Prenatal Diagnosis

WILEY

PRENATAL DIAGNOSIS

ORIGINAL ARTICLE OPEN ACCESS

Clinical Grading of Artificial Intelligence-Based 3D Fetal Brain Segmentations: A Cross-Vendor Evaluation of Deep Learning in Fetal Neuroimaging

Moska Aliasi[1] | Linde S. Hesse[2] | Madeleine K. Wyburd[2] | Maartje C. Snoep[1] | Renee M. Smit[1] | Ana I. L. Namburete[2] | Monique C. Haak[1]

[1]Department of Obstetrics and Fetal Medicine, Leiden University Medical Center, Leiden, the Netherlands | [2]Department of Computer Science, University of Oxford, Oxford, UK

Correspondence: Moska Aliasi (m.aliasi@lumc.nl)

Received: 6 June 2025 | **Revised:** 11 February 2026 | **Accepted:** 17 February 2026

ABSTRACT

Objective: To evaluate the performance of automated (sub)cortical fetal brain segmentation methods on a novel 3D ultrasound dataset acquired from a different vendor, and to introduce a clinician-focused visual evaluation framework complementary to the widely used Dice Similarity Coefficient (DSC).

Method: This cohort study included 270 volumes (141 fetuses, 19–26 + 6 weeks gestation). Deep learning models were applied to segment the cavum septum pellucidum et vergae (CSPV), lateral posterior ventricle horn (LPVH), choroid plexus (ChP), cerebellum (CBM) and cortical plate (CoP) on a new dataset acquired by a different ultrasound vendor. Segmentations were visually graded (1–4 = high to poor quality) based on predefined criteria. Grades were analyzed as "adequate" (1 + 2) or "inadequate" (3 + 4).

Results: CSPV, ChP and CBM showed the best segmentation grades (> 83.1% grade 1, > 90.5% adequate) and were robust across gestation. LPVH showed the lowest performance (73.9% adequate). Overall segmentation quality across all structures was high (87.2% adequate). Intra- and interobserver agreement was 90.1% and 82.1%–92.7%, respectively.

Conclusion: These deep-learning methods can reliably segment (sub)cortical structures when applied to a novel dataset acquired with a different ultrasound vendor, demonstrating robustness. Incorporating visual assessment alongside quantitative metrics provides insight into anatomical accuracy and clinical usability.

1 | Introduction

Central nervous system (CNS) abnormalities are among the most frequently diagnosed congenital malformations and are associated with significant long-term neurodevelopmental consequences [1]. Accurate prenatal detection is essential for counseling, pregnancy management, and postnatal planning. The International Society of Ultrasound in Obstetrics and Gynecology (ISUOG) recommends detailed assessment of multiple fetal brain structures during routine and targeted anomaly scans [2, 3], including evaluation of cortical folding patterns, as disrupted gyrification may signal underlying neurological impairment [4–7].

Three-dimensional (3D) ultrasound plays an important role in advanced neurosonography by enabling multiplanar image reconstruction and volumetric analysis. Compared to traditional two-dimensional (2D) imaging, 3D ultrasound offers more comprehensive visualization of fetal brain anatomy by allowing retrospective navigation through any desired imaging plane [3].

Summary

- What is known about this topic?
  - 3D ultrasound plays an important role in advanced neurosonography, but manual segmentation is time-consuming and requires significant expertise
  - Automated segmentation methods offer the potential to reduce both time and expertise required for detailed 3D analysis
  - Most models are trained and validated on single datasets, and their generalizability to a different ultrasound vendor or scanning protocol remains uncertain
- What does this study add?
  - Deep-learning methods can reliably segment fetal brain structures when applied to a novel dataset acquired with a different ultrasound vendor, demonstrating robustness across vendors.
  - By combining traditional quantitative metrics with clinician-focused visual grading, this study offers a more practical and anatomically meaningful evaluation of segmentation quality, highlighting readiness for real-world implementation in fetal neuroimaging workflows.

This is particularly advantageous when optimal acquisition of a specific 2D plane is hindered by fetal position, maternal habitus, or acoustic shadowing. As a result, 3D imaging enhances diagnostic confidence and consistency in evaluating fetal neuroanatomy. However, a critical limitation persists: manual segmentation of anatomical structures in 3D volumes is time-consuming and demands significant expertise, making it impractical for routine clinical use.

Artificial intelligence (AI), particularly deep learning, offers a promising avenue for automating segmentation of fetal brain structures [8–19]. These algorithms learn to recognize complex anatomical patterns from large imaging datasets and have demonstrated strong performance in research settings [20–23]. Segmentation accuracy is often evaluated using the Dice Similarity Coefficient (DSC), a widely used metric that quantifies the degree of overlap between the automated segmentation and manually labeled reference (ground-truth) [21, 24]. Automated segmentation has the potential to significantly reduce the time and expertise required for detailed 3D analysis, thereby streamlining clinical workflow and broadening access to advanced neursonographic assessments. However, most models are trained and validated on homogeneous datasets typically acquired from a single site, and their generalizability to different imaging conditions (such as alternative ultrasound vendors or acquisition protocols) remains largely untested, limiting their clinical adoption [19].

This study evaluates the performance of two automated 3D segmentation methods for fetal (sub)cortical brain structures on an independent dataset acquired using a different ultrasound system than those used for model development. In addition to conventional DSC-based evaluation, we introduce clinician-focused visual grading criteria to better capture anatomical accuracy and practical utility. Although DSC is a widely used measure, its limitations in reflecting anatomical correctness and clinical usability have been discussed in previous studies [25–27]. By combining quantitative and qualitative assessments, our aim is to explore real-world clinical applicability and highlight key considerations for integrating AI tools into prenatal neuroimaging workflows.

2 | Materials and Methods

This cohort study was conducted at the Leiden University Medical Center (LUMC), a tertiary care hospital in the Netherlands, and included two groups. The first group consisted of pregnancies with fetuses prenatally diagnosed with an isolated structural congenital heart disease (CHD), defined as the absence of other major malformations (including abnormalities of the brain). The second group was composed of uncomplicated healthy pregnant women recruited from midwifery practices. In the LUMC, all fetuses with a CHD are included in a fetal surveillance program , and as part of this program, a detailed neurosonography (including 3D evaluation) is performed every 4 weeks (Biobank Obstetrics—Congenital Heart Disease study). The second group consisted of healthy fetuses, selected based on a normal second-trimester anomaly scan and the absence of congenital abnormalities or dysmorphic features at birth. In this group, a detailed evaluation of the brain was performed every 4 weeks, serving as a control group for the CHD cases (Fetal Neurodevelopment study). Data from these studies are presented in previous publications by our research group [28–31]. Both the CHD and the control cases were grouped together for this study. Cases were included if at least one US examination was performed between 19 and 26 + 6 weeks of gestation. This age range was selected as the network was developed and trained on 3D-US images during this gestational age (GA) [21, 24].

All scans were performed on an APLIO i800 (Canon Medical Systems) using a PVT-675MVS three-dimensional abdominal transducer (frequency range 2–9 MHz). During each scanning session, detailed neurosonography was performed assessing four planes (axial, coronal, sagittal and parasagittal) by an expert sonographer (MA/MS/RS). Multiple 3D volumes were acquired at the transventricular and transcerebellar planes (axial) as defined by the ISUOG Practice Guidelines [2]. We awaited a period of absence of fetal movements before the acquisition of the 3D volumes. GA was based on the first-trimester US [32].

Numerical variables are presented as mean (standard deviation [SD]) or median (interquartile range [IQR]) and categorical data as frequencies (%).

