Advances in Automated Fetal Brain MRI Segmentation and Biometry: Insights from the FeTA 2024 Challenge

Vladyslav Zalevskyi[1a,b], Thomas Sanchez[a,b], Misha Kaandorp[e,1], Margaux Roulet[a,b], Diego Fajardo-Rojas[c], Liu Li[g], Jana Hutter[c,d], Hongwei Bran Li[i,j], Matthew Barkovich[h], Hui Ji[e,f], Luca Wilhelmi[e], Aline Dändliker[e], Céline Steger[e,f], Mériam Koob[a], Yvan Gomez[ai,aj], Anton Jakovčić[aq], Melita Klaić[aq], Ana Adžić[aq], Pavel Marković[aq], Gracia Grabarić[aq], Milan Rados[aq], Jordina Aviles Verdera[c], Gregor Kasprian[as], Gregor Dovjak[as], Raphael Gaubert-Rachmühl[e], Maurice Aschwanden[e], Qi Zeng[k], Davood Karimi[k], Denis Peruzzo[l], Tommaso Ciceri[l], Giorgio Longari[m], Rachika E. Hamadache[n], Amina Bouzid[n], Xavier Lladó[n], Simone Chiarella[o], Gerard Martí-Juan[p], Miguel Ángel González Ballester[p,ak], Marco Castellaro[q], Marco Pinamonti[q], Valentina Visani[q], Robin Cremese[r], Keïn Sam[r], Fleur Gaudfernau[s], Param Ahir[t], Mehul Parikh[t], Maximilian Zenk[u,al], Michael Baumgartner[u,al], Klaus Maier-Hein[u,al,am,an,ao], Li Tianhong[v], Yang Hong[v], Zhao Longfei[v], Domen Preloznik[w], Žiga Špiclin[w], Jae Won Choi[x], Muyang Li[y], Jia Fu[y], Guotai Wang[y], Jingwen Jiang[z], Lyuyang Tong[z], Bo Du[z], Andrea Gondova[aa,at], Sungmin You[aa,at], Kiho Im[aa,at,au], Abdul Qayyum[g], Moona Mazher[ab], Steven A Niederer[g], Maya Yanko[af], Bella Specktor-Fadida[ag], Dafna Ben Bashat[ah], Andras Jakab[e,ae], Roxane Licandro[ac,ad], Kelly Payette[†c,d,e], Meritxell Bach Cuadra[†b,a]

[a] *Department of Radiology, Lausanne University Hospital and University of Lausanne, Lausanne, Switzerland*
[b] *CIBM Center for Biomedical Imaging, Lausanne, Switzerland*
[c] *Department of Early Life Imaging, School of Biomedical Engineering & Imaging Sciences, King's College London, London, UK*
[d] *Department of Imaging Physics and Engineering, School of Biomedical Engineering & Imaging Sciences, King's College London, London, UK*
[e] *Center for MR-Research, University Children's Hospital Zurich, University of Zurich, Zurich, Switzerland*
[f] *Neuroscience Center Zurich, University of Zurich, Zurich, Switzerland*
[g] *National Heart & Lung Institute, Imperial College London, London, UK*
[h] *University of California, San Francisco; UCSF Benioff Children's Hospital, San Francisco, California, USA*
[i] *Department of Quantitative Biomedicine, University of Zurich, Zurich, Switzerland*
[j] *Department of Informatics, Technical University of Munich, Munich, Germany*
[k] *Boston Children's Hospital, Harvard Medical School, Boston, Massachusetts, USA*
[l] *Neuroimaging Unit, Scientific Institute IRCCS E. Medea, Bosisio Parini, Italy*
[m] *Department of Informatics, Systems and Communication, University of Milano Bicocca, Milan, Italy*

[n]Research Institute of Computer Vision and Robotics (ViCOROB), Universitat de Girona, Girona, Spain
[o]Università di Bologna, Bologna, Italy
[p]BCN MedTech, Department of Engineering, Universitat Pompeu Fabra, Barcelona, Spain
[q]Department of Information Engineering, University of Padova, Padova, Italy
[r]Institut Pasteur, Université Paris Cité, CNRS UMR 3571, Decision and Bayesian Computation, Paris, France
[s]Inria, HeKA, PariSantéCampus, Paris, France
[t]L. D. College of Engineering, Gujarat, India
[u]Medical Faculty Heidelberg, Heidelberg University, Pattern Analysis and Learning Group, Department of Radiation Oncology, Heidelberg University Hospital, Heidelberg, Germany
[v]Canon Medical Systems (China) Co., Ltd, , China
[w]Faculty of Electrical Engineering, University of Ljubljana, Ljubljana, Slovenia
[x]Department of Radiology, Seoul National University Hospital, Seoul, South Korea
[y]School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, Chengdu, China
[z]School of Computer Science, Wuhan University, Wuhan, China
[aa]Fetal Neonatal Neuroimaging and Developmental Science Center, Boston Children's Hospital, Harvard Medical School, Boston, Massachusetts, USA
[ab]Hawkes Institute, Department of Computer Science, University College London, London, UK
[ac]Laboratory for Computational Neuroimaging, Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital/Harvard Medical School, Charlestown, Massachusetts, USA
[ad]Department of Biomedical Imaging and Image-guided Therapy, Computational Imaging Research Lab (CIR), Early Life Image Analysis Group, Medical University of Vienna, Vienna, Austria
[ae]University Research Priority Project Adaptive Brain Circuits in Development and Learning (AdaBD), University of Zurich, Zurich, Switzerland
[af]Sagol Brain Institute, Tel Aviv Sourasky Medical Center and School of EE, Tel-Aviv University, Tel-Aviv, Israel
[ag]Department of Medical Imaging Sciences, The Faculty of Social Welfare and Health Sciences, University of Haifa, Haifa, Israel
[ah]Sagol Brain Institute, Tel Aviv Sourasky Medical Center and Faculty of Medicine and Sagol School of Neuroscience, Tel-Aviv University, Tel-Aviv, Israel
[ai]Department Woman-Mother-Child, CHUV, Lausanne, Switzerland
[aj]BCNatal Fetal Medicine Research Center (Hospital Clínic and Hospital Sant Joan de Déu), Universitat de Barcelona, Barcelona, Spain
[ak]ICREA, Barcelona, Spain
[al]German Cancer Research Center (DKFZ) Heidelberg, Division of Medical Image Computing, Heidelberg, Germany
[am]Pattern Analysis and Learning Group, Department of Radiation Oncology, Heidelberg University Hospital, Heidelberg, Germany
[an]Helmholtz Imaging, German Cancer Research Center (DKFZ), Heidelberg, Germany
[ao]Faculty of Mathematics and Computer Science, Heidelberg University, Heidelberg, Germany
[ap]University of Zurich, Zurich, Switzerland
[aq]Croatian Institute for Brain Research, School of Medicine, University of Zagreb, Zagreb, Croatia

[ar] *Department of Biomedical Engineering, School of Biomedical Engineering & Imaging Sciences, King's College, London, United Kingdom*
[as] *Department of Biomedical Imaging and Image-Guided Therapy, Division of Neuroradiology and Musculoskeletal Radiology, Medical University of Vienna, Vienna, Austria*
[at] *Division of Newborn Medicine, Boston Children's Hospital, Harvard Medical School, Boston, Massachusetts, USA*
[au] *Department of Radiology, Boston Children's Hospital, Harvard Medical School, Boston, Massachusetts, USA*

Abstract

Accurate segmentation and biometric analysis are essential for studying the developing fetal brain *in utero*. The Fetal Brain Tissue Annotation (FeTA) Challenge 2024 builds upon previous editions to further advance the clinical relevance and robustness of automated fetal brain MRI analysis. This year's challenge introduced biometry prediction as a new task complementing the usual segmentation task. The segmentation task also included a new low-field (0.55T) MRI testing set and used Euler characteristic difference (ED) as a topology-aware metric for ranking, extending the traditional overlap or distance-based measures.

A total of 16 teams submitted segmentation methods for evaluation. Segmentation performance across top teams was highly consistent across both standard and low-field MRI data. Longitudinal analysis over past FeTA editions revealed minimal improvement in accuracy over time, suggesting a potential performance plateau, particularly as results now approach or surpass reported levels of inter-rater variability. However, the introduction of the ED metric revealed topological differences that were not captured by conventional metrics, underscoring its value in assessing segmentation quality. Notably, the curated low-field MRI dataset achieved the highest segmentation performance, illustrating the potential of affordable imaging systems when combined with high-quality preprocessing and reconstruction.

A total of 7 teams submitted automated biometry methods for evaluation. While promising, this task exposed a critical limitation: most submitted methods failed to outperform a simple baseline that predicted measurements based solely on gestational age, without using image data. Performance varied widely across biometric measurements and between teams, indicating both current challenges and opportunities for improvement in this area. These findings highlight the need for better integration of volumetric context and stronger modeling strategies needed for the clinical adoption of automated fetal biometry estimation.

In addition, we analyzed different dimensions of domain shifts within our data

and observed that image quality was the most influential factor affecting model generalization, with Dice score differences of up to 0.10 between low- and high-quality scans. The choice of super-resolution reconstruction pipeline also had a substantial impact on segmentation performance. Other factors—such as gestational age, pathology, and acquisition site—also contributed to performance variability, but their effects were comparatively smaller.

Overall, FeTA 2024 provides a rigorous, multi-faceted benchmark for evaluating multi-class segmentation and biometry estimation in fetal brain MRI. It emphasizes the need for data-centric approaches, improved topological modeling, and greater dataset diversity to develop clinically reliable and generalizable AI tools for fetal neuroimaging.

Keywords: Fetal Brain, MRI, Low-field, Segmentation, Topology, Biometry, Domain Shift, Challenge results

---

† Equal contributions.

1. Introduction

The fetal brain undergoes rapid and complex development throughout gestation, influenced by both genetic and environmental factors. Understanding this dynamic process is critical in both clinical and research domains, as neurodevelopmental disruptions are linked to congenital anomalies and long-term cognitive or physiological impairments (Griffiths et al., 2017; Ciceri et al., 2024; Van den Bergh et al., 2018). In vivo imaging biomarkers derived from ultrasonography (US) or magnetic resonance imaging (MRI) provide non-invasive and quantifiable metrics to monitor prenatal brain development. Deviations from normative patterns in these biomarkers have been associated with a range of pathologies, including corpus callosum (Marathu et al., 2024; Lamon et al., 2024) and posterior fossa malformations (Dovjak et al., 2020; Mahalingam et al., 2021), ventriculomegaly (Chen et al., 2024), and have been shown to correlate with neurodevelopmental outcomes in conditions such as congenital heart disease (Sadhwani et al., 2022), intrauterine growth restriction (Egaña-Ugrinovic et al., 2015; Meijerink et al., 2023), and preterm birth (Story et al., 2021; Hall et al., 2024).

Fetal brain MRI has emerged as an important non-invasive tool for studying neurodevelopment in utero and diagnosing congenital disorders, complementing ultrasonography (Griffiths et al., 2017; Alamo et al., 2010). Accurate and automatic segmentation of fetal brain tissues in MRI is critical for quantitative analysis and biomarker extraction, including tissue volumetry, cortical morphometry (Payette

et al., 2023), and biometric measurements (She et al., 2023). Manual segmentation, however, remains labor-intensive, error-prone, and susceptible to inter-observer variability, underscoring the necessity of reliable automated techniques.

While clinical US and 2D MRI are the standard techniques for assessing fetal development (Tilea et al., 2009), the use of super-resolution reconstruction (SRR) techniques to generate 3D fetal brain reconstructions has emerged as a powerful advancement. SRR methods fuse multiple 2D MRI slices (often motion-corrupted) into a single, enhanced 3D motion-corrected volume, significantly improving brain analysis (Gafner et al., 2020; Avisdris et al., 2021; Matthew et al., 2024). Recent studies have shown that biometric measurements derived from 3D SRR volumes correlate strongly with those from ultrasound, while offering greater rater confidence than using 2D MRI series (Lamon et al., 2024; Khawam et al., 2021; Gafner et al., 2020; Sanchez et al., 2024b; Kyriakopoulou et al., 2016; Ciceri et al., 2023).

The Fetal Tissue Annotation (FeTA) challenges, held in 2021 (Payette et al., 2023) and 2022 (Payette et al., 2024), have significantly advanced fetal brain MRI analysis by providing public datasets and standardized evaluation protocols for brain tissue segmentation. The **FeTA 2024 challenge** builds on previous editions, retaining the core **brain tissue segmentation** task and introducing a new clinically relevant objective: **biometry extraction**, alongside several other key innovations.

Firstly, FeTA 2024 introduces a new low-field (LF, 0.55T) MRI testing dataset. LF MRI offers a low-cost alternative to 1.5–3T systems, making it especially valuable in resource-limited settings (Arnold et al., 2023; Marques et al., 2019). This affordability supports research in low- and middle-income countries with large pediatric populations, where access to high-field MRI is limited, hindering studies on brain development under normal and adverse conditions (Murali et al., 2023; Aviles Verdera et al., 2023).

Secondly, we introduced the Euler characteristic difference as an additional ranking metric for segmentation (Taha and Hanbury, 2015). Unlike overlap- or distance-based metrics, it captures topological correctness, offering a complementary view of performance (Maier-Hein et al., 2024). This is especially relevant for downstream tasks like cortical surface extraction or morphometric analyses (e.g., sulcal folding, cortical maturation, or structural abnormality assessment) (Yehuda et al., 2023; Clouchoux et al., 2011).

In FeTA 2024, we promote the development of generalizable, fully automated methods for fetal brain analysis, enabling the extraction of key imaging biomarkers via multi-class tissue segmentation and biometry across diverse acquisition and reconstruction settings. This paper offers **a comprehensive overview of the challenge**, covering its organization, submitted algorithms, performance assessment,

benchmarking, and evaluation using the BIAS reporting framework, which emphasizes transparency, reproducibility, and fairness (Maier-Hein et al., 2020). We also analyze performance trends over time, tracking improvements in state-of-the-art segmentation accuracy across FeTA editions. Finally, we assess **data quality** in both training and testing sets to examine its impact on the generalization of submitted methods. Combined with other domain shifts—such as gestational age, pathology, super-resolution reconstruction, and acquisition site—our work provides a deep **overview of how domain shifts affect deep learning models for fetal brain analysis** and informs strategies to mitigate their impact.

2. Methods

2.1. Challenge organization

Context. The FeTA 2024 challenge was held as a thematic event within the Perinatal, Preterm, and Pediatric Image Analysis (PIPPI) workshop[2], part of the Medical Image Computing and Computer-Assisted Intervention (MICCAI) 2024 conference. The challenge was run through a custom platform, available at `https://fetachallenge.github.io/`, which provided participants with all the necessary information on the organization, time frame, and submission instructions.

Data, participation and submission. Challenge participation required submission of **fully automated** segmentation and/or biometry algorithms. A training set of 3D super-resolution fetal brain MRI from two institutions was provided; no validation set was released, and test data remained private for evaluation. Participants could use publicly available external datasets and pre-trained models, provided these were public and fully documented in the algorithm description, as well as use both 2D and 3D models.

Participants submitted their algorithms as Docker containers with a command-line interface for test data evaluation[3]. Any programming language was allowed, provided the input/output followed the evaluation utility specifications. Each team was allowed one submission, except in cases of technical errors (e.g., Docker issues), which could be corrected upon notification by the organizers. Evaluation on test data was performed by the organizers using publicly available code[4]. To promote transparency and reproducibility, FeTA 2024 encouraged participants to share

[2] `https://pippiworkshop.github.io/`

[3] Instructions: `https://github.com/fetachallenge/fetachallengesubmission`.

[4] Available at `https://fetachallenge.github.io/pages/Evaluation`.

their code publicly. A Docker Hub page (`https://hub.docker.com/repositories/fetachallenge2024`) was created to host containers from teams who agreed to release their Docker images.

Timetable, rewards and results paper. The challenge followed a predefined schedule: training data was released on May 21, 2024; registration opened after challenge acceptance. The Docker submission deadline was extended to August 4, 2024, and algorithm descriptions were due by August 12. On August 23, the top five teams were invited to prepare 2-minute pitch presentations for the challenge day and, along with all participants, were invited to present posters at the dedicated conference session. The challenge took place in person on October 6, 2024, during MICCAI. Results were announced live and later published on the challenge website, along with top teams' presentations (with their consent). The top three teams in each task received certificates and small gifts, including a 3D-printed fetal brain keychain for in-person attendees. The highest-ranking team in each task also received a box of artisanal Swiss-made cookies. Organizers could participate but were not eligible for awards.

All teams with valid submissions and interest in the publication were included in this results paper, with up to three members per team listed as co-authors. Teams were free to publish their algorithms and results independently after the challenge, without embargo, provided they cited both the data publication (Payette et al., 2021) and this summary paper.

Data usage terms and conflicts of interest. The training data from the University Children's Hospital Zürich (**Kispi**) and General Hospital Vienna/Medical University of Vienna were provided with specific licensing conditions. Kispi data, hosted on the Synapse platform[5], **is for non-commercial use only**. **Vienna data** is governed by a custom Data Transfer Agreement, **allowing use for challenge purposes only**. Participants could modify the data, including generating synthetic data through augmentation, as long as modifications were documented and synthetic data could be provided to the organizers upon request. None of the organizers participated in this year's challenge or have conflicts of interest to disclose. The challenge awards were funded by the institutional budget (Kispi), and none of the participants were involved in funding. Only organizers at Kispi had full access to the testing dataset, as they managed data transfer agreements with all providers.

[5] `https://www.synapse.org/Synapse:syn25649159/wiki/610007`

2.2. Challenge tasks

The FeTA challenge presents two primary tasks (see Figure 1). Participants could choose to compete in either or both tasks.

Task 1. Fetal brain tissue segmentation. This task aims to develop algorithms that automatically delineate different tissues in SRR fetal brain MRI. The 3D semantic segmentation involves classifying each voxel into one of seven predefined classes: Background, External CSF, Grey Matter (GM), White Matter (WM), Ventricles including cavum (VM), Cerebellum (CBM), Deep Grey Matter (SGM), and Brainstem (BSM). Reference annotation procedures and inter-rater variability analyses for all datasets (except the new LF set) are detailed in Payette et al. (2021) and Payette et al. (2024). The LF dataset followed the same annotation protocol, with seven annotators (AJ, CS, RG, VZ, YG, MA, MR) each segmenting a specific label map. These were merged into a single reference annotation, reviewed, and corrected by two fetal MRI experts (KP, AJ).

Task 2. Biometric measurements prediction. The goal of this task is to develop algorithms that automatically and accurately estimate key fetal brain biometry from MRI. The selected measurements—length of the corpus callosum (LCC), height of the vermis (HV), brain biparietal diameter (bBIP), skull biparietal diameter (sBIP), and transverse cerebellar diameter (TCD)—were chosen to **minimize annotation burden** while providing **complementary anatomical and diagnostic value**. Four raters contributed: YG (5 years' fetal MRI experience), MKo (16 years), and junior raters RG and MA (reviewed by AJ, 12 years)[6]. Not all measurements were available for all cases: in the test set, 15 cases lacked LCC, one lacked HV, and one lacked TCD due to annotator uncertainty. In the training set, 102 of 120 cases had complete annotations—10 Kispi cases were excluded for poor quality; 5 Kispi and 3 Vienna cases had partial annotations. While the main goal was to predict biometry values, the training set also included 3D **landmark annotations**—single-voxel labels marking anatomical structures used to derive each measurement. Clinicians identified these landmarks during annotation, and the actual biometry values were computed via organizer-provided scripts. Both the landmarks and scripts were shared, allowing participants to either regress biometry directly or predict landmarks, followed by automated biometry measurement.

[6]Biometry annotation protocol is described on our website: `https://fetachallenge.github.io/pages/Data_description`

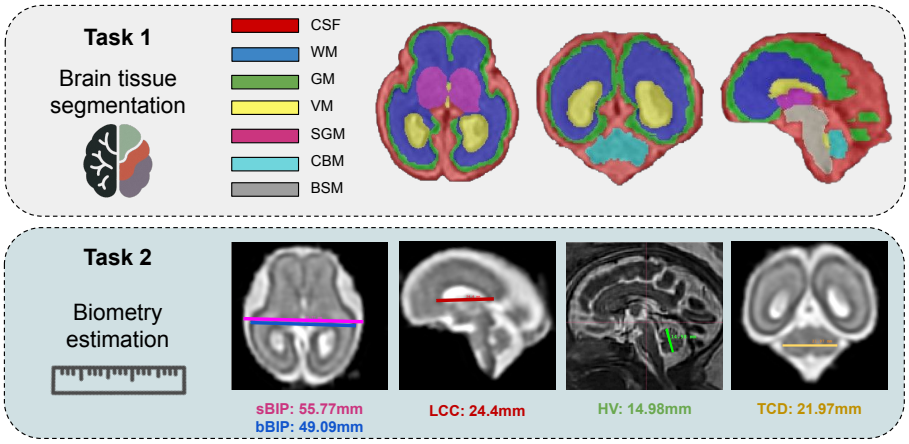


Figure 1: FeTA 2024 Challenge Tasks. Task 1 involves segmentation of fetal brain tissues into seven classes, while Task 2 focuses on estimating five biometric measurements, both illustrated in the figure.

2.3. Challenge data sets

Subject selection aimed to ensure a representative cohort spanning 18–35 weeks of gestation, including both neurotypical and pathological cases (e.g., spina bifida, ventriculomegaly, corpus callosum malformations) to reflect clinical practice. UCSF and CHUV data were acquired during clinical fetal MRI scans following ultrasound referral, performed by trained medical staff. Data from KCL, Kispi, and Vienna were collected using research protocols. All cohorts had approval by the local ethics committee for use in the challenge after anonymization[7].

Each case included a 3D fetal brain MRI reconstruction, manual brain tissue segmentation, and biometry annotations. Metadata included gestational age (GA) and a binary label indicating neurotypical or pathological status. To preserve anonymity, gender was excluded and GA was randomly offset by ± 3 days.

The challenge dataset comprises of 120 training and 180 test cases. The test set was split into *in-domain* (from the same institutions and protocols as training data) and *out-of-domain* cases. To ensure balance, both subsets were similar in size to the training set. Demographic characteristics, including GA and pathology distribution,

[7]**KISPI:** Ethical Committee of the Canton of Zurich, Switzerland (Decision numbers: 2017 00885, 2016 01019, 2017 00167). **CHUV**: Ethics Committee of the Canton de Vaud, Switzerland (CER-VD 2021 00124). **Vienna:** Approved by the ethics review board and data clearing department at the Medical University of Vienna. **UCSF:** institutional review board (IRB 16 20619). **KCL**: Ethics Committee Dulwich (Ethics code 19 LO 0852).

