely labels to fail to be segmented (that is, where the algorithm was unable to detect any voxels with the specific tissue) were the brainstem and the cerebellum, in particular in the pathological cases. This could potentially be explained by unclear demarcations of the brainstem and cerebellum in pathological groups which contained some cases of the Chiari-II malformation. Overall, the most challenging labels to segment were cortical GM, deep GM, and the brainstem. This can be seen in Figs. 8-10, where these three tissue labels have worse performances than the other tissues, along with a larger distribution of evaluation metrics in each team. The potential reasons for this are multifold. The lateral and ventral borders of the deep GM and ventral portion of the brainstem are not well defined and are challenging for experienced radiologists to delineate. In the GM, the contrast between WM and GM changes throughout gestation due to neuronal migration and axonal outgrowth, while the surface pattern of the cortical GM becomes increasingly complex.

In general, the pathological brains were more challenging to segment than the non-pathological brains due to the larger variations in neuroanatomy. Selective data augmentation on these pathological cases could be a potential solution to this. The results of the image quality and SR reconstruction methods are related to each other, as the majority of the low quality images were done with the mialSR method, and the excellent quality brain volumes included were reconstructed with the Simple IRTK method. We would like to emphasize this is not a comment on the SR methods themselves, only a reflection of what cases were chosen for each reconstruction method. As expected, the low quality images, and therefore also the mialSR reconstructions were more challenging to accurately segment than the high quality and IRTK SR reconstructions, with lower DSC scores and a wider range of variability as can be seen in Fig. 13.

4.3. Clinical applications

Potential applications of the fully automatic and highly accurate fetal brain MRI segmentation algorithms are broad and span from neuroscience (characterizing spatio-temporal lateralization of the cortex (Kasprian et al., 2011; Vasung et al., 2020), virtopsies (identification and analysis of the details of demise (Rüegger et al., 2014)), surgery (clinical guidelines for early fetal surgery (Clewell et al., 1982; Meuli et al., 1997; Meuli and Moehrlen, 2013)), medicine (identification of biomarkers of outcome needed for stratification tools and development of early interventions (Rollins et al., 2021)), volumetric studies (Polat et al., 2017; Sadhwani et al., 2022) and development of new public health policies (prenatal programs focused on reduction of stress during pregnancy (van den Heuvel et al., 2021; Wu et al., 2020)).

Fetal MRI offers a unique possibility to study the human specific aspects of neurodevelopment. It remains the only non-invasive in vivo imaging modality to study connectivity, function, and structural anatomy of the fetal brain in a single session (Jakab et al., 2021; A. 2015). From the perspective of neuroscience, it is critically important to study the relationship between brain structure and function. However, this requires parcellation of the brain and cortex into regions or areas (e.g.

Table 4
Complete Rankings (DSC, HD95, VS combined) for each label; * indicates a tie, with the rank in brackets following the team name.

Rank	eCSF	GM	WM	Ventricle	Cerebel-lum	Deep GM	Brainstem
1	ichilove-axe	pengyy	NVAUTO	NVAUTO	NVAUTO	SJTU_EIEE_2–426Lab	SJTU_EIEE_2–426Lab
2	ichilove-combi	Hilab	Hilab* (2)	Hilab	SJTU_EIEE_2–426Lab	TRABIT	NVAUTO
3	Neurophet* (3)	NVAUTO	pengyy* (2)	pengyy	TRABIT	2Ai	xlab
4	Davoodkarimi* (3)	2Ai	TRABIT	Neurophet	davoodkarimi	NVAUTO* (4)	Hilab
5	NVAUTO* (5)	Physense-UPF Team	xlab	SJTU_EIEE_2–426Lab	Neurophet	Hilab* (4)	pengyy
6	muw_dsobotka* (5)	davoodkarimi	Physense-UPF Team	Davoodkarimi* (6)	ichilove-axe	ichilove-combi* (6)	Neurophet
7	MIAL	xlab	SJTU_EIEE_2–426Lab	TRABIT* (6)	2Ai	pengyy* (6)	Physense-UPF Team
8	A3	SJTU_EIEE_2–426Lab* (8)	ichilove-combi	xlab* (6)	muw_dsobotka	xlab	2Ai
9	pengyy	Neurophet* (8)	Neurophet	ichilove-axe	Hilab	MIAL	davoodkarimi
10	SingleNets	TRABIT	SingleNets* (10)	Physense-UPF Team	ichilove-combi	Neurophet	muw_dsobotka
11	Moona_Mazher* (11)	ichilove-axe	Davoodkarimi* (10)	2Ai	pengyy	muw_dsobotka	ichilove-combi
12	BIT_LILAB* (11)	ichilove-combi* (12)	ichilove-axe	muw_dsobotka	BIT_LILAB	Physense-UPF Team	ichilove-axe
13	xlab	muw_dsobotka* (12)	2Ai	ichilove-combi	Physense-UPF Team* (13)	Davoodkarimi* (13)	TRABIT
14	Hilab	BIT_LILAB* (12)	BIT_LILAB	Moona_Mazher* (14)	xlab* (13)	BIT_LILAB* (13)	SingleNets* (14)
15	SJTU_EIEE_2–426Lab* (15)	SingleNets	muw_dsobotka	BIT_LILAB* (14)	Moona_Mazher	SingleNets* (15)	Moona_Mazher* (14)
16	ZJUWULAB* (15)	ZJUWULAB	Moona_Mazher	SingleNets	SingleNets	Moona_Mazher* (15)	BIT_LILAB
17	FeVer	Moona_Mazher	ZJUWULAB	MIAL* (17)	ZJUWULAB	A3	FeVer
18	2Ai	MIAL	MIAL	ZJUWULAB* (17)	FeVer* (18)	ichilove-axe	MIAL
19	TRABIT	FeVer	Anonymous	FeVer	MIAL* (18)	ZJUWULAB	ZJUWULAB
20	Physense-UPF Team	Anonymous	FeVer	A3	Anonymous	Anonymous* (20)	Anonymous* (20)
21	Anonymous	A3	A3	Anonymous	A3	FeVer* (20)	A3* (20)

(Amunts et al., 2020; Desikan et al., 2006; Klein and Tourville, 2012)). A first crucial step toward this is to perform reliable segmentation of the developing cerebral cortex, which was the objective of this FeTA Challenge.

Furthermore, normative charts showing age-related changes in volume of different brain structures throughout the lifespan, similar to head circumference in the pediatric population, have just started to emerge (Bethlehem et al., 2021)https://paperpile.com/c/igTxNz/ikp0. Nonetheless, in addition to obvious challenges of fetal MRI acquisition, the harmonization of MRI acquisition protocols across sites and the development of robust and automatic algorithms for accurate and precise segmentation of fetal brain remain prerequisites for any future clinical application.

4.4. Limitations, lessons learned, and future considerations

Some limitations of the challenge include the fact that all images included were acquired from a single center, and therefore algorithms developed with this dataset are unlikely to be generalizable to other centers. In addition, while the total number of cases included is relatively large for the type of dataset, it is relatively small when compared to other datasets used for training neural networks (Bakas et al., 2019; Menze et al., 2015). The manual segmentations included in both the training and testing dataset were not perfect, and therefore there are mislabeled voxels. Annotations were made mainly in the axial plane, leading to some noisy labels and discontinuity in the annotations in the coronal and sagittal planes. The manual annotations were especially challenging in the mialSR reconstructions, as it was the low-resolution scans that underwent reorientation rather than the final reconstructed volume, resulting in a final reconstruction that was not exactly 'in plane' according to standard fetal atlases. This led to the phenomenon of participants' algorithms performing quite well visually but receiving mid-range evaluation metrics. One team even performed their own revisions on the manual segmentations using their own in house experts, and then used them in their training dataset (L. Fidon et al., 2021). While organizing the challenge we were aware of these errors, and therefore included three different metrics in order to reduce the reliance on any one metric. Future work includes improving the manual segmentations included in the FeTA Dataset. Further research into inter-rater variability in fetal brain segmentations is also required to understand what values of evaluation metrics are considered 'good enough'. Preliminary research has been conducted using a small sample set (Payette et al., 2021a), and the results showed that when the quality of the SR reconstruction was good, the inter-rater agreement was very high. With decreasing quality of the reconstructions, the inter-rater agreement also decreased, especially in the external CSF, brainstem, and deep GM tissues. A more extensive study with more samples and raters should be performed to truly understand the inter-rater agreement of fetal brain segmentations.