2.1 | Selection Process and Alignment

At the outset of the analysis, a neurosonography expert (MA) conducted an initial visual review to select the highest-quality volume for each target brain structure from every scanning session. During each scanning session, multiple 3D ultrasound volumes were acquired, often capturing the same anatomical structures with varying image quality due to fetal position, movement, or acoustic shadowing. For further analysis, the

volume with the best visualization of the structure of interest was selected per scanning session. When optimal visualization of different structures was achieved in separate volumes, more than one volume from the same session was included. This approach reflects routine clinical practice, in which multiple acquisitions are obtained to optimize anatomical assessment within a single examination. Importantly, this selection strategy does not represent preferential inclusion of only the highest quality volumes across the full dataset. Instead, at least one representative volume per scanning session (and thus per fetus) was included, ensuring that each case contributed to the evaluation. Each selected volume underwent automated alignment using rigid registration to ensure a consistent anatomical orientation across datasets [33]. Manual corrections were made as needed to refine the accuracy of brain orientation. For each step of the workflow, the approximate expert time required, was recorded to provide insight into the practical feasibility of the evaluation process. Automated segmentation.

Two convolutional neural networks (CNNs) were applied to perform automated segmentation. For both CNNs, the input consisted of a 3D ultrasound volume. The models generated a corresponding 3D segmentation output, in which the target anatomical structures were delineated within the volume. Both models were originally trained and tested on datasets acquired using Philips ultrasound systems in fetuses between 18 and 26 + 6 weeks' gestation, collected as part of the Intergrowth-21st Fetal Growth Longitudinal Study [34].

2.2 | CNN 1: Subcortical Structures

This model developed by Hesse et al. [24], used a few-shot learning strategy to segment key subcortical structures from 3D US images. Specifically, the model segmented the cavum septum pellucidum et vergae (CSPV), the cerebrospinal fluid of the lateral posterior ventricle horn (LPVH), the choroid plexus (ChP) and the cerebellum (CBM), using only nine manually annotated volumes for training. Few-shot learning enables effective model generalization from minimal labeled data, which is particularly beneficial in fetal ultrasound where annotated 3D datasets are scarce. These structures were chosen based on both clinical importance in second trimester anomaly screening [2] and their distinct anatomical boundaries, which facilitate model learning. The CSP and its posterior extension, the cavum vergae (CV), were segmented as a combined structure due to their similar fluid-filled appearance on ultrasound [35]. The ChP, located within the posterior horn of the lateral ventricles, was included as both a landmark and a part of the total ventricular volume. This model achieved DSC scores of 0.78 for CSPV, 0.85 for LPVH and ChP, and 0.90 for CBM [24]. The DSC metric quantifies spatial overlap between predicted and ground-truth segmentations, ranging from 0 (no overlap) to 1 (perfect overlap) [36].

2.3 | CNN 2: Cortical Plate

The second model, developed by Wyburd et al., performed cortical plate (CoP) segmentation [21]. This CNN achieved a DSC of 0.81 ± 0.06 on an independent test set, enabling automated assessment of cortical development from 3D ultrasound volumes.

In this study, we evaluated the performance of both models on volumes acquired using Canon Medical Systems equipment, enabling assessment of model robustness when applied to data from a different ultrasound vendor. To maximize structural visibility, the LPVH, ChP and CoP were segmented in the hemisphere distal to the ultrasound transducer, where signal attenuation and acoustic shadowing artifacts from the fetal skull were reduced.

All 3D visualization and alignment procedures were performed using MATLAB (version 9.8) and MITK-Workbench. Preprocessing included rigid alignment, intensity normalization, and inspection of orientation consistency [33, 37].

2.4 | Evaluation Protocol

Following automated segmentation, each model's output was visually evaluated in 3D by an experienced ultrasound specialist (MA). The segmentation output consisted of a 3D volume with segmentation labels. Visual evaluation was performed by navigating through the complete 3D volume, allowing assessment of segmentation quality across all planes rather than on isolated 2D images. Segmentation quality grades were based on the overall anatomical correctness of the segmented structure throughout the entire 3D volume. The percentage deviation thresholds (< 20%, 20%–50%, > 50%) reflected qualitative estimates of the proportion of the segmented structure deviating from the expected anatomical boundaries across the volume, rather than measurements derived from a single slice or plane.

Each segmented structure within the selected volume was assessed according to predefined scoring criteria, resulting in four segmentation quality grades:

1. High quality: accurate delineation of the structure and its anatomical boundaries
2. Intermediate quality: minor boundary errors, defined as segmentation deviations of < 20% from the true tissue boundary
3. Low quality: major boundary errors present, with deviations of 20%–50% from the true tissue boundary
4. Poor quality: severe deviation (> 50%), incorrect structure identification, or complete failure to recognize and segment the target structure.

These scoring criteria were uniformly applied across all structures, including the CoP. For CoP segmentation to be considered correct, inclusion of the Sylvian fissure was necessary, as this is considered to be a key landmark. Figure 1 provides an overview of representative examples of each segmentation grade per structure, including anatomical definition. Segmentations rated as Grades 1 or 2 were deemed "adequate," while Grades 3 and 4 were classified as "inadequate".