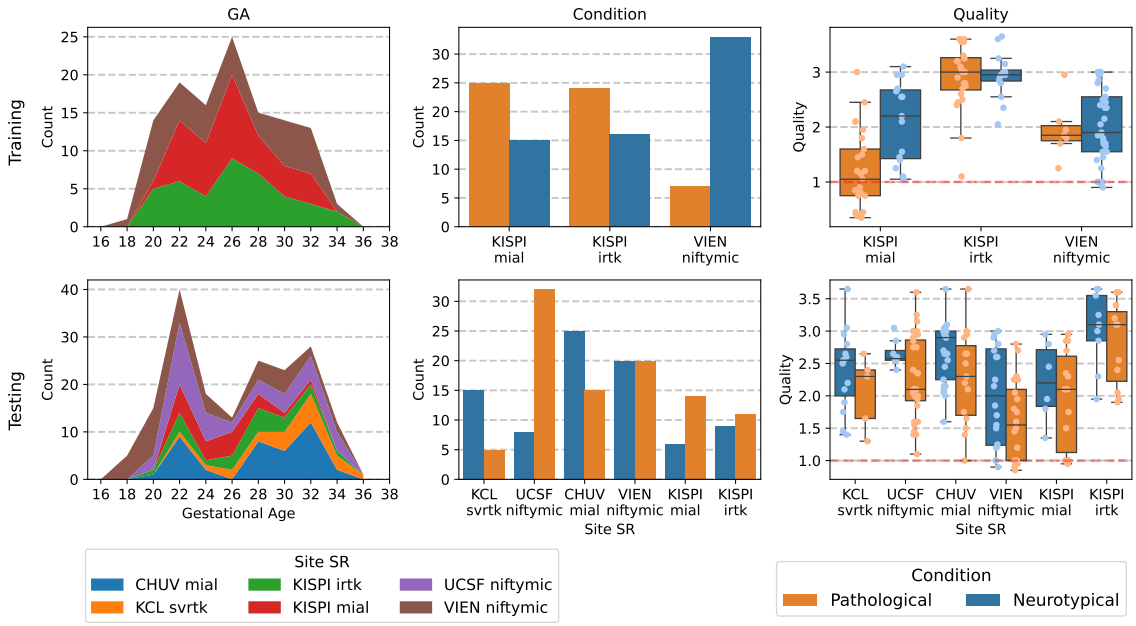


Figure 2: FeTA 2024 data distribution by GA (weeks), condition, and image quality (0 = lowest, 4 = highest; 1 = minimum acceptable), stratified by Site and SR method for training (top) and testing (bottom) sets. In the image quality plots, the red dotted line marks the threshold score (1.0); images with a score below this value are classified as poor quality.

were matched across training and testing cohorts (see Figure 2). For this year's challenge, manual quality control was performed on all training and testing cases following the protocol by Sanchez et al. (2024a), ensuring comparable data quality between training and testing sets (Figure 2).

The FeTA 2024 training and testing datasets are identical to those used in FeTA 2022, with two additions: a new low-field out-of-domain test set from King's College London (KCL) and manual biometry annotations.

All cases were acquired using T2-weighted single-shot fast spin-echo sequences[8], the standard for structural fetal MRI due to their high signal-to-noise ratio and reduced sensitivity to fetal motion. To further mitigate motion artifacts, multiple stacks were acquired in various orientations (axial, sagittal, coronal, and off-plane). Manual selection of 2D stacks was done at each site and then combined into a single high-resolution, isotropic 3D image via super-resolution reconstruction. The result-

[8]Also known as HASTE (Siemens), SSTSE (Philips), or SSFSE (GE), depending on the scanner manufacturer.

Table 1: FeTA 2024 datasets properties. N_n - number of neurotypical subjects, N_p - number of pathological subjects. "+" indicates the minimum TE value

Used for	Testing domain	Institution	Scanner	N	SR method	SR res. (mm^3)	TR/TE (ms)	GA (weeks)	N_n/N_p
Training	In domain	KISPI	GE Signa Discovery MR450/MR750 (1.5T/3T respectively)*	80	MIALSRTK (40) IRTK-simple (40)	$(0.5)^3$	2000-3500/ 120+	20-34.4	49/31
		Vienna	Philips Ingenia/Intera (1.5T) Philips Achieva (3T)*	40	NiftyMIC	$(1.0)^3$	6000-22000/ 80-140	19.3-34.4	33/7
Testing	In domain	KISPI	GE Signa Discovery MR450/MR750 (1.5T/3T respectively)*	40	MIALSRTK (20) IRTK-simple (20)	$(0.5)^3$	2000-3500/ 120+	21.3-34.6	15/25
		Vienna	Philips Ingenia/Intera (1.5T) Philips Achieva (3T)*	40	NiftyMIC	$(1.0)^3$	6000-22000/ 80-140	18.1-35.5	20/20
	Out of domain	CHUV	Siemens MAGNETOM Aera (1.5T)	40	MIALSRTK	$(1.125)^3$	1200/90	21.0-35.0	25/15
		UCSF	GE Signa Discovery MR750/MR750W (3T)	40	NiftyMIC	$(0.8)^3$	200-3500/ 100+	20.0-35.1	8/32
		KCL	Siemens MAGNETOM Free.Max (0.55T)	20	SVRTK	$(0.8)^3$	2500/106	21.0-35.0	15/5

*The training dataset contained data from both 1.5T and 3T scanners. However, which cases belonged to which scanner were not provided to the participants as it was part of the data anonymization process. Therefore, the breakdown of number of cases per scanner is not provided here.

ing 3D volumes were zero-padded to 256x256x256 and reoriented to a standard radiological plane. A summary of acquisition parameters, demographic characteristics, and reconstruction methods in all sites is provided in Table 1. Additional details about the new KCL dataset are provided below. For further detailed information on the FeTA 2022 acquisitions, please refer to Payette et al. (2024).

Data from **KCL** was collected using a 0.55T low-field MRI scanner (Siemens MAGNETOM Free.Max) with a HASTE sequence as part of a prospective single-center study and fully anonymized following local procedures (Ethics Committee Dulwich 19 LO 0852). The acquired stacks had a resolution of 1.5mm x 1.5mm x 4.5mm, which were then reconstructed into a high-resolution volume of 0.8mm x 0.8mm x 0.8mm using SVRTK (Uus et al., 2020). Key acquisition parameters include a flip angle of 180°, a field-of-view of $450 \times 450 mm^2$, and a base resolution of 304x304 pixels, yielding a voxel size of $1.5 \times 1.5 \times 4.5 mm^3$. The acquisition time ranged from 64 to 122 seconds. Data collection took place at St Thomas' Hospital in London, United Kingdom, without the use of maternal or fetal sedation. All acquisitions were performed using the contour L coil and the integrated spine coil while the mother was in a supine position. This dataset is used only in the testing set.

2.4. Evaluation Metrics

We provide a short recall of the ranking metrics. The detailed mathematical formulation is available in supplementary materials A1.

Task 1. Segmentation. Performance of segmentation algorithms is comprehensively assessed through complementary metrics of spatial overlap, volume, shape, and topological correctness:

- **Dice Similarity Coefficient (Dice; ↑)[9]:** measures voxel-wise correspondence between the predicted and ground truth (GT) segmentations.
- **Volume Similarity (VS; ↑):** measures the similarity of the volumes between the predicted and GT segmentations.
- **Hausdorff Distance (HD95; ↓):** quantifies the distance between contours of the predicted and GT segmentations with robustness to outliers.
- **Euler Characteristics Difference (ED; ↓)**: evaluates the topological similarity between the predicted and GT segmentations.

As ED is included in the ranking for the first time, we describe it further. It is based on the Euler characteristic (EC):

$$EC = \mathrm{BN}_0 - \mathrm{BN}_1 + \mathrm{BN}_2$$

where Betti Number BN_0 represents the number of connected components (i.e., regions), BN_1 represents the number of loops or holes and BN_2 represents the number of voids or cavities. The ED difference is then computed as $|EC_{pred} - EC_{GT}|$. Smaller differences indicate better topological alignment. The Betti number values of GT are: for all brain tissue labels, $\mathrm{BN}_1 = 0$ and $\mathrm{BN}_2 = 0$. For the eCSF, WM, ventricles, cerebellum, dGM, and brainstem, $\mathrm{BN}_0 = 1$, while for GM, $\mathrm{BN}_0 = 2$.

Task 2: Biometry Estimation. The primary metric for evaluating biometry estimation algorithms is **mean average percentage error (MAPE; ↓)** , which quantifies the error in the estimated biometric measurements relative to the actual measurements:

$$MAPE = \frac{1}{N} \sum_{i=1}^{N} \frac{|y_i - \hat{y}_i|}{y_i} \times 100,$$

where y_i and $\hat{y}_i$ are GT and predicted measurements respectively, and N is the total number of measurements. This metric accounts for variable sizes of the target structures and is used to assess the accuracy of the estimated biometric measurements.

[9] ↑ means that a higher score is better and ↓ that a lower score is better

2.5. Ranking

Submissions are ranked based on metrics computed for each brain tissue label (or biometric measurement) in the predicted maps of the fetal brain volumes. For segmentation, the final rank is the average of all 4 metrics: Dice, HD95, VS, and ED. For biometry, the final rank is based on MAPE. For metrics where higher values are better (Dice, VS), the algorithm with the highest value ranks best. For metrics where lower values are better (like HD95 and ED for segmentation and MAPE for biometry tasks), the algorithm with the lowest value ranks best. The individual label rankings are summed, and the algorithm with the highest combined rank is considered the best.

In cases of missing results (e.g., if an algorithm fails to detect a label or if the entire label map is empty), the worst possible values will be assigned to the algorithm. For example, if a label is missing in the label map, it will receive a Dice and VS of 0. For HD95, EC, and MAPE, the missing values are set to double the maximum value of other algorithms for that sub-ranking. This ensures that algorithms with missing results are ranked last for that specific task/brain tissue.

2.5.1. Biometry baselines

In the ranking of Task 2, two additional baseline models representing lower and upper performance limits were incorporated as separate submissions. These entries, intended solely for benchmarking purposes, were not considered in the formal determination of the challenge competition ranking.

Lower bound: Gestational age regression model. This *model*, referred to as [**GA**] in the result's Table 5, is a simple univariate linear regression baseline. For each biometric measurement y, the model predicts its value $\hat{y}$ using the gestational age (GA) as the sole explanatory variable, mathematically:

$$\hat{y} = \beta_0 + \beta_1 \cdot \mathrm{GA},$$

where β_0 is the intercept and β_1 is the regression coefficient learned from the training data. This baseline does not rely on the image and aims at quantifying how strongly the GA can account for the size of a given structure.

Upper bound: Inter-rater variability. The upper bound is set by averaging inter-rater variability, further denoted as [**inter-rater**]. This reflects the best-expected accuracy, accounting for measurement errors and uncertainties between manual raters. For each biometric measurement in the test dataset, annotations from two independent observers are used by comparing one observer's measurement to the other's, with the result averaged across all test cases.

2.6. Statistical analysis

The non-parametric **Wilcoxon signed-rank** test was used to assess performance differences between algorithms, as the Shapiro-Wilk test indicated non-normal distribution. To evaluate performance differences across subsets (e.g., neurotypical vs. pathological cohorts), we applied the **Mann-Whitney U test** (Wilcoxon rank-sum test). For all tests, statistical significance was set at $p < 0.05$. For multiple comparisons, such as between sites or labels, we applied **Bonferroni correction**.

2.7. Further analysis

FeTA 2024, as the third edition of the challenge, provides an opportunity to assess progress and unsolved challenges. We report two additional analyses: **(i)** the evolution of top-performing segmentation models over the last three editions, and **(ii)** the impact of different domain shift sources on model performance.

2.7.1. Insights from three years of competition: progress or plateau?

To assess progress in fetal brain tissue segmentation, we analyze the evolution of top-performing algorithms over time. Specifically, we compare the performance of the highest-ranked teams from the FeTA challenges in 2021, 2022, and 2024, evaluating segmentation accuracy across the dataset splits available in each respective year.

To extend the longitudinal comparison, we perform a retrospective evaluation of the 2022 winning method on the KCL dataset, first introduced as a test set in 2024. This is enabled by the 2022 winning team's release of their Docker container[10], allowing us to assess the generalization of a previously state-of-the-art solution to new, unseen data, and to identify both progress and persistent limitations.

2.7.2. Quantifying domain shifts

Domain shifts remain a key obstacle in fetal brain MRI analysis, often undermining model generalizability. These shifts arise from variations in subject demographics, imaging protocols, scanner types, and reconstruction methods (Dockès et al., 2021). In fetal imaging, GA notably affects brain morphology and contrast, while pathologies such as ventriculomegaly, for example, can significantly alter anatomical structure. Beyond biological and acquisition-related variability, low contrast or motion artifacts can degrade reconstruction quality, adversely affecting segmentation and biometry.

[10] `https://hub.docker.com/r/fetachallenge22/feta-imperial-tum-2022-nnunet`

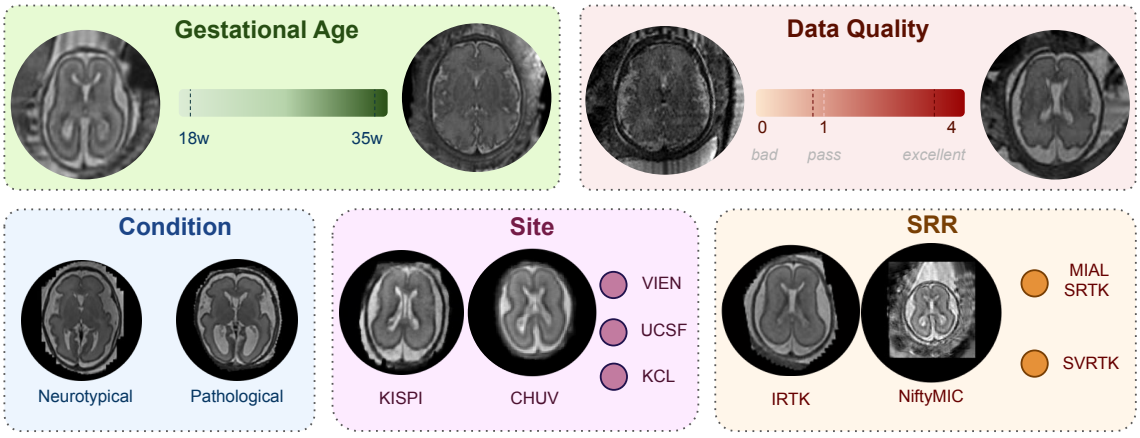


Figure 3: Illustration of the sources of domain shifts in fetal brain MRI datasets of FeTA 2024. Demonstrated across gestational age (18 vs. 35 weeks), data quality (0.9 vs. 3.64), clinical condition (neurotypical vs. pathological), acquisition site (KISPI vs. CHUV), and SRR methods (IRTK vs. NiftyMIC). In each comparison, only the indicated domain is varied, while all other domains remain constant. Additional domains within each source, not shown here, are represented by circles.

Is image quality a domain in itself? To assess whether image quality impacts model generalization, we manually rated the quality of all 180 test volumes using the protocol from Sanchez et al. (2024a) and explored the interaction of data quality with the performance of the submitted algorithms across the test data.

Comparing the impact of domain shift factors. To assess how domain shifts influence segmentation performance, we examine six key sources of variability: image quality, GA, condition (neurotypical or pathological), acquisition site, testing domain (seen vs. unseen during training), and the SRR method. These factors are summarized in Figure 3. To evaluate the influence of domain shift factors on segmentation performance, we trained a random forest regressor for each metric of interest (Dice, HD95, Volume Similarity, Euler Difference), using six dataset-level variables as input features. Target values were defined as the average metric scores across the top 3 teams. To estimate feature importance, we applied SHapley Additive exPlanations (SHAP) (Lundberg and Lee, 2017), which quantify the contribution of each feature by computing its average marginal effect across all possible feature combinations. This approach provides a unified and interpretable measure of how each factor affects performance.

3. Results

3.1. Algorithm description

We received 176 access requests for the KISPI training cohort hosted on Synapse during the challenge active period (May–July 2024). However, not all of these requests were related to the FeTA challenge, as the dataset is also available for broader research purposes. For the Vienna dataset, 53 data access applications were submitted, but only 30 applicants completed the data transfer agreement process and successfully received the data.

For the segmentation task, we received 16 valid submissions, all evaluated on the full test set. One team declined participation in this paper and was excluded from the analysis; results from the remaining 15 teams are presented. For the biometry task, 7 teams submitted results. One team (`falcons`) failed to generate valid outputs for all test cases and was penalized accordingly, as described in the Section 2.5. Notably, all biometry participants also submitted segmentation entries, leveraging segmentation outputs either as a preprocessing step or direct input for biometry estimation. A detailed description of each algorithm is provided in the supplementary materials (appendix A2) and summarized in Tables 2 and 3.

3.1.1. Common data and model augmentation strategies

Participants adopted a variety of approaches, with the majority utilizing 3D architectures—14 out of 16 for the segmentation task and 6 out of 7 for the biometry task. Across both tasks, two strategies were commonly used: data augmentation and model ensembling.

Data augmentation was universally applied, with all segmentation (16 teams) and biometry (7 teams) models using it. Standard transformations like flipping, rotation, scaling, and intensity shifts were common, while advanced methods, such as SynthSeg (Billot et al., 2023) or global intensity non-linear augmentations (GIN) (Ouyang et al., 2022), were used by 3 teams. Some teams also simulated domain-specific artifacts, including fetal motion and bias field.

Ensembling was a key approach in segmentation, used by 14 out of 16 teams. This included combining models trained on different cross-validation splits (4 teams) or using varied architectures, training setups, data orientations, or augmentation schemes (8 teams). Some also integrated pre- or post-processing models, like denoising autoencoders or skull-stripping (2 teams). Ensembling was less common in the biometry task, with only 2 out of 7 teams employing it, as most models built biometry predictions in a single pipeline on top of segmentation outputs.

3.1.2. Segmentation models

Among the 16 submissions, the most common architectures were **nnU-Net** Isensee et al. (2018) (9 teams) and **U-Net** (Ronneberger et al., 2015; Çiçek et al., 2016) (6 teams), often used as baselines. Many teams enhanced these models with **attention mechanisms** (Vaswani, 2017), **residual connections** (He et al., 2016), or **ensembling**. Others explored alternatives such as **Swin Transformers** (Liu et al., 2021), custom U-Net variants, or hybrid CNN–Transformer designs. Most models were developed in **PyTorch** (12 teams), with parameter counts ranging from 5M to 140M (median: 31M, mean: 44.8M).

Use of **external data** was limited to 5 teams, primarily leveraging dHCP data (Hughes et al., 2016; Edwards et al., 2022), fetal brain atlases (Gholipour et al., 2017; Uus et al., 2023a), or foundation models pretrained on large-scale image datasets (Roy et al., 2023).

3.1.3. Biometry models

All biometry models leveraged segmentation outputs, either as pre-processing, auxiliary, or core input. Two teams employed nnU-Net or U-Net variants for direct regression, while others used custom CNNs (1/7) or more complex architectures integrating attention mechanisms or hybrid designs (4/7). Prediction strategies varied across teams: two teams directly regressed biometry values; three teams predicted 3D landmark coordinates; and two teams generated 3D landmark heatmaps. In the latter two approaches, biometry values were subsequently computed using scripts provided by the organizers. Most teams used 3D models (6/7), implemented primarily in **PyTorch** (6/7), with one using TensorFlow. Three teams leveraged **external data**, such as dHCP and fetal brain atlases, or employed foundation models pretrained on large-scale datasets.

Table 2: Summary of the algorithms submitted for the fetal brain tissue segmentation task.

Team name	Model Architecture	Deep Learning Framework	Dim	Data Augmentation	Cross-Validation	External Data	Ensembling	Original Aspects
cemrg	Hybrid Cross Attention Swin Transformer and CNN	PyTorch, nnUNet	3D	Horizontal Flipping, Vertical flipping, scaling, normalization	5-fold	No	No	The Cross Attention Transformer (CAT) block design.
CeSNE-DiGAIR	3D UNet	MONAI	3D	Deformable (SyN) registrations between couples of neurotypicall and pathological scans from the preprocessed training dataset. Skull-stripping with BOUNTI	Not specified	No	Use of models for post-processing	Denoising autoencoder for segmentation accuracy enhancement.
falcons	2D Attention Gated U-Net	TensorFlow	2D	Rotation, width/height shift, vertical/horizontal flip, zooming, brightness, gaussian noise, gaussian blurring	Not specified	70 images from dHCP)	Models with different architecture and (or) training data	Series of preprocessing steps including brain extraction, alignment, and non-uniform intensity correction. Ensembling of models trained on different orientations (axial, sagital, coronal)
feta_sigma	UxLSTMEnc, UNet	nnUNet	3D	Rotation, Scaling, Translation, Gaussian Noise, Mirror Transform.	5-fold	No	Models with different architecture and (or) training data	Use of UxLSTM and ensembling with nnUnet, Background masking.
hilab	nnU-Net	PyTorch, nnUNet	3D	Default nnU-Net augmentations, histogram equalization, differentiated probabilities for sample selection in random copy-paste augmentations, replication of challenging cases.	5-fold	No	Models with different architecture and (or) training data	Applying histogram equalisation to 3D images, differentiated probabilities for sample selection in random copy-paste augmentations, strategically replicating challenging cases in the training data. Enesemble of 5 models with different hyperparameters and pre-processing settings.
jwcrad	Residual-USE-Net	PyTorch, MONAI	3D	Rotation, scaling, translation, intensity shift, low resolution simulation.	5-fold	No	Model trained on differnt CV splits	Custom auxiliary loss function based on transformation consistency.
LIT	Attention UNet, nnUNet ResidualEncoderUNet	PyTorch, nnUNet	3D	Rotation, Scaling, Gaussian Noise, Gaussian Blur, Brightness Alteration, Contrast Adjustment, Low Resolution Simulation, Gamma Adjustment, Mirroring	6-fold	No	Model trained on differnt CV splits	Custom brain mask detection with Attention Unet
lmrcmc	nnUNet, SegVol	nnUNet, MONAI	3D	nnUNet: default, SegVol: flip, ScaleIntensity, ShiftIntensity, GibbsNoise, BiasField, KSpaceSpikeNoise and Affine augmentation; with SLAug.	Not specified	No	Models with different architecture and (or) training data	Ensemble of U-Net and a foundation model, use of the SegVol model in fetal brain segmentation.
mic-dkfz	U-Net (nnUNet), U-Net with Residual encoder	nnUNet	3D	randomized; blur, gaussian noise, spatial (rotation, scaling, flipping), brightness, contrast, low-resolution simulation, gamma, sharpening, blank rectangle	5-fold	Yes (pre-training in MultiTalent)	Models with different architecture and (or) training data	Pretraining with MultiTalent on a collection of publicly available datasets. Ensemble of 3 nn-Unet configurations with different data augmentations .
paramahir_2023	3D UNet (segmentation), custom UNet-based (biometry)	MONAI	3D	Random Flipping, Random Rotation, Random Intensity Shifts	Not specified	No	No	Combination of segmentation and biometry prediction in a unified pipeline.
pasteurdbc	MedNeXt_L and nnUNet	nnUNet	3D	RandomScaling, RandomRoatation, RandomAdjustContrast, RandFlip	5-fold	Multi-modal multi-organ medical image datasets used in the pre-trained MedNeXt_L foundational model	Models with different architecture and (or) training data	Used additional datasets with CT and brain MRI images for model pre-training
qd_neuroincyte	Swin UNETR	MONAI	3D	Random sliding window, flipping, 1% gaussian noise, rigid rotation of ± 25° around all axes, random shifting ± 5 mm along all axes.	Not specified	No	Use of models for post-processing	Brain masking for vienna
unipd-sum-aug	2D Swin-UMamba	PyTorch, nnUNetv2, Monai	2D	TorchIO transforms and GIN techniques, pair-wise co-registration, affine and rigid transforms using the Advanced Nomalization Tools.	5-fold	Model pre-trained on ImageNet	Model trained on differnt CV splits	Pretrained on imageNet repository and using GIN.
UPFetal24	nnU-Net ResEncL	nnUNetv2	3D	Default nnU-Net augmentations; differentiated by specific data augmentations for each of the three models.	5-fold for config	dHCP and fetal atlasses	Models with different architecture and (or) training data	Data augmentation strategies and ensembling of models
ViCOROB	nnUNet	nnUNet, PyTorch	3D	Random bias field, motion artifacts, low-resolution simulation, SynthSeg-inspired T2w image synthesizer	3-fold	No	Model trained on differnt CV splits	SynthSeg-inspired T2w image synthesizer, Sharpness-Aware Minimization (SAM) optimizer

Table 3: Summary of the algorithms submitted for the biometry estimation task.