Differences in the manual annotation protocols to other, publicly available resources (Gholipour et al., 2017; Makropoulos et al., 2018) further limit the generalizability of our results. Compared to these resources, our ground truth annotations did not include the amygdalae and hippocampi as separate labels. We decided not to include these structures as separate labels since their visibility in younger fetuses and pathological cases was deemed low in our data, which would have likely led to a high inter-rater variability. Furthermore, the ventricle label in our dataset comprised the lateral, 3rd and 4th ventricles, while the dHCP newborn release defined these as part of the eCSF label. Unifying anatomical annotation rules remains a challenging task.

In the future, we aim to expand the FeTA Dataset to include data from multiple centers in order to increase the generalizability of algorithms trained using this dataset. We also hope to extend the number of different pathologies included, and to increase the number of cases at the outer range of the gestational ages, especially at older gestational ages. At the moment, the FeTA dataset does not provide demographic

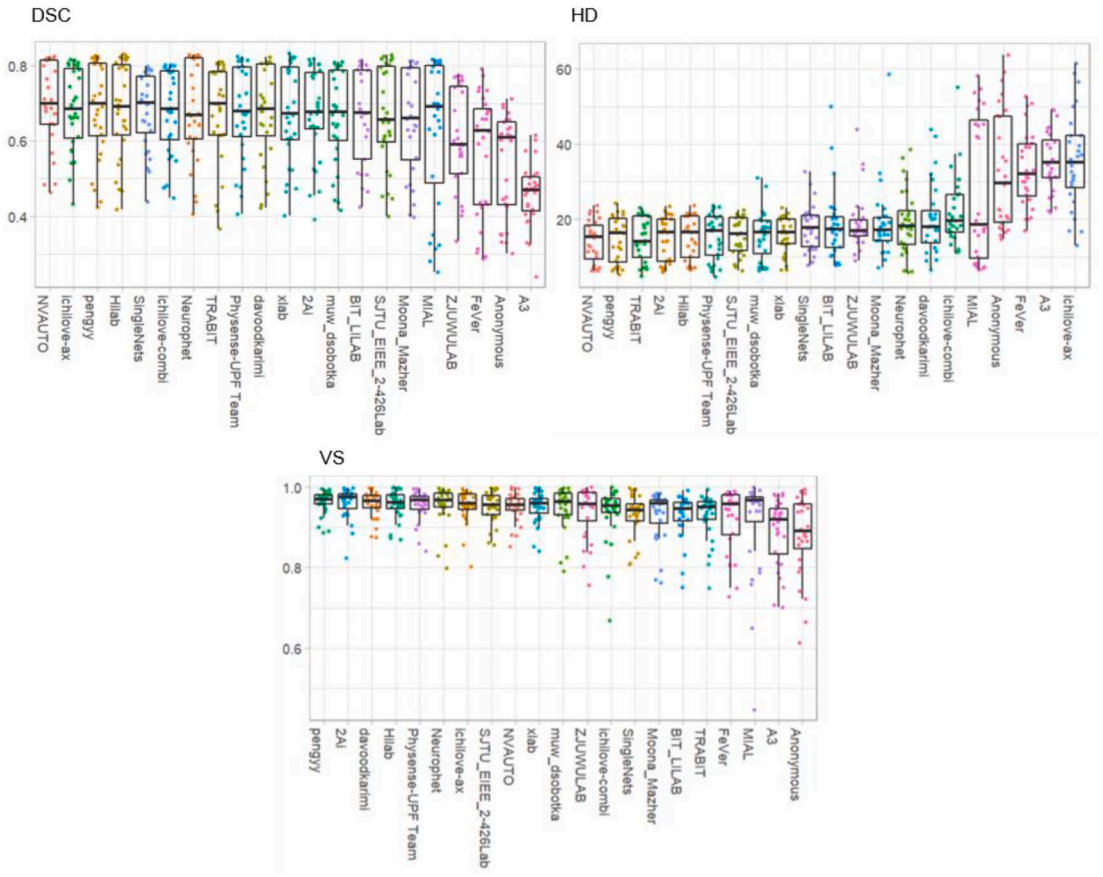

Fig. 16. Evaluation metrics (DSC, HD95, VS) of the GM label from younger GA fetuses (21-28GA).

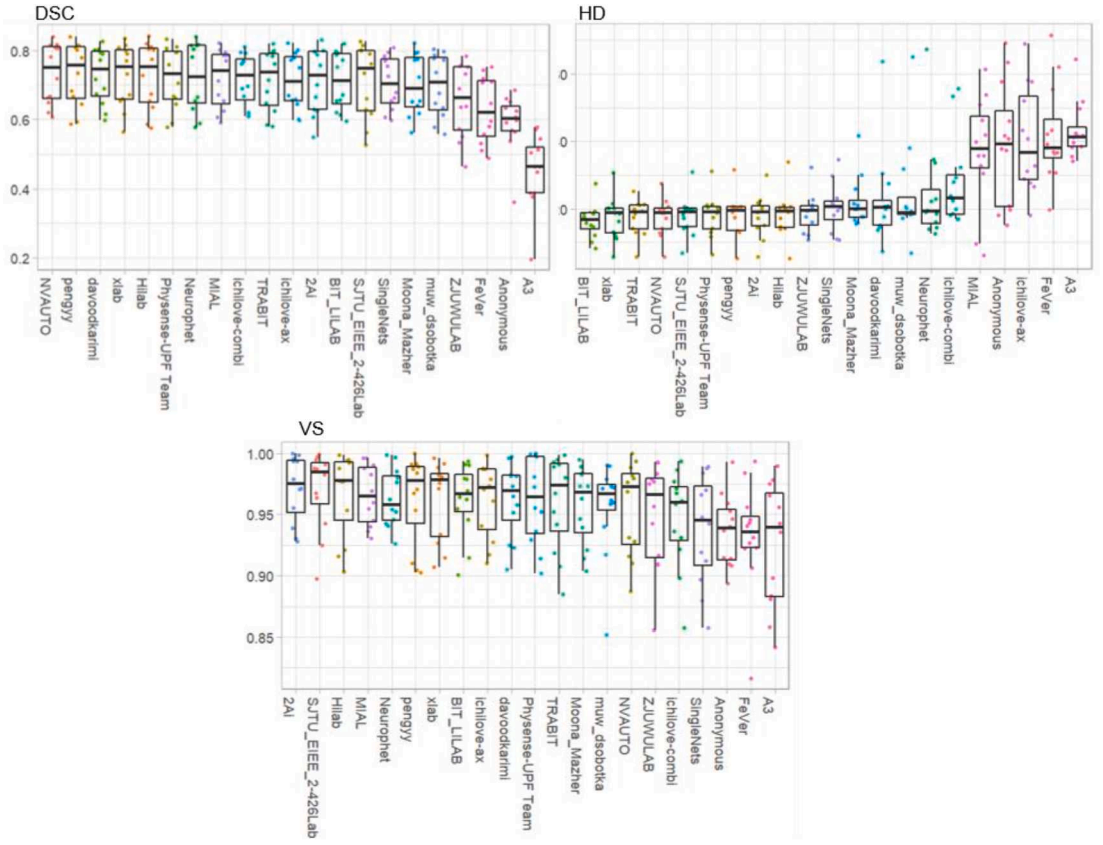

Fig. 17. Evaluation metrics (DSC, HD95, VS) of the GM label from older GA fetuses (29-35GA).

information on the participants. This is potentially something to include in the future, as recent research has demonstrated that ethnicity may be important in fetal brain imaging (Wu et al., 2021a). We also hope to streamline the manual segmentation procedure in order to allow a larger training dataset, and to re-evaluate the evaluation metrics used for future challenges. In addition, a FeTA 2022 Challenge is planned, focusing on generalization of automatic segmentation algorithms across data from different hospitals (Payette et al., 2022).