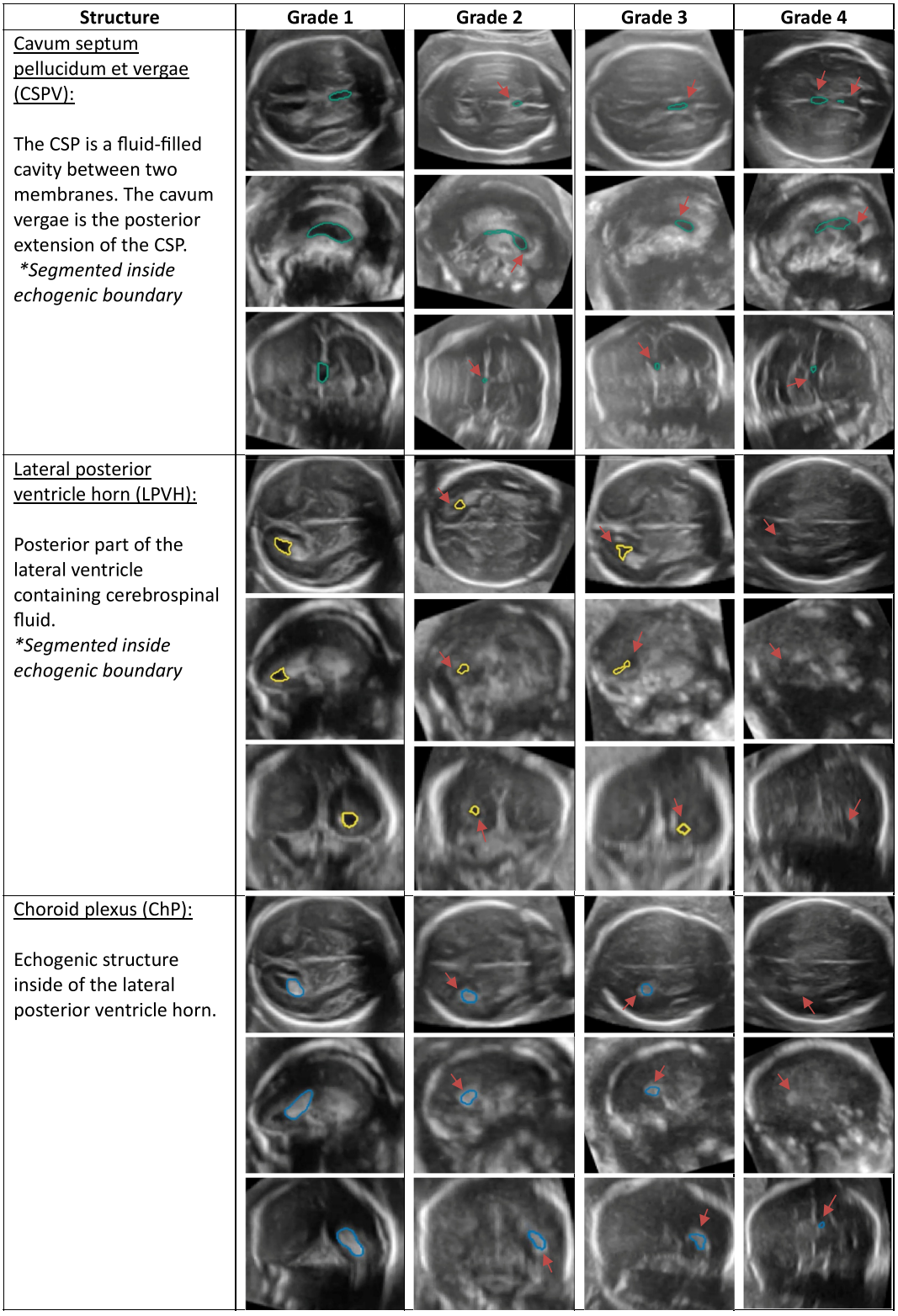

FIGURE 1 | Overview segmentation grades per structure, including anatomical definition The red arrow indicates where the algorithm over-segmented, under-segmented or failed to detect the structure of interest at all. All images are transabdominal scans. Three different planes are displayed for each grade and structure (from top to bottom: transventricular, sagittal, and coronal planes).

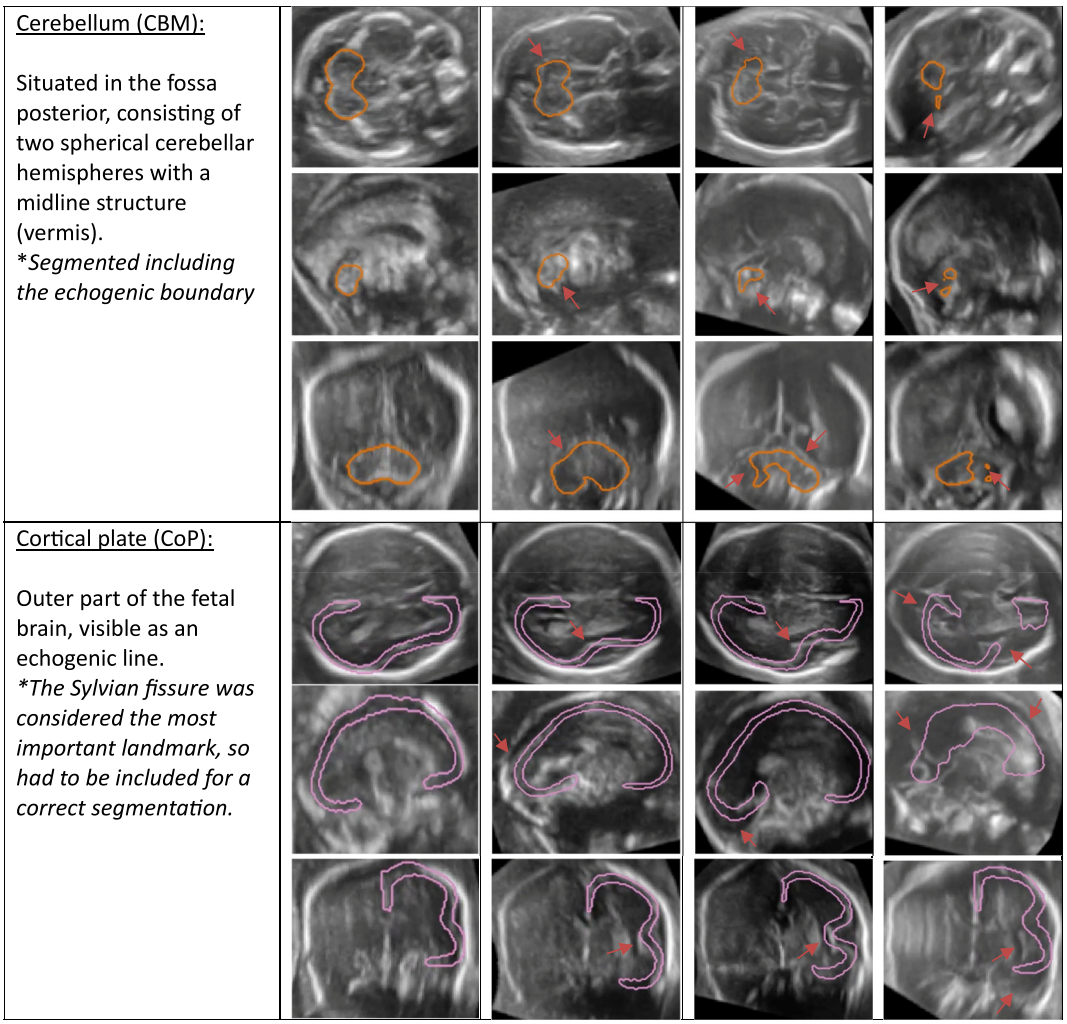

FIGURE 1 | (Continued)

To evaluate reproducibility, a randomly selected subset comprising 30% of all volumes was used to calculate the intra- and inter-observer agreement. Intra-observer agreement was determined by having the same observer (MA) reassess this subset twice at two separate time points. Subsequently, inter-observer agreement was assessed by two independent neurosonography experts (MS and RS), who independently graded the same set of sample volumes. Their assessments were compared with the initial grading performed by MA, resulting in two inter-observer agreement measures (MS vs. MA, and RS vs. MA). Agreement was defined as the percentage of volumes receiving identical segmentation scores. The sample included proportional representation of all (sub)cortical structures. Furthermore, the time required to complete each step of the segmentation and evaluation process was recorded.

3 | Results

A total of 1508 3D volumes (= 249 scanning sessions) were collected during the period January 2020—August 2022. Thirteen scanning sessions had to be excluded as none of the structures of interest were visible in the acquired volumes due to poor image quality (high maternal BMI, fetal movements). After initial visual assessment, the best volume(s) per scanning session was selected for further analysis. Ultimately, 270 volumes arising from 236 scanning sessions (= 141 subjects) were included in this study. In the majority of cases, all the structures of interest were visible in 1 volume (85.7%). After automated alignment of the volumes, manual realignment was performed in 158 volumes to ensure equivalent plane orientation. Figure 2 shows the inclusion process including the time required for each workflow step. The data distribution of the volumes per GA week is shown in Figure 3. Details of the baseline characteristics of the included subjects are depicted in Table 1. The majority of the included subjects (68.1%) had two US visits during pregnancy.