Team name	Architecture	Dimensionality	Original Aspects	External datasets	Framework/languange
qd_neuroincyte	SwinUnetr	3D	Relies on segmentation. Predict landmark heat maps only using the segmentation maps and then calculate biometry.	No additional data was used	Pytorch 2.2.2
CeSNE-DiGAIR	CNN	3D	Relies on segmentation. Predict the keypoints given the segmentation.	No additional data was used	PyTorch Version 2.4.0
jwcrad	Residual-USE-Net	3D	Relies on segmentation. Uses the segmentation maps to localize and preprocess the input images by masking and cropping the original 3D image. Predict landmark heat maps using the preprocessed images and then calculate biometry.	No additional data was used	PyTorch 2.2.2
pasteurdbc	MedNeXt_L nnUNet	3D	Use of a pre-trained foundational model.	Yes (for the pre-trained MedNeXt_L foundational model, multi-modal multi-organ medical image datasets)	
falcons	Attention Gated U-Net	2D	Relies on segmentation. Predict the biometry values directly	Yes (+70 images from dHCP)	Tensorflow(2.10.0) FMRIB Software Library(FSL 6.0), CIVET(2.1.0), Advanced Normalization Tools(ANTs), Scikit-learn (1.5.1)
feta_sigma	nnUNet, UxLSTMEnc	3D	Ensemble network of nnUnet and UxLSTMEnc.	No additional data was used	PyTorch
paramahir_2023	UNet	3D	Relies on segmentation. Predict the biometry values by regressing the U-Net features.	No additional data was used	PyTorch 2.3 -

3.2. FeTA 2024 results

3.2.1. Brain tissue segmentation ranking

Segmentation performance overview. Figure 4 highlights performance across sites and metrics, revealing a general **performance plateau** among top methods. For most teams, average Dice scores stabilized around **0.8-0.82**, HD95 around **2.8-2.1**, and VS around **0.9-0.92**, while the ED showed wider variability (ranging from **20** to **40**), highlighting its sensitivity to topological inaccuracies.

Site-specific trends. Despite being introduced in this edition as a new low-field, out-of-domain dataset, KCL showed the best segmentation performance. In contrast, KISPI yielded the lowest performance, even though it was part of the previous editions' training and testing data. Across metrics, UCSF and KISPI displayed higher interquartile ranges, particularly for Dice, HD95, and VS, reflecting greater variability across methods. Some teams (`falcons`, `qd_neuroincyte`) experienced performance drops on sites that use NiftyMIC SRR, like UCSF or VIEN, with Dice scores dropping to 0.38–0.44 compared to 0.76–0.83 on other sites.

Label-specific trends. SGM, GM, and BS were consistently the most challenging labels to segment across all teams, as shown by lower performance metrics in Supplementary Materials A4. Among the top three models, Dice scores dropped from an

average (across all labels) of 0.82 to 0.80 for SGM, 0.79 for BS, and 0.74 for GM. HD95 increased from 2.24 to 3.6 for BS and 3.0 for SGM, while VS declined from 0.92 to 0.86 for SGM and 0.88 for BS. GM also showed a marked increase in ED, from 33.14 to 137, reflecting a significant loss in topological accuracy.

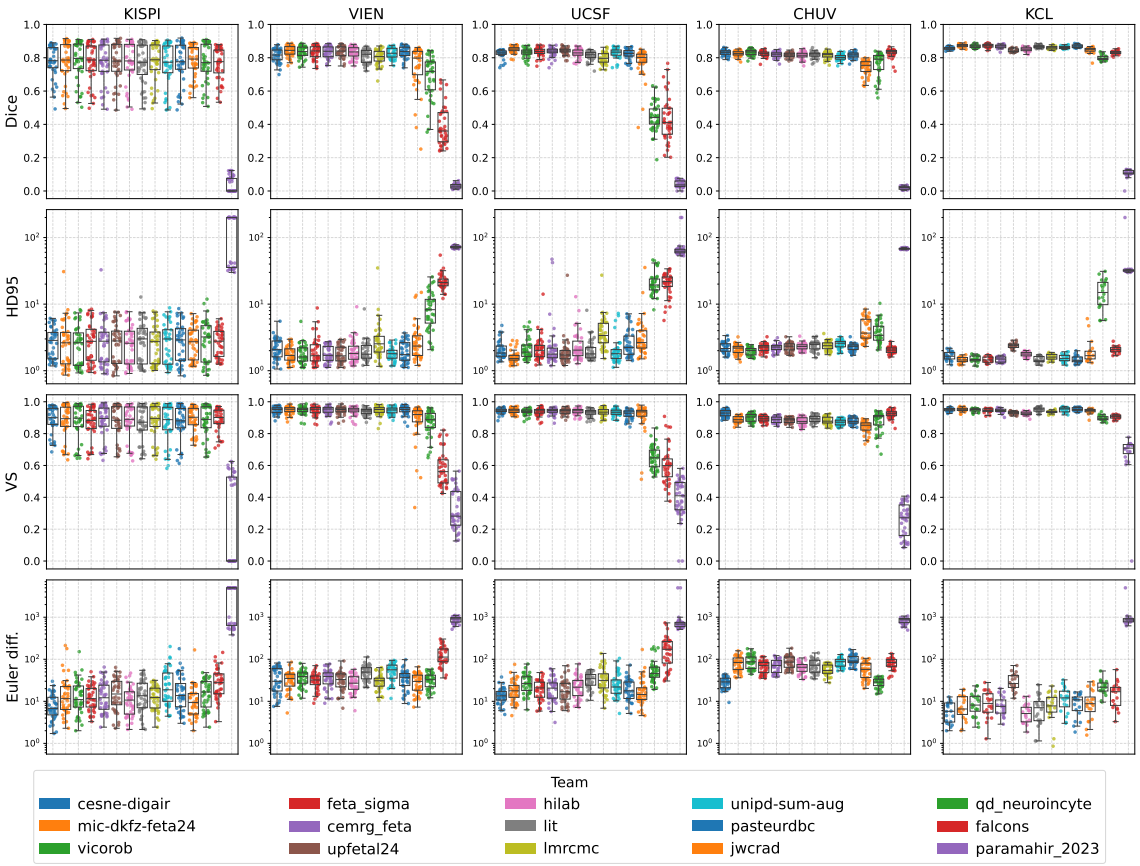


Figure 4: Segmentation performance by site and evaluation metric. In each subplot, teams are ranked from left to right based on their average performance across all labels for the given metric (best to worst). Team colors are consistent across plots and correspond to the legend.

Ranking summary. Table 4 presents the aggregated average metrics and rankings per team [11]. Qualitative examples of the segmentations are provided in Supplementary

[11]The rankings originally published on the website and announced during the MICCAI challenge differ from those presented in this paper due to a change in the ED estimation method. Specifically,

Appendix A9. Notable rank discrepancies across metrics highlight their complementary nature. Figure 5 provides a more granular view, showing single-metric rankings across different sites and tissue labels. Dice score rankings remained relatively consistent across submissions and anatomical regions, while ED rankings showed greater variability, both across tissues and sites, reinforcing the importance of using multiple metrics to capture distinct aspects of segmentation quality.

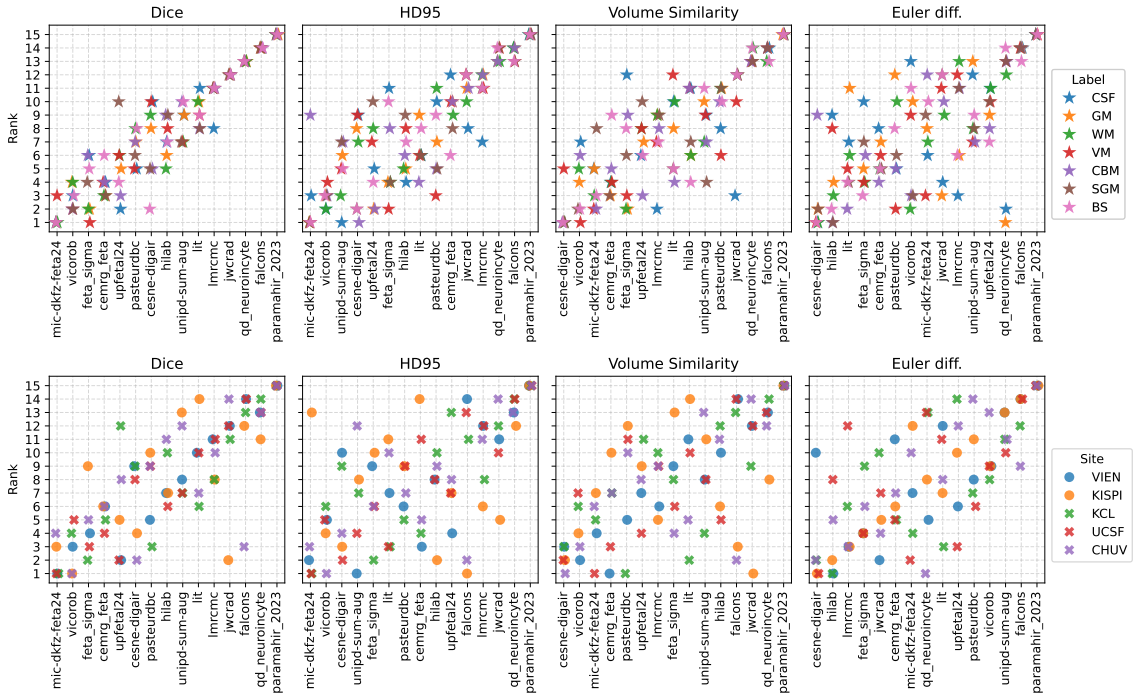


Figure 5: Detailed submissions rankings. Top: Across labels. Bottom: Across sites (circles indicated in-domain splits and crosses out-of-domain splits). Teams are sorted in each subplot by their overall ranking in the segmentation task, from the best to the worst.

Figure 6 further illustrates the added value of topological metrics. In a comparative example, `mic-dkfz-feta24` achieves similar Dice and lower HD95 scores but

we updated the way the ground-truth Euler characteristic is determined. In the original rankings, it was computed based on the manual segmentations. In the current results, it is calculated from manually defined topological properties (see 2.4). Manual segmentations were created via interpolation and because many structures were not segmented on every slice, the resulting ground-truth segmentations contained numerous topological errors (e.g., holes, disconnected components). As a result, they did not reliably represent the expected topological properties of the anatomical structures. To address this, we now use manually specified topological values to calculate the ground-truth EC.

poorer ED and VS, suggesting that voxel-level agreement alone may not suffice for tasks requiring topologically accurate surfaces, such as morphological analysis.

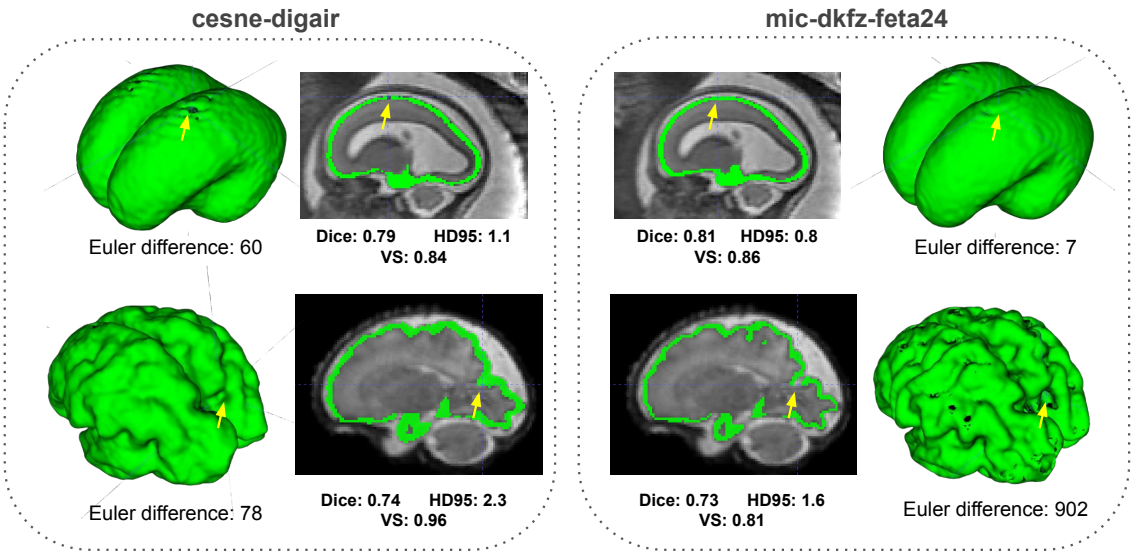


Figure 6: Segmentation results and reconstructed cortical GM surfaces for two representative fetal cases from `cesne-digair` and `mic-dkfz-feta24`, visualized using ITK-SNAP (Yushkevich et al., 2006). The top row (fetus at 22 weeks GA) illustrates that higher Dice scores sometimes correspond to smaller topological errors. However, the bottom row (fetus at 30 weeks GA) demonstrates significant topological issues (e.g., holes, fragmentation) in the `mic-dkfz-feta24` surface, despite comparable Dice and HD95 values. This underscores the need for additional topological and structural metrics, such as ED, to comprehensively evaluate segmentation quality, as metrics like Dice or HD95 alone are insufficient to capture topological accuracy.

Per-tissue and condition analysis. Extended performance results split by site, tissue label, and pathology status are available in supplementary materials (sections A3, A4, and A5, respectively).

3.2.2. Biometry ranking

Performance across sites and measurement. Figure 7 summarizes model performance per site and biometric measurement, with detailed values available in the supplementary materials section A6. VIEN was the most challenging site, where no method outperformed the **[GA]** baseline (MAPE: 0.106±0.112), including the best-performing teams `cesne-digair` and `jwcrad`, which reached similar error levels. In contrast, KISPI emerged as the least challenging, with all three top teams exceeding the baseline. Across KCL, UCSF, and CHUV, only two teams per site (out of the top

Table 4: Segmentation ranking and average metrics

Team	Dice		HD95		VS		ED		Mean rank	Final rank
	Rank	Value	Rank	Value	Rank	Value	Rank	Value		
cesne-digair	8	0.816	3	2.317	1	0.929	1	20.921	3.25	**1**
mic-dkfz-feta24	1	0.828	2	2.224	3	0.918	8	37.206	3.50	**2**
vicorob	2	0.825	1	2.187	2	0.920	11	41.293	4.00	**3**
feta_sigma	3	0.822	7	2.430	5	0.914	4	31.710	4.75	**4**
cemrg_feta	4	0.822	10	2.836	4	0.916	7	34.382	6.25	**5**
upfetal24	5	0.820	6	2.412	6	0.913	9	39.967	6.50	**6**
hilab	7	0.816	8	2.434	9	0.911	3	30.123	6.75	**7**
lit	10	0.808	5	2.391	8	0.911	10	40.085	8.25	**8**
lmrcmc	11	0.805	11	3.179	7	0.913	5	32.751	8.50	**9**
unipd-sum-aug	9	0.811	4	2.332	10	0.909	13	46.668	9.00	**10**
pasteurdbc	6	0.817	9	2.474	11	0.909	12	41.521	9.50	**11**
jwcrad	12	0.769	12	3.569	12	0.886	2	29.744	9.50	**11**
qd_neuroincyte	13	0.681	13	10.441	13	0.827	6	34.295	11.25	**13**
falcons	14	0.628	14	11.040	14	0.765	14	100.729	14.00	**14**
paramahir_2023	15	0.040	15	80.757	15	0.337	15	1416.515	15.00	**15**

3: jwcrad, feta_sigma, cesne_digair) achieved better-than-baseline performance. Measurement-wise, LCC, HV, and TCD were consistently more difficult, with HV and LCC showing the highest MAPE across all teams and raters. In contrast, sBIP and bBIP were among the best estimated. Notably, only jwcrad surpassed the baseline across all measurements, while a few others, including feta_sigma and pasteurdbc, did so on selected metrics.

Although multiple teams performed comparably on individual metrics, the clear winner in the ranking (see Table 5) was jwcrad, demonstrating consistent superiority across both site and measurement variations.

Robustness in pathological vs. neurotypical condition. To assess model generalizability, we compared biometry performance between neurotypical and pathological brains (see Figure 8). While most measurements did not reveal statistically significant differences between groups, bBIP showed better accuracy in the healthy cohort, particularly at VIEN. Conversely, UCSF results suggested slightly better performance for pathological subjects.

In summary, the best-performing method, jwcrad, came within 9% of expert agreement for some measurements (e.g., TCD). However, for others like bBIP, its results differed from the expert range by as much as 60%. This reveals important gaps where automated biometry methods still fall short, especially in pathological cases.

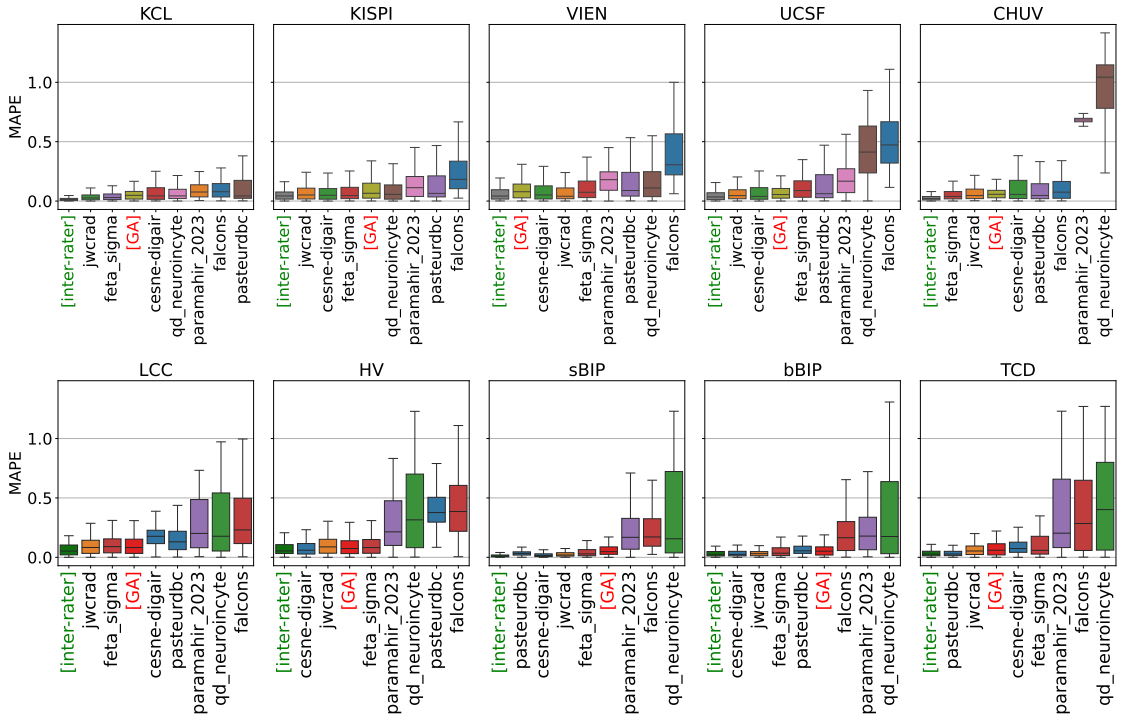


Figure 7: Biometry results per site (top) and label (bottom) for all teams participating in the Task 2 together with the GA baseline model ([GA]) and the inter-rater variability [inter-rater]. Teams are sorted in ascending order for each subplot independently, based on their mean MAPE for a given site or label.

3.3. Segmentation performance across challenge editions (2021, 2022 and 2024)

Over the three editions of the FeTA challenge, the segmentation task has expanded both in terms of dataset size (from 40 to 180 test cases) and site diversity (from 1 to 5 imaging centers). To evaluate progress over time, we compared segmentation performance across the years 2021, 2022, and 2024, focusing on common testing sites. Table 6 summarizes aggregated metrics for the top-performing teams each year (`cesne-digair` for 2024, `FIT_1` for 2022 and `NVAUTO` for 2021), and Figure 9 provides a visual overview of mean scores per label and site across years, with markers for statistically significant differences.

KISPI split (2021–2024). This is the only site included in all three editions. No statistically significant improvement over the years was observed across the tracked metrics: Dice ($0.79\pm0.16 \rightarrow 0.77\pm0.18 \rightarrow 0.78\pm0.15$), HD95 ($2.81\pm3.43 \rightarrow 3.17\pm4.16 \rightarrow 2.95\pm2.86$), and VS ($0.89\pm0.16 \rightarrow 0.87\pm0.18 \rightarrow 0.89\pm0.14$). The only statistically

Table 5: Metrics and ranking for the biometry estimation task sorted by the final MAPE. [GA] and [inter-rater] entries do not represent participating models, thus their rank is marked as *

Team	LCC		HV		bBIP		sBIP		TCD		Final MAPE	Final rank
	MAPE	Rank	MAPE	Rank	MAPE	Rank	MAPE	Rank	MAPE	Rank		
[inter-rater]	*9.59*	*	*8.04*	*	*3.28*	*	*1.49*	*	*4.89*	*	*5.38*	*
jwcrad	**11.15**	**1**	10.32	2	5.43	2	4.78	3	7.21	2	**7.72**	**1**
[GA]	*12.75*	*3*	*11.26*	*3*	*6.82*	*5*	*6.47*	*5*	*10.80*	*3*	*9.56*	*
cesne-digair	17.72	4	**9.82**	**1**	**4.02**	**1**	4.74	2	12.34	4	9.59	2
feta_sigma	12.59	2	11.55	4	5.74	3	5.54	4	13.66	5	9.76	3
pasteurdbc	20.47	5	43.48	7	6.51	4	**3.74**	**1**	**5.43**	**1**	15.83	4
paramahir_2023	28.48	6	29.35	5	26.13	7	25.46	6	30.78	6	28.03	5
falcons	34.88	8	46.25	8	24.62	6	28.13	7	36.72	7	34.09	6
qd_neuroincyte	32.78	7	42.84	6	38.41	8	37.83	8	47.92	8	40.07	7

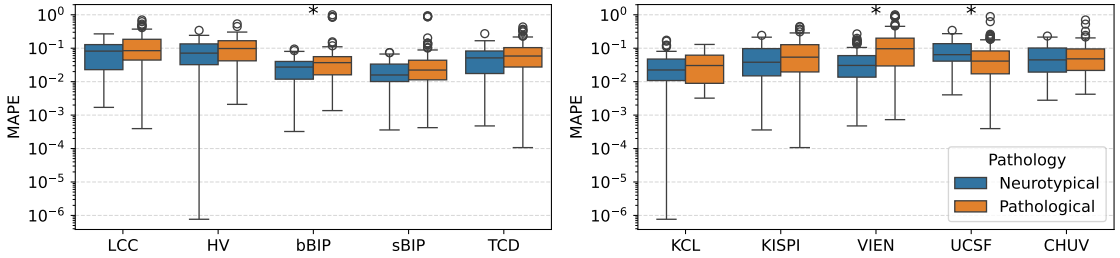


Figure 8: Biometry results for healthy and pathological subjects across labels and sites for the winning team `jwcrad`. Asterisks above the boxplot indicate statistically significant differences between the two groups ($p < 0.05$, Mann-Whitney test).

significant change occurred in the VS metric for the GM label between 2021 and 2022 ($0.96 \rightarrow 0.94$), but no consistent improvement was found in subsequent years or for other labels.