5. Conclusion

The algorithms developed as part of the FeTA Challenge provide a benchmark for future segmentation algorithms and can already be used to research fetal neurodevelopment. Our study found that most groups working on segmentation methods are using U-Nets, and that 3D U-Nets seem to be superior to 2D based on the evaluation metrics. In a dataset with large variation, such as the FeTA Dataset, the variation plays a role in the success of the algorithm. There was not one algorithm that was the 'best' when specific subsets of the data was analyzed, although there was a 'best' algorithm when the testing dataset was assessed as a whole. With the current networks, it appears as if a plateau of performance has been reached when the super-resolution reconstructions are of good quality. There is still room for improvement in low quality reconstructions, and in pathological cases. The use of trustworthy AI methods as demonstrated in Fidon et al. (2022) could also be used to improve performance, or pathology-specific atlases (Lucas Fidon et al., 2021). There are still many opportunities for improvement in developing multi-class segmentation techniques for the fetal brain throughout gestation, and therefore this challenge is the starting point for further development of such algorithms.

CRediT authorship contribution statement

Kelly Payette: Conceptualization, Writing – original draft, Data curation, Visualization, Software, Validation, Investigation. **Hongwei Bran Li:** Writing – review & editing, Software, Validation, Conceptualization. **Priscille de Dumast:** Writing – review & editing, Conceptualization. **Roxane Licandro:** Writing – review & editing, Conceptualization. **Hui Ji:** Data curation. **Md Mahfuzur Rahman Siddiquee:** Data curation, Writing – original draft, Writing – review & editing. **Daguang Xu:** Data curation, Writing – original draft, Writing – review & editing. **Andriy Myronenko:** Data curation, Writing – original draft, Writing – review & editing. **Hao Liu:** Data curation, Writing – original draft, Writing – review & editing. **Yuchen Pei:** Data curation, Writing – original draft, Writing – review & editing. **Lisheng Wang:** Data curation, Writing – original draft, Writing – review & editing. **Ying Peng:** Data curation, Writing – original draft, Writing – review & editing. **Juanying Xie:** Data curation, Writing – original draft, Writing – review & editing. **Huiquan Zhang:** Data curation, Writing – original draft, Writing – review & editing. **Guiming Dong:** Data curation, Writing – original draft, Writing – review & editing. **Hao Fu:** Data curation, Writing – original draft, Writing – review & editing. **Guotai Wang:** Data curation, Writing – original draft, Writing – review & editing. **ZunHyan Rieu:** Data curation, Writing – original draft, Writing – review & editing. **Donghyeon Kim:** Data curation, Writing – original draft, Writing – review & editing. **Hyun Gi Kim:** Data curation, Writing – original draft, Writing – review & editing. **Davood Karimi:** Data curation, Writing – original draft, Writing – review & editing. **Ali Gholipour:** Data curation, Writing – original draft, Writing – review & editing. **Helena R. Torres:** Data curation, Writing – original draft, Writing – review & editing. **Bruno Oliveira:** . **João L. Vilaça:** Data curation, Writing – original draft, Writing – review & editing. **Yang Lin:** Data curation, Writing – original draft, Writing – review & editing. **Netanell Avisdris:** Data curation, Writing – original draft, Writing – review & editing. **Ori Ben-Zvi:** . **Dafna Ben Bashat:** Data curation, Writing – original draft, Writing – review & editing. **Lucas Fidon:** Data curation, Writing – original draft, Writing – review & editing. **Michael Aertsen:** Data curation, Writing – original draft, Writing – review & editing. **Tom Vercauteren:** Data curation, Writing – original draft, Writing – review & editing. **Daniel Sobotka:** Data curation, Writing – original draft, Writing – review & editing. **Georg Langs:** Data curation, Writing – original draft, Writing – review & editing. **Mireia Alenyà:** Data curation, Writing – original draft, Writing – review & editing. **Maria Inmaculada Villanueva:** Data curation, Writing – original draft, Writing – review & editing. **Oscar Camara:** Data curation, Writing – original draft, Writing – review & editing. **Bella Specktor Fadida:** Data curation, Writing – original draft, Writing – review & editing. **Leo Joskowicz:** Data curation, Writing – original draft, Writing – review & editing. **Liao Weibin:** Data curation, Writing – original draft, Writing – review & editing. **Lv Yi:** Data curation, Writing – original draft, Writing – review & editing. **Li Xuesong:** Data curation, Writing – original draft, Writing – review & editing. **Moona Mazher:** Data curation, Writing – original draft, Writing – review & editing. **Abdul Qayyum:** Data curation, Writing – original draft, Writing – review & editing. **Domenec Puig:** Data curation, Writing – original draft, Writing – review & editing. **Hamza Kebiri:** Data curation, Writing – original draft, Writing – review & editing. **Zelin Zhang:** Data curation, Writing – original draft, Writing – review & editing. **Xinyi Xu:** Data curation, Writing – original draft, Writing – review & editing. **Dan Wu:** Data curation, Writing – original draft, Writing – review & editing. **Kuanlun Liao:** Data curation, Writing – original draft, Writing – review & editing. **Yixuan Wu:** Data curation, Writing – original draft, Writing – review & editing. **Jintai Chen:** Data curation, Writing – original draft, Writing – review & editing. **Yunzhi Xu:** Data curation, Writing – original draft, Writing – review & editing. **Li Zhao:** Data curation, Writing – original draft, Writing – review & editing. **Lana Vasung:** Writing – review & editing, Conceptualization. **Bjoern Menze:** Supervision. **Meritxell Bach Cuadra:** Conceptualization, Writing – review & editing, Supervision, Investigation. **Andras Jakab:** Conceptualization, Writing – review & editing, Data curation, Visualization, Supervision, Investigation.

Declaration of Competing Interest

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

Data availability

The challenge data is available on Synapse, as described in the paper.

Acknowledgments

The authors would like to acknowledge funding from the following funding sources: the OPO Foundation, the University Research Priority Project Adaptive Brain Circuits in Development and Learning (AdaBD) of the University of Zürich, the Prof. Dr. Max Cloetta Foundation, the Anna Müller Grocholski Foundation, the Foundation for Research in Science and the Humanities at the UZH, the EMDO Foundation, the Hasler Foundation, the FZK Grant, the Swiss National Science Foundation (project 205321-182602), the Forschungskredit (Grant NO. FK-21-125) from University of Zurich, the ZNZ PhD Grant, the EU H2020 Marie Sklodowska-Curie [765148], Austrian Science Fund FWF [P 35189], Vienna Science and Technology Fund WWTF [LS20-065], and the Austrian Research Fund Grant I3925-B27 in collaboration with the French National Research Agency (ANR). We acknowledge access to the expertise of the CIBM Center for Biomedical Imaging, a Swiss research center of excellence founded and supported by Lausanne University Hospital (CHUV), University of Lausanne (UNIL), Ecole polytechnique fédérale de Lausanne (EPFL), University of Geneva (UNIGE) and Geneva University Hospitals (HUG). We would also like to acknowledge funding

from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement TRABIT No 765148, as well as from core and project funding from the Wellcome [203148/Z/16/Z; 203145Z/16/Z; WT101957], and EP-SRC [NS/A000049/1; NS/A000050/1; NS/A000027/1]. TV is supported by a Medtronic / RAEng Research Chair [RCSRF1819\7\34]. HBL is supported by an Nvidia Academic GPU grant and Forschungskredit (grant No. K-74851-01-01) from the University of Zurich. The authors would also like to thank NVIDIA for providing access to computing resources.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi 10.1016/j.media.2023.102833.