The CSPV, ChP and CBM showed the best segmentation performance (segmentation score 1 > 83.1% and adequate segmentation > 90.5%), as shown in Table 2. These segmentation scores were robust to advancing GA, even at the end of the second trimester with adequate segmentation > 82.6% (Figure 4). The LPVH had the lowest segmentation overall performance (score 1 63.0%, score 2 10.9%), which declined throughout gestation with the lowest performance at 25 weeks

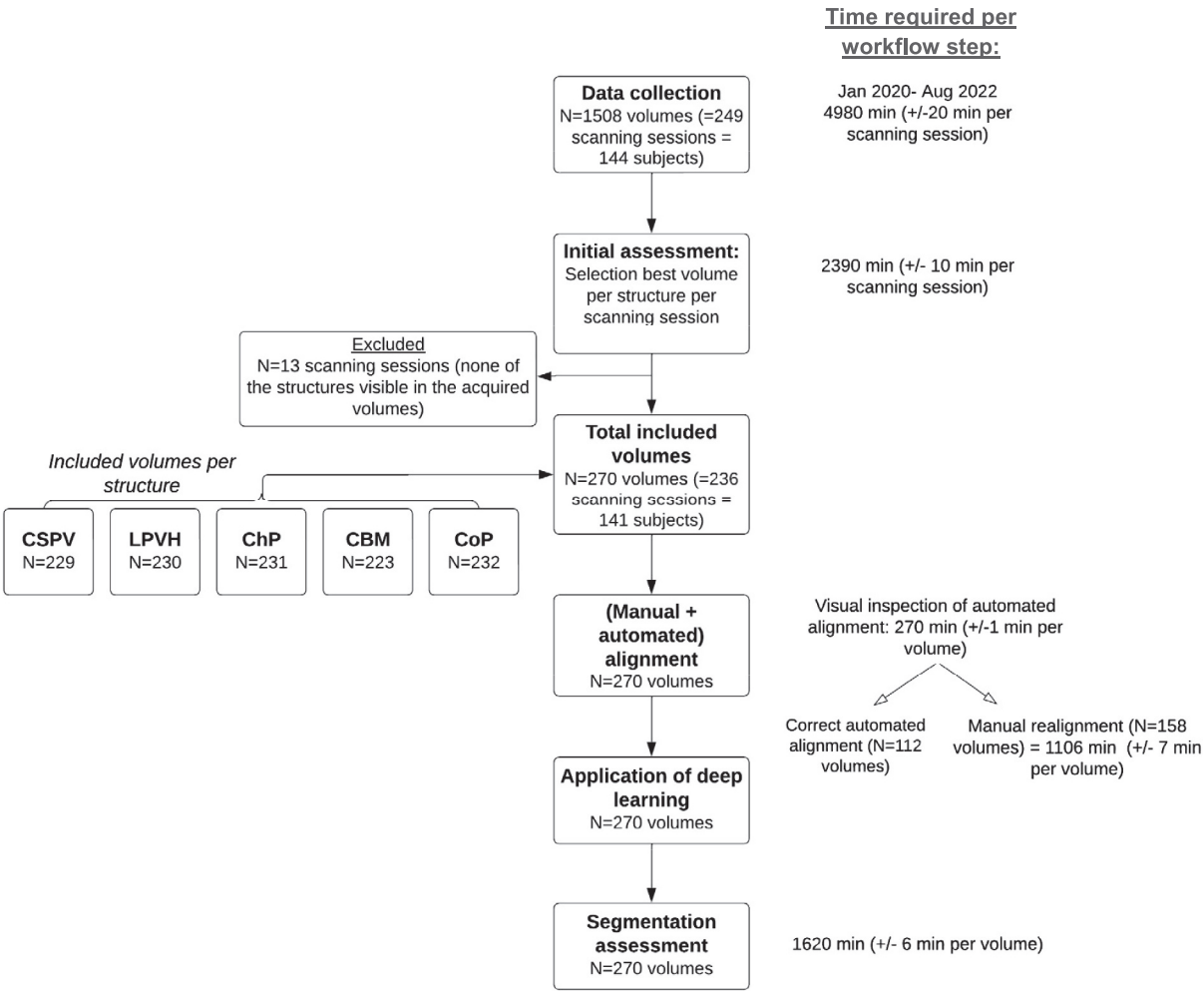

FIGURE 2 | Flowchart of the included volumes. * Time required indicates the approximate expert time investment needed for each step of the workflow.

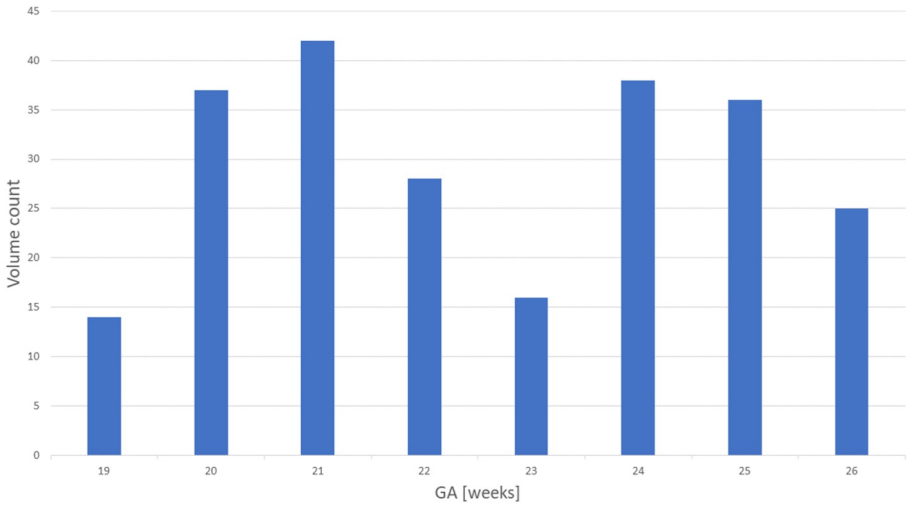

FIGURE 3 | Data distribution per GA (weeks). The x-axis represents gestational age (weeks), and the y-axis indicates the number of evaluated volumes per gestational age.

gestation (adequate segmentation 56.8%). Upon visual inspection of the data, it became apparent that the algorithm encountered challenges in distinguishing between cerebrospinal fluid and the ChP within the lateral ventricle, as illustrated in Figure 1. The automated segmentation analysis of the CoP showed high segmentation performance especially early in the pregnancy with 100% adequate segmentation rates at 20 and 22 weeks of gestation. However, the performance dropped

slightly toward the end of the gestational period we examined. The overall adequate segmentation percentage of all the structures combined, was high (87.2%); this was consistent during pregnancy.

The intra-observer agreement was 90.1% for all the structures combined (range 83.1%–94.8%). The inter-observer agreement was 82.7% and 82.1% (range 70.1%–90.1% depending on the structure) (Table 3).

4 | Discussion

This second-trimester post hoc clinical evaluation study demonstrates that deep learning-based automated segmentation methods achieve high performance in segmenting (sub)cortical fetal brain structures. While these algorithms showed high DSC in previous studies [21, 24], this study addressed its known limitations, such as insensitivity to anatomical shape errors, by incorporating and proposing a visual grading. The fact that these models could be applied to a novel, unseen dataset acquired with a different vendor and operator team than those used during the training phase, is promising with regard to future clinical evaluation, as it provides insight into model robustness beyond the original training setting. While this study demonstrates robustness of the segmentation models when applied to an ultrasound system from a different vendor, broader generalizability across ultrasound systems requires further validation.

AI is increasingly explored in fetal imaging, particularly for detecting CHD and central nervous system abnormalities and improving automatic biometry measurements [38–43]. However, despite these advancements, these AI models are not yet implemented in routine clinical practice. As a result, (sub)cortical structures are still predominantly evaluated using 2D ultrasound, limiting the accuracy and efficiency of neurosonographic evaluations. Few studies have reported on automated subcortical US segmentation methods, as ground-truth labels are needed for training, which are scarce and not easy to obtain due to the challenges of manual annotations. Some of these studies face limiting factors such as the automated analysis being restricted to only one subcortical structure [44], segmentation only on 2D US planes [45, 46], or the use of weakly labeled volumes obtained from MRI atlas-based segmentation for testing and training [47]. Additionally, many studies analyzing 3D US subcortical brain development used the multiplanar or Virtual Organ Computer-aided Analysis (VOCAL) segmentation software (GE Healthcare), which still requires manual measurements for calculating volumetric data and is therefore not considered an automated method [48–54].

