POLITECNICO DI TORINO
Repository ISTITUZIONALE

Automated Brain Segmentation in 3D Cranial Ultrasound Using Deep Learning: Toward Scalable Bedside Monitoring of Neonatal Neurodevelopment

Original
Automated Brain Segmentation in 3D Cranial Ultrasound Using Deep Learning: Toward Scalable Bedside Monitoring of Neonatal Neurodevelopment / Khaled, R., Pizarro, J., Benavente-Fernández, I., Lubián-López, S.P., Shah, S.T.H., Shah, S.A.H., Deriu, M.A., Gontard, L.C.. - In: MACHINE LEARNING AND KNOWLEDGE EXTRACTION. - ISSN 2504-4990. - 8:8(2026), pp. 1-30. [10.3390/make8080221]

Availability:
This version is available at: 11583/3013567 since: 2026-07-27T08:27:12Z

Publisher:
MDPI

Published
DOI:10.3390/make8080221

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

26 November 2026

machine learning & knowledge extraction

MDPI

Article

Automated Brain Segmentation in 3D Cranial Ultrasound Using Deep Learning: Toward Scalable Bedside Monitoring of Neonatal Neurodevelopment

Roa'a Khaled [1,2,*], Joaquín Pizarro [2], Isabel Benavente-Fernández [3,4,5], Simón P. Lubián-López [3,4,5], Syed Taimoor Hussain Shah [6], Syed Adil Hussain Shah [6,7], Marco Agostino Deriu [6] and Lionel C. Gontard [1,8]

1 Department of Condensed Matter Physics, University of Cadiz, 11519 Puerto Real, Spain
2 Department of Computer Engineering, University of Cadiz, 11519 Puerto Real, Spain; joaquin.pizarro@uca.es
3 Division of Neonatology, Department of Pediatrics, Puerta del Mar University Hospital, 11009 Cadiz, Spain; isabel.benavente@uca.es (I.B.-F.); simon.lubian@uca.es (S.P.L.-L.)
4 Biomedical Research and Innovation Institute of Cadiz (INiBICA) Research Unit, Puerta del Mar University Hospital, 11009 Cadiz, Spain
5 Area of Pediatrics, Department of Child and Mother Health and Radiology, Medical School, University of Cadiz, 11003 Cadiz, Spain
6 PolitoBIOMed Lab, Department of Mechanical and Aerospace Engineering, Politecnico di Torino University, 10129 Turin, Italy; taimoor.shah@polito.it (S.T.H.S.); syed_shah@polito.it (S.A.H.S.); marco.deriu@polito.it (M.A.D.)
7 Department of Research and Development (R&D), GPI SpA, 38123 Trento, Italy
8 Institute of Electron Microscopy and Materials Research (IMEYMAT), University of Cadiz, 11519 Puerto Real, Spain
* Correspondence: roaa.khaled@uca.es

Abstract

Preterm infants are at high risk of neurodevelopmental disorders (NDDs), yet current brain assessments using 2D cranial ultrasound (cUS) remain subjective and limited. 3D cUS offers richer anatomical details but is underutilized due to its complexity and lack of automated tools. We developed and evaluated Deep Learning (DL) methods for total brain (TB) segmentation from 3D cUS of preterm neonates, emphasizing accuracy, reproducibility, and clinical deployability. We compared 2D UNet models, a contextual UNet, and self-configuring 2D/3D nnUNet variants, evaluating accuracy, efficiency, and GPU/CPU feasibility. Clinical relevance was examined through longitudinal total brain volume (TBV) trajectories and their association with 2-year neurodevelopmental outcomes. The nnUNet ensemble achieved the best performance (Dice: 0.97, Volumetric Difference Error: 2.32%), while contextual UNet offered a favorable accuracy-efficiency trade-off on low-resource hardware. Longitudinal analysis showed significantly slower TBV growth in infants with adverse outcomes ($p = 0.007$), supporting the prognostic value of automated volumetric measurements. We developed a clinician-facing web application to illustrate clinical integration. This work demonstrates the feasibility of DL-based TB segmentation from 3D cUS and provides a scalable reproducible framework supporting early quantitative assessment of brain development, laying the groundwork for future lightweight, privacy-aware AI solutions for Neonatal Intensive Care Unit integration.

Keywords: Deep Learning; cranial Ultrasound; brain segmentation; preterm neonates; nnUNet; neurodevelopmental outcomes; NICU clinical decision support; low-resource inference

Check for updates

Academic Editor: Enea Parimbelli

Received: 1 June 2026
Revised: 17 July 2026
Accepted: 21 July 2026
Published: 25 July 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Preterm infants are highly vulnerable to injury; thus, early assessment is essential to detect potential neurodevelopmental disorders (NDDs) and enable timely interventions that can improve long-term outcomes [1,2]. However, current routine bedside assessment remains limited by visual inspection and manual measurements of 2D cranial ultrasound (cUS), which is subjective and operator-dependent [3–5]. MRI provides superior contrast but is often impractical for repeated monitoring of neonates due to logistical and safety constraints [6]. Consequently, cUS remains the standard modality at Neonatal Intensive Care Units (NICUs) due to its portability, cost-effectiveness, and widespread availability, despite its restricted field of view and variability in manual measurements [7].

3D cUS offers volumetric visualization that can improve assessment [8–11], but its clinical adoption is limited by the time and expertise needed to interpret volumetric data [10]. Automated segmentation is therefore essential for scalable bedside use, offering fast, objective, and reproducible measurements of brain structures [12,13].

Beyond structural assessment, total brain size is itself a clinically important biomarker of brain growth and has been associated with later neurodevelopmental outcomes in preterm infants. Today in routine NICU practice, head circumference (HC) remains the bedside standard that serves as an indirect linear surrogate for longitudinal monitoring of brain growth [14]. Although some studies have shown that manually extracted total brain volume (TBV) values were strongly associated with 2-year neurodevelopmental outcome [15,16], quantitative volumetric assessment of the total brain (TB) remains difficult to obtain at the bedside. Reflecting this gap, efforts have been made to approximate TBV from cUS using simplified 2D linear-axis formulas to enable more routine volumetric monitoring during NICU admission [11], but such approaches remain manual, operator-dependent, and not readily scalable. This underscores the clinical value of an automated, reproducible method for extracting TBV directly from 3D cUS, independent of whether regional structures are additionally segmented. In addition, automated TB segmentation provides a whole-brain mask that serves as a preprocessing step for future automated regional brain segmentation by isolating the brain from the surrounding skull and background tissues.

Nevertheless, existing automated cUS methods remain rare compared with MRI methods, and primarily target specific structures, such as cerebral ventricles, or focus on fetal rather than neonatal imaging. No prior work has addressed TB segmentation from 3D cUS, despite its importance for quantitative brain monitoring [17–21].

Beyond algorithmic accuracy, reproducibility and deployability in resource-limited environments are critical for clinical translation, particularly in NICUs where hardware capabilities vary and real-time inference is often constrained [22]. Despite substantial progress in AI-based medical image analysis, translation to routine bedside practice remains limited [23], with fewer than 2% of proposed models progressing beyond the prototyping stage and many exhibiting a high risk of bias [24,25]. These challenges highlight the importance of evaluating not only segmentation performance but also computational feasibility under realistic clinical conditions.

This study proposes a fully automated Deep Learning (DL)-based method for segmenting the TB from 3D cUS images of preterm neonates. We assessed real-world deployment in typical NICUs clinical settings, lacking high-end hardware to support the broader goal of integrating DL into NICU workflows for reliable and scalable early neurodevelopmental assessment. Although this work does not propose a new DL architecture, its scientific contribution lies in establishing, to the best of our knowledge, the first automated framework for TB segmentation from 3D cUS images of preterm neonates. Rather than introducing a new architecture, the study demonstrates the clinical translation of state-of-the-art DL methods

to a challenging application for which no automated benchmark previously existed. The main contributions of this study are summarized as follows:

- We propose a fully automated pipeline for TB segmentation from 3D cUS images of preterm neonates, achieving high segmentation accuracy and computational efficiency.
- We comprehensively evaluated conventional UNet models and nnUNet with respect to segmentation accuracy, computational efficiency, and deployment feasibility on both GPU and CPU-only hardware representative of NICU settings.
- We performed clinical validation through analysis of longitudinal TBV growth and its association with 2-year neurodevelopmental outcomes.
- We provide an open-source implementation, pretrained models, sample data, and a clinician-oriented web application to facilitate research reproducibility and clinical adoption.

2. Related Work

Early efforts to automate neonatal brain segmentation used classical image processing methods, including atlas-based, statistical, and Machine Learning (ML) methods [26]. Atlas-based methods rely on predefined anatomical templates and tissue priors, but require heavy preprocessing and often fail to capture the substantial anatomical variability seen in preterm infants [26,27]. Statistical methods, such as thresholding, region growing [28] and morphology-based segmentation [29] are computationally efficient but highly sensitive to noise, poor tissue contrast, and intensity inhomogeneity, common limitations in neonatal cUS [30]. ML approaches, including Support Vector Machine (SVM), Random Forests (RF), and fuzzy clustering variants [17,26,31,32], offer more flexibility but still rely on handcrafted features and show limited robustness and generalization [27,29]. These challenges, compounded by motion artifacts and the inherently low-contrast nature of neonatal US, have motivated the transition toward DL-based solutions [33].

DL, particularly UNet-based architectures, has become dominant because of their effectiveness in biomedical image segmentation [17,34–36]. Such methods have also been used for neonatal brain images [33,37]; however, they focus mainly on MRI [17]. Beyond handcrafted UNet variants, the nnUNet framework has emerged as a strong baseline in biomedical image segmentation due to its self-configuring design and consistent performance across diverse datasets [38]. Despite its widespread adoption, nnUNet has not been applied to neonatal cUS.

A recent systematic review identified eight DL-based methods for cUS segmentation (specifically ventricles segmentation), published between 2018 and 2022, with reported Dice Similarity Coefficient (DSC) values ranging from 0.72 to 0.91 [12,17,39–45]. Only a few new studies on cUS have since been published. For instance, Regalado et al. performed Sylvian fissure segmentation from 2D cUS (DSC 0.777 ± 0.072) [46], while Szentimrey et al. performed ventricular segmentation (DSC 0.757 ± 0.055) [13]. These studies demonstrate the feasibility of DL for neonatal cUS but focus on localized structures rather than TB segmentation. To date, no study has addressed automated TB segmentation from 3D neonatal cUS, despite its importance for quantifying TBV, a key biomarker of early brain development, and for supporting downstream tasks by providing a reliable brain mask that improves anatomical focus and computational efficiency.

Computational performance is also under-reported in cUS DL studies. Most prior work benchmarks inference only on high-end GPUs, with reported inference times ranging from 0.01 to 70.4 s for 2D methods and 2 to <60 s for 3D methods [12,39–45]. Only a few studies have assessed low-resource performance. Martin et al. reported 70.4 s per 2D slice on a Quadro M1000M GPU [43], and Gontard et al. achieved under 60 s per 3D volume on an entry-level Nvidia 920MX [39]. This lack of benchmarking on realistic clinical hardware

motivates our inclusion of CPU-only evaluation. In this study, we define real-time as 3D volume-level segmentation completed within sub-second to few-second latency, such that it does not delay routine bedside review in the NICU, and near-real-time as processing completed within a clinically acceptable delay of seconds to a few minutes.

Recent DL advances have introduced architectures beyond UNet variants, including Transformer-based models that capture long-range dependencies via self-attention [47]. Hybrid models such as TransUNet [48] and UNetFormer [49] have demonstrated strong segmentation performance across a variety of medical imaging applications. Nevertheless, recent reviews indicate that UNet-based architectures remain among the most widely adopted and competitive approaches because they offer an effective balance between segmentation accuracy, computational efficiency, and generalizability across diverse medical imaging tasks [50]. In this work, we therefore focus on UNet-based models, as they provide a suitable compromise between segmentation performance, computational efficiency, and deployment feasibility for bedside neonatal applications on standard hospital hardware.

Beyond methodological advances in automated analysis of cUS, numerous studies have linked early brain volumetric measures, both global and regional, to later neurodevelopmental outcomes in preterm infants. At the global level, Mayrink et al. reported that larger head circumference at discharge and at five months of corrected age was positively correlated with cognitive, language, and motor Bayley-III scores at 12 and 18 months [51], whereas Ruiz-González et al. demonstrated that manually measured TBV from cUS predicted abnormal term MRI findings and that slower TBV growth was correlated with adverse cognitive and language outcomes at two years of age [15,16].

At the regional level, slower thalamic growth trajectory from early MRI to term-equivalent age was associated with lower Bayley-III cognitive, motor, and language scores at 2 years [52], and larger ventricular volume at both early and term-equivalent age MRI was independently associated with lower IQ and poorer motor outcomes at 4.5 years [53]. In another study, larger cortical gray matter volume and lower CSF and ventricular volumes at early MRI were associated with better Bayley-III motor and cognitive scores at two years with these findings extended to term-equivalent age by the same group, additionally identifying white matter, deep gray matter, and cerebellar volumes as predictors of motor, cognitive, and language outcomes at two years [54,55]. These findings motivated our use of 2-year Bayley-III scores to assess the clinical relevance of automatically obtained TBV trajectories.

Together, these limitations underscore the need for automated, clinically feasible TB segmentation from 3D neonatal cUS, a gap that this study aims to address.

3. Materials and Methods

3.1. Data Cohort

The dataset used in this study is part of a private, ongoing longitudinal cohort of preterm infants recruited at Puerta del Mar University Hospital (Cádiz, Spain) since May 2018. All procedures received institutional ethics approval and written informed consent (PI0052/2017, ITI-0019-2019 and PID2019-110484RB-I00). Clinical and imaging data were anonymized before analysis, and infants were eligible if born at $\leq$32 weeks of gestation (in medical terms: $\leq$32 weeks of gestation + 6 days) and/or with a birth weight $\leq$1500 g.

3D cUS volumes were acquired for each infant at least once shortly after birth and, when possible, at more time points during hospitalization using the Voluson i and Voluson S8 systems (GE Healthcare). The S8 provides higher image quality, resulting in a dataset of mixed-quality volumes. Acquisition parameters and export workflow are provided in Appendix A.

At two years corrected age, or postmenstrual age (PMA), neurodevelopment was assessed using the Bayley Scales of Infant and Toddler Development, Third Edition (BSID-III) [56]. This assessment provides standardized scores (mean 100, SD 15) across cognitive, motor, and language domains. Scores were analyzed both quantitatively and qualitatively, and scores $\geq$85 were classified as favorable and $<$85 as adverse, following established practice [16,57]. An overall outcome was defined as adverse if any of the three domains had a score $<$85.

The DL models were trained and evaluated on 397 3D cUS volumes from 63 preterm infants. Cohort characteristics are summarized in Table 1.

Table 1. Characteristics of the cohort used for model development, including gestational age (GA) and birth weight (mean ± SD), gender distribution, and the number of volumes acquired with each ultrasound system.

Description	Value
Mean GA at birth (weeks)	28.70 $\pm$ 2.44
Mean weight at birth (grams)	1133.41 $\pm$ 352.21
Gender (Female:Male)	29:34
Volumes from Voluson i system	368 (from 48 patients, 76% of cohort)
Volumes from Voluson S8 system	29 (from 15 patients, 24% of cohort)

For longitudinal TBV and outcome analysis, we expanded our cohort from the original 63 infants (397 volumes) to a total of 109 infants (684 volumes)and retained only those with $\geq$2 valid scans and without Post-Hemorrhagic Ventricular Dilatation (PHVD) to enable reliable growth modeling and reduce confounding.

To ensure reliable longitudinal modeling, two scan-level filters were applied. First, scans acquired after 45 weeks of PMA were excluded because they were sparse, clinically heterogeneous, and disproportionately represented in the adverse outcome group, likely due to prolonged hospitalization in more complex cases, which could otherwise bias slope estimation. Second, TBV outliers were removed using an unbiased statistical rule (the Interquartile Range (IQR) method within 2-week PMA windows) to eliminate extreme values attributable to measurement noise or biological anomalies. This procedure relied solely on statistical deviation and was independent of segmentation quality or model performance. After filtering, all 109 infants remained in the cohort, while the total number of 3D volumes decreased modestly (from 684 to 659 volumes) with cohort demographics unchanged (Table 2).

Table 2. Clinical and scanning characteristics of the extended cohort used for longitudinal study, including gestational age (GA) in weeks and weight at birth in grams (mean $\pm$ SD), gender and outcome distribution, and the range and median number of 3D volumes per infant, where each volume represents one longitudinal time point.

Description	Value
Mean GA at birth in weeks	29.14 $\pm$ 2.32
Mean weight at birth in grams	1150.37 $\pm$ 354.73
Gender (Female:Male)	55:54
Outcome (Favorable:Adverse)	82:27
3D Volumes per patient (min–max)	2–33
3D Volumes per patient (median)	5

Data Annotation

The 3D cUS volumes from the 63 infants used for model development were manually annotated by two engineers trained in neonatal neuroimaging under the supervision of a

senior neonatologist (15 years of experience) at Puerta del Mar University Hospital, Cádiz, Spain. Pixels were labeled as *brain* or *non-brain* using the MELAGE toolbox [58], and the dataset was divided among annotators. One performed slice-by-slice segmentation, while the other annotated key slices and interpolated intermediate slices, both following a standardized protocol established with the neonatologist. Due to time constraints, the neonatologist reviewed only a subset of the annotations.

These differing annotation strategies may introduce some variability in the smoothness and continuity of Ground Truth (GT) surfaces. A formal inter- or intra-operator variability analysis was not feasible given the high annotation burden, although standardized training and partial expert review helped mitigate variability. Quantifying annotation variability remains an important direction for future work.

3.2. Development of DL-Based Segmentation Models

General Preprocessing, Data Splitting, and Model Development

Raw 3D cUS volumes had variable spacings (0.366–0.528 mm) and heterogeneous volume dimensions ranging from 157 to 397 (X: coronal), 155 to 572 (Y: sagittal), and 209 to 515 voxels (Z: axial) (median 185 × 251 × 217). To standardize resolution, all volumes were resampled to isotropic 0.5 mm spacing (computed as the average spacing). This applies to all DL models in this study to ensure fair comparisons. After resampling, volume dimensions varied based on the physical size of each volume, ranging from 126–238 (X), 123–343 (Y), and 145–309 (Z) voxels (median 172 × 231 × 201).

Data splitting was performed at the patient level (i.e., all longitudinal volumes from each infant were kept within the same split) to avoid data leakage; 80% of infants were assigned to the training set and 20% to the testing set, preserving similar gestational age (GA) and birth weight distributions (Table 3). The overall data partitioning strategy is illustrated in Figure 1.

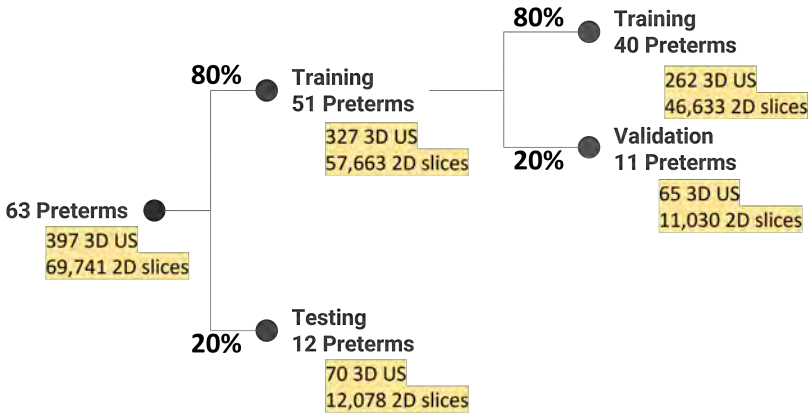


Figure 1. Patient-level data splitting strategy showing the distribution of infants, 3D cUS volumes, and axial 2D slices across the training, validation, and testing sets. All data from each infant were kept within a single subset.

Model development followed a stratified 5-fold cross-validation protocol using the training set. For each fold, the training set was further divided into training and validation subsets using an 80:20 patient-level split, ensuring that all longitudinal scans from the same infant remained within a single subset. The independent testing cohort was kept completely separate and used only for the final performance evaluation. This protocol was applied consistently to all evaluated models to ensure a fair comparison.

Model-specific preprocessing, training, and postprocessing are described below.

Table 3. Summary of gestational age (GA) in weeks, birth weight in grams, and gender distribution for the full cohort and for the training and testing subsets used for model development (reported as mean ± SD and Min.–Max. ranges).

Statistic	Total	Training Set	Testing Set
GA (mean ± SD)	28.70 ± 2.44	28.71 ± 2.40	28.67 ± 2.58
GA (Min.–Max.)	23–32.86	23–32.86	24–31.71
Weight (mean ± SD)	1133.41 ± 352.21	1107.75 ± 326.38	1242.5 ± 428.77
Weight (Min.–Max.)	470–1880	470–1650	550–1880
Females	29 (46%)	26 (51%)	3 (25%)
Males	34 (54%)	25 (49%)	9 (75%)

3.2.1. Handcrafted Models (2D and Contextual UNet)

Preprocessing

Each resampled 3D volume was intensity-normalized and sliced into 2D axial, coronal, and sagittal images.

Training

A UNet architecture (depth = 3, defined as the number of levels in the encoder and decoder) was trained separately on slices from each anatomical plane [34]. Axial slices achieved the best performance and were used to train a Contextual UNet, which incorporates three consecutive axial slices as a 3-channel input. The architectural depth and training hyperparameters were selected based on commonly used UNet configurations in the segmentation literature and are detailed in Table 4.

Table 4. Architectural configurations and hyperparameters of the 2D UNet, Contextual UNet, and 2D/3D nnUNet models. BCE = Binary Cross-Entropy; Dice+CE = combination of soft Dice and Cross-Entropy; SGD = Stochastic Gradient Descent with Nesterov momentum.

	2D UNet	Contextual UNet	2D nnUNet	3D nnUNet
Depth	3	3	6	6
Layers	13	13	32	32
Parameters	116,753	117,041	20,619,082	30,785,994
Input patch size	1 × 256 × 256	3 × 256 × 256	1 × 256 × 224	112 × 160 × 128
Batch size	64	64	57	2
Loss function	BCE	BCE	Dice+CE	Dice+CE
Optimizer	Adam	Adam	SGD	SGD
Initial LR	0.001	0.001	0.01	0.01
Epochs (max.)	40	40	1000	1000

Each fold was trained independently using the corresponding training and validation partitions.

Postprocessing

Reconstruction of 3D volumes from 2D predictions was followed by binary hole filling (structuring element 3 × 3 × 3, three iterations), largest-connected-component selection, and anisotropic-diffusion smoothing (1 iteration, Perona-Malik scheme) to reduce noise and enforce volumetric consistency.

3.2.2. Fully Automatic Approach with nnUNet Framework

Unlike manually tuned architectures, nnUNet is a self-configuring framework that automatically determines the preprocessing, architecture, and training hyperparameters based on dataset statistics [38]. It selects resampling strategies, patch sizes, batch sizes, and network topologies according to spacing, intensity distribution, target anatomical shape,

and GPU memory, and evaluates multiple configurations (2D, 3D, and 3D cascade) via five-fold cross-validation to automatically select the best-performing model or ensemble. nnUNet also applies empirical postprocessing, such as connected-component filtering, only when beneficial.

Preprocessing

In our case, the Default Preprocessor was automatically selected by nnUNet, which performs resampling to isotropic voxel spacing (already applied in our general preprocessing), intensity normalization based on percentile clipping, cropping, and padding, and automatic selection of the input patch size and batch size. The 2D configuration used axial slicing after RAS (Right–Anterior–Superior) reorientation.

Training

Training followed the default nnUNet pipeline using the same patient-level 5-fold cross-validation protocol described earlier. nnUNet trained both 2D and full-resolution 3D models, and automatically selected a 2D–3D ensemble as the best-performing configuration. All training settings followed the default nnUNet pipeline and are summarized in Table 4.

Postprocessing

nnUNet did not select any postprocessing, as it did not improve validation performance.

3.3. Evaluation

Segmentation performance was evaluated using four metrics:

- Dice Similarity Coefficient (DSC): measures spatial overlap.
- Volumetric Difference Error (VDE): percentage deviation between predicted and GT volumes.
- Volumetric Similarity Index (VSI): assesses agreement in volumetric size.
- Average Surface Distance (ASD): mean distance between predicted and ground-truth surfaces.

We additionally performed a qualitative visual assessment of anatomical accuracy and post hoc explainability analysis using feature-activation attention, Grad-CAM, LRP-inspired relevance maps, saliency maps, and fold-wise uncertainty visualizations. These explainability methods were used to qualitatively assess whether model responses focused on anatomically meaningful brain regions and to identify areas of higher prediction uncertainty. Computational efficiency was also evaluated, including training duration, inference time (per slice and per 3D volume), epochs to convergence, postprocessing time, and model complexity in Floating-Point Operations Per Second (FLOPs).

3.4. Implementation

Image preprocessing and model development were performed in Python 3.9 and PyTorch 2.0 on an Ubuntu 18.04 server equipped with 125 GB usable RAM, two NVIDIA RTX A5000 GPUs (24 GB each), and an Intel Xeon Gold 5220R CPU.

Additionally, we implemented a prototype web application to demonstrate how segmentation models can be integrated into a clinician-facing workflow. The application uses a Flask backend and a lightweight HTML/JavaScript front-end that enables users to upload volumes, select models, view synchronized axial, coronal, and sagittal slices with mask overlays, and download outputs with volumetric measurements. Additional details on the clinician-facing web application and its privacy-preserving design are provided in Appendix B.

The complete implementation, including model training codes, inference codes, trained models, sample 3D cUS volumes, and the web application, are available in a publicly accessible repository; the link has been temporarily withheld to preserve anonymity during peer review.

4. Results Analysis and Discussion

This section reports segmentation accuracy and computational performance of the proposed models, contextualizes the results with respect to existing DL approaches for neonatal cUS, and explores the clinical relevance of automated brain volumetric measurements in relation to neurodevelopmental outcomes.

4.1. Segmentation Performance and Explainability

We first compared the three handcrafted 2D UNet models trained on axial, coronal, and sagittal planes. As summarized in Table 5 and illustrated by the box plots in Appendix C (Figure A2), the axial model consistently outperformed the others across all metrics (DSC = 0.92 ± 0.04), with lower volumetric and surface errors. Qualitative examples in Figure 2 further illustrate the improved anatomical alignment of axial predictions with the GT. Performance differences across planes highlight the importance of slice orientation in 2D model training. The axial plane provides more consistent anatomical coverage through the anterior fontanel, whereas coronal and sagittal planes suffer from partial coverage and less reliable landmarks, which may also affect GT fidelity.

Table 5. Quantitative segmentation performance of the developed models obtained using 5-fold cross-validation. Plane-specific 2D UNets, a Contextual UNet, and nnUNet variants (2D, 3D, and ensemble) were compared using DSC, VDE, VSI, and ASD (mean ± SD).

Model	DSC	VDE (%)	VSI	ASD
Sagittal 2D UNet	0.82 ± 0.08	34.48 ± 25.11	0.86 ± 0.09	5.17 ± 2.61
Coronal 2D UNet	0.89 ± 0.05	12.11 ± 9.31	0.94 ± 0.04	3.02 ± 1.41
Axial 2D UNet	0.91 ± 0.04	10.68 ± 10.03	0.95 ± 0.04	2.00 ± 0.90
Contextual UNet	0.93 ± 0.03	4.45 ± 4.47	0.98 ± 0.02	1.37 ± 0.63
2D nnUNet	0.96 ± 0.02	2.66 ± 3.61	0.99 ± 0.02	0.83 ± 0.50
3D nnUNet	**0.97 ± 0.02**	2.49 ± 2.43	**0.99 ± 0.01**	**0.79 ± 0.40**
nnUNet ensemble	**0.97 ± 0.02**	**2.34 ± 2.37**	**0.99 ± 0.01**	0.83 ± 0.38

Bold values indicate the best performance for each metric.

To incorporate spatial context, we trained a Contextual UNet using three adjacent axial slices. This model improved volumetric and surface accuracy over the axial 2D UNet (DSC = 0.93 ± 0.03, Table 5, Appendix C (Figure A3)), confirming the benefit of contextual information. Qualitative results show smoother and more complete reconstructions (Figure 2).

Seeking additional performance gains and automation, we further evaluated the nnUNet framework, which automatically selects optimized 2D, 3D, and ensemble configurations. The selected 2D–3D ensemble achieved the highest segmentation accuracy overall (DSC = 0.97 ± 0.02; Table 5, Appendix C (Figure A3)) and produced the most consistent volumetric predictions (Figure 3).

Assessing the contribution of each component in the ensemble, an independent evaluation of the 2D and 3D nnUNet models showed comparable quantitative performance (Table 5), with the 3D variant producing smoother surfaces and fewer artifacts on visual inspection (Figure 3), supporting the complementary nature of the ensemble.

Overall, albeit at the cost of higher computational demand, performance gains of nnUNet over handcrafted models reflect increased architectural depth, richer skip connection strategies, and richer spatial context as the 3D nnUNet operates directly on volumetric patches.

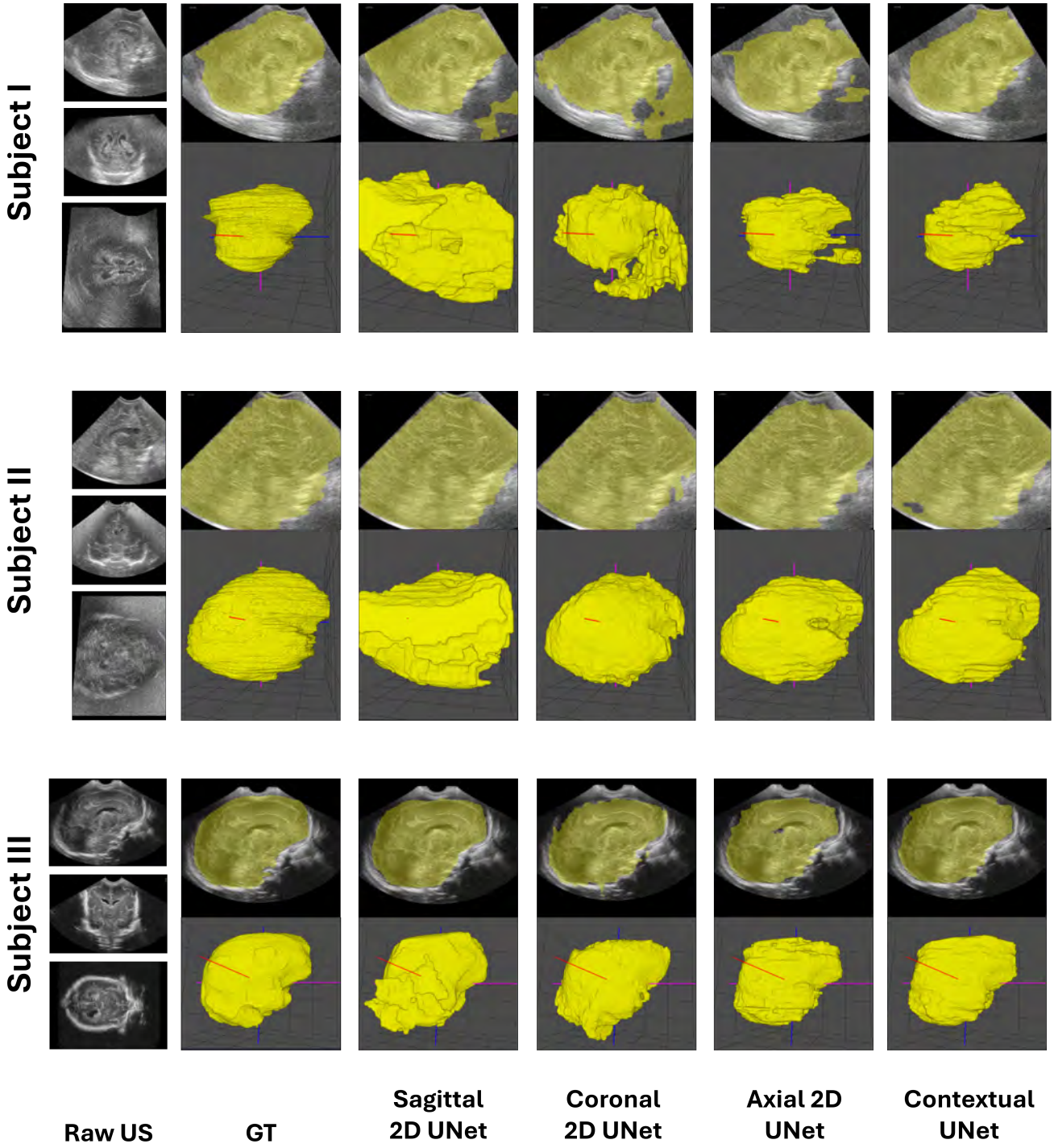


Figure 2. Qualitative comparison of neonatal brain segmentation using plane-specific 2D UNet and Contextual UNet models. Each row corresponds to a different patient. From left to right: raw US volume in three anatomical views, Ground Truth (GT), sagittal, coronal, axial 2D UNet, and Contextual UNet models. Segmentations (GT and model outputs) are displayed in yellow as sagittal slice overlays and 3D renderings in standard anatomical orientation. Colored axes in the 3D renderings denote anatomical directions (red = left–right, blue = anterior–posterior, magenta = inferior–superior).