References

Amunts, K., Mohlberg, H., Bludau, S., Zilles, K., 2020. Julich-Brain: a 3D probabilistic atlas of the human brain's cytoarchitecture. Science 369, 988–992. https://doi.org/10.1126/science.abb4588.

Bach Cuadra, M., Schaer, M., Andre, A., Guibaud, L., Eliez, S., Thiran, J.-P., 2009. Brain tissue segmentation of fetal MR images. In: Workshop on Image Analysis for Developing Brain, in 12th International Conference on Medical Image Computing and Computer Assisted Intervention. Presented at the Workshop on Image Analysis for Developing Brain, in 12th International Conference on Medical Image Computing and Computer Assisted Intervention. London, UK.

Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., Shinohara, R.T., Berger, C., Ha, S.M., Rozycki, M., Prastawa, M., Alberts, E., Lipkova, J., Freymann, J., Kirby, J., Bilello, M., Fathallah-Shaykh, H., Wiest, R., Kirschke, J., Wiestler, B., Colen, R., Kotrotsou, A., Lamontagne, P., Marcus, D., Milchenko, M., Nazeri, A., Weber, M.-A., Mahajan, A., Baid, U., Gerstner, E., Kwon, D., Acharya, G., Agarwal, M., Alam, M., Albiol, Alberto, Albiol, Antonio, Albiol, F.J., Alex, V., Allinson, N., Amorim, P.H.A., Amrutkar, A., Anand, G., Andermatt, S., Arbel, T., Arbelaez, P., Avery, A., Azmat, M., B., P., Bai, W., Banerjee, S., Barth, B., Batchelder, T., Batmanghelich, K., Battistella, E., Beers, A., Belyaev, M., Bendszus, M., Benson, E., Bernal, J., Bharath, H.N., Biros, G., Bisdas, S., Brown, J., Cabezas, M., Cao, S., Cardoso, J.M., Carver, E.N., Casamitjana, A., Castillo, L.S., Catà, M., Cattin, P., Cerigues, A., Chagas, V.S., Chandra, S., Chang, Y.-J., Chang, S., Chang, K., Chazalon, J., Chen, S., Chen, W., Chen, J.W., Chen, Z., Cheng, K., Choudhury, A.R., Chylla, R., Clérigues, A., Colleman, S., Colmeiro, R.G.R., Combalia, M., Costa, A., Cui, X., Dai, Z., Dai, L., Daza, L.A., Deutsch, E., Ding, C., Dong, C., Dong, S., Dudzik, W., Eaton-Rosen, Z., Egan, G., Escudero, G., Estienne, T., Everson, R., Fabrizio, J., Fan, Y., Fang, L., Feng, X., Ferrante, E., Fidon, L., Fischer, M., French, A.P., Fridman, N., Fu, H., Fuentes, D., Gao, Y., Gates, E., Gering, D., Gholami, A., Gierke, W., Glocker, B., Gong, M., González-Villá, S., Grosges, T., Guan, Y., Guo, S., Gupta, S., Han, W.-S., Han, I.S., Harmuth, K., He, H., Hernández-Sabaté, A., Herrmann, E., Himthani, N., Hsu, W., Hsu, C., Hu, Xiaojun, Hu, Xiaobin, Hu, Yan, Hu, Yifan, Hua, R., Huang, T.-Y., Huang, W., Van Huffel, S., Huo, Q., HV, V., Iftekharuddin, K.M., Isensee, F., Islam, M., Jackson, A.S., Jambawalikar, S.R., Jesson, A., Jian, W., Jin, P., Jose, V.J.M., Jungo, A., Kainz, B., Kamnitsas, K., Kao, P.-Y., Karnawat, A., Kellermeier, T., Kermi, A., Keutzer, K., Khadir, M.T., Khened, M., Kickingereder, P., Kim, G., King, N., Knapp, H., Knecht, U., Kohli, L., Kong, D., Kong, X., Koppers, S., Kori, A., Krishnamurthi, G., Krivov, E., Kumar, P., Kushibar, K., Lachinov, D., Lambrou, T., Lee, J., Lee, C., Lee, Y., Lee, M., Lefkovits, S., Lefkovits, L., Levitt, J., Li, T., Li, Hongwei, Li, W., Li, Hongyang, Li, Xiaochuan, Li, Y., Li, Heng, Li, Zhenye, Li, Xiaoyu, Li, Zeju, Li, XiaoGang, Li, W., Lin, Z.-S., Lin, F., Lio, P., Liu, C., Liu, B., Liu, X., Liu, M., Liu, J., Liu, L., Llado, X., Lopez, M.M., Lorenzo, P.R., Lu, Z., Luo, L., Luo, Z., Ma, J., Ma, K., Mackie, T., Madabushi, A., Mahmoudi, I., Maier-Hein, K.H., Maji, P., Mammen, C.P., Mang, A., Manjunath, B.S., Marcinkiewicz, M., McDonagh, S., McKenna, S., McKinley, R., Mehl, M., Mehta, S., Mehta, R., Meier, R., Meinel, C., Merhof, D., Meyer, C., Miller, R., Mitra, S., Moiyadi, A., Molina-Garcia, D., Monteiro, M.A.B., Mrukwa, G., Myronenko, A., Nalepa, J., Ngo, T., Nie, D., Ning, H., Niu, C., Nuechterlein, N.K., Oermann, E., Oliveira, A., Oliveira, D.D.C., Oliver, A., Osman, A.F.I., Ou, Y.-N., Ourselin, S., Paragios, N., Park, M.S., Paschke, B., Pauloski, J.G., Pawar, K., Pawlowski, N., Pei, L., Peng, S., Pereira, S.M., Perez-Beteta, J., Perez-Garcia, V.M., Pezold, S., Pham, B., Phophalia, A., Piella, G., Pillai, G.N., Piraud, M., Pisov, M., Popli, A., Pound, M.P., Pourreza, R., Prasanna, P., Prkovska, V., Pridmore, T.P., Puch, S., Puybareau, É., Qian, B., Qiao, X., Rajchl, M., Rane, S., Rebsamen, M., Ren, H., Ren, X., Revanuru, K., Rezaei, M., Rippel, O., Rivera, L.C., Robert, C., Rosen, B., Rueckert, D., Safwan, M., Salem, M., Salvi, J., Sanchez, I., Sánchez, I., Santos, H.M., Sartor, E., Schellingerhout, D., Scheufele, K., Scott, M.R., Scussel, A.A., Sedlar, S., Serrano-Rubio, J.P., Shah, N.J., Shah, N., Shaikh, M., Shankar, B.U., Shboul, Z., Shen, Haipeng, Shen, D., Shen, L., Shen, Haocheng, Shenoy, V., Shi, F., Shin, H.E., Shu, H., Sima, D., Sinclair, M., Smedby, O., Snyder, J.M., Soltaninejad, M., Song, G., Soni, M., Stawiaski, J., Subramanian, S., Sun, L., Sun, R., Sun, J., Sun, K., Sun, Y., Sun, G., Sun, S., Suter, Y.R., Szilagyi, L., Talbar, S., Tao, D., Tao, D., Teng, Z., Thakur, S., Thakur, M.H., Tharakan, S., Tiwari, P., Tochon, G., Tran, T., Tsai, Y.M., Tseng, K.-.L., Tuan, T.A., Turlapov, V., Tustison, N., Vakalopoulou, M., Valverde, S., Vanguri, R., Vasiliev, E., Ventura, J., Vera, L., Vercauteren, T., Verrastro, C.A., Vidyaratne, L., Vilaplana, V., Vivekanandan, A., Wang, G., Wang, Q., Wang, C.J., Wang, W., Wang, D., Wang, R., Wang, Y., Wang, C., Wang, G., Wen, N., Wen, X., Weninger, L., Wick, W., Wu, S., Wu, Q., Wu, Y., Xia, Y., Xu, Y., Xu, X., Xu, P., Yang, T.-.L., Yang, X., Yang, H.-.Y., Yang, J., Yang, H., Yang, G., Yao, H., Ye, X., Yin, C., Young-Moxon, B., Yu, J., Yue, X., Zhang, S., Zhang, A., Zhang, K., Zhang, Xuejie, Zhang, Lichi, Zhang, Xiaoyue, Zhang, Y., Zhang, Lei, Zhang, J., Zhang, Xiang, Zhang, T., Zhao, S., Zhao, Y., Zhao, X., Zhao, L., Zheng, Y., Zhong, L., Zhou, C., Zhou, X., Zhou, F., Zhu, H., Zhu, J., Zhuge, Y., Zong, W., Kalpathy-Cramer, J., Farahani, K., Davatzikos, C., van Leemput, K., Menze, B., 2019. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. arXiv:1811.02629 [cs, stat].