This study illustrates that these automated methods can accurately assess anatomical structures when applied to a new dataset of a different US manufacturer, demonstrating their potential to detect errors in anatomical shapes beyond mere volumetric data. For example, segmentation of the cortical plate may achieve a high DSC overlap, but not represent the shape accurately. As each volume was independently visually assessed for correctness, this study underscores the robustness of these algorithms in providing detailed and precise anatomical segmentation, which is crucial for clinical applications. Our proposed method to evaluate deep learning algorithms requires no new manual labeling, making our approach relatively simple and quick to implement. Manual labeling is extremely time-consuming, requires expert knowledge and is highly user-dependent. Therefore, methods that rely on manual labels to

TABLE 1 | Baseline characteristics of the included subjects.

***N* = 141 subjects; 236 scanning sessions**	
Maternal age, mean (sD)	32.0 (4.1)
BMI (kg/m^2), median (IQR)	23.1 (21.6–26.0)
CHD, *n* (%)	77 (54.6%)
Controls, *n* (%)	64 (45.4%)
Scanning sessions per subject, *n* (%)	
1	44 (31.2%)
2	96 (68.1%)
Included volumes per scanning sessions, *n* (%)	
1	202 (85.7%)
2	34 (14.3%)

Abbreviations: BMI, body mass index; CHD, congenital heart disease.

TABLE 2 | Overview of assessment of the included volumes per (sub)cortical structure.

	CSPV *N* = 229	**LPVH *N* = 230**	**ChP *N* = 231**	**CBM *N* = 223**	**CoP *N* = 232**	**Total *N* = 1145**
Visual assessment automated segmentation, *n* = volume (%)						
Grade:						
1.High (adequate)	193 (84.3)	145 (63.0)	192 (83.1)	187 (83.9)	180 (77.6)	897 (78.4)
2.Intermediate (adequate)	15 (6.5)	25 (10.9)	17 (7.3)	20 (9.0)	25 (10.8)	102 (8.9)
3.Low (inadequate)	14 (6.1)	27 (11.7)	11 (4.8)	13 (5.8)	20 (8.6)	85 (7.4)
4.Bad (inadequate)	7 (3.1)	33 (14.3)	11 (4.8)	3 (1.3)	7 (3.0)	61 (5.3)
Adequate segmentation, %	90.8	73.9	90.5	92.9	88.4	87.2
Inadequate segmentation, %	9.2	26.4	9.5	7.1	11.6	12.8

Abbreviations: CBM, cerebellum; ChP, choroid plexus; CoP, cortical plate; CSPV, cavum septum pellucidum et vergae; LPVH, lateral posterior ventricle horn.

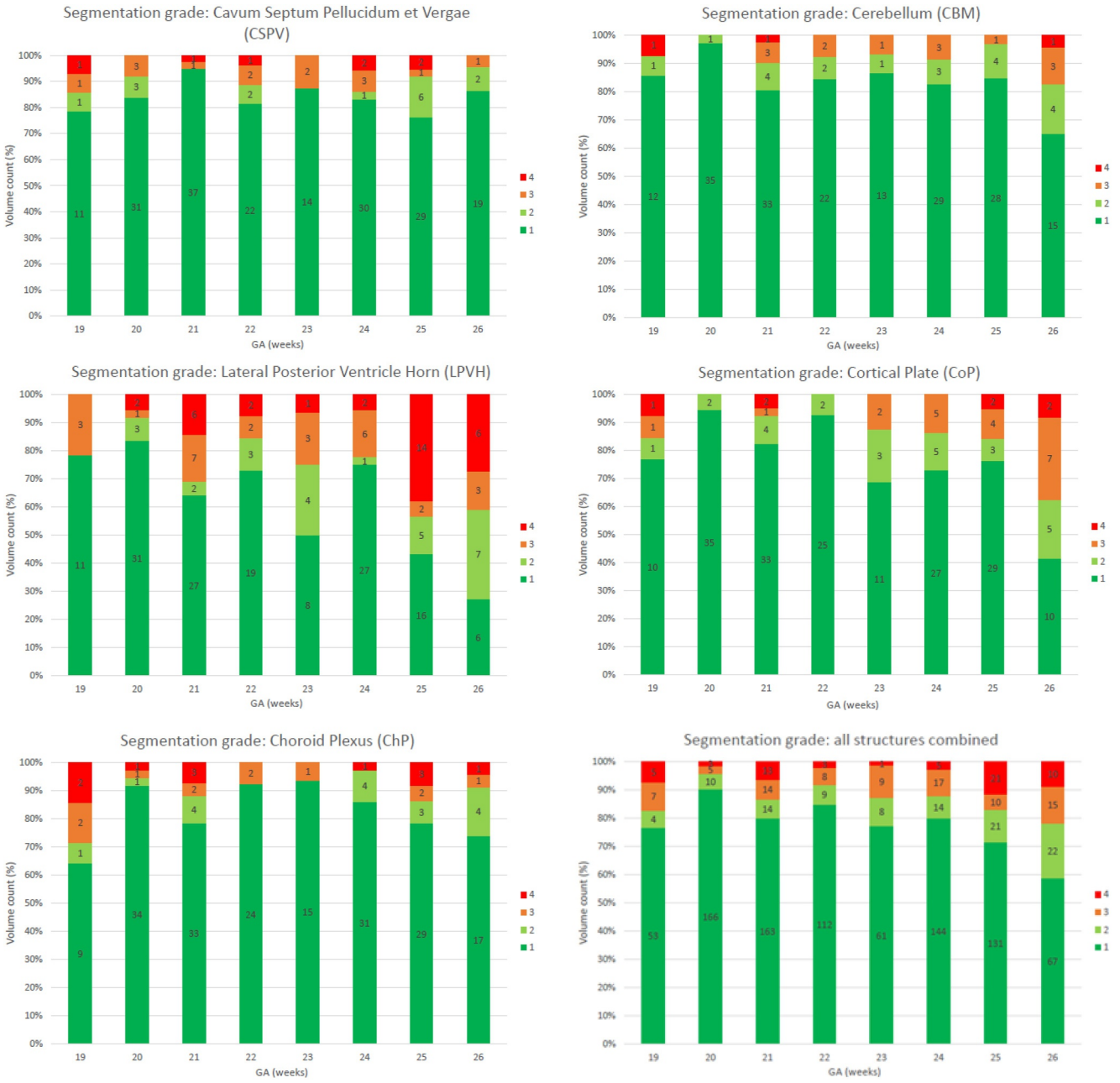

FIGURE 4 | Segmentation grade of the automated (sub)cortical segmentation per individual structure and all structures combined. Segmentation quality was visually graded on a four-point scale (Grade 1: high quality; Grade 2: intermediate quality; Grade 3: low quality; Grade 4: poor quality). The x-axis represents gestational age (weeks), and the y-axis indicates the proportion of volumes assigned to each segmentation grade.

TABLE 3 | Overview of the intra- and inter-observer agreement per (sub) cortical structure.