Other sites (2022–2024). For sites such as CHUV, UCSF, and VIEN—which were included in both 2022 and 2024—no consistent improvement was observed across metrics or tissue labels. While some metrics showed statistically significant changes, these were isolated and not consistent across sites, making it difficult to interpret them as evidence of overall progress in segmentation performance. Notably, ED improved substantially for CHUV and KCL. However, this improvement may be partially influenced by the inclusion of ED in the 2024 ranking, which favored algorithms with better topological performance. At CHUV, both VS and ED were significantly better than in 2022, but the other two metrics did not show similar trends.

Overall, although methods have become more sophisticated and the data more diverse, performance has not consistently improved across editions.

Table 6: Mean and standard deviation (mean±std) for different metrics across years and splits over all labels. Bold and * highlight the years that have statistically significant improvement in the values compared to the previous year. ED was not estimated in FeTA 2021

Year	Site	Dice	HD95	VS	ED
2021	KISPI	0.79±0.16	2.81±3.43	0.89±0.16	not estimated
2022	CHUV	0.81±0.09	2.33±1.68	0.88±0.10	77.43±168.91
	KCL	0.87±0.05	1.46±0.57	0.95±0.05	28.61±53.89
	KISPI	0.77±0.18	3.17±4.16	0.87±0.18	18.98±54.56
	UCSF	0.84±0.06	2.02±1.44	0.95±0.05	18.73±44.52
	VIEN	0.84±0.08	1.87±1.46	0.95±0.06	32.20±67.05
2024	CHUV	0.83±0.06	2.23±1.40	**0.93±0.06***	**29.10±51.56***
	KCL	0.86±0.05	1.69±0.52	0.95±0.04	**6.26±13.21***
	KISPI	0.78±0.15	2.95±2.86	0.89±0.14	9.21±17.84
	UCSF	0.82±0.07	2.13±1.31	0.94±0.05	14.57±25.90
	VIEN	0.81±0.09	2.27±1.69	0.95±0.05	38.13±93.48

3.4. Domain shifts evaluation

3.4.1. Impact of image quality on performance

The impact of image quality on model performance as determined by computing the conditional mean across quality ratings ($\mathbb{E}[f(x)|\text{Quality}]$) is shown in the rightmost column in Figure 10. We see a clear effect of image quality on Dice, with a generally increasing Dice with the increasing image quality, amounting to a change from 0.75 Dice on average for the lowest quality data (with scores close to 1) and an average quality close to 0.85 for the highest quality data. Results using HD95 and VS generally align with the ones from Dice, except for GA and quality. The relationship is, however, not as clear for ED, although best quality images tend to yield the smallest ED.

A more detailed analysis of the correlation between quality and the difference scores in the supplementary materials A7 showed a generally high Pearson correlation between quality and Dice (r = 0.5-0.7) for all sites except KISPI-mial (r=0.4) and USCF-nmic (r=0.06 – no correlation). The same trends, although weaker, were observed for HD95 and VS, except CHUV-mial and HD95, which had virtually no correlation (r=0.05). Results for ED showed no clear pattern, and larger correlations (r=0.3-0.4) were not statistically significant.

3.4.2. Relative contribution of domain-shift sources

Figure 10 displays the conditional means across different factors. The analysis revealed a pronounced site-SR effect: for example, the KISPI-mial site produced notably lower Dice scores, whereas the CHUV-mial site was associated with higher ED values. In addition, gestational age (GA) significantly affected both Dice and

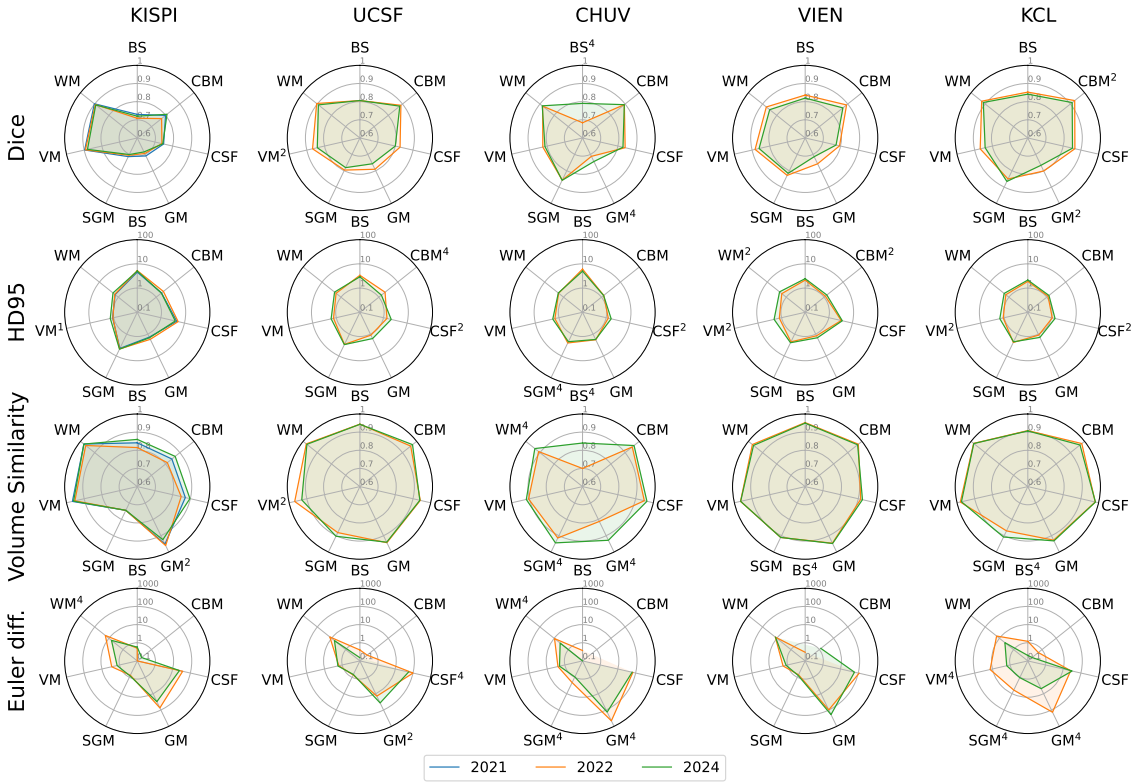


Figure 9: Segmentation performance improvement over the three editions of the FeTA Challenge. The superscript number above each label indicates whether the performance for a particular year and site-label-metric combination was statistically significantly better compared to all other available years. A superscript "2" indicates that the results for 2022 were the best, while a superscript "4" indicates that the results for 2024 were the best.

ED scores. Similar trends can be found in the supplementary materials (appendix A7), although HD95 scores appear to be less influenced by GA.

Figure 11 presents a SHAP analysis for all metrics using only image-level descriptors—namely, image quality, subject condition, and gestational age. (We excluded Site-SR from this analysis because its dependence on the other variables could lead to misleading SHAP values under the assumption of feature independence (Mase et al., 2019).) Overall, the SHAP analysis summarizes how these factors influence the Dice and ED scores: image quality generally has the largest impact, followed by GA (except for HD95). Although the pathological status of a subject generally has a small effect, we observed that severely pathological cases often have lower GA, which might introduce confounding. The plot's color coding further confirms that higher

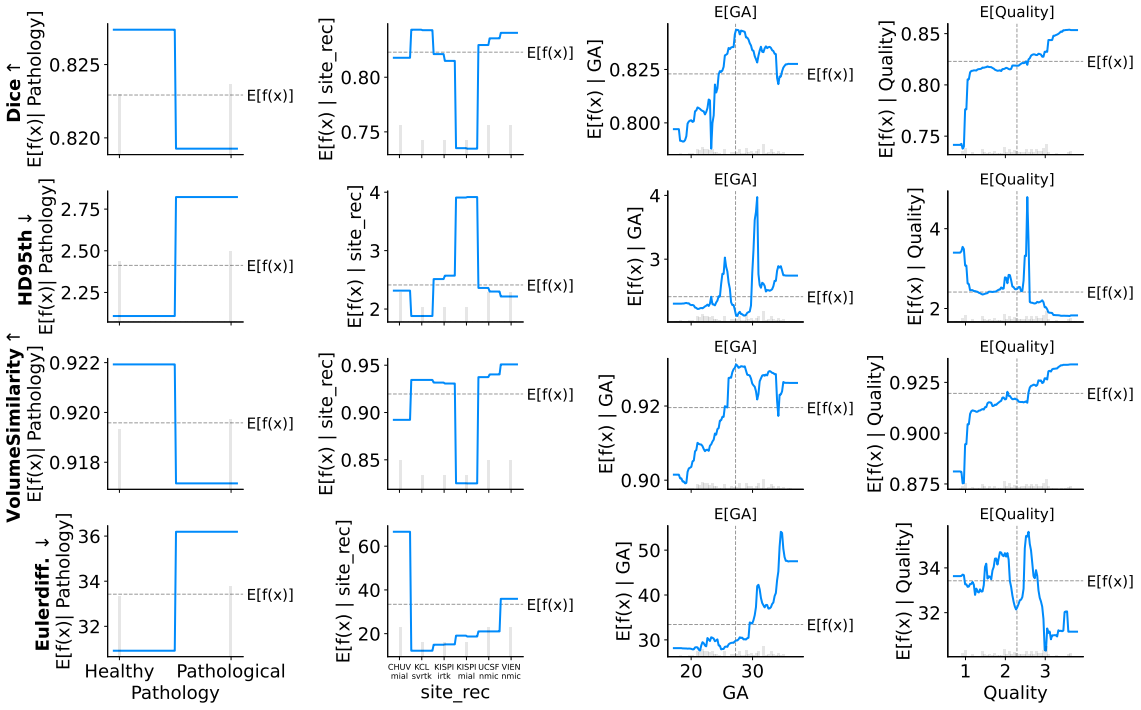


Figure 10: **Conditional mean** plots for Dice, HD95, VS and ED, for different domain shift factors: pathology, site and SR, GA and image quality. The conditional mean shows how a given metric deviates from the global expected performance ($\mathbb{E}[f(x)]$) when a specific variable is used for conditioning.

image quality and GA are associated with increased Dice scores—for example, poor quality data may result in about –0.1 Dice, compared to an average of +0.05 Dice for good quality data.

4. Discussion

4.1. FeTA 2024 results and ranking

The multi-site, multi-task design of this challenge offered a unique opportunity to evaluate the progress and robustness of fetal brain image analysis algorithms. We summarize the main observations below.

Submitted methods overview. Analyzing solutions of the best 3 teams for the segmentation task, we see that all of them used 3D models, specifically U-Nets or nnU-Nets. External data did not play a major role, with two of the top three teams relying solely on the provided dataset. Notably, the first-place team, `cesne-digair`,

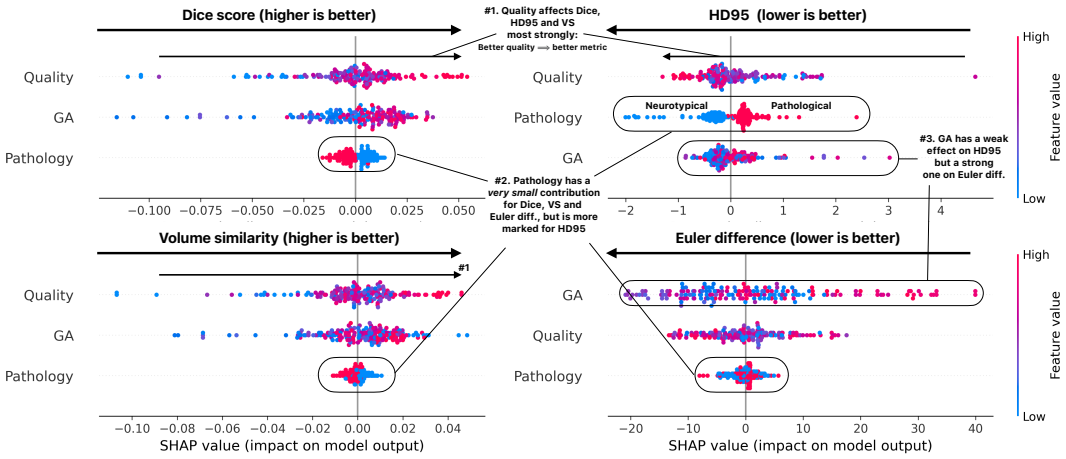


Figure 11: **SHAP value distribution** across the data for segmentation metrics. Compared to conditional means (Fig. 10), SHAP values are the attribution of the impact on Dice or ED of different factors. Blue dots correspond to lower values of a variable (i.e. low GA, low quality), and red ones to higher values.

uniquely incorporated a denoising autoencoder (Larrazabal et al., 2020) to enhance segmentation accuracy. This approach significantly boosted their performance, leading to a 50% improvement in the ED metric compared to the second-best team. This gain was particularly important for topological correctness, as no other metric showed such a large performance gap. All top-three teams applied extensive data augmentation, including techniques like SynthSeg (Billot et al., 2023), combinations of standard augmentations, and deformable registration between pathological and healthy subjects to simulate greater anatomical diversity, as well as model ensembling.

The top three biometry teams each approached estimation differently—using keypoint regression, landmark heatmap segmentation, and other methods—but all relied on the results of the tissues segmentation and 3D models. Their model architectures varied (CNN, U-Net, and Transformer), indicating that no standard baseline has yet emerged for this task. The diversity in methods suggests that the field is still exploring the most effective strategies for biometry estimation.

Topology metric is a valuable add-on. In the brain tissue segmentation task, the introduction of the topology-aware metric provided meaningful complementary insights beyond traditional overlap-based measures. Despite the architectural diversity

and growing methodological sophistication of the submitted approaches, the performance differences among the top teams were minimal, with Dice scores showing tight clustering, suggesting that gains in segmentation accuracy may be reaching a plateau. Differences in team rankings across evaluation metrics (Dice, HD95, VS, ED) highlight the need to consider complementary metrics beyond voxel-wise overlap. Introducing ED as a ranking metric provided a more nuanced assessment of the segmentation quality. This is reflected in Table 6, where we see a marked improvement in ED. While teams did not specifically optimize their models for topological consistency, the new ranking scheme allowed us to discriminate between methods that otherwise had very similar performances (Table 4). Evaluation noise, where performance variations across testing sets is larger than the difference across top-performing methods, is a well-known problem in medical imaging challenges (Varoquaux and Cheplygina, 2022), and the introduction of an additional ranking metric allowed for selection methods with desirable properties. Further validation on clinical tasks leveraging surface extraction(Clouchoux et al., 2011; Yehuda et al., 2023) would be needed to truly see the potential of encouraging topological consistency in the FeTA challenge.

Low-field MRI tissue segmentation quality is encouraging. The newly introduced 0.55T data from KCL provided an unexpected insight: it consistently achieved the highest segmentation accuracy across all sites. However, it is important to note that, to reduce domain shifts, we retrospectively selected high-quality reconstructions using a version of SVRTK specifically tailored for low-field MRI data(Uus et al., 2020, 2024). This careful case selection, paired with a very recent SR pipeline, might have positively biased the performance for this cohort. As such, performance on more challenging low-field cases remains to be fully assessed. Nevertheless, these results are encouraging for two reasons: progress in SR pipelines (Uus et al., 2024; Xu et al., 2023) means that more challenging cases will be successfully reconstructed, and that image quality will generally increase (Sanchez et al., 2024a; Uus et al., 2023b). Low-field MRI systems hold significant promise for expanding access to prenatal imaging, particularly in low- and middle-income countries. When combined with good image quality and advances in automatic fetal exam planning (Neves Silva et al., 2024), they could meaningfully enhance prenatal care in resource-limited settings.

Automated biometry needs strong baselines to ensure meaningful progress. To reflect current clinical practice and bridge the gap between routine 2D fetal brain assessments and emerging 3D imaging techniques, we introduced a new biometry task focused on 2D brain measurements—key clinical indicators traditionally used to assess fetal neurodevelopmental status (Tilea et al., 2009; Lamon et al., 2024). One of

the striking results of this first edition is that most submissions did not manage to outperform a simple model, that predicts biometry solely based on gestational age, completely ignoring image information. Most teams built biometry estimators atop segmentation outputs, potentially propagating segmentation errors, particularly in smaller and more complex structures such as vermis and corpus callosum measurements. Importantly, the biometric measurements used in this challenge were derived from clinical 2D protocols. While these can be estimated from 3D volumes, they were originally designed for manual 2D evaluation. This suggests that alternative measurements, specifically tailored to leverage the full spatial context of 3D SRR images, may offer even more informative and robust indicators of fetal neurodevelopment, though such approaches remain largely unexplored. Nevertheless, this first competition confirms, once more, the need to have strong baseline models and validation procedures (Eisenmann et al., 2023), and that deep learning might not always be the optimal solution (Grinsztajn et al., 2022).

4.2. FeTA challenge in perspective

FeTA across the years. A retrospective analysis of FeTA challenge results over the years revealed no statistically significant improvements in performance metrics, with the exception of ED at two out of five sites. Similarly, no notable improvement was seen at the label level, with GM, SGM, and BS consistently remaining the most challenging structures to segment. GM is particularly difficult due to its very thin appearance in fetal brains, where partial volume effects and complex surface morphology make it especially prone to topological segmentation errors, leading to significantly higher ED values compared to other labels. Moreover, both GM and SGM have inherently low tissue contrast in MRI, making them harder to distinguish accurately (Prayer et al., 2006). These challenges are further illustrated in Supplementary Materials A9.1, which provide qualitative examples showing that most segmentation errors occur in regions corresponding to GM, SGM, and BS. This outcome is not entirely unexpected, as most top-performing teams relied on similar 3D architectures—primarily 3D U-Net (Çiçek et al., 2016) and nnU-Net (Isensee et al., 2018)—enhanced with extensive data augmentation and model ensembling. These findings suggest that incremental architectural modifications or model engineering alone are unlikely to yield substantial gains, aligning with trends observed in other challenges where U-Net-based approaches often outperform more complex alternatives (Eisenmann et al., 2023). While these techniques help mitigate certain domain shifts related to scanner differences or pathological variations, some cases remain persistently difficult across all methods. Addressing these harder cases may require deeper domain expertise and a shift toward a more data-centric approach, prioritiz-

ing data quality, annotation consistency, and dataset diversity as core components of model development (Sambasivan et al., 2021; Zha et al., 2023).

Sources of domain shifts. Domain shifts are widely recognized as a key challenge for deep learning methods in medical imaging (Dockès et al., 2021; Wiles et al., 2021; Richiardi et al., 2025), yet the specific sources of these shifts are rarely disentangled. In our analysis, though not causal, we observed that image quality had the strongest impact on generalization performance: moving from the lowest to the highest quality levels resulted in an average Dice score difference of approximately 0.10. In contrast, gestational age had a more modest effect, influencing Dice scores by about 0.05, while the scanning site contributed a difference of around 0.075 between the best- and worst-performing centers. Interestingly, pathology was the least influential factor, accounting for only about 0.008 in Dice variation. Additionally, because Dice is known to be biased toward larger structures (Maier-Hein et al., 2024), we also evaluated performance using a normalized Dice metric that accounts for label volume (Raina et al., 2023). As detailed in the supplementary materials section A8, the normalized Dice scores yielded rankings nearly identical to those based on standard Dice, indicating that structure size did not significantly distort the comparative performance of the reviewed algorithms, suggesting that while the size bias exists, its effect was uniform across methods.

Our results show that, despite pathological cases making up only about one-third of the training data, models were still able to generalize to pathological examples in the test set. While performance was slightly lower for pathological subjects in some datasets compared to healthy subjects, submitted models demonstrated the ability to correctly handle both healthy and pathological data. Given the rarity and wide variability of fetal pathologies (Attallah et al., 2019), expanding pathological datasets—whether through additional real cases or synthetic data (Dannecker et al., 2024; Kaandorp et al., 2025b)—will be crucial to narrowing this performance gap and improving overall model robustness, which is an important step toward real-world clinical deployment.

Overall, our findings suggest that technical and acquisition-related factors may play a more significant role in out-of-domain generalization than subject-level clinical variables. Still, further causal investigations (Castro et al., 2020) are needed to confirm these patterns and to avoid misinterpretation due to confounding factors.

4.3. Roadmap for future advancements in fetal brain MRI analysis

While many proposed solutions appear to be reaching a performance plateau, model-centric innovations still play an important role. That said, incorporating domain-specific augmentations and auxiliary learning objectives may lead to more

impactful improvements than simply refining model architectures. For example, enforcing *topological consistency* within the loss function, as demonstrated by de Dumast et al. (2022); Li et al. (2023); Lux et al. (2024)—can help maintain anatomical plausibility in the predictions. Similarly, integrating *uncertainty estimation* provides a powerful way of identifying low-confidence predictions, which is particularly relevant in clinical decision-support systems. Several studies (Zenk et al., 2025; Molchanova et al., 2025) have demonstrated the utility of uncertainty-aware models for quality control in medical image segmentation.

Beyond model architecture, data-driven strategies hold substantial potential for improvement. A notable limitation of current solutions is the relatively modest use of *external data*, which has been largely limited to *healthy* subjects from datasets like dHCP or fetal brain atlases (Gholipour et al., 2017; Uus et al., 2023a; Price et al., 2019). Leveraging broader, more diverse datasets — especially those capturing rare or pathological conditions — could support more robust and clinically useful models, though curating and annotating pathological datasets is a huge endeavor.

Manually segmenting the fetal brain is a time-consuming and tedious task, susceptible to inter-rater variability (Payette et al., 2021), and the FeTA challenge data are not exempt from this issue. When comparing model performance to inter-rater variability, we observe that top-performing teams—achieving Dice scores around 0.82, HD95 around 2.2, and VS around 0.92—are approaching the best observed human agreement levels, previously estimated on a subset of data as 0.73 ± 0.15 for Dice, 3.45 ± 2.34 for HD95, and 0.86 ± 0.10 for VS (Payette et al., 2023). This raises the intriguing possibility that some predictions may be more faithful to the underlying anatomy than the ground truth annotations, potentially leading to penalization of high-performing models (Valabregue et al., 2024a,b).

A promising direction to address the limitations of data diversity and annotation availability is data synthesis (Zalevskyi et al., 2024), particularly for generating rare or pathological fetal brain appearances. Recent work (Dannecker et al., 2024; Liu et al., 2024; Kaandorp et al., 2025a) has highlighted the potential of synthetic data to augment training and improve sensitivity to abnormal anatomy. Moreover, the strong influence of image quality on generalization performance underscores the need for better modeling of artifacts specific to fetal brain SR pipelines (Sanchez et al., 2024a). These efforts, in combination with foundation models and domain adaptation techniques, offer exciting prospects for enhancing model generalization across scanners, domains, and populations, ultimately helping to mitigate model drift and support the development of more trustworthy AI systems.

5. Conclusion

The FeTA 2024 challenge provided a valuable opportunity to evaluate the progress made in fetal brain segmentation since previous editions and to expand the scope toward new, clinically relevant tasks such as biometry. Our additional validation using the Euler difference metric showed that some existing methods can already produce topologically consistent segmentations. However, achieving this consistency more reliably, particularly through improved segmentation losses, remains an open area for further development. Likewise, the successful application of models to low-field data, with surprisingly strong performance, highlights both the advancements in recent super-resolution methods and the models' capacity to generalize across diverse imaging settings.