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 Clin. 18, 103–114. https://doi.org/10.1016/j.nicl.2018.01.006.

Bethlehem, R.a.I., Seidlitz, J., White, S.R., Vogel, J.W., Anderson, K.M., Adamson, C., Adler, S., Alexopoulos, G.S., Anagnostou, E., Areces-Gonzalez, A., Astle, D.E., Auyeung, B., Ayub, M., Ball, G., Baron-Cohen, S., Beare, R., Bedford, S.A., Benegal, V., Beyer, F., Bae, J.B., Blangero, J., Cábez, M.B., Boardman, J.P., Borzage, M., Bosch-Bayard, J.F., Bourke, N., Calhoun, V.D., Chakravarty, M.M., Chen, C., Chertavian, C., Chetelat, G., Chong, Y.S., Cole, J.H., Corvin, A., Courchesne, E., Crivello, F., Cropley, V.L., Crosbie, J., Crossley, N., Delarue, M., Desrivieres, S., Devenyi, G., Biase, M.A.D., Dolan, R., Donald, K.A., Donohoe, G., Dunlop, K., Edwards, A.D., Elison, J.T., Ellis, C.T., Elman, J.A., Eyler, L., Fair, D.A., Fletcher, P.C., Fonagy, P., Franz, C.E., Galan-Garcia, L., Gholipour, A., Giedd, J., Gilmore, J.H., Glahn, D.C., Goodyer, I., Grant, P.E., Groenewold, N.A., Gunning, F.M., Gur, R.E., Gur, R.C., Hammill, C.F., Hansson, O., Hedden, T., Heinz, A., Henson, R., Heuer, K., Hoare, J., Holla, B., Holmes, A.J., Holt, R., Huang, H., Im, K., Ipser, J., Jack, C.R., Jackowski, A.P., Jia, T., Johnson, K.A., Jones, P.B., Jones, D.T., Kahn, R., Karlsson, H., Karlsson, L., Kawashima, R., Kelley, E.A., Kern, S., Kim, K., Kitzbichler, M.G., Kremen, W.S., Lalonde, F., Landeau, B., Lee, S., Lerch, J., Lewis, J.D., Li, J., Liao, W., Linares, D.P., Liston, C., Lombardo, M.V., Lv, J., Lynch, C., Mallard, T.T., Marcelis, M., Markello, R.D., Mazoyer, B., McGuire, P., Meaney, M.J., Mechelli, A., Medic, N., Misic, B., Morgan, S.E., Mothersill, D., Nigg, J., Ong, M.Q.W., Ortinau, C., Ossenkoppele, R., Ouyang, M., Palaniyappan, L., Paly, L., Pan, P.M., Pantelis, C., Park, M.M., Paus, T., Pausova, Z., Binette, A.P., Pierce, K., Qian, X., Qiu, J., Qiu, A., Raznahan, A., Rittman, T., Rollins, C.K., Romero-Garcia, R., Ronan, L., Rosenberg, M.D., Rowitch, D.H., Salum, G.A., Satterthwaite, T.D., Schaare, H.L., Schachar, R.J., Schultz, A.P., Schumann, G., Schöll, M., Sharp, D., Shinohara, R.T., Skoog, I., Smyser, C.D., Sperling, R.A., Stein, D.J., Stolicyn, A., Suckling, J., Sullivan, G., Taki, Y., Thyreau, B., Toro, R., Tsvetanov, K.A., Turk-Browne, N.B., Tuulari, J.J., Tzourio, C., Vachon-Presseau, É., Valdes-Sosa, M.J., Valdes-Sosa, P.A., Valk, S.L., Amelsvoort, T. van, Vandekar, S.N., Vasung, L., Victoria, L.W., Villeneuve, S., Villringer, A., Vértes, P.E., Wagstyl, K., Wang, Y.S., Warfield, S.K., Warrier, V., Westman, E., Westwater, M.L., Whalley, H.C., Witte, A.V., Yang, N., Yeo, B.T.T., Yun, H.J., Zalesky, A., Zar, H.J., Zettergren, A., Zhou, J.H., Ziauddeen, H., Zugman, A., Zuo, X.N., Aibl, Initiative, A.D.N., Investigators, A.D.R.W.B., Asrb, Team, C., Cam-CAN, Ccnp, 3r-Brain, Cobre, Group, E.D.B.A. working, FinnBrain, Study, H.A.B., Imagen, K., Nspn, Oasis-3, Project, O., Pond, The PREVENT-AD Research Group, V., Alexander-Bloch, A.F., 2021. Brain charts for the human lifespan. 10.1101/2021.06.08.447489.

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

Clewell, W.H., Johnson, M.L., Meier, P.R., Newkirk, J.B., Zide, S.L., Hendee, R.W., Bowes, W.A., Hecht, F., O'Keeffe, D., Henry, G.P., Shikes, R.H., 1982. A Surgical Approach to the Treatment of Fetal Hydrocephalus. N. Engl. J. Med. 306, 1320–1325. https://doi.org/10.1056/NEJM198206033062202.

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. Cereb. Cortex 23, 2932–2943. https://doi.org/10.1093/cercor/bhs281.

De Asis-Cruz, J., Andescavage, N., Limperopoulos, C., 2021. Adverse prenatal exposures and fetal brain development: insights from advanced fetal magnetic resonance imaging. Biol. Psychiatry: Cognit. Neurosci. Neuroimaging. https://doi.org/10.1016/j.bpsc.2021.11.009.

de Dumast, P., Kebiri, H., Atat, C., Dunet, V., Koob, M., Cuadra, M.B., 2020. Segmentation of the cortical plate in fetal brain MRI with a topological loss. arXiv:2010.12391 [cs, eess].