	CSPV $N = 76$[a]	LPVH $N = 77$[a]	ChP $N = 77$[a]	CBM $N = 68$[a]	CoP $N = 77$[a]	Total
Intra-observer agreement	94.7%	83.1%	94.8%	98.7%	88.3%	90.1%
Inter-observer agreement (1)	85.5%	77.9%	90.1%	89.7%	70.1%	82.7%
Inter-observer agreement (2)	84.2%	72.7%	88.3%	88.2%	77.9%	82.1%
Overall agreement	88.1%	77.9%	91.1%	92.2%	78.7%	85.0%

Note: Intra-observer agreement reflects grading by the same observer (MA) of the same subset at two separate time points. Inter-observer agreement reflects agreement of the same subset between the primary observer (MA) and two additional observers (MS and RS), reported separately as inter-observer agreement (1) and (2), respectively.
Abbreviations: CBM, cerebellum; ChP, choroid plexus; CoP, cortical plate; CSPV, cavum septum pellucidum et vergae; LPVH, lateral posterior ventricle horn.
[a]N = total volumes in which the individual structure was visible (total included volumes for the intra- and inter-observer agreement, $n = 81$).

assess performance (e.g., using DSC overlap) are ultimately limited by these factors. As manual annotations are not required, time can be saved as data selection and manual inspection of the volumes took only a few minutes on average. Possibly, in the future, this process can be fully automated and ideally integrated into US machines, requiring minimal human interaction. Although the evaluated segmentation models demonstrate robust performance on second-trimester datasets, their direct use for detecting subtle cortical anomalies or for clinical decision-making in cases of pathology remains untested. Potential use cases include integration into structured checklists for quality control of image acquisition, visual comparison of automated segmentations with raw ultrasound data to assist anatomical interpretation, and decision-support prompts that highlight regions requiring closer expert inspection. Such applications may improve consistency, but their use remains dependent on expert clinical judgment. Future studies should specifically evaluate these use cases in both normal and abnormal fetal brain development to determine their practical value in clinical settings.

The segmentation of the CSPV, ChP, CBM and CoP showed high performance. This can readily be explained by the fact that all three structures have a clear boundary due to the contrast between the structure of interest and the surrounding tissues. In contrast to the LPVH, which showed the lowest performance due to difficulties in distinguishing between cerebrospinal liquor and ChP in the lateral ventricle, resulting in larger prediction errors. Differences in BMI and algorithm performance were not assessed in this study; however, such differences can typically explain variations in performance. All included structures were derived from the same volumes/cases, and thus similar BMI distributions.

Since this study evaluated deep learning model performance from both anatomical and clinical perspectives, segmentation grading was based solely on visual inspection. While this could be considered a limitation, visual assessment provides critical insight into segmentation usability that quantitative metrics alone cannot capture. While we acknowledge the importance of quantitative metrics for segmentation evaluation, this study offers an alternative method that is more time-efficient and provides insights from a clinical point of view and therefore goes beyond mere pixel-level evaluation. Segmentation performance was evaluated on selected volumes with adequate image quality, as multiple acquisitions were obtained per scanning session and the best volume per structure within each session was chosen for analysis. The grading of the segmentations in this study was standardized by the use of pre-defined criteria; however, as evaluation was based on visual assessment, the subjectivity of the expert may have been attributed to the scoring. However, as the inter-observer agreement was relatively high and equivalent (82.1%–82.7%) between the two sonographers that performed the inter-observer analyses, the role of subjectivity seems limited. As only isolated CHD cases and healthy fetuses were included in this study, the performance of these algorithms in abnormal cases was not assessed. The next step would be to evaluate these automated methods in abnormal neurodevelopmental cases to comprehensively determine their clinical utility.

5 | Conclusion

This study demonstrates that novel deep learning 3D models can successfully segment (sub)cortical structures in a novel, unseen dataset from a different US vendor, accurately capturing anatomical boundaries. By integrating visual grading criteria with traditional quantitative metrics, we provide an evaluation framework for assessing segmentation quality from a clinician's perspective. This approach bridges the gap between technical performance metrics and clinical relevance, as anatomical correctness and shape are critical for clinical usability and real-world application. These automated segmentation methods may support prenatal neurosonography workflows in the future through visual comparison, support of anatomical interpretation and decision-support. Moving forward, these automated methods should be compared across multiple ultrasound vendors and evaluated in abnormal neurodevelopmental cases to assess their universality and performance differences. We have presented a methodology for performing such evaluations without additional manual labels. The remaining aspects to be addressed include the need for expert assessment for data selection and segmentation quality evaluation, indicating that these algorithms cannot yet be applied independently without expert visual anatomical review. Future studies should address imaging and acquisition challenges accompanying third trimester ultrasonography, as crucial neurodevelopmental changes occur during this phase. Efforts should be made to facilitate the integration of segmentation algorithms into ultrasound devices, leading to the utilization of AI in daily clinical practice.

Acknowledgments

The authors have nothing to report.

Funding

This work was funded by Canon Medical Systems Nederland. The funder played no role in study design, data collection, analysis, decision to publish, or preparation of the manuscript.

Ethics Statement

Appropriate ethical standards were followed while conducting this study.

Consent

All participants provided informed consent and could therefore be enrolled in this study. The local ethics committee Leiden-Den Haag-Delft approved both the Biobank Obstetrics - Congenital Heart Disease study (data of approval: 14–04-2020, reference number: B19.060) and Fetal Neurodevelopment study (data of approval: 27–07-2020, reference number: P20.037).

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

References

1. H. Dolk, M. Loane, and E. Garne, "The Prevalence of Congenital Anomalies in Europe," *Advances in Experimental Medicine and Biology* 686 (2010): 349–364, https://doi.org/10.1007/978-90-481-9485-8_20.

2. G. Malinger, D. Paladini, K. K. Haratz, A. Monteagudo, G. L. Pilu, and I. E. Timor-Tritsch, "ISUOG Practice Guidelines (Updated): Sonographic Examination of the Fetal Central Nervous System. Part 1: Performance of Screening Examination and Indications for Targeted Neurosonography," *Ultrasound in Obstetrics and Gynecology* 56, no. 3 (2020): 476–484, https://doi.org/10.1002/uog.22145.

3. D. Paladini, G. Malinger, R. Birnbaum, et al., "ISUOG Practice Guidelines (Updated): Sonographic Examination of the Fetal Central Nervous System. Part 2: Performance of Targeted Neurosonography," *Ultrasound in Obstetrics and Gynecology* 57, no. 4 (2021): 661–671, https://doi.org/10.1002/uog.23616.

4. G. Malinger, D. Lev, and T. Lerman-Sagie, "Abnormal Sulcation as an Early Sign for Migration Disorders," *Ultrasound in Obstetrics and Gynecology* 24, no. 7 (2004): 704–705, https://doi.org/10.1002/uog.1795.

5. R. Guerrini, W. B. Dobyns, and A. J. Barkovich, "Abnormal Development of the Human Cerebral Cortex: Genetics, Functional Consequences and Treatment Options," *Trends in Neurosciences* 31, no. 3 (2008): 154–162, https://doi.org/10.1016/j.tins.2007.12.004.

6. D. M. Andrade, "Genetic Basis in Epilepsies Caused by Malformations of Cortical Development and in Those With Structurally Normal Brain," *Human Genetics* 126, no. 1 (2009): 173–193, https://doi.org/10.1007/s00439-009-0702-1.

7. V. Fernández, C. Llinares-Benadero, and V. Borrell, "Cerebral Cortex Expansion and Folding: What Have We Learned?," *EMBO Journal* 35, no. 10 (2016): 1021–1044, https://doi.org/10.15252/embj.201593701.