In the biometry task, this first edition offered key insights, particularly on the importance of providing simple baseline models to guide participants. It also led to the emergence of a promising approach for automated biometry prediction.

As the field of fetal brain MRI analysis continues to evolve, FeTA 2024 emphasizes the need not only for more powerful and innovative models but also for building reliable and generalizable tools that can support real-world clinical adoption.

Declaration of competing interests

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

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

Acknowledgments

This research was funded by the Swiss National Science Foundation (182602, 215641, IZKSZ3_218590), ERA-NET Neuron MULTI-FACT project (SNSF 31NE30 203977), UKRI FLF (MR/T018119/1), the Adaptive Brain Circuits in Development and Learning Project, University Research Priority Program of the University of Zürich; by the Vontobel Foundation; by the Anna Müller Grocholski Foundation and the Prof. Max Cloetta Foundation and DFG Heisenberg funding (502024488); we acknowledge the Leenaards and Jeantet Foundations as well as CIBM Center for

Biomedical Imaging, a Swiss research center of excellence founded and supported by CHUV, UNIL, EPFL, UNIGE and HUG.

Diego Fajardo-Rojas would like to acknowledge funding from the EPSRC Centre for Doctoral Training in Smart Medical Imaging (EP/S022104/1).

Lyuyang Tong, Bo Du and Jingwen Jiang would like to acknowledge funding from the National Key Research and Development Program of China under Grants 2023YFC2705700, the National Natural Science Foundation of China under Grants 62306217 and 62225113, the Postdoctoral Fellowship Program of CPSF under Grant Number GZC20231987, the China Postdoctoral Science Foundation under Grant Number 2024T170686 and 2024M752471, the Major Program (JD) of Hubei Province (2023BAA017), the Innovative Research Group Project of Hubei Province under Grants 2024AFA017.

Gerard Martí-Juan is supported by the project PCI2021-122044-2A, funded by the project ERA-NET NEURON Cofund2, by MCIN/AEI/10.13039/501100011033/ and by the European Union NextGenerationEU/PRTR.

R. Hamadache holds an IFUdG PhD grant from the University of Girona. R. Hamadache and X. Lladó received support by the PID2023-146187OB-I00 from the Ministerio de Ciencia e Innovación, Spain.

M.O. Candela-Leal, A. Gondova, and S. You received support from the National Institute of Neurological Disorders and Stroke (R01NS114087) and National Institute of Biomedical Imaging and Bioengineering (R01EB031170) of the National Institutes of Health (NIH).

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used Grammarly and ChatGPT (GPT 4o) to assist with spell checking, grammar refinement, and language clarity improvements. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

Alamo, L., Laswad, T., Schnyder, P., Meuli, R., Vial, Y., Osterheld, M.C., Gudinchet, F., 2010. Fetal mri as complement to us in the diagnosis and characterization of anomalies of the genito-urinary tract. European journal of radiology 76, 258–264.

Arnold, T.C., Freeman, C.W., Litt, B., Stein, J.M., 2023. Low-field mri: clinical promise and challenges. Journal of Magnetic Resonance Imaging 57, 25–44.

Attallah, O., Sharkas, M.A., Gadelkarim, H., 2019. Fetal brain abnormality classification from mri images of different gestational age. Brain sciences 9, 231.

Aviles Verdera, J., Story, L., Hall, M., Finck, T., Egloff, A., Seed, P.T., Malik, S.J., Rutherford, M.A., Hajnal, J.V., Tomi-Tricot, R., Hutter, J., 2023. Reliability and feasibility of low-field-strength fetal mri at 0.55 t during pregnancy. Radiology 309. URL: http://dx.doi.org/10.1148/radiol.223050, doi:10.1148/radiol.223050.

Avisdris, N., Yehuda, B., Ben-Zvi, O., Link-Sourani, D., Ben-Sira, L., Miller, E., Zharkov, E., Ben Bashat, D., Joskowicz, L., 2021. Automatic linear measurements of the fetal brain on mri with deep neural networks. International Journal of Computer Assisted Radiology and Surgery 16, 1481–1492. URL: http://dx.doi.org/10.1007/s11548-021-02436-8, doi:10.1007/s11548-021-02436-8.

Van den Bergh, B.R.H., Dahnke, R., Mennes, M., 2018. Prenatal stress and the developing brain: Risks for neurodevelopmental disorders. Development and Psychopathology 30, 743–762. URL: http://dx.doi.org/10.1017/S0954579418000342, doi:10.1017/s0954579418000342.

Billot, B., Greve, D.N., Puonti, O., Thielscher, A., Van Leemput, K., Fischl, B., Dalca, A.V., Iglesias, J.E., et al., 2023. Synthseg: Segmentation of brain mri scans of any contrast and resolution without retraining. Medical image analysis 86, 102789.

Castro, D.C., Walker, I., Glocker, B., 2020. Causality matters in medical imaging. Nature Communications 11, 3673.

Chen, X., Xu, D., Gu, X., Li, Z., Zhang, Y., Wu, P., Huang, Z., Zhang, J., Li, Y., 2024. Machine learning in prenatal mri predicts postnatal ventricular abnormalities in fetuses with isolated ventriculomegaly. European Radiology , 1–10.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O., 2016. 3d U-Net: learning dense volumetric segmentation from sparse annotation, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19, Springer. pp. 424–432.

Ciceri, T., Casartelli, L., Montano, F., Conte, S., Squarcina, L., Bertoldo, A., Agarwal, N., Brambilla, P., Peruzzo, D., 2024. Fetal brain mri atlases and datasets: a review. NeuroImage , 120603.

Ciceri, T., Squarcina, L., Pigoni, A., Ferro, A., Montano, F., Bertoldo, A., Persico, N., Boito, S., Triulzi, F.M., Conte, G., Brambilla, P., Peruzzo, D., 2023. Geometric reliability of super-resolution reconstructed images from clinical fetal mri in the second trimester. Neuroinformatics 21, 549–563. URL: http://dx.doi.org/10.1007/s12021-023-09635-5, doi:10.1007/s12021-023-09635-5.

Clouchoux, C., Kudelski, D., Gholipour, A., Warfield, S.K., Viseur, S., Bouyssi-Kobar, M., Mari, J.L., Evans, A.C., du Plessis, A.J., Limperopoulos, C., 2011. Quantitative in vivo mri measurement of cortical development in the fetus. Brain Structure and Function 217, 127–139. URL: http://dx.doi.org/10.1007/s00429-011-0325-x, doi:10.1007/s00429-011-0325-x.

Dannecker, M., Kyriakopoulou, V., Cordero-Grande, L., Price, A.N., Hajnal, J.V., Rueckert, D., 2024. Cina: Conditional implicit neural atlas for spatio-temporal representation of fetal brains, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 181–191.

Dockès, J., Varoquaux, G., Poline, J.B., 2021. Preventing dataset shift from breaking machine-learning biomarkers. GigaScience 10, giab055. doi:10.1093/gigascience/giab055.

Dovjak, G.O., Diogo, M.C., Brugger, P.C., Gruber, G.M., Weber, M., Glatter, S., Seidl, R., Bettelheim, D., Prayer, D., Kasprian, G.J., 2020. Quantitative fetal magnetic resonance imaging assessment of cystic posterior fossa malformations. Ultrasound in Obstetrics and Gynecology 56, 78–85. URL: http://dx.doi.org/10.1002/uog.21890, doi:10.1002/uog.21890.

de Dumast, P., Kebiri, H., Dunet, V., Koob, M., Cuadra, M.B., 2022. Multi-dimensional topological loss for cortical plate segmentation in fetal brain mri. arXiv preprint arXiv:2208.07566 .

Edwards, A.D., Rueckert, D., Smith, S.M., Abo Seada, S., Alansary, A., Almalbis, J., Allsop, J., Andersson, J., Arichi, T., Arulkumaran, S., Bastiani, M., Batalle, D., Baxter, L., Bozek, J., Braithwaite, E., Brandon, J., Carney, O., Chew, A., Christiaens, D., Chung, R., Colford, K., Cordero-Grande, L., Counsell, S.J., Cullen, H., Cupitt, J., Curtis, C., Davidson, A., Deprez, M., Dillon, L., Dimitrakopoulou, K., Dimitrova, R., Duff, E., Falconer, S., Farahibozorg, S.R.,

Fitzgibbon, S.P., Gao, J., Gaspar, A., Harper, N., Harrison, S.J., Hughes, E.J., Hutter, J., Jenkinson, M., Jbabdi, S., Jones, E., Karolis, V., Kyriakopoulou, V., Lenz, G., Makropoulos, A., Malik, S., Mason, L., Mortari, F., Nosarti, C., Nunes, R.G., O'Keeffe, C., O'Muircheartaigh, J., Patel, H., Passerat-Palmbach, J., Pietsch, M., Price, A.N., Robinson, E.C., Rutherford, M.A., Schuh, A., Sotiropoulos, S., Steinweg, J., Teixeira, R.P.A.G., Tenev, T., Tournier, J.D., Tusor, N., Uus, A., Vecchiato, K., Williams, L.Z.J., Wright, R., Wurie, J., Hajnal, J.V., 2022. The developing human connectome project neonatal data release. Frontiers in Neuroscience 16. URL: http://dx.doi.org/10.3389/fnins.2022.886772, doi:10.3389/fnins.2022.886772.

Egaña-Ugrinovic, G., Savchev, S., Bazán-Arcos, C., Puerto, B., Gratacós, E., Sanz-Cortés, M., 2015. Neurosonographic assessment of the corpus callosum as imaging biomarker of abnormal neurodevelopment in late-onset fetal growth restriction. Fetal Diagnosis and Therapy 37, 281–288. URL: http://dx.doi.org/10.1159/000366160, doi:10.1159/000366160.

Eisenmann, M., Reinke, A., Weru, V., Tizabi, M.D., Isensee, F., Adler, T.J., Ali, S., Andrearczyk, V., Aubreville, M., Baid, U., et al., 2023. Why is the winner the best?, in: Proceedings of the IEEE/CVF conference on computer vision and Pattern recognition, pp. 19955–19966.

Gafner, M., Fried, S., Gosher, N., Jeddah, D., Sade, E.K., Barzilay, E., Mayer, A., Katorza, E., 2020. Fetal brain biometry: Is there an agreement among ultrasound, mri and the measurements at birth? European Journal of Radiology 133, 109369. URL: https://www.sciencedirect.com/science/article/pii/S0720048X20305593, doi:https://doi.org/10.1016/j.ejrad.2020.109369.

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

Griffiths, P.D., Bradburn, M., Campbell, M.J., Cooper, C.L., Graham, R., Jarvis, D., Kilby, M.D., Mason, G., Mooney, C., Robson, S.C., et al., 2017. Use of mri in the diagnosis of fetal brain abnormalities in utero (meridian): a multicentre, prospective cohort study. The Lancet 389, 538–546.

Grinsztajn, L., Oyallon, E., Varoquaux, G., 2022. Why do tree-based models still outperform deep learning on typical tabular data? Advances in neural information processing systems 35, 507–520.

Hall, M., de Marvao, A., Schweitzer, R., Cromb, D., Colford, K., Jandu, P., O'Regan, D.P., Ho, A., Price, A., Chappell, L.C., Rutherford, M.A., Story, L., Lamata, P., Hutter, J., 2024. Preeclampsia associated differences in the placenta, fetal brain, and maternal heart can be demonstrated antenatally: An observational cohort study using mri. Hypertension 81, 836–847. URL: http://dx.doi.org/10.1161/HYPERTENSIONAHA.123.22442, doi:10.1161/hypertensionaha.123.22442.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Identity mappings in deep residual networks, in: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14, Springer. pp. 630–645.

Hughes, E.J., Winchman, T., Padormo, F., Teixeira, R., Wurie, J., Sharma, M., Fox, M., Hutter, J., Cordero-Grande, L., Price, A.N., Allsop, J., Bueno-Conde, J., Tusor, N., Arichi, T., Edwards, A.D., Rutherford, M.A., Counsell, S.J., Hajnal, J.V., 2016. A dedicated neonatal brain imaging system: A dedicated neonatal brain imaging system. Magnetic Resonance in Medicine 78, 794–804. URL: http://dx.doi.org/10.1002/mrm.26462, doi:10.1002/mrm.26462.

Isensee, F., Petersen, J., Klein, A., Zimmerer, D., Jaeger, P.F., Kohl, S., Wasserthal, J., Koehler, G., Norajitra, T., Wirkert, S., et al., 2018. nnu-net: Self-adapting framework for u-net-based medical image segmentation. arXiv preprint arXiv:1809.10486 .

Kaandorp, M.P., Agbelese, D., Asma-ull, H., Kim, H.G., Payette, K., Grehten, P., Giulio, G.A., Lánczi, L.I., Jakab, A., 2025a. Pathological mri segmentation by synthetic pathological data generation in fetuses and neonates. arXiv preprint arXiv:2501.19338 .

Kaandorp, M.P.T., Agbelese, D., Asma-ull, H., Kim, H.G., Payette, K., Grehten, P., Giulio, G.A., Lánczi, L.I., Jakab, A., 2025b. Pathological mri segmentation by synthetic pathological data generation in fetuses and neonates. URL: https://arxiv.org/abs/2501.19338, doi:10.48550/ARXIV.2501.19338.

Khawam, M., de Dumast, P., Deman, P., Kebiri, H., Yu, T., Tourbier, S., Lajous, H., Hagmann, P., Maeder, P., Thiran, J.P., Meuli, R., Dunet, V., Bach Cuadra, M.,

Koob, M., 2021. Fetal brain biometric measurements on 3d super-resolution reconstructed t2-weighted mri: An intra- and inter-observer agreement study. Frontiers in Pediatrics 9. URL: http://dx.doi.org/10.3389/fped.2021.639746, doi:10.3389/fped.2021.639746.

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., 2016. Normative biometry of the fetal brain using magnetic resonance imaging. Brain Structure and Function 222, 2295–2307. URL: http://dx.doi.org/10.1007/s00429-016-1342-6, doi:10.1007/s00429-016-1342-6.

Lamon, S., de Dumast, P., Sanchez, T., Dunet, V., Pomar, L., Vial, Y., Koob, M., Bach Cuadra, M., 2024. Assessment of fetal corpus callosum biometry by 3d super-resolution reconstructed t2-weighted magnetic resonance imaging. Frontiers in Neurology 15. URL: http://dx.doi.org/10.3389/fneur.2024.1358741, doi:10.3389/fneur.2024.1358741.

Larrazabal, A.J., Martínez, C., Glocker, B., Ferrante, E., 2020. Post-dae: anatomically plausible segmentation via post-processing with denoising autoencoders. IEEE transactions on medical imaging 39, 3813–3820.

Li, L., Ma, Q., Ouyang, C., Li, Z., Meng, Q., Zhang, W., Qiao, M., Kyriakopoulou, V., Hajnal, J.V., Rueckert, D., et al., 2023. Robust segmentation via topology violation detection and feature synthesis, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 67–77.

Liu, P., Puonti, O., Sorby-Adams, A., Kimberly, W.T., Iglesias, J.E., 2024. Pepsi: Pathology-enhanced pulse-sequence-invariant representations for brain mri, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 676–686.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF international conference on computer vision, pp. 10012–10022.

Lundberg, S.M., Lee, S.I., 2017. A unified approach to interpreting model predictions, in: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 30.

Curran Associates, Inc., pp. 4765–4774. URL: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

Lux, L., Berger, A.H., Weers, A., Stucki, N., Rueckert, D., Bauer, U., Paetzold, J.C., 2024. Topograph: An efficient graph-based framework for strictly topology preserving image segmentation. arXiv preprint arXiv:2411.03228 .

Mahalingam, H.V., Rangasami, R., Seshadri, S., Suresh, I., 2021. Imaging spectrum of posterior fossa anomalies on foetal magnetic resonance imaging with an algorithmic approach to diagnosis. Polish Journal of Radiology 86, 183–194. URL: http://dx.doi.org/10.5114/pjr.2021.105014, doi:10.5114/pjr.2021.105014.

Maier-Hein, L., Reinke, A., Godau, P., Tizabi, M.D., Buettner, F., Christodoulou, E., Glocker, B., Isensee, F., Kleesiek, J., Kozubek, M., et al., 2024. Metrics reloaded: recommendations for image analysis validation. Nature methods 21, 195–212.

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. Medical Image Analysis 66, 101796. URL: https://www.sciencedirect.com/science/article/pii/S1361841520301602, doi:https://doi.org/10.1016/j.media.2020.101796.

Marathu, K.K., Vahedifard, F., Kocak, M., Liu, X., Adepoju, J.O., Bowker, R.M., Supanich, M., Cosme-Cruz, R.M., Byrd, S., 2024. Fetal mri analysis of corpus callosal abnormalities: Classification, and associated anomalies. Diagnostics 14, 430. URL: http://dx.doi.org/10.3390/diagnostics14040430, doi:10.3390/diagnostics14040430.

Marques, J.P., Simonis, F.F., Webb, A.G., 2019. Low-field mri: An mr physics perspective. Journal of magnetic resonance imaging 49, 1528–1542.

Mase, M., Owen, A.B., Seiler, B., 2019. Explaining black box decisions by shapley cohort refinement. arXiv preprint arXiv:1911.00467 .

Matthew, J., Uus, A., Egloff Collado, A., Luis, A., Arulkumaran, S., Fukami-Gartner, A., Kyriakopoulou, V., Cromb, D., Wright, R., Colford, K., Deprez, M., Hutter, J., O'Muircheartaigh, J., Malamateniou, C., Razavi, R., Story, L., Hajnal, J.V., Rutherford, M.A., 2024. Automated craniofacial biometry with 3d t2w fetal mri. PLOS Digital Health 3, e0000663. URL: http://dx.doi.org/10.1371/journal.pdig.0000663, doi:10.1371/journal.pdig.0000663.

Meijerink, L., van Ooijen, I.M., Alderliesten, T., Terstappen, F., Benders, M., Bekker, M.N., 2023. Ep15.06: Fetal brain development in fetal growth restriction using mri: a systematic review. Ultrasound in Obstetrics &; Gynecology 62, 189–189. URL: http://dx.doi.org/10.1002/uog.26875, doi:10.1002/uog.26875.

Molchanova, N., Raina, V., Malinin, A., La Rosa, F., Depeursinge, A., Gales, M., Granziera, C., Müller, H., Graziani, M., Cuadra, M.B., 2025. Structural-based uncertainty in deep learning across anatomical scales: Analysis in white matter lesion segmentation. Computers in biology and medicine 184, 109336.

Murali, S., Ding, H., Adedeji, F., Qin, C., Obungoloch, J., Asllani, I., Anazodo, U., Ntusi, N.A.B., Mammen, R., Niendorf, T., Adeleke, S., 2023. Bringing mri to low- and middle-income countries: Directions, challenges and potential solutions. NMR in Biomedicine 37. URL: http://dx.doi.org/10.1002/nbm.4992, doi:10.1002/nbm.4992.

Neves Silva, S., McElroy, S., Aviles Verdera, J., Colford, K., St Clair, K., Tomi-Tricot, R., Uus, A., Ozenne, V., Hall, M., Story, L., et al., 2024. Fully automated planning for anatomical fetal brain mri on 0.55 t. Magnetic Resonance in Medicine 92, 1263–1276.

Ouyang, C., Chen, C., Li, S., Li, Z., Qin, C., Bai, W., Rueckert, D., 2022. Causality-inspired single-source domain generalization for medical image segmentation. IEEE Transactions on Medical Imaging 42, 1095–1106.

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., 2021. An automatic multi-tissue human fetal brain segmentation benchmark using the fetal tissue annotation dataset. Scientific Data 8. URL: http://dx.doi.org/10.1038/s41597-021-00946-3, doi:10.1038/s41597-021-00946-3.

Payette, K., Li, H.B., de Dumast, P., Licandro, R., Ji, H., Siddiquee, M.M.R., Xu, D., Myronenko, A., Liu, H., Pei, Y., et al., 2023. Fetal brain tissue annotation and segmentation challenge results. Medical image analysis 88, 102833.

Payette, K., Steger, C., Licandro, R., De Dumast, P., Li, H.B., Barkovich, M., Li, L., Dannecker, M., Chen, C., Ouyang, C., McConnell, N., Miron, A., Li, Y., Uus, A., Grigorescu, I., Gilliland, P.R., Siddiquee, M.M.R., Xu, D., Myronenko, A., Wang, H., Huang, Z., Ye, J., Alenyà, M., Comte, V., Camara, O., Masson, J.B., Nilsson,

A., Godard, C., Mazher, M., Qayyum, A., Gao, Y., Zhou, H., Gao, S., Fu, J., Dong, G., Wang, G., Rieu, Z., Yang, H., Lee, M., Płotka, S., Grzeszczyk, M.K., Sitek, A., Daza, L.V., Usma, S., Arbelaez, P., Lu, W., Zhang, W., Liang, J., Valabregue, R., Joshi, A.A., Nayak, K.N., Leahy, R.M., Wilhelmi, L., Dändliker, A., Ji, H., Gennari, A.G., Jakovčić, A., Klaić, M., Adžić, A., Marković, P., Grabarić, G., Kasprian, G., Dovjak, G., Rados, M., Vasung, L., Cuadra, M.B., Jakab, A., 2024. Multi-center fetal brain tissue annotation (feta) challenge 2022 results. IEEE Transactions on Medical Imaging , 1–1doi:10.1109/TMI.2024.3485554.

Prayer, D., Kasprian, G., Krampl, E., Ulm, B., Witzani, L., Prayer, L., Brugger, P.C., 2006. Mri of normal fetal brain development. European Journal of Radiology 57, 199–216. URL: https://www.sciencedirect.com/science/article/pii/S0720048X05003852, doi:https://doi.org/10.1016/j.ejrad.2005.11.020. fetal Imaging.

Price, A.N., Cordero-Grande, L., Hughes, E., Hiscocks, S., Green, E., McCabe, L., Hutter, J., Ferrazzi, G., Deprez, M., Roberts, T., et al., 2019. The developing human connectome project (dhcp): fetal acquisition protocol, in: Proceedings of the annual meeting of the International Society of Magnetic Resonance in Medicine (ISMRM), International Society for Magnetic Resonance in Medicine (ISMRM).

Raina, V., Molchanova, N., Graziani, M., Malinin, A., Muller, H., Cuadra, M.B., Gales, M., 2023. Tackling bias in the dice similarity coefficient: introducing ndsc for white matter lesion segmentation, in: 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI), IEEE. pp. 1–5.

Richiardi, J., Ravano, V., Molchanova, N., Gordaliza, P.M., Kober, T., Cuadra, M.B., 2025. Domain shift, domain adaptation, and generalization: A focus on mri, in: Trustworthy AI in Medical Imaging. Elsevier, pp. 127–151.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18, Springer. pp. 234–241.

Roy, S., Koehler, G., Ulrich, C., Baumgartner, M., Petersen, J., Isensee, F., Jaeger, P.F., Maier-Hein, K., 2023. Mednext: Transformer-driven scaling of convnets for medical image segmentation. URL: https://arxiv.org/abs/2303.09975, doi:10.48550/ARXIV.2303.09975.

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 145, 1108–1119. URL: http://dx.doi.org/10.1161/CIRCULATIONAHA.121.056305, doi:10.1161/circulationaha.121.056305.

Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., Aroyo, L.M., 2021. "everyone wants to do the model work, not the data work": Data cascades in high-stakes ai, in: proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, pp. 1–15.