Desikan, R.S., Ségonne, F., Fischl, B., Quinn, B.T., Dickerson, B.C., Blacker, D., Buckner, R.L., Dale, A.M., Maguire, R.P., Hyman, B.T., Albert, M.S., Killiany, R.J., 2006. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. Neuroimage 31, 968–980. https://doi.org/10.1016/j.neuroimage.2006.01.021.

Dice, L.R., 1945. Measures of the amount of ecologic association between species. Ecology 26, 297–302. https://doi.org/10.2307/1932409.

Dittrich, E., Kasprian, G.J., Prayer, D., Langs, G., 2011. Atlas learning in fetal brain development. Top. Magn. Reson. Imaging. 22, 107–111.

Egaña-Ugrinovic, G., Sanz-Cortes, M., Figueras, F., Bargalló, N., Gratacós, E., 2013. Differences in cortical development assessed by fetal MRI in late-onset intrauterine growth restriction. Am. J. Obstet. Gynecol. 209 https://doi.org/10.1016/j.ajog.2013.04.008, 126.e1-126.e8.

Fetit, A.E., Alansary, A., Cordero-Grande, L., Cupitt, J., Davidson, A.B., Edwards, A.D., Hajnal, J.V., Hughes, E., Kamnitsas, K., Kyriakopoulou, V., Makropoulos, A., Patkee, P.A., Price, A.N., Rutherford, M.A., Rueckert, D., 2020. A deep learning approach to segmentation of the developing cortex in fetal brain MRI with minimal

manual labeling. In: Arbel, T., Ben Ayed, I., de Bruijne, M., Descoteaux, M., Lombaert, H., Pal, C. (Eds.), Proceedings of the Third Conference on Medical Imaging with Deep Learning, Proceedings of Machine Learning Research. PMLR, pp. 241–261.

Fidon, Lucas, Aertsen, M., Emam, D., Mufti, N., Guffens, F., Deprest, T., Demaerel, P., David, A.L., Melbourne, A., Ourselin, S., Deprest, J., Vercauteren, T., de Bruijne, M., Cattin, P.C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y., 2021a. Label-set loss functions for partial supervision: application to fetal brain 3D MRI parcellation. In: Essert, C. (Ed.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. Springer International Publishing, Cham, pp. 647–657. https://doi.org/10.1007/978-3-030-87196-3_60.

Fidon, L., Aertsen, M., Kofler, F., Bink, A., David, A.L., Deprest, T., Emam, D., Guffens, F., Jakab, A., Kasprian, G., Kienast, P., Melbourne, A., Menze, B., Mufti, N., Pogledic, I., Prayer, D., Stuempflen, M., Van Elslander, E., Ourselin, S., Deprest, J., Vercauteren, T., 2022. A Dempster-Shafer approach to trustworthy AI with application to fetal brain MRI segmentation. 10.48550/arXiv.2204.02779.

Fidon, L., Viola, E., Mufti, N., David, A.L., Melbourne, A., Demaerel, P., Ourselin, S., Vercauteren, T., Deprest, J., Aertsen, M., 2021b. A spatio-temporal atlas of the developing fetal brain with spina bifida aperta. Open Res. Europe 1.

Gholipour, A., Estroff, J.A., Barnewolt, C.E., Robertson, R.L., Grant, P.E., Gagoski, B., Warfield, S.K., Afacan, O., Connolly, S.A., Neil, J.J., Wolfberg, A., Mulkern, R.V., 2014. Fetal MRI: a technical update with educational aspirations. Concepts Magn. Reson. Part A Bridg. Educ. Res. 43, 237–266. https://doi.org/10.1002/cmr.a.21321.

Gholipour, A., Rollins, C., Velasco-Annis, C., Ouaalam, A., Akhondi-Asl, A., Afacan, O., Ortinau, C., Clancy, S., Limperopoulos, C., Yang, E., Estroff, J., Warfield, S., 2017. A normative spatiotemporal MRI atlas of the fetal brain for automatic segmentation and analysis of early brain growth. Sci. Rep. 7 https://doi.org/10.1038/s41598-017-00525-w.

Glenn, O.A., 2010. MR imaging of the fetal brain. Pediatr. Radiol. 40, 68–81. https://doi.org/10.1007/s00247-009-1459-3.

Gowland, P., 2011. Safety of fetal MRI scanning. In: Prayer, D. (Ed.), Fetal MRI. Springer, Berlin, Heidelberg, pp. 49–54. https://doi.org/10.1007/174_2010_122.

Griffiths, P.D., Bradburn, M., Campbell, M.J., Cooper, C.L., Embleton, N., Graham, R., Hart, A.R., Jarvis, D., Kilby, M.D., Lie, M., Mason, G., Mandefield, L., Mooney, C., Pennington, R., Robson, S.C., Wailoo, A., 2019. MRI in the diagnosis of fetal developmental brain abnormalities: the MERIDIAN diagnostic accuracy study. Health Technol. Assess. 23, 1–144. https://doi.org/10.3310/hta23490.

Gwet, K.L., 2019. irrCAC: computing chance-corrected agreement coefficients (CAC), R Package version 1.0.

Hart, A.R., Embleton, N.D., Bradburn, M., Connolly, D.J.A., Mandefield, L., Mooney, C., Griffiths, P.D., 2020. Accuracy of in-utero MRI to detect fetal brain abnormalities and prognosticate developmental outcome: postnatal follow-up of the MERIDIAN cohort. Lancet Child Adolesc. Health 4, 131–140. https://doi.org/10.1016/S2352-4642(19)30349-9.

Hausdorff, F., 1991. Set Theory. American Mathematical Society, RI.

Hong, J., Yun, H.J., Park, G., Kim, S., Laurentys, C.T., Siqueira, L.C., Tarui, T., Rollins, C. K., Ortinau, C.M., Grant, P.E., Lee, J.-M., Im, K., 2020. Fetal cortical plate segmentation using fully convolutional networks with multiple plane aggregation. Front. Neurosci. 14, 1226. https://doi.org/10.3389/fnins.2020.591683.

Hosny, I.A., Elghawabi, H.S., 2010. Ultrafast MRI of the fetus: an increasingly important tool in prenatal diagnosis of congenital anomalies. Magn. Reson. Imaging 28, 1431–1439. https://doi.org/10.1016/j.mri.2010.06.024.

Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H., 2021. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods 18, 203–211. https://doi.org/10.1038/s41592-020-01008-z.

Jakab, A., Payette, K., Mazzone, L., Schauer, S., Muller, C.O., Kottke, R., Ochsenbein-Kölble, N., Tuura, R., Moehrlen, U., Meuli, M., 2021. Emerging magnetic resonance imaging techniques in open spina bifida in utero. Eur. Radiol. Exp. 5, 23. https://doi.org/10.1186/s41747-021-00219-z.

Jakab, A., Pogledic, I., Schwartz, E., Gruber, G., Mitter, C., Brugger, P.C., Langs, G., Schöpf, V., Kasprian, G., Prayer, D., 2015. Fetal cerebral magnetic resonance imaging beyond morphology. Semin. Ultrasound CT MR 36, 465–475. https://doi.org/10.1053/j.sult.2015.06.003.

Jarvis, D.A., Finney, C.R., Griffiths, P.D., 2019. Normative volume measurements of the fetal intra-cranial compartments using 3D volume in utero MR imaging. Eur. Radiol. 29, 3488–3495. https://doi.org/10.1007/s00330-018-5938-5.

Karimi, D., Rollins, C.K., Velasco-Annis, C., Ouaalam, A., Gholipour, A., 2023. Learning to segment fetal brain tissue from noisy annotations. Med Image Anal. 85, 102731 https://doi.org/10.1016/j.media.2022.102731.