8. A. Esteva, A. Robicquet, B. Ramsundar, et al., "A Guide to Deep Learning in Healthcare," *Nature Medicine* 25, no. 1 (2019): 24–29, https://doi.org/10.1038/s41591-018-0316-z.

9. R. Ramirez Zegarra and T. Ghi, "Use of Artificial Intelligence and Deep Learning in Fetal Ultrasound Imaging," *Ultrasound in Obstetrics and Gynecology* 62, no. 2 (2023): 185–194, https://doi.org/10.1002/uog.26130.

10. S. Liu, Y. Wang, X. Yang, et al., "Deep Learning in Medical Ultrasound Analysis: A Review," *Engineering* 5, no. 2 (2019): 261–275, https://doi.org/10.1016/j.eng.2018.11.020.

11. J. Egger, C. Gsaxner, A. Pepe, et al., "Medical Deep Learning—A Systematic Meta-Review," *Computer Methods and Programs in Biomedicine* 221 (2022): 106874, https://doi.org/10.1016/j.cmpb.2022.106874.

12. A. Esteva, B. Kuprel, R. A. Novoa, et al., "Dermatologist-Level Classification of Skin Cancer With Deep Neural Networks," *Nature* 542, no. 7639 (2017): 115–118, https://doi.org/10.1038/nature21056.

13. F. Dhombres, J. Bonnard, K. Bailly, P. Maurice, A. T. Papageorghiou, and J. M. Jouannic, "Contributions of Artificial Intelligence Reported in Obstetrics and Gynecology Journals: Systematic Review," *Journal of Medical Internet Research* 24, no. 4 (2022): e35465, https://doi.org/10.2196/35465.

14. R. M. Jones, A. Sharma, R. Hotchkiss, et al., "Assessment of a Deep-Learning System for Fracture Detection in Musculoskeletal Radiographs," *npj Digital Medicine* 3, no. 1 (2020): 144, https://doi.org/10.1038/s41746-020-00352-w.

15. L. Jin, J. Yang, K. Kuang, et al., "Deep-Learning-Assisted Detection and Segmentation of Rib Fractures From CT Scans: Development and Validation of FracNet," *EBioMedicine* 62 (2020): 103106, https://doi.org/10.1016/j.ebiom.2020.103106.

16. M. Madani, M. M. Behzadi, and S. Nabavi, "The Role of Deep Learning in Advancing Breast Cancer Detection Using Different Imaging Modalities: A Systematic Review," *Cancers (Basel)* 14, no. 21 (2022): 5334, https://doi.org/10.3390/cancers14215334.

17. C. Zhang, X. Sun, K. Dang, et al., "Toward an Expert Level of Lung Cancer Detection and Classification Using a Deep Convolutional Neural Network," *Oncologist* 24, no. 9 (2019): 1159–1165, https://doi.org/10.1634/theoncologist.2018-0908.

18. J. G. Nam, S. Park, E. J. Hwang, et al., "Development and Validation of Deep Learning-Based Automatic Detection Algorithm for Malignant Pulmonary Nodules on Chest Radiographs," *Radiology* 290, no. 1 (2019): 218–228, https://doi.org/10.1148/radiol.2018180237.

19. L. Drukker, J. A. Noble, and A. T. Papageorghiou, "Introduction to Artificial Intelligence in Ultrasound Imaging in Obstetrics and Gynecology," *Ultrasound in Obstetrics and Gynecology* 56, no. 4 (2020): 498–505, https://doi.org/10.1002/uog.22122.

20. L. S. Hesse and A. I. Namburete, eds., "Improving u-net Segmentation With Active Contour Based Label Correction," in *Medical Image Understanding and Analysis: 24Th Annual Conference, MIUA 2020* (Springer, 2020): July 15-17, 2020, Proceedings 24.

21. M. K. Wyburd, M. Jenkinson, and A. I. L. Namburete, eds., Cortical Plate Segmentation Using CNNs in 3D Fetal Ultrasound2020 (Springer International Publishing.

22. F. Moser, R. Huang, A. T. Papageorghiou, B. W. Papież, and A. I. Namburete, eds., "Automated Fetal Brain Extraction From Clinical Ultrasound Volumes Using 3D Convolutional Neural Networks," in *Medical Image Understanding and Analysis: 23Rd Conference, MIUA 2019, Liverpool, UK, July 24–26, 2019, Proceedings 23* (Springer, 2020).

23. L. Venturini, A. T. Papageorghiou, J. A. Noble, and A. I. Namburete, eds., *Uncertainty Estimates as Data Selection Criteria to Boost omni-supervised Learning. Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23Rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23* (Springer, 2020).

24. L. S. Hesse, M. Aliasi, F. Moser, et al., "Subcortical Segmentation of the Fetal Brain in 3D Ultrasound Using Deep Learning," *NeuroImage* 254 (2022): 119117, https://doi.org/10.1016/j.neuroimage.2022.119117.

25. D. Müller, I. Soto-Rey, and F. Kramer, "Towards a Guideline for Evaluation Metrics in Medical Image Segmentation," *BMC Research Notes* 15, no. 1 (2022): 210, https://doi.org/10.1186/s13104-022-06096-y.

26. M. Yeung, L. Rundo, Y. Nan, E. Sala, C.-B. Schönlieb, and G. Yang, "Calibrating the Dice Loss to Handle Neural Network Overconfidence for Biomedical Image Segmentation," *Journal of Digital Imaging* 36, no. 2 (2023): 739–752, https://doi.org/10.1007/s10278-022-00735-3.

27. J. Bertels, D. Robben, D. Vandermeulen, and P. Suetens, eds., Optimization with Soft Dice Can Lead to a Volumetric Bias. Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries; 2020 2020//, Springer International Publishing.

28. S. M. P. Everwijn, A. I. L. Namburete, N. van Geloven, et al., "The Association Between Flow and Oxygenation and Cortical Development in Fetuses With Congenital Heart Defects Using a Brain-Age Prediction Algorithm," *Prenatal Diagnosis* 41, no. 1 (2021): 43–51, https://doi.org/10.1002/pd.5813.

29. S. M. P. Everwijn, A. I. L. Namburete, N. van Geloven, et al., "Cortical Development in Fetuses With Congenital Heart Defects Using an Automated brain-age Prediction Algorithm," *Acta Obstetricia et Gynecologica Scandinavica* 98, no. 12 (2019): 1595–1602, https://doi.org/10.1111/aogs.13687.

30. S. M. Everwijn, J. F. van Bohemen, N. van Geloven, et al., "Serial Neurosonography in Fetuses With Congenital Heart Defects Shows Mild Delays in Cortical Development," *Prenatal Diagnosis* 41, no. 13 (2021): 1649–1657, https://doi.org/10.1002/pd.6038.

31. F. A. R. Jansen, E. W. van Zwet, S. M. P. Everwijn, et al., "Fetuses With Isolated Congenital Heart Defects Show Normal Cerebral and Extracerebral Fluid Volume Growth: A 3D Sonographic Study in the Second and Third Trimester," *Fetal Diagnosis and Therapy* 45, no. 4 (2019): 212–220, https://doi.org/10.1159/000488674.

32. L. J. Salomon, Z. Alfirevic, C. M. Bilardo, et al., "ISUOG Practice Guidelines: Performance of first-trimester Fetal Ultrasound Scan," *Ultrasound in Obstetrics and Gynecology* 41, no. 1 (2013): 102–113, https://doi.org/10.1002/uog.12342.