Sanchez, T., Mihailov, A., Gomez, Y., Juan, G.M., Eixarch, E., Jakab, A., Dunet, V., Koob, M., Auzias, G., Cuadra, M.B., 2024a. Assessing data quality on fetal brain mri reconstruction: a multi-site and multi-rater study, in: International Workshop on Preterm, Perinatal and Paediatric Image Analysis, Springer. pp. 46–56.

Sanchez, T., Mihailov, A., Koob, M., Girard, N., Manchon, A., Valenzuela, I., Gómez-Chiari, M., Martí Juan, G., Pron, A., Eixarch, E., et al., 2024b. Biometry and volumetry in multi-centric fetal brain mri: assessing the bias of super-resolution reconstruction. medRxiv , 2024–09.

She, J., Huang, H., Ye, Z., Huang, W., Sun, Y., Liu, C., Yang, W., Wang, J., Ye, P., Zhang, L., et al., 2023. Automatic biometry of fetal brain mris using deep and machine learning techniques. Scientific Reports 13, 17860.

Story, L., Davidson, A., Patkee, P., Fleiss, B., Kyriakopoulou, V., Colford, K., Sankaran, S., Seed, P., Jones, A., Hutter, J., Shennan, A., Rutherford, M., 2021. Brain volumetry in fetuses that deliver very preterm: An mri pilot study. NeuroImage: Clinical 30, 102650. URL: http://dx.doi.org/10.1016/j.nicl.2021.102650, doi:10.1016/j.nicl.2021.102650.

Taha, A.A., Hanbury, A., 2015. Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool. BMC medical imaging 15, 1–28.

Tilea, B., Alberti, C., Adamsbaum, C., Armoogum, P., Oury, J., Cabrol, D., Sebag, G., Kalifa, G., Garel, C., 2009. Cerebral biometry in fetal magnetic resonance imaging: new reference data. Ultrasound in Obstetrics and Gynecology 33, 173–181.

Uus, A., Kyriakopoulou, V., Cordero Grande, L., Christiaens, D., Pietsch, M., Price, A., Wilson, S., Patkee, P., Karolis, S., Schuh, A., Gartner, A., Williams, L., Hughes, E., Arichi, T., O'Muircheartaigh, J., Hutter, J., Robinson, E., Tournier, J.D., Rueckert, D., Counsell, S., Rutherford, M., Deprez, M., Hajnal, J.V., Edwards, A.D., 2023a. Multi-channel spatio-temporal mri atlas of the normal fetal brain development from the developing human connectome project. URL: https://doi.gin.g-node.org/10.12751/g-node.ysgsy1, doi:10.12751/G-NODE.YSGSY1.

Uus, A., Zhang, T., Jackson, L.H., Roberts, T.A., Rutherford, M.A., Hajnal, J.V., Deprez, M., 2020. Deformable slice-to-volume registration for motion correction of fetal body and placenta mri. IEEE Transactions on Medical Imaging 39, 2750–2759. URL: http://dx.doi.org/10.1109/TMI.2020.2974844, doi:10.1109/tmi.2020.2974844.

Uus, A.U., Egloff Collado, A., Roberts, T.A., Hajnal, J.V., Rutherford, M.A., Deprez, M., 2023b. Retrospective motion correction in foetal mri for clinical applications: existing methods, applications and integration into clinical practice. The British journal of radiology 96, 20220071.

Uus, A.U., Neves Silva, S., Aviles Verdera, J., Payette, K., Hall, M., Colford, K., Luis, A., Sousa, H.S., Ning, Z., Roberts, T., et al., 2024. Scanner-based real-time 3d brain+ body slice-to-volume reconstruction for t2-weighted 0.55 t low field fetal mri. medRxiv , 2024–04.

Valabregue, R., Girka, F., Pron, A., Rousseau, F., Auzias, G., 2024a. Comprehensive analysis of synthetic learning applied to neonatal brain mri segmentation. Human Brain Mapping 45, e26674.

Valabregue, R., Khemir, I., Auzias, G., Rousseau, F., Ounissi, M., 2024b. Unraveling systematic biases in brain segmentation: Insights from synthetic training, in: MIDL 2024-Medical Imaging with Deep Learning.

Varoquaux, G., Cheplygina, V., 2022. Machine learning for medical imaging: methodological failures and recommendations for the future. NPJ digital medicine 5, 48.

Vaswani, A., 2017. Attention is all you need. Advances in Neural Information Processing Systems .

Wiles, O., Gowal, S., Stimberg, F., Alvise-Rebuffi, S., Ktena, I., Dvijotham, K., Cemgil, T., 2021. A fine-grained analysis on distribution shift. arXiv preprint arXiv:2110.11328 .

Xu, J., Moyer, D., Gagoski, B., Iglesias, J.E., Grant, P.E., Golland, P., Adalsteinsson, E., 2023. Nesvor: implicit neural representation for slice-to-volume reconstruction in mri. IEEE transactions on medical imaging 42, 1707–1719.

Yehuda, B., Rabinowich, A., Link-Sourani, D., Avisdris, N., Ben-Zvi, O., Specktor-Fadida, B., Joskowicz, L., Ben-Sira, L., Miller, E., Ben Bashat, D., 2023. Automatic quantification of normal brain gyrification patterns and changes in fetuses with polymicrogyria and lissencephaly based on mri. American Journal of Neuroradiology 44, 1432–1439. URL: http://dx.doi.org/10.3174/ajnr.A8046, doi:10.3174/ajnr.a8046.

Yushkevich, P.A., Piven, J., Cody Hazlett, H., Gimpel Smith, R., Ho, S., Gee, J.C., Gerig, G., 2006. User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. Neuroimage 31, 1116–1128.

Zalevskyi, V., Sanchez, T., Roulet, M., Lajous, H., Verdera, J.A., Hutter, J., Kebiri, H., Cuadra, M.B., 2024. Maximizing domain generalization in fetal brain tissue segmentation: the role of synthetic data generation, intensity clustering and real image fine-tuning. URL: https://arxiv.org/abs/2411.06842, doi:10.48550/ARXIV.2411.06842.

Zenk, M., Zimmerer, D., Isensee, F., Traub, J., Norajitra, T., Jäger, P.F., Maier-Hein, K., 2025. Comparative benchmarking of failure detection methods in medical image segmentation: unveiling the role of confidence aggregation. Medical image analysis 101, 103392.

Zha, D., Bhat, Z.P., Lai, K.H., Yang, F., Hu, X., 2023. Data-centric ai: Perspectives and challenges, in: Proceedings of the 2023 SIAM international conference on data mining (SDM), SIAM. pp. 945–948.

Supplementary Materials

Contents

Appendix A1. Evaluation metrics description.

Task 1: Segmentation

To evaluate the performance of the segmentation algorithms, we used four complementary metrics:

1. **Dice Similarity Coefficient (Dice)**: Measures voxel-wise correspondence between the predicted and ground truth (GT) segmentations. It is computed as:

$$Dice = \frac{2 \cdot |A \cap B|}{|A| + |B|}$$

where A and B represent the predicted and GT segmentation sets per label, respectively. Higher values of DSC indicate better overlap.

2. **Volume Similarity (VS)**: Assesses similarity of volumes between predicted and GT segmentations, defined as:

$$VS = 1 - \frac{|V_{pred} - V_{GT}|}{|V_{pred} + V_{GT}|}$$

where V_{pred} and V_{GT} are the volumes of the predicted and GT regions, respectively. A value close to 1 indicates high similarity.

3. **Hausdorff Distance (HD95)**: Quantifies contour similarity between predicted and GT segmentations using the 95th-percentile Hausdorff distance:

$$HD95 = \max\left(\max_{x \in A} \min_{y \in B} \|x - y\|, \max_{y \in B} \min_{x \in A} \|x - y\|\right)$$

where A and B are boundary points of the predicted and GT segmentations, and $\|\cdot\|$ denotes the Euclidean distance. Lower HD95 values indicate better contour agreement.

4. **Euler Characteristics (ED) Difference**: Evaluates topological similarity between predicted and GT segmentations, based on Betti numbers. The Euler characteristic is:

$$EC = \text{Betti}_0 - \text{Betti}_1 + \text{Betti}_2$$

where Betti_0 is the number of connected components, Betti_1 the number of loops, and Betti_2 the number of voids. The EC difference is computed as:

$$ED = |EC_{pred} - EC_{GT}|$$

Smaller differences indicate better topological alignment. The expected GT values for Betti numbers are: $\text{BN}_1 = 0$ and $\text{BN}_2 = 0$ for all brain tissues. For the eCSF, WM, ventricles, cerebellum, dGM, and brainstem, $\text{BN}_0 = 1$; for GM, $\text{BN}_0 = 2$.

These metrics together provide a comprehensive evaluation of segmentation accuracy, considering spatial overlap, volume, shape, and topology.

Task 2: Biometry Estimation

The primary metric for evaluating biometry estimation algorithms is **mean average percentage error (MAPE)**, quantifying error relative to actual measurements:

$$MAPE = \frac{1}{N} \sum_{i=1}^{N} \frac{|y_i - \hat{y}_i|}{y_i} \times 100$$

where y_i and $\hat{y}_i$ are the ground truth and predicted measurements, respectively, and N is the total number of measurements.

This metric accounts for variable target structure sizes and assesses the accuracy of the estimated biometric measurements.

Appendix A2. Algorithm descriptions

Team Algorithm Descriptions

The algorithm descriptions are presented for all participating teams in any of the tasks, in the following order:

1. CEMRG
2. falcons
3. feta_sigma
4. hilab
5. Jwcrad
6. lmrcmc
7. LIT
8. mic-dkfz
9. paramahir_2023
10. pasteurdbc
11. unipd-sum-aug
12. UPFetal24
13. vicorob
14. qd_neuroincyte
15. CeSNE-DiGAIR

Appendix A3. FeTA 2024 segmentation results by site

Table A3: Segmentation results for FeTA 2024 by site, presented as mean ± standard deviation.

Team	Site	Dice	HD95	Volume Similarity	Euler diff.
cemrg_feta	CHUV	0.819 +- 0.079	2.239 +- 1.641	0.888 +- 0.091	73.271 +- 157.842
	KCL	0.868 +- 0.048	1.482 +- 0.575	0.944 +- 0.050	8.729 +- 17.511
	KISPI	0.775 +- 0.178	3.756 +- 12.339	0.875 +- 0.175	17.150 +- 47.150
	UCSF	0.835 +- 0.064	4.180 +- 10.689	0.941 +- 0.056	21.732 +- 48.451
	VIEN	0.834 +- 0.079	1.846 +- 1.174	0.947 +- 0.056	38.200 +- 76.002
cesne-digair	CHUV	0.830 +- 0.060	2.234 +- 1.397	0.927 +- 0.058	29.104 +- 51.562
	KCL	0.856 +- 0.052	1.694 +- 0.524	0.950 +- 0.041	6.257 +- 13.210
	KISPI	0.776 +- 0.154	2.954 +- 2.863	0.890 +- 0.138	9.211 +- 17.841
	UCSF	0.823 +- 0.069	2.126 +- 1.307	0.942 +- 0.050	14.571 +- 25.900
	VIEN	0.814 +- 0.087	2.266 +- 1.687	0.947 +- 0.050	38.132 +- 93.476
falcons	CHUV	0.828 +- 0.063	2.054 +- 1.271	0.922 +- 0.056	82.214 +- 166.732
	KCL	0.832 +- 0.061	2.076 +- 1.080	0.906 +- 0.065	17.450 +- 33.388
	KISPI	0.763 +- 0.151	2.890 +- 2.490	0.887 +- 0.139	35.557 +- 80.026
	UCSF	0.430 +- 0.253	21.320 +- 14.077	0.602 +- 0.281	192.339 +- 318.110
	VIEN	0.389 +- 0.275	22.377 +- 17.046	0.580 +- 0.296	134.443 +- 158.795
feta_sigma	CHUV	0.823 +- 0.075	2.248 +- 1.659	0.892 +- 0.085	68.571 +- 143.222
	KCL	0.871 +- 0.049	1.469 +- 0.576	0.944 +- 0.051	10.521 +- 23.896
	KISPI	0.772 +- 0.178	3.311 +- 4.349	0.867 +- 0.177	14.814 +- 36.054
	UCSF	0.835 +- 0.064	2.427 +- 3.941	0.937 +- 0.059	21.279 +- 47.773
	VIEN	0.835 +- 0.084	2.216 +- 3.005	0.943 +- 0.062	32.771 +- 65.733
hilab	CHUV	0.813 +- 0.079	2.311 +- 1.464	0.878 +- 0.083	66.004 +- 147.774
	KCL	0.851 +- 0.058	1.765 +- 0.823	0.926 +- 0.066	5.807 +- 11.417
	KISPI	0.775 +- 0.176	2.937 +- 3.457	0.878 +- 0.169	13.029 +- 30.945
	UCSF	0.828 +- 0.068	2.677 +- 3.871	0.939 +- 0.053	24.786 +- 62.808
	VIEN	0.831 +- 0.081	2.145 +- 2.654	0.941 +- 0.062	28.832 +- 61.805
jwcrad	CHUV	0.741 +- 0.126	4.495 +- 4.241	0.840 +- 0.156	60.668 +- 119.543
	KCL	0.845 +- 0.060	2.101 +- 2.275	0.942 +- 0.059	10.857 +- 25.192
	KISPI	0.780 +- 0.154	3.001 +- 3.122	0.892 +- 0.142	15.054 +- 63.731
	UCSF	0.775 +- 0.129	4.139 +- 12.272	0.908 +- 0.132	22.382 +- 72.242
	VIEN	0.743 +- 0.173	3.375 +- 3.568	0.876 +- 0.174	30.314 +- 48.812
lit	CHUV	0.817 +- 0.074	2.397 +- 1.632	0.892 +- 0.078	73.400 +- 150.808
	KCL	0.867 +- 0.049	1.478 +- 0.511	0.950 +- 0.041	7.957 +- 16.179
	KISPI	0.761 +- 0.180	3.400 +- 4.158	0.866 +- 0.176	17.854 +- 49.090
	UCSF	0.812 +- 0.075	2.139 +- 1.852	0.931 +- 0.062	33.982 +- 59.647
	VIEN	0.810 +- 0.093	2.086 +- 2.610	0.937 +- 0.059	51.168 +- 93.347
lmrcmc	CHUV	0.814 +- 0.080	2.406 +- 1.409	0.881 +- 0.084	58.054 +- 126.355
	KCL	0.860 +- 0.052	1.611 +- 0.728	0.936 +- 0.058	9.429 +- 21.316
	KISPI	0.774 +- 0.171	3.013 +- 3.697	0.881 +- 0.168	14.675 +- 29.453
	UCSF	0.798 +- 0.084	4.399 +- 6.528	0.933 +- 0.066	38.914 +- 66.956
	VIEN	0.805 +- 0.093	3.682 +- 8.311	0.945 +- 0.057	31.021 +- 50.330
mic-dkfz-feta24	CHUV	0.824 +- 0.074	2.125 +- 1.520	0.889 +- 0.083	83.104 +- 190.117
	KCL	0.872 +- 0.046	1.456 +- 0.540	0.950 +- 0.048	8.471 +- 17.095
	KISPI	0.778 +- 0.176	3.667 +- 12.334	0.878 +- 0.175	24.968 +- 98.618
	UCSF	0.849 +- 0.060	1.648 +- 0.929	0.943 +- 0.059	20.679 +- 48.929
	VIEN	0.839 +- 0.077	1.842 +- 1.361	0.947 +- 0.055	34.443 +- 70.139
paramahir_2023	CHUV	0.019 +- 0.016	67.984 +- 9.712	0.257 +- 0.191	836.771 +- 514.176
	KCL	0.105 +- 0.114	40.203 +- 38.827	0.665 +- 0.278	1053.286 +- 1194.848
	KISPI	0.036 +- 0.069	134.320 +- 80.816	0.220 +- 0.303	3243.929 +- 2176.151
	UCSF	0.042 +- 0.044	69.115 +- 33.709	0.392 +- 0.246	891.479 +- 1040.067
	VIEN	0.029 +- 0.040	71.887 +- 14.464	0.317 +- 0.263	875.496 +- 803.573
pasteurdbc	CHUV	0.816 +- 0.080	2.272 +- 1.574	0.879 +- 0.087	98.582 +- 221.114
	KCL	0.870 +- 0.048	1.494 +- 0.557	0.953 +- 0.043	9.714 +- 19.600
	KISPI	0.770 +- 0.178	3.247 +- 4.094	0.868 +- 0.177	24.864 +- 82.402
	UCSF	0.820 +- 0.076	2.826 +- 3.592	0.920 +- 0.075	21.893 +- 49.927
	VIEN	0.835 +- 0.081	2.040 +- 1.813	0.946 +- 0.060	36.650 +- 71.768
qd_neuroincyte	CHUV	0.762 +- 0.127	3.973 +- 3.116	0.877 +- 0.125	27.993 +- 45.480
	KCL	0.800 +- 0.071	15.895 +- 15.467	0.896 +- 0.073	22.821 +- 24.303
	KISPI	0.769 +- 0.174	3.436 +- 4.185	0.878 +- 0.170	19.061 +- 41.108
	UCSF	0.444 +- 0.244	22.082 +- 26.679	0.648 +- 0.288	62.807 +- 133.521
	VIEN	0.689 +- 0.194	9.546 +- 9.233	0.869 +- 0.132	33.057 +- 49.988
unipd-sum-aug	CHUV	0.802 +- 0.084	2.517 +- 1.798	0.871 +- 0.093	85.850 +- 191.883
	KCL	0.863 +- 0.047	1.553 +- 0.587	0.950 +- 0.041	13.136 +- 22.792
	KISPI	0.762 +- 0.184	3.169 +- 3.723	0.868 +- 0.180	33.868 +- 108.254
	UCSF	0.827 +- 0.069	2.199 +- 2.283	0.934 +- 0.059	29.396 +- 62.257
	VIEN	0.826 +- 0.077	1.835 +- 0.987	0.945 +- 0.054	54.325 +- 108.289
upfetal24	CHUV	0.816 +- 0.080	2.296 +- 1.703	0.882 +- 0.091	89.425 +- 204.043
	KCL	0.844 +- 0.061	2.391 +- 1.490	0.931 +- 0.072	31.729 +- 45.318
	KISPI	0.776 +- 0.175	3.055 +- 3.537	0.876 +- 0.171	19.861 +- 44.504
	UCSF	0.840 +- 0.060	2.452 +- 5.282	0.940 +- 0.056	20.836 +- 51.070
	VIEN	0.837 +- 0.078	1.855 +- 1.405	0.945 +- 0.056	33.868 +- 68.319
vicorob	CHUV	0.831 +- 0.070	1.969 +- 1.266	0.899 +- 0.077	93.789 +- 206.532
	KCL	0.869 +- 0.051	1.499 +- 0.659	0.947 +- 0.051	9.986 +- 19.663
	KISPI	0.782 +- 0.170	2.982 +- 3.709	0.881 +- 0.169	19.579 +- 71.933
	UCSF	0.830 +- 0.067	2.276 +- 1.965	0.937 +- 0.056	29.343 +- 68.239

Appendix A4. FeTA 2024 segmentation results by label

Table A4: Segmentation results for FeTA 2024 by label, presented as mean ± standard deviation.

Team	Label	Dice	HD95	Volume Similarity	Euler diff.
cemrg_feta	BS	0.773 +- 0.109	4.066 +- 4.480	0.864 +- 0.128	0.561 +- 0.749
	CBM	0.869 +- 0.121	3.097 +- 15.680	0.933 +- 0.124	0.178 +- 0.718
	CSF	0.805 +- 0.126	3.090 +- 6.242	0.922 +- 0.126	79.172 +- 77.460
	GM	0.748 +- 0.075	1.885 +- 4.882	0.921 +- 0.062	144.156 +- 183.114
	SGM	0.801 +- 0.109	3.417 +- 4.473	0.863 +- 0.124	1.056 +- 1.166
	VM	0.864 +- 0.061	2.028 +- 6.517	0.951 +- 0.046	2.389 +- 2.444
	WM	0.892 +- 0.038	2.269 +- 5.881	0.961 +- 0.032	13.161 +- 14.716
cesne-digair	BS	0.790 +- 0.091	3.665 +- 2.187	0.899 +- 0.091	0.206 +- 0.525
	CBM	0.868 +- 0.088	1.466 +- 0.716	0.945 +- 0.083	0.367 +- 1.624
	CSF	0.794 +- 0.108	2.726 +- 2.860	0.937 +- 0.080	49.006 +- 35.876
	GM	0.730 +- 0.077	1.557 +- 0.786	0.935 +- 0.043	87.644 +- 109.380
	SGM	0.798 +- 0.105	3.061 +- 1.881	0.879 +- 0.125	0.833 +- 0.647
	VM	0.848 +- 0.066	1.713 +- 1.228	0.946 +- 0.045	1.578 +- 1.468
	WM	0.883 +- 0.043	2.030 +- 0.835	0.964 +- 0.030	6.817 +- 9.445
falcons	BS	0.519 +- 0.310	11.287 +- 10.191	0.688 +- 0.288	28.778 +- 41.150
	CBM	0.570 +- 0.361	19.103 +- 22.806	0.721 +- 0.288	42.172 +- 94.277
	CSF	0.608 +- 0.244	12.110 +- 11.915	0.719 +- 0.265	175.044 +- 316.827
	GM	0.604 +- 0.177	10.279 +- 11.330	0.845 +- 0.145	289.128 +- 272.968
	SGM	0.542 +- 0.324	9.231 +- 16.580	0.617 +- 0.334	17.350 +- 25.692
	VM	0.762 +- 0.158	5.269 +- 5.874	0.863 +- 0.144	44.506 +- 74.144
	WM	0.791 +- 0.145	10.000 +- 11.203	0.903 +- 0.116	108.122 +- 164.206
feta_sigma	BS	0.776 +- 0.104	4.263 +- 4.436	0.865 +- 0.118	0.483 +- 0.751
	CBM	0.867 +- 0.124	2.042 +- 3.452	0.925 +- 0.128	0.244 +- 0.759
	CSF	0.802 +- 0.132	3.058 +- 5.381	0.911 +- 0.136	83.772 +- 73.078
	GM	0.750 +- 0.076	1.438 +- 0.766	0.922 +- 0.063	121.089 +- 165.583
	SGM	0.800 +- 0.112	3.196 +- 2.267	0.860 +- 0.128	1.083 +- 1.143
	VM	0.869 +- 0.057	1.293 +- 0.696	0.951 +- 0.040	2.261 +- 2.175
	WM	0.893 +- 0.038	1.722 +- 0.536	0.962 +- 0.032	13.039 +- 12.755
hilab	BS	0.772 +- 0.098	4.155 +- 3.437	0.874 +- 0.108	0.350 +- 0.523
	CBM	0.866 +- 0.121	1.931 +- 3.170	0.924 +- 0.123	0.067 +- 0.310
	CSF	0.802 +- 0.121	2.506 +- 3.554	0.916 +- 0.125	79.889 +- 89.962
	GM	0.740 +- 0.086	1.492 +- 0.743	0.917 +- 0.068	114.972 +- 170.136
	SGM	0.790 +- 0.107	3.513 +- 3.317	0.860 +- 0.124	0.750 +- 0.838
	VM	0.853 +- 0.071	1.672 +- 2.266	0.931 +- 0.067	2.439 +- 2.077
	WM	0.889 +- 0.039	1.768 +- 0.550	0.955 +- 0.039	12.394 +- 13.574
jwcrad	BS	0.676 +- 0.175	5.253 +- 3.670	0.779 +- 0.212	7.417 +- 74.544
	CBM	0.799 +- 0.153	3.189 +- 3.409	0.878 +- 0.164	2.989 +- 9.602
	CSF	0.774 +- 0.121	2.693 +- 3.000	0.926 +- 0.093	62.778 +- 55.944
	GM	0.705 +- 0.088	1.931 +- 1.834	0.908 +- 0.085	105.406 +- 138.418
	SGM	0.737 +- 0.155	6.690 +- 15.354	0.827 +- 0.176	9.600 +- 74.649
	VM	0.829 +- 0.098	2.808 +- 3.244	0.931 +- 0.088	4.700 +- 6.119
	WM	0.866 +- 0.081	2.420 +- 2.004	0.950 +- 0.064	15.317 +- 16.491
lit	BS	0.770 +- 0.103	4.159 +- 2.815	0.872 +- 0.119	0.511 +- 0.639
	CBM	0.866 +- 0.120	1.787 +- 3.629	0.935 +- 0.118	0.083 +- 0.394
	CSF	0.782 +- 0.131	2.615 +- 3.683	0.917 +- 0.123	78.561 +- 69.369
	GM	0.718 +- 0.081	1.652 +- 1.190	0.916 +- 0.067	184.161 +- 166.182