Kasprian, G., Langs, G., Brugger, P.C., Bittner, M., Weber, M., Arantes, M., Prayer, D., 2011. The prenatal origin of hemispheric asymmetry: an in utero neuroimaging study. Cereb. Cortex 21, 1076–1083. https://doi.org/10.1093/cercor/bhq179.

Khalili, N., Lessmann, N., Turk, E., Claessens, N., Heus, R.de, Kolk, T., Viergever, M.A., Benders, M.J.N.L., Išgum, I., 2019. Automatic brain tissue segmentation in fetal MRI using convolutional neural networks. Magn. Reson. Imaging. https://doi.org/10.1016/j.mri.2019.05.020.

Kikinis, R., Pieper, S.D., Vosburgh, K.G., 2014. 3D slicer: a platform for subject-specific image analysis, visualization, and clinical support. Intraoperative Imaging and Image-Guided Therapy. Springer, New York, NY, pp. 277–289. https://doi.org/10.1007/978-1-4614-7657-3_19.

Klein, A., Tourville, J., 2012. 101 labeled brain images and a consistent human cortical labeling protocol. Front. Neurosci. 6, 171. https://doi.org/10.3389/fnins.2012.00171.

Kuklisova-Murgasova, M., Quaghebeur, G., Rutherford, M.A., Hajnal, J.V., Schnabel, J. A., 2012. Reconstruction of fetal brain MRI with intensity matching and complete outlier removal. Med. Image Anal. 16, 1550–1564. https://doi.org/10.1016/j.media.2012.07.004.

Kyriakopoulou, V., Vatansever, D., Davidson, A., Patkee, P., Elkommos, S., Chew, A., Martinez-Biarge, M., Hagberg, B., Damodaram, M., Allsop, J., Fox, M., Hajnal, J.V., Rutherford, M.A., 2017. Normative biometry of the fetal brain using magnetic resonance imaging. Brain Struct. Funct. 222, 2295–2307. https://doi.org/10.1007/s00429-016-1342-6.

Licandro, R., Langs, G., Kasprian, G.J., Sablatnig, R., Prayer, D., Schwartz, E., 2016. A Longitudinal Diffeomorphic Atlas-Based Tissue Labeling Framework for Fetal Brains Using Geodesic Regression. Presented at the 21st Computer Vision Winter Workshop. Rimske Toplice, Slovenia.

Maier-Hein, L., Reinke, A., Kozubek, M., Martel, A.L., Arbel, T., Eisenmann, M., Hanbury, A., Jannin, P., Müller, H., Onogur, S., Saez-Rodriguez, J., van Ginneken, B., Kopp-Schneider, A., Landman, B.A., 2020. BIAS: transparent reporting of biomedical image analysis challenges. Med. Image Anal. 66, 101796 https://doi.org/10.1016/j.media.2020.101796.

Makropoulos, A., Robinson, E.C., Schuh, A., Wright, R., Fitzgibbon, S., Bozek, J., Counsell, S.J., Steinweg, J., Vecchiato, K., Passerat-Palmbach, J., Lenz, G., Mortari, F., Tenev, T., Duff, E.P., Bastiani, M., Cordero-Grande, L., Hughes, E., Tusor, N., Tournier, J.-.D., Hutter, J., Price, A.N., Teixeira, R.P.A.G., Murgasova, M., Victor, S., Kelly, C., Rutherford, M.A., Smith, S.M., Edwards, A.D., Hajnal, J.V., Jenkinson, M., Rueckert, D., 2018. The developing human connectome project: a minimal processing pipeline for neonatal cortical surface reconstruction. Neuroimage 173, 88–112. https://doi.org/10.1016/j.neuroimage.2018.01.054.

Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., Lanczi, L., Gerstner, E., Weber, M.-.A., Arbel, T., Avants, B.B., Ayache, N., Buendia, P., Collins, D.L., Cordier, N., Corso, J.J., Criminisi, A., Das, T., Delingette, H., Demiralp, Ç., Durst, C.R., Dojat, M., Doyle, S., Festa, J., Forbes, F., Geremia, E., Glocker, B., Golland, P., Guo, X., Hamamci, A., Iftekharuddin, K.M., Jena, R., John, N.M., Konukoglu, E., Lashkari, D., Mariz, J.A., Meier, R., Pereira, S., Precup, D., Price, S.J., Raviv, T.R., Reza, S.M.S., Ryan, M., Sarikaya, D., Schwartz, L., Shin, H.-.C., Shotton, J., Silva, C.A., Sousa, N., Subbanna, N.K., Szekely, G., Taylor, T.J., Thomas, O.M., Tustison, N.J., Unal, G., Vasseur, F., Wintermark, M., Ye, D.H., Zhao, L., Zhao, B., Zikic, D., Prastawa, M., Reyes, M., Van Leemput, K., 2015. The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans. Med. Imaging 34, 1993–2024. https://doi.org/10.1109/TMI.2014.2377694.

Meuli, M., Meuli-Simmen, C., Hutchins, G.M., Seller, M.J., Harrison, M.R., Adzick, N.S., 1997. The spinal cord lesion in human fetuses with myelomeningocele: implications for fetal surgery. J. Pediatr. Surg., Pap. Presented 43rd Annu. Int. Congress 32, 448–452. https://doi.org/10.1016/S0022-3468(97)90603-5.

Meuli, M., Moehrlen, U., 2013. Fetal surgery for myelomeningocele: a critical appraisal. Eur. J. Pediatr. Surg. 23, 103–109. https://doi.org/10.1055/s-0033-1343082.

MONAI Consortium, 2020. MONAI: medical Open Network for AI. 10.5281/zenodo.4323058.

Nagaraj, U.D., Venkatesan, C., Bierbrauer, K.S., Kline-Fath, B.M., 2022. Value of pre- and postnatal magnetic resonance imaging in the evaluation of congenital central nervous system anomalies. Pediatr. Radiol. 52, 802–816. https://doi.org/10.1007/s00247-021-05137-1.

Payette, K., de Dumast, P., Kebiri, H., Ezhov, I., Paetzold, J.C., Shit, S., Iqbal, A., Khan, R., Kottke, R., Grehten, P., Ji, H., Lanczi, L., Nagy, M., Beresova, M., Nguyen, T.D., Natalucci, G., Karayannis, T., Menze, B., Bach Cuadra, M., Jakab, A., 2021a. An automatic multi-tissue human fetal brain segmentation benchmark using the fetal tissue annotation dataset. Sci. Data 8, 167. https://doi.org/10.1038/s41597-021-00946-3.

Payette, K., Dumast, P.de, Jakab, A., Cuadra, M.B., Vasung, L., Licandro, R., Menze, B., Zurich, H.L., 2021b. Fetal brain tissue annotation and segmentation challenge. 10.5281/zenodo.4573144.

Payette, K., Jakab, A., 2021. Fetal tissue annotation challenge - FeTA MICCAI 2021 [WWW Document]. URL 10.7303/syn25649159 (accessed 2.23.22).

Payette, K., Steger, C., Dumast, P.de, Jakab, A., Cuadra, M.B., Vasung, L., Licandro, R., Barkovich, M., Li, H., 2022. Fetal tissue annotation challenge. 10.5281/zenodo.6362587.

Pierre Deman, Sebastien Tourbier, Reto Meuli, Meritxell Bach Cuadra, 2020. Meribach/mevislabFetalMRI: MEVISLAB MIAL super-resolution reconstruction of fetal brain MRI v1.0. 10.5281/zenodo.3878564.

Polat, A., Barlow, S., Ber, R., Achiron, R., Katorza, E., 2017. Volumetric MRI study of the intrauterine growth restriction fetal brain. Eur. Radiol. 27, 2110–2118. https://doi.org/10.1007/s00330-016-4502-4.