33. F. Moser, R. Huang, B. W. Papież, and A. I. L. Namburete, "BEAN: Brain Extraction and Alignment Network for 3D Fetal Neurosonography," *NeuroImage* 258 (2022): 119341, https://doi.org/10.1016/j.neuroimage.2022.119341.

34. A. T. Papageorghiou, E. O. Ohuma, D. G. Altman, et al., "International Standards for Fetal Growth Based on Serial Ultrasound Measurements: The Fetal Growth Longitudinal Study of the INTERGROWTH-21st Project," *Lancet (London, England)* 384, no. 9946 (2014): 869–879, https://doi.org/10.1016/s0140-6736(14)61490-2.

35. D. M. Sherer, M. Sokolovski, M. Dalloul, P. Santoso, J. Curcio, and O. Abulafia, "Prenatal Diagnosis of Dilated Cavum Septum Pellucidum Et Vergae," *American Journal of Perinatology* 21, no. 5 (2004): 247–251, https://doi.org/10.1055/s-2004-829869.

36. K. H. Zou, S. K. Warfield, A. Bharatha, et al., "Statistical Validation of Image Segmentation Quality Based on a Spatial Overlap Index," *Academic Radiology* 11, no. 2 (2004): 178–189, https://doi.org/10.1016/s1076-6332(03)00671-8.

37. A. I. L. Namburete, R. van Kampen, A. T. Papageorghiou, and B. W. Papież, Multi-Channel Groupwise Registration to Construct an Ultrasound-specific Fetal Brain Atlas. Data Driven Treatment Response Assessment and Preterm, Perinatal, and Paediatric Image Analysis; 2018 2018//(Springer International Publishing.

38. S. Płotka, T. Włodarczyk, A. Klasa, M. Lipa, A. Sitek, and T. Trzciński, eds., *Fetalnet: Multi-Task Deep Learning Framework for Fetal Ultrasound Biometric Measurements. Neural Information Processing* (Springer International Publishing, 2021).

39. J. C. Prieto, H. Shah, A. J. Rosenbaum, et al., "An Automated Framework for Image Classification and Segmentation of Fetal Ultrasound Images for Gestational Age Estimation," *Proceedings of SPIE - The International Society for Optical Engineering* 11596 (2021): 11596, https://doi.org/10.1117/12.2582243.

40. R. Yasrab, H. Zhao, Z. Fu, L. Drukker, A. T. Papageorghiou, and J. A. Noble, "Automating the Human Action of First-Trimester Biometry Measurement From Real-World Freehand Ultrasound," *Ultrasound in Medicine and Biology* (2024).

41. R. Arnaout, L. Curran, Y. Zhao, J. C. Levine, E. Chinn, and A. J. Moon-Grady, "An Ensemble of Neural Networks Provides expert-level Prenatal Detection of Complex Congenital Heart Disease," *Nature Medicine* 27, no. 5 (2021): 882–891, https://doi.org/10.1038/s41591-021-01342-5.

42. J. Tan, A. Au, Q. Meng, S. FinesilverSmith, J. Simpson, D. Rueckert, Eds, "Automated Detection of Congenital Heart Disease in Fetal Ultrasound Screening. Medical Ultrasound, and Preterm, Perinatal and Paediatric Image Analysis," in *First International Workshop, ASMUS 2020, and 5th International Workshop, PIPPI 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4-8, 2020, Proceedings 1* (Springer, 2020).

43. M. Lin, X. He, H. Guo, et al., "Use of real-time Artificial Intelligence in Detection of Abnormal Image Patterns in Standard Sonographic Reference Planes in Screening for Fetal Intracranial Malformations," *Ultrasound in Obstetrics and Gynecology* 59, no. 3 (2022): 304–316, https://doi.org/10.1002/uog.24843.

44. B. Gutiérrez-Becker, F. Arámbula Cosío, M. E. Guzmán Huerta, J. A. Benavides-Serralde, L. Camargo-Marín, and V. Medina Bañuelos, "Automatic Segmentation of the Fetal Cerebellum on Ultrasound Volumes, Using a 3D Statistical Shape Model," *Medical, & Biological Engineering & Computing* 51 (2013): 1021–1030, https://doi.org/10.1007/s11517-013-1082-1.

45. R. Huang, A. Namburete, and A. Noble, Data for Paper'Learning to Segment Key Clinical Anatomical Structures in Fetal Neurosonography Informed by a Region-based Descriptor, 2018).

46. Y. Wu, K. Shen, Z. Chen, and J. Wu, eds., "Automatic Measurement of Fetal Cavum Septum Pellucidum From Ultrasound Images Using Deep Attention Network," in *2020 IEEE International Conference on Image Processing* (ICIP), 2020).

47. A. Gholipour, C. K. Rollins, C. Velasco-Annis, et al., "A Normative Spatiotemporal MRI Atlas of the Fetal Brain for Automatic Segmentation and Analysis of Early Brain Growth," *Scientific Reports* 7, no. 1 (2017): 476, https://doi.org/10.1038/s41598-017-00525-w.

48. G. Babucci, K. Rosen, B. Cappuccini, and G. Clerici, "3D Evaluation of Fetal Brain Structures: Reference Values and Growth Curves," *Journal of Maternal-Fetal and Neonatal Medicine* 34, no. 21 (2021): 3546–3551, https://doi.org/10.1080/14767058.2019.1686477.

49. A. Benavides-Serralde, E. Hernández-Andrade, J. Fernández-Delgado, et al., "Three-Dimensional Sonographic Calculation of the Volume of Intracranial Structures in Growth-Restricted and Appropriate-For-Gestational Age Fetuses," *Ultrasound in Obstetrics and Gynecology: The Official Journal of the International Society of Ultrasound in Obstetrics and Gynecology* 33, no. 5 (2009): 530–537, https://doi.org/10.1002/uog.6343.

50. C.-H. Chang, F.-M. Chang, C.-H. Yu, H.-C. Ko, and H.-Y. Chen, "Assessment of Fetal Cerebellar Volume Using three-dimensional Ultrasound," *Ultrasound in Medicine and Biology* 26, no. 6 (2000): 981–988, https://doi.org/10.1016/s0301-5629(00)00225-8.

51. J. E. Araujo, H. A. Guimarães Filho, C. R. Pires, L. M. Nardozza, A. F. Moron, and R. Mattar, "Validation of Fetal Cerebellar Volume by three-dimensional Ultrasonography in Brazilian Population," *Archives of Gynecology and Obstetrics* 275, no. 1 (2007): 5–11, https://doi.org/10.1007/s00404-006-0192-5.

52. M. Rutten, L. Pistorius, E. Mulder, P. Stoutenbeek, L. De Vries, and G. Visser, "Fetal Cerebellar Volume and Symmetry on 3-d Ultrasound: Volume Measurement With Multiplanar and Vocal Techniques," *Ultrasound in Medicine and Biology* 35, no. 8 (2009): 1284–1289, https://doi.org/10.1016/j.ultrasmedbio.2009.03.016.

53. C. K. Rollins, C. G. Watson, L. A. Asaro, et al., "White Matter Microstructure and Cognition in Adolescents With Congenital Heart Disease," *Journal of Pediatrics* 165, no. 5 (2014): 936, https://doi.org/10.1016/j.jpeds.2014.07.028.

54. S. Zeng, Q. Zhou, J. Zhou, M. Li, C. Long, and Q. Peng, "Volume of Intracranial Structures on Three-Dimensional Ultrasound in Fetuses With Congenital Heart Disease," *Ultrasound in Obstetrics and Gynecology* 46, no. 2 (2015): 174–181, https://doi.org/10.1002/uog.14677.