	SGM	0.793 +- 0.116	3.250 +- 2.395	0.862 +- 0.136	0.933 +- 1.122
	VM	0.848 +- 0.063	1.487 +- 1.074	0.924 +- 0.060	2.361 +- 2.044
	WM	0.878 +- 0.054	1.790 +- 0.867	0.951 +- 0.036	13.983 +- 21.532
lmrcmc	BS	0.761 +- 0.112	4.559 +- 5.464	0.871 +- 0.116	1.333 +- 2.422
	CBM	0.855 +- 0.108	3.499 +- 7.590	0.927 +- 0.111	3.850 +- 11.014
	CSF	0.799 +- 0.120	2.689 +- 5.592	0.932 +- 0.115	52.617 +- 48.115
	GM	0.713 +- 0.076	2.053 +- 2.993	0.910 +- 0.077	124.883 +- 138.935
	SGM	0.782 +- 0.116	4.384 +- 3.510	0.852 +- 0.132	3.722 +- 8.246
	VM	0.846 +- 0.077	2.597 +- 6.345	0.945 +- 0.053	5.006 +- 7.943
	WM	0.876 +- 0.044	2.474 +- 4.290	0.953 +- 0.036	37.844 +- 72.378
mic-dkfz-feta24	BS	0.792 +- 0.102	3.430 +- 2.265	0.877 +- 0.116	6.250 +- 74.563
	CBM	0.875 +- 0.121	2.452 +- 14.827	0.936 +- 0.124	5.589 +- 74.609
	CSF	0.811 +- 0.123	2.492 +- 3.960	0.921 +- 0.124	76.667 +- 79.789
	GM	0.754 +- 0.078	1.342 +- 0.664	0.919 +- 0.068	153.933 +- 222.976
	SGM	0.803 +- 0.112	2.981 +- 1.915	0.859 +- 0.128	1.278 +- 1.303
	VM	0.867 +- 0.060	1.275 +- 0.655	0.953 +- 0.046	2.144 +- 2.387
	WM	0.895 +- 0.036	1.598 +- 0.464	0.961 +- 0.032	14.583 +- 15.262
paramahir_2023	BS	0.002 +- 0.004	80.114 +- 52.805	0.439 +- 0.300	1315.700 +- 1572.594
	CBM	0.003 +- 0.011	98.476 +- 46.129	0.471 +- 0.319	1182.661 +- 1613.381
	CSF	0.057 +- 0.068	75.774 +- 55.679	0.364 +- 0.283	1668.756 +- 1510.524
	GM	0.040 +- 0.038	77.447 +- 54.473	0.284 +- 0.275	1239.539 +- 1637.509
	SGM	0.034 +- 0.039	81.118 +- 51.993	0.238 +- 0.230	1424.194 +- 1523.328
	VM	0.049 +- 0.056	74.136 +- 54.399	0.328 +- 0.273	2102.550 +- 1503.443
	WM	0.092 +- 0.105	78.237 +- 53.946	0.238 +- 0.238	982.206 +- 1700.717
pasteurdbc	BS	0.771 +- 0.110	4.197 +- 2.852	0.863 +- 0.124	5.939 +- 74.584
	CBM	0.867 +- 0.109	1.809 +- 1.441	0.925 +- 0.112	0.217 +- 1.135
	CSF	0.804 +- 0.132	2.797 +- 4.533	0.918 +- 0.136	75.700 +- 75.225
	GM	0.735 +- 0.082	1.517 +- 0.790	0.909 +- 0.079	184.989 +- 256.428
	SGM	0.795 +- 0.116	3.224 +- 2.146	0.850 +- 0.132	1.072 +- 1.237
	VM	0.862 +- 0.064	1.333 +- 0.902	0.945 +- 0.055	2.056 +- 1.902
	WM	0.884 +- 0.050	2.439 +- 3.815	0.951 +- 0.048	20.678 +- 21.555
qd_neuroincyte	BS	0.564 +- 0.322	16.358 +- 33.573	0.679 +- 0.343	31.933 +- 164.379
	CBM	0.699 +- 0.287	12.927 +- 18.047	0.813 +- 0.238	6.317 +- 11.709
	CSF	0.642 +- 0.164	7.771 +- 4.291	0.826 +- 0.140	50.983 +- 42.466
	GM	0.604 +- 0.156	6.012 +- 5.081	0.882 +- 0.104	85.433 +- 61.912
	SGM	0.693 +- 0.171	10.169 +- 9.490	0.791 +- 0.174	13.939 +- 25.437
	VM	0.765 +- 0.161	11.839 +- 11.728	0.894 +- 0.104	26.578 +- 37.313
	WM	0.797 +- 0.120	8.013 +- 6.939	0.902 +- 0.106	24.883 +- 26.451
unipd-sum-aug	BS	0.762 +- 0.119	4.033 +- 2.908	0.857 +- 0.133	6.178 +- 74.569
	CBM	0.857 +- 0.102	2.000 +- 1.808	0.929 +- 0.107	0.278 +- 0.702
	CSF	0.798 +- 0.138	2.282 +- 3.628	0.917 +- 0.139	85.600 +- 90.364
	GM	0.726 +- 0.086	1.499 +- 0.758	0.910 +- 0.078	215.856 +- 226.509
	SGM	0.794 +- 0.111	3.363 +- 2.449	0.863 +- 0.132	1.222 +- 1.194
	VM	0.854 +- 0.067	1.443 +- 0.958	0.937 +- 0.055	2.400 +- 2.590
	WM	0.885 +- 0.038	1.708 +- 0.486	0.954 +- 0.038	15.144 +- 16.221
upfetal24	BS	0.780 +- 0.113	4.028 +- 2.620	0.871 +- 0.132	1.611 +- 3.007
	CBM	0.872 +- 0.108	1.529 +- 1.116	0.935 +- 0.109	0.350 +- 1.339
	CSF	0.810 +- 0.121	2.556 +- 4.235	0.921 +- 0.122	84.133 +- 80.372
	GM	0.742 +- 0.077	1.401 +- 0.708	0.917 +- 0.074	164.483 +- 233.666
	SGM	0.790 +- 0.108	3.652 +- 4.141	0.850 +- 0.128	1.611 +- 3.941

	VM	0.861 +- 0.064	1.636 +- 3.494	0.942 +- 0.051	3.061 +- 4.046
	WM	0.889 +- 0.035	2.082 +- 3.293	0.954 +- 0.036	24.522 +- 35.431
vicorob	BS	0.788 +- 0.104	3.744 +- 2.639	0.886 +- 0.109	6.372 +- 74.563
	CBM	0.873 +- 0.101	1.616 +- 1.192	0.932 +- 0.106	0.167 +- 0.780
	CSF	0.807 +- 0.124	2.370 +- 3.839	0.919 +- 0.127	97.094 +- 91.240
	GM	0.745 +- 0.076	1.415 +- 0.774	0.919 +- 0.070	169.622 +- 238.677
	SGM	0.801 +- 0.112	3.079 +- 2.062	0.868 +- 0.131	0.928 +- 1.041
	VM	0.868 +- 0.059	1.405 +- 1.703	0.956 +- 0.042	2.706 +- 2.315
	WM	0.891 +- 0.039	1.677 +- 0.504	0.957 +- 0.035	12.161 +- 12.048

Appendix A5. FeTA 2024 segmentation results by pathology

Table A5: Segmentation results for FeTA 2024 by pathology, presented as mean ± standard deviation.

Team	Pathology	Dice	HD95	Volume Similarity	Euler diff.
cemrg_feta	Neurotypical	0.838 +- 0.078	1.949 +- 1.685	0.923 +- 0.083	36.990 +- 104.012
	Pathological	0.808 +- 0.127	3.596 +- 10.477	0.910 +- 0.122	32.150 +- 79.136
cesne-digair	Neurotypical	0.834 +- 0.072	2.037 +- 1.401	0.939 +- 0.063	17.255 +- 36.003
	Pathological	0.801 +- 0.114	2.557 +- 2.138	0.921 +- 0.095	24.059 +- 65.344
falcons	Neurotypical	0.695 +- 0.233	9.431 +- 15.061	0.819 +- 0.212	89.800 +- 155.566
	Pathological	0.571 +- 0.297	12.417 +- 13.435	0.720 +- 0.285	110.080 +- 229.547
feta_sigma	Neurotypical	0.838 +- 0.079	1.944 +- 1.509	0.922 +- 0.082	35.382 +- 95.466
	Pathological	0.809 +- 0.128	2.846 +- 4.160	0.907 +- 0.125	28.568 +- 69.351
hilab	Neurotypical	0.832 +- 0.079	2.029 +- 1.465	0.917 +- 0.080	33.253 +- 97.439
	Pathological	0.802 +- 0.126	2.781 +- 3.629	0.906 +- 0.120	27.445 +- 72.466
jwcrad	Neurotypical	0.796 +- 0.110	3.051 +- 3.409	0.901 +- 0.121	33.539 +- 88.625
	Pathological	0.746 +- 0.163	4.013 +- 8.390	0.872 +- 0.166	26.496 +- 68.179
lit	Neurotypical	0.830 +- 0.080	1.971 +- 1.471	0.922 +- 0.079	40.859 +- 102.670
	Pathological	0.789 +- 0.132	2.751 +- 3.317	0.902 +- 0.122	39.423 +- 85.891
lmrcmc	Neurotypical	0.822 +- 0.086	2.884 +- 5.872	0.921 +- 0.082	31.589 +- 78.737
	Pathological	0.790 +- 0.126	3.432 +- 4.936	0.906 +- 0.118	33.745 +- 71.625
mic-dkfz-feta24	Neurotypical	0.842 +- 0.076	1.844 +- 1.368	0.925 +- 0.080	41.797 +- 130.019
	Pathological	0.817 +- 0.126	2.550 +- 8.005	0.912 +- 0.121	33.278 +- 93.104
paramahir_2023	Neurotypical	0.041 +- 0.066	75.381 +- 49.828	0.365 +- 0.293	1316.442 +- 1491.284
	Pathological	0.038 +- 0.060	85.358 +- 55.674	0.314 +- 0.282	1502.144 +- 1709.552
pasteurdbc	Neurotypical	0.834 +- 0.082	2.000 +- 1.522	0.918 +- 0.085	46.558 +- 146.162
	Pathological	0.802 +- 0.129	2.879 +- 3.591	0.901 +- 0.126	37.212 +- 98.278
qd_neuroincyte	Neurotypical	0.744 +- 0.178	9.321 +- 15.672	0.867 +- 0.156	31.234 +- 71.161
	Pathological	0.626 +- 0.242	11.399 +- 16.616	0.792 +- 0.233	36.915 +- 78.492
unipd-sum-aug	Neurotypical	0.827 +- 0.083	2.041 +- 1.614	0.917 +- 0.087	49.062 +- 132.939
	Pathological	0.797 +- 0.131	2.582 +- 2.797	0.903 +- 0.125	44.620 +- 111.600
upfetal24	Neurotypical	0.832 +- 0.080	2.096 +- 1.579	0.918 +- 0.087	45.657 +- 131.983
	Pathological	0.811 +- 0.124	2.682 +- 4.136	0.909 +- 0.120	35.099 +- 88.130
vicorob	Neurotypical	0.841 +- 0.077	1.870 +- 1.600	0.929 +- 0.076	46.210 +- 138.108
	Pathological	0.811 +- 0.122	2.457 +- 2.685	0.911 +- 0.117	37.085 +- 97.032

Appendix A6. FeTA 2024 biometry results by site, label and pathology

Biometry Estimation In this section of the supplementary materials, we present detailed results for the second task of the FeTA 2024 challenge, which focuses on fetal brain biometry estimation. To complement the primary evaluation metrics, we also report the **Mean Absolute Error (MAE)**, which offers an intuitive interpretation of the prediction errors in physical units (millimeters).

The MAE is defined as:

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} |\hat{y}_i - y_i|$$

where N is the number of samples, $\hat{y}_i$ is the predicted measurement for the i-th subject, and y_i is the corresponding ground truth measurement.

All MAE values are reported in millimeters (mm), which facilitates a direct understanding of the magnitude of the errors made by the algorithms in the context of fetal brain structure dimensions.

Table A6.1. Biometry results for all teams participating in the FeTA 2024 biometry challenge together with the baseline model and inter-rater variability stratified by label. The values for each team are sorted by MAPE in the increasing order.

Team	Pathology	MAE (mm)	MAPE
GA	Pathological	3.468+-4.504	0.090+-0.120
	Neurotypical	2.948+-3.057	0.103+-0.137
inter-rater	Neurotypical	1.333+-1.781	0.041+-0.057
	Pathological	1.625+-2.040	0.065+-0.105
cesne-digair	Neurotypical	3.283+-6.208	0.092+-0.126
	Pathological	2.896+-4.831	0.099+-0.139
falcons	Neurotypical	10.262+-14.883	0.262+-0.281
	Pathological	13.525+-14.425	0.411+-0.421
feta_sigma	Neurotypical	2.447+-2.567	0.069+-0.074
	Pathological	3.499+-4.260	0.123+-0.168
jwcrad	Neurotypical	1.948+-1.779	0.057+-0.057
	Pathological	3.123+-6.939	0.095+-0.143
paramahir_2023	Pathological	9.564+-11.359	0.257+-0.235
	Neurotypical	13.438+-15.592	0.306+-0.262
pasteurdbc	Neurotypical	3.704+-3.551	0.139+-0.177
	Pathological	3.852+-3.698	0.176+-0.254
qd_neuroincyte	Pathological	14.208+-19.129	0.384+-0.452
	Neurotypical	19.497+-26.977	0.419+-0.433

Table A6.2. Biometry results for all teams participating in the FeTA 2024 biometry challenge together with the baseline model and inter-rater variability stratified by label

Team	Team	MAE (mm)	MAPE
[GA]	HV	1.409±1.388	0.113±0.131
	LCC	3.406±3.302	0.127±0.161
	TCD	2.787±3.415	0.108±0.167
	bBIP	4.156±4.864	0.068±0.074
	sBIP	4.363±4.691	0.065±0.064
[inter-rater]	HV	0.993±0.885	0.080±0.088
	LCC	2.542±3.541	0.096±0.142
	TCD	1.189±1.233	0.049±0.072
	bBIP	1.852±1.388	0.033±0.028
	sBIP	0.951±0.950	0.015±0.016
cesne-digair	HV	1.197±1.199	0.098±0.120
	LCC	5.704±3.329	0.177±0.102
	TCD	3.238±3.914	0.123±0.162
	bBIP	2.381±3.250	0.040±0.055
	sBIP	3.079±10.121	0.047±0.147
falcons	HV	5.747±4.488	0.463±0.507
	LCC	9.804±8.722	0.349±0.349
	TCD	10.294±10.779	0.367±0.342
	bBIP	14.906±17.945	0.246±0.274
	sBIP	18.980±20.905	0.281±0.293
feta_sigma	HV	1.429±1.273	0.116±0.121
	LCC	3.540±3.073	0.126±0.145
	TCD	3.213±3.489	0.137±0.203
	bBIP	3.369±3.453	0.057±0.060
	sBIP	3.507±5.144	0.055±0.075
jwcrad	HV	1.377±1.087	0.103±0.083
	LCC	3.241±3.047	0.112±0.117
	TCD	1.983±1.735	0.072±0.072
	bBIP	3.285±7.850	0.054±0.131
	sBIP	3.013±7.579	0.048±0.132
paramahir_2023	HV	4.148±3.911	0.294±0.243
	LCC	8.936±8.877	0.285±0.244
	TCD	9.459±9.225	0.308±0.270
	bBIP	16.397±16.698	0.261±0.242
	sBIP	17.728±18.415	0.255±0.244
pasteurdbc	HV	5.694±2.572	0.435±0.262
	LCC	5.833±5.816	0.205±0.230
	TCD	1.261±1.381	0.054±0.099
	bBIP	3.820±2.663	0.065±0.046
	sBIP	2.470±1.637	0.037±0.027
qd_neuroincyte	HV	5.762±5.638	0.428±0.528
	LCC	9.946±10.714	0.328±0.379
	TCD	15.352±15.694	0.479±0.413
	bBIP	24.821±29.450	0.384±0.432
	sBIP	26.958±32.977	0.378±0.436

Table A6.3. Biometry results for all teams participating in the FeTA 2024 biometry challenge together with the baseline model and inter-rater variability stratifies by site

Team	Site	MAE (mm)	MAPE
[GA]	KCL	2.365±2.297	0.055±0.045
	CHUV	3.027±2.790	0.083±0.121
	UCSF	3.280±4.375	0.094±0.121
	VIEN	3.788±5.361	0.106±0.112
	KISPI	3.262±3.219	0.121±0.172
[inter-rater]	KCL	0.747±0.645	0.020±0.024
	CHUV	1.019±0.924	0.029±0.029
	UCSF	1.421±1.513	0.050±0.059
	KISPI	1.753±1.814	0.069±0.095
	VIEN	2.168±3.070	0.086±0.135
cesne-digair	UCSF	2.487±3.767	0.079±0.094
	KCL	3.109±3.544	0.085±0.116
	KISPI	2.610±4.410	0.096±0.147
	VIEN	3.665±9.285	0.106±0.157
	CHUV	3.551±3.629	0.109±0.133
falcons	KCL	4.885±6.207	0.109±0.108
	CHUV	11.379±22.126	0.262±0.427
	KISPI	7.156±5.066	0.282±0.311
	VIEN	14.906±13.615	0.435±0.457
	UCSF	18.260±12.581	0.508±0.230
feta_sigma	KCL	1.779±1.603	0.046±0.047
	CHUV	2.164±2.126	0.069±0.116
	KISPI	2.567±2.734	0.111±0.189
	VIEN	3.540±3.598	0.117±0.122
	UCSF	4.402±5.357	0.121±0.119
jwcrad	KCL	1.305±1.025	0.035±0.035
	CHUV	2.392±1.896	0.071±0.079
	UCSF	2.499±5.267	0.074±0.096
	KISPI	2.047±1.704	0.078±0.085
	VIEN	4.051±9.407	0.108±0.182
paramahir_2023	KCL	4.153±4.023	0.107±0.105
	KISPI	4.483±3.553	0.161±0.184
	VIEN	7.354±6.228	0.177±0.102
	UCSF	6.523±5.287	0.185±0.137
	CHUV	30.396±16.278	0.677±0.044
pasteurdbc	KCL	3.515±2.607	0.119±0.139
	CHUV	3.129±2.686	0.127±0.175
	UCSF	3.663±3.287	0.167±0.224
	KISPI	3.749±2.977	0.179±0.283
	VIEN	4.774±5.341	0.182±0.221
qd_neuroincyte	KCL	5.122±8.736	0.094±0.132
	KISPI	3.080±3.803	0.122±0.208
	VIEN	7.525±10.180	0.234±0.461
	UCSF	15.855±13.145	0.429±0.262
	CHUV	45.380±30.413	0.958±0.296

Appendix A7. Correlation of quality and challenge metrics

Correlation plots of visual quality scores and metrics across different sites and super-resolution methods for the best teams in FeTA 2024 Each dot represents an average metric value(of the top 3 teams in FeTA2024 `cesne-digair`, `mic-dkfz-feta24`, `vicorob`) for a given subject, with blue indicating quality scores (left axis) and red indicating a given metric (right axis). Sites and methods are grouped on the x-axis. Dashed lines connect data points for individual subjects across metrics. Pearson correlation coefficients (r) between quality and Dice are shown above each group

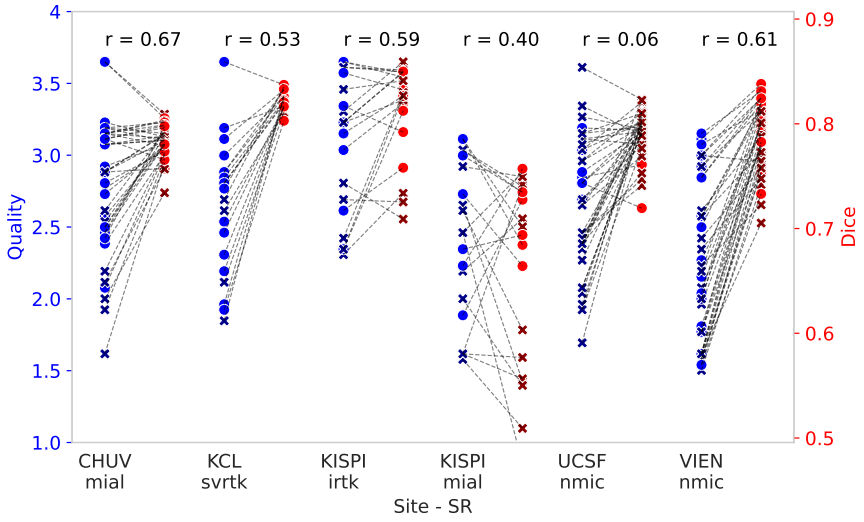


Figure A7.1: Correlation between visual quality scores and Dice across different sites and super-resolution methods.

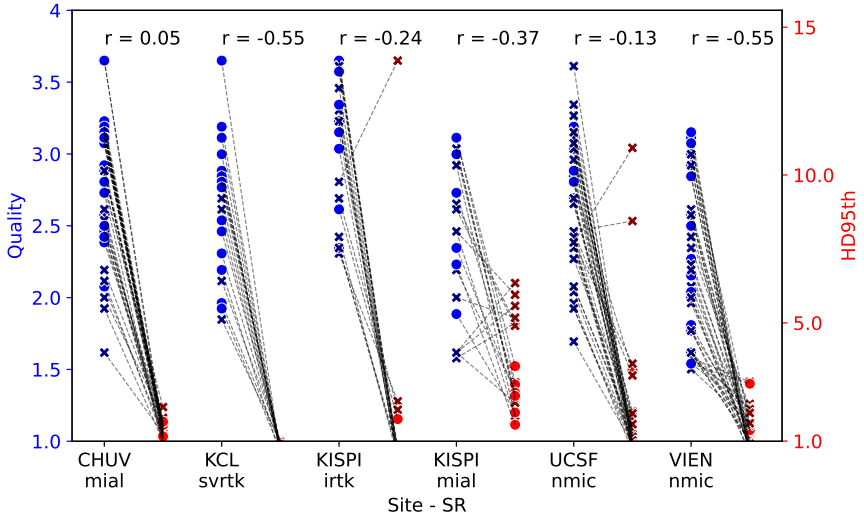


Figure A7.2: Correlation between visual quality scores and HD95 across different sites and super-resolution methods.

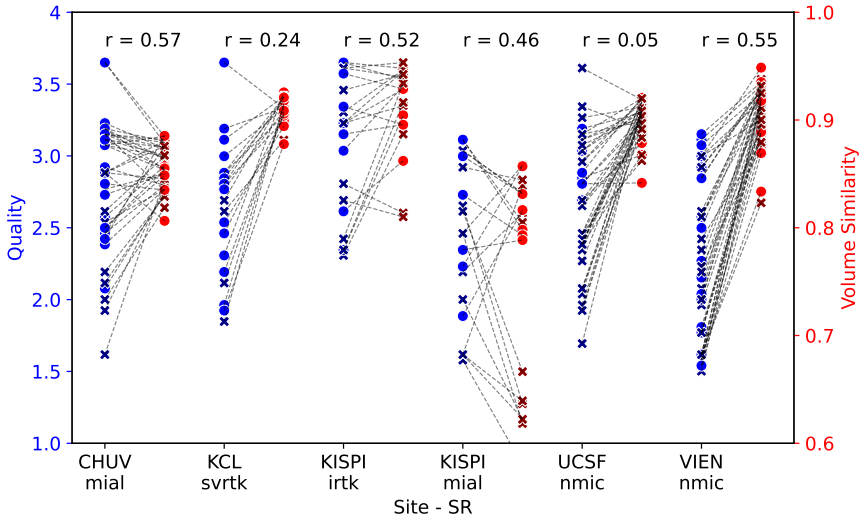


Figure A7.3: Correlation between visual quality scores and VS across different sites and super-resolution methods.

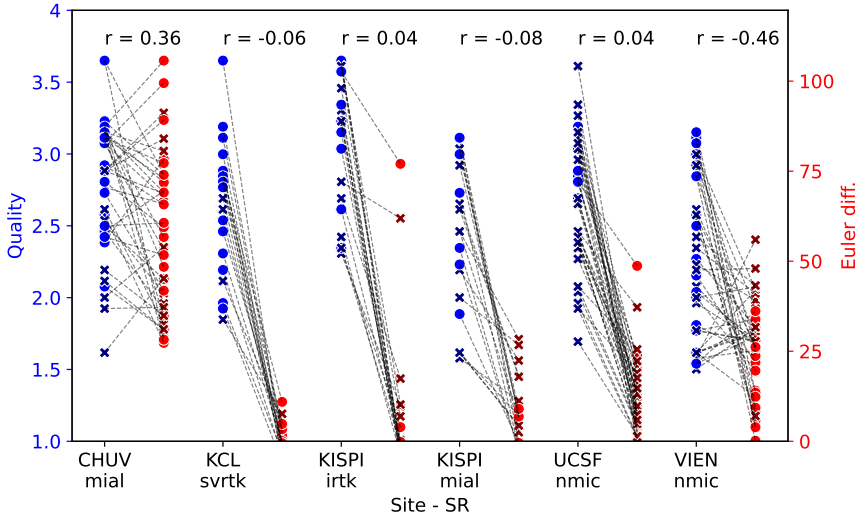
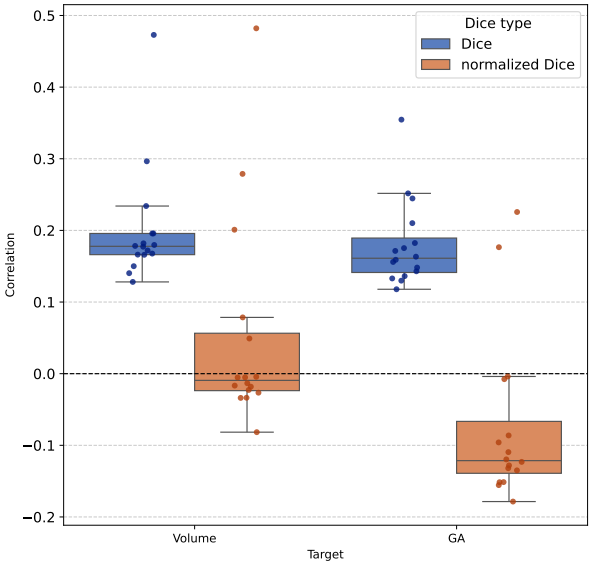


Figure A7.4: Correlation between visual quality scores and ED across different sites and super-resolution methods.

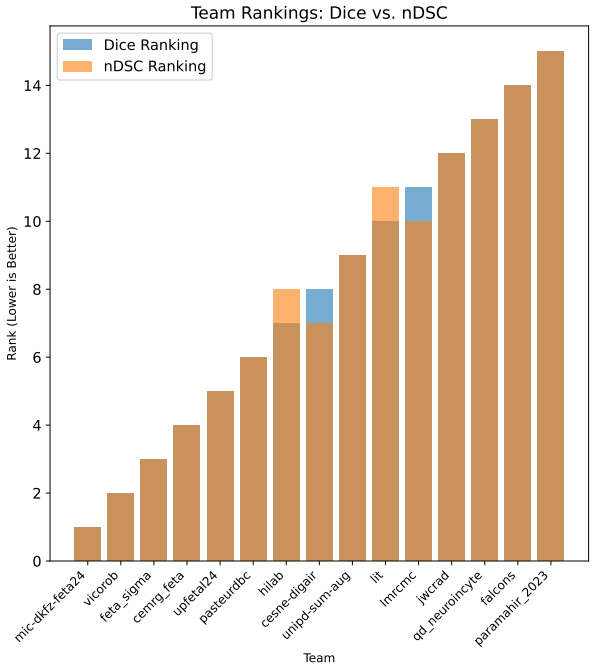
Appendix A8. Exploring normalized Dice coefficient

Correlation between Dice, Normalized Dice, Volume, and GA In Figure A8.1 (a), we present the distribution of Pearson correlation coefficients between each team's Dice score and the volume of the segmented label, averaged across all subjects for the 15 participating teams in the FeTA 2024 challenge. We observe that using the normalized Dice coefficient mitigates the known bias of the standard Dice metric toward larger volumes. However, normalized Dice does not eliminate the correlation with gestational age (GA). This residual association may be due to GA acting as a complex confounding factor, potentially interacting with other variables such as acquisition site and pathology, rather than directly influencing the metric.

Impact of Using Normalized Dice on Team Rankings In Figure A8.1 (b), we assess the stability of team rankings when using normalized Dice instead of the standard Dice score. Although normalized Dice reduces the volume-related bias, the correlation structure across teams appears relatively stable. Rankings remain largely consistent, with only two pairs of teams swapping positions. Notably, the top six teams maintain their original ranking when switching from Dice to normalized Dice, suggesting that the relative performance differences are robust to this normalization.



(a) Correlation coefficients between Dice/normalized Dice with volume and GA.



(b) Team-ranking stability using Dice vs. normalized Dice.

Figure A8.1: (a) Per-team Dice vs. volume/GA correlations. (b) Impact of normalization on team rankings.

Appendix A9. Qualitative examples of predictions

Qualitative Results This section of the supplementary materials presents qualitative comparisons of the predictions made by the top four teams in the FeTA 2024 challenge.

Figure A9.1 shows representative segmentation outputs for five subjects from the testing set, while Figure A9.2 highlights the corresponding segmentation errors. Overall, the visual differences between the top-performing models are minimal, indicating that all four algorithms deliver highly consistent and accurate segmentations in well-defined cases.

Figures A9.3 and A9.4 illustrate cases with the lowest Dice scores across the evaluated models. These difficult cases are typically associated with poor image quality (e.g., Subjects 2 and 3), or signs of abnormal fetal brain development (e.g., Subjects 0, 1, and 4). Such challenges underscore the limitations of current methods when dealing with low-quality inputs or atypical anatomy.

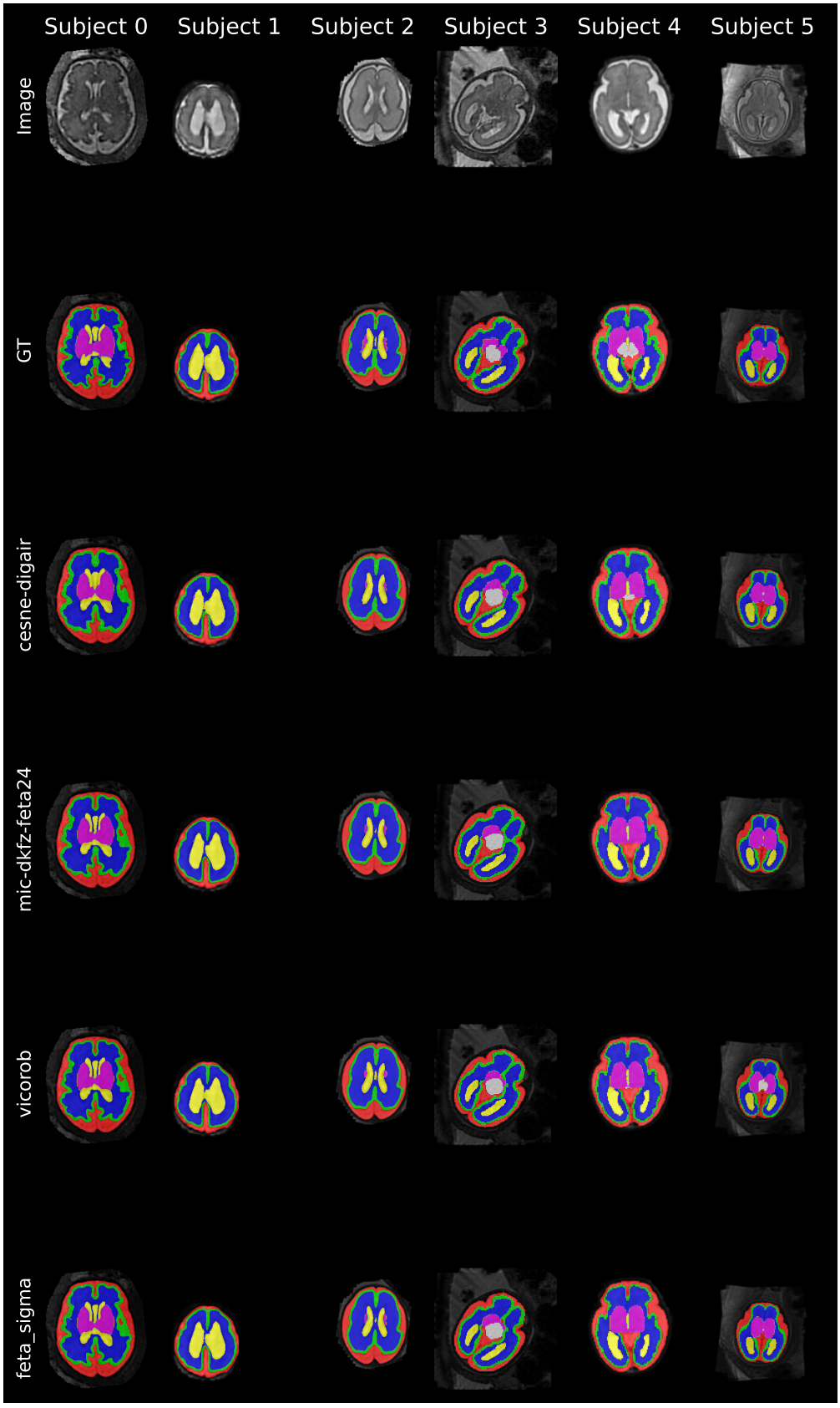


Figure A9.1: Segmentation results for five testing subjects produced by the top four teams in the FeTA 2024 challenge.

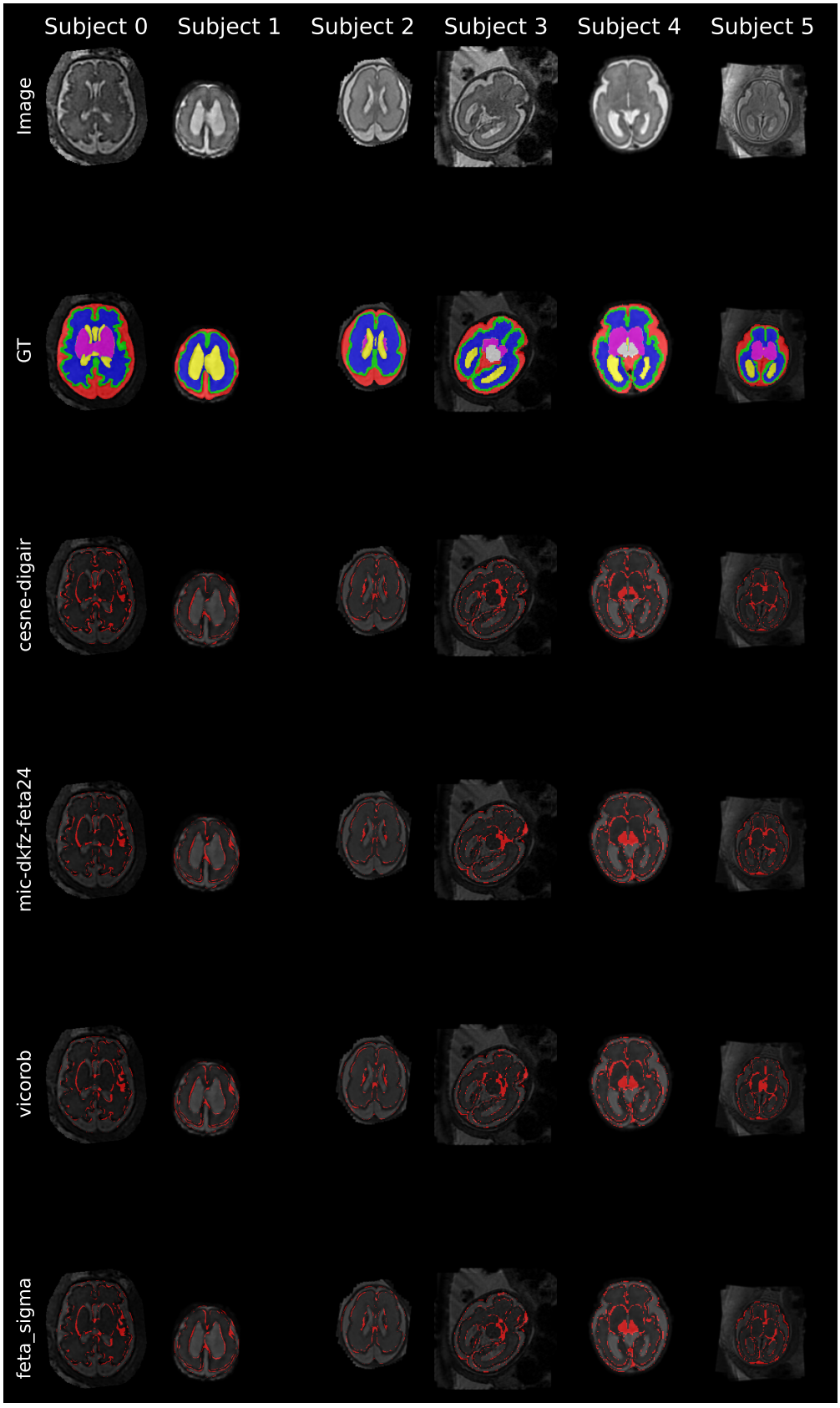


Figure A9.2: Segmentation errors for five testing subjects produced by the top four teams in the FeTA 2024 challenge. Red regions indicate voxels where predicted labels differ from the ground truth.

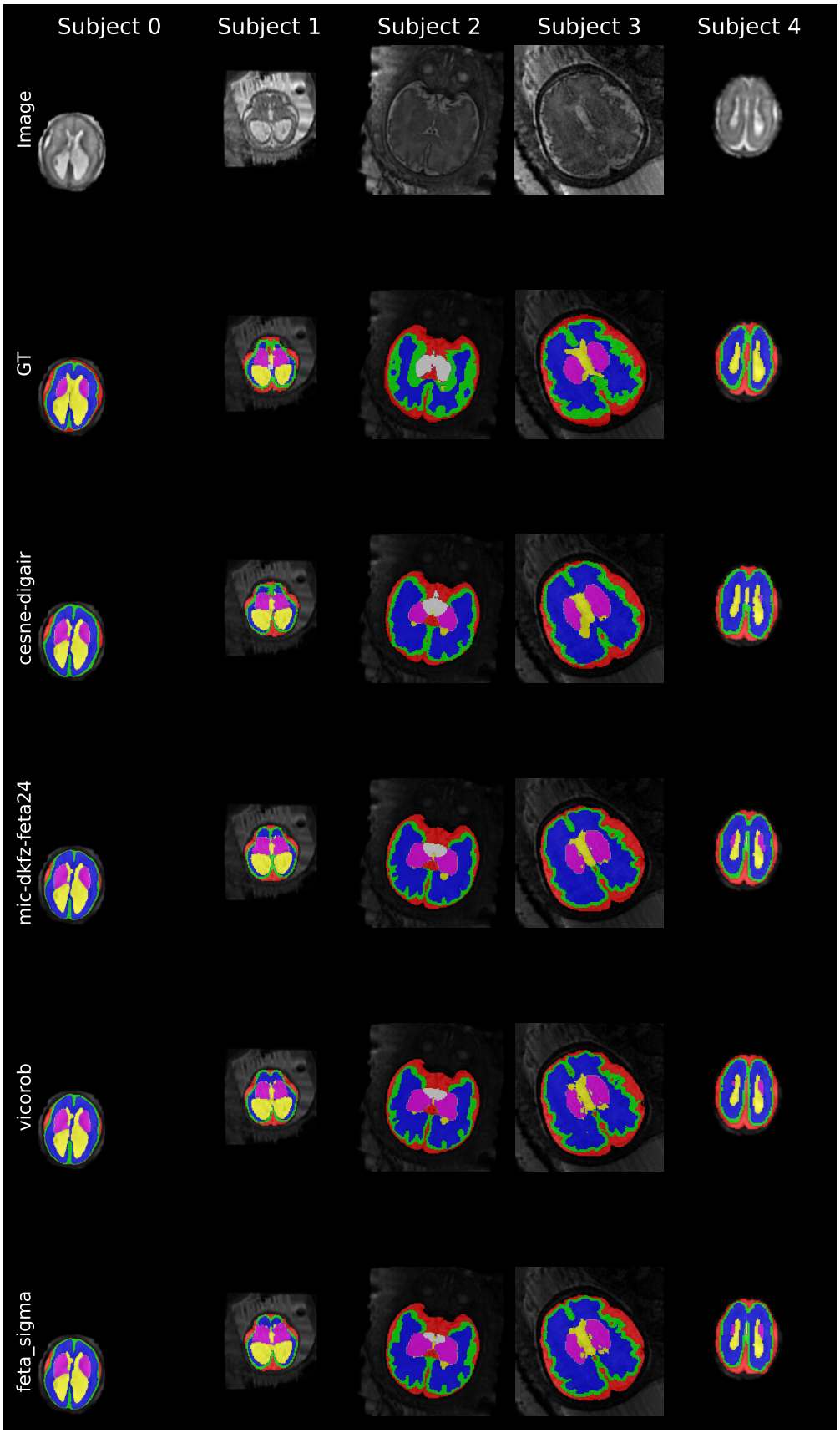


Figure A9.3: Segmentation results for five challenging testing subjects with the lowest Dice scores, produced by the top four teams in the FeTA 2024 challenge.

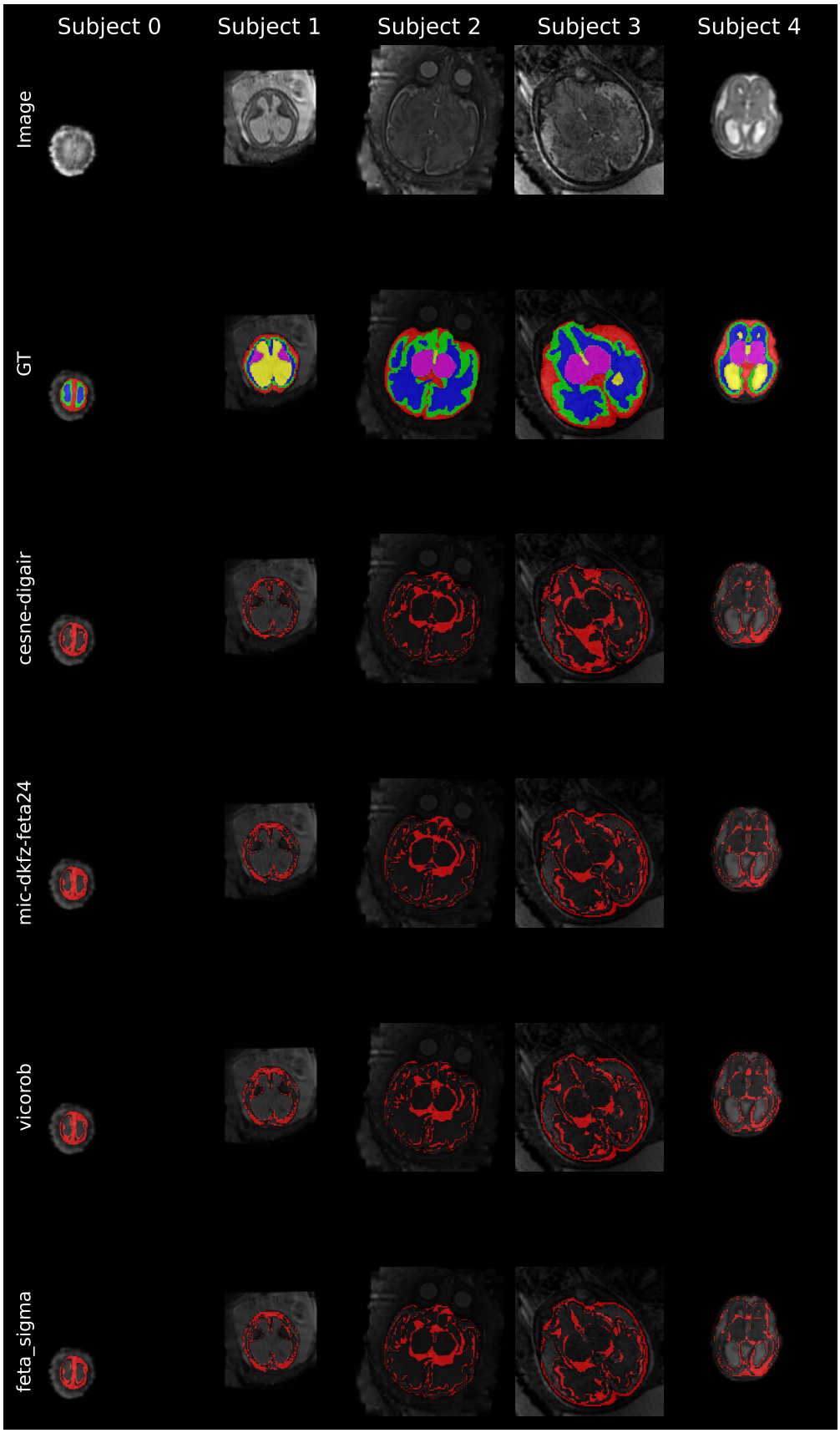


Figure A9.4: Segmentation errors for the five challenging testing subjects shown in Figure A9.3. Red voxels indicate mismatches between the predicted labels and ground truth.