Prayer, D., Kasprian, G., Krampl, E., Ulm, B., Witzani, L., Prayer, L., Brugger, P.C., 2006. MRI of normal fetal brain development. Eur. J. Radiol., Fetal Imaging 57, 199–216. https://doi.org/10.1016/j.ejrad.2005.11.020.

Core Team, R., 2020. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Rollins, C.K., Ortinau, C.M., Stopp, C., Friedman, K.G., Tworetzky, W., Gagoski, B., Velasco-Annis, C., Afacan, O., Vasung, L., Beaute, J.I., Rofeberg, V., Estroff, J.A., Grant, P.E., Soul, J.S., Yang, E., Wypij, D., Gholipour, A., Warfield, S.K., Newburger, J.W., 2021. Regional brain growth trajectories in fetuses with congenital heart disease. Ann. Neurol. 89, 143–157. https://doi.org/10.1002/ana.25940.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: convolutional Networks for Biomedical Image Segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 234–241. https://doi.org/10.1007/978-3-319-24574-4_28.

Rüegger, C.M., Bartsch, C., Martinez, R.M., Ross, S., Bolliger, S.A., Koller, B., Held, L., Bruder, E., Bode, P.K., Caduff, R., Frey, B., Schäffer, L., Bucher, H.U., 2014.

Minimally invasive, imaging guided virtual autopsy compared to conventional autopsy in foetal, newborn and infant cases: study protocol for the paediatric virtual autopsy trial. BMC Pediatr. 14, 15. https://doi.org/10.1186/1471-2431-14-15.

Sadhwani, A., Wypij, D., Rofeberg, V., Gholipour, A., Mittleman, M., Rohde, J., Velasco-Annis, C., Calderon, J., Friedman, K.G., Tworetzky, W., Grant, P.E., Soul, J.S., Warfield, S.K., Newburger, J.W., Ortinau, C.M., Rollins, C.K., 2022. Fetal brain volume predicts neurodevelopment in congenital heart disease. Circulation. https://doi.org/10.1161/CIRCULATIONAHA.121.056305.

Sanroma, G., Benkarim, O.M., Piella, G., Lekadir, K., Hahner, N., Eixarch, E., González Ballester, M.A., 2018. Learning to combine complementary segmentation methods for fetal and 6-month infant brain MRI segmentation. Comput. Med. Imaging Graph. 69, 52–59. https://doi.org/10.1016/j.compmedimag.2018.08.007.

Serag, A., Aljabar, P., Ball, G., Counsell, S.J., Boardman, J.P., Rutherford, M.A., Edwards, A.D., Hajnal, J.V., Rueckert, D., 2012. Construction of a consistent high-definition spatio-temporal atlas of the developing brain using adaptive kernel regression. Neuroimage 59, 2255–2265. https://doi.org/10.1016/j.neuroimage.2011.09.062.

Skotting, M.B., Eskildsen, S.F., Ovesen, A.S., Fonov, V.S., Ringgaard, S., Hjortdal, V.E., Lauridsen, M.H., 2021. Infants with congenital heart defects have reduced brain volumes. Sci. Rep. 11, 4191. https://doi.org/10.1038/s41598-021-83690-3.

Taha, A.A., Hanbury, A., 2015. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC Med. Imaging 15. https://doi.org/10.1186/s12880-015-0068-x.

Tourbier, S., Bresson, X., Hagmann, P., Meuli, R., Bach Cuadra, M., 2019. sebastientourbier/mialsuperresolutiontoolkit: MIAL Super-Resolution Toolkit v1.0. 10.5281/zenodo.2598448.

Tourbier, S., Bresson, X., Hagmann, P., Thiran, J.-P., Meuli, R., Cuadra, M.B., 2015. An efficient total variation algorithm for super-resolution in fetal brain MRI with adaptive regularization. Neuroimage 118, 584–597. https://doi.org/10.1016/j.neuroimage.2015.06.018.

van den Heuvel, M.I., Hect, J.L., Smarr, B.L., Qawasmeh, T., Kriegsfeld, L.J., Barcelona, J., Hijazi, K.E., Thomason, M.E., 2021. Maternal stress during pregnancy alters fetal cortico-cerebellar connectivity in utero and increases child sleep problems after birth. Sci. Rep. 11, 2228. https://doi.org/10.1038/s41598-021-81681-y.

Vasung, L., Abaci Turk, E., Ferradal, S.L., Sutin, J., Stout, J.N., Ahtam, B., Lin, P.-Y., Ellen Grant, P., 2019. Exploring early human brain development with structural and physiological neuroimaging. Neuroimage 187, 226–254. https://doi.org/10.1016/j.neuroimage.2018.07.041.

Vasung, L., Rollins, C.K., Yun, H.J., Velasco-Annis, C., Zhang, J., Wagstyl, K., Evans, A., Warfield, S.K., Feldman, H.A., Grant, P.E., Gholipour, A., 2020. Quantitative in vivo MRI assessment of structural asymmetries and sexual dimorphism of transient fetal compartments in the human brain. Cereb. Cortex 30, 1752–1767. https://doi.org/10.1093/cercor/bhz200.

Wiesenfarth, M., Reinke, A., Landman, B.A., Eisenmann, M., Saiz, L.A., Cardoso, M.J., Maier-Hein, L., Kopp-Schneider, A., 2021. Methods and open-source toolkit for analyzing and visualizing challenge results. Sci. Rep. 11, 2369. https://doi.org/10.1038/s41598-021-82017-6.

Wu, J., Sun, T., Yu, B., Li, Z., Wu, Q., Wang, Y., Qian, Z., Zhang, Y., Jiang, L., Wei, H., 2021a. Age-specific structural fetal brain atlases construction and cortical development quantification for chinese population. Neuroimage 241, 118412. https://doi.org/10.1016/j.neuroimage.2021.118412.

Wu, J., Yu, B., Wang, L., Yang, Q., Zhang, Y., 2021b. Longitudinal Chinese population structural fetal brain atlases construction: toward precise fetal brain segmentation. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. 2021, 2745–2749. https://doi.org/10.1109/EMBC46164.2021.9630514.

Wu, Y., Lu, Y.-C., Jacobs, M., Pradhan, S., Kapse, K., Zhao, L., Niforatos-Andescavage, N., Vezina, G., du Plessis, A.J., Limperopoulos, C., 2020. Association of prenatal maternal psychological distress with fetal brain growth, metabolism, and cortical maturation. JAMA Netw. Open 3, e1919940. https://doi.org/10.1001/jamanetworkopen.2019.19940.

Xu, X., Sun, C., Sun, J., Shi, W., Shen, Y., Zhao, R., Luo, W., Li, M., Wang, G., Wu, D., 2022. Spatiotemporal atlas of the fetal brain depicts cortical developmental gradient. J. Neurosci. 42, 9435–9449. https://doi.org/10.1523/JNEUROSCI.1285-22.2022.

Zhao, L., Asis-Cruz, J.D., Feng, X., Wu, Y., Kapse, K., Largent, A., Quistorff, J., Lopez, C., Wu, D., Qing, K., Meyer, C., Limperopoulos, C., 2022. Automated 3D fetal brain segmentation using an optimized deep learning approach. Am. J. Neuroradiol. https://doi.org/10.3174/ajnr.A7419.

Zvi, E., Shemer, A., Toussia-Cohen, S., Zvi, D., Bashan, Y., Hirschfeld-Dicker, L., Oselka, N., Amitai, M.-M., Ezra, O., Bar-Yosef, O., Katorza, E., 2020. Fetal exposure to MR imaging: long-term neurodevelopmental outcome. AJNR Am. J. Neuroradiol. 41, 1989–1992. https://doi.org/10.3174/ajnr.A6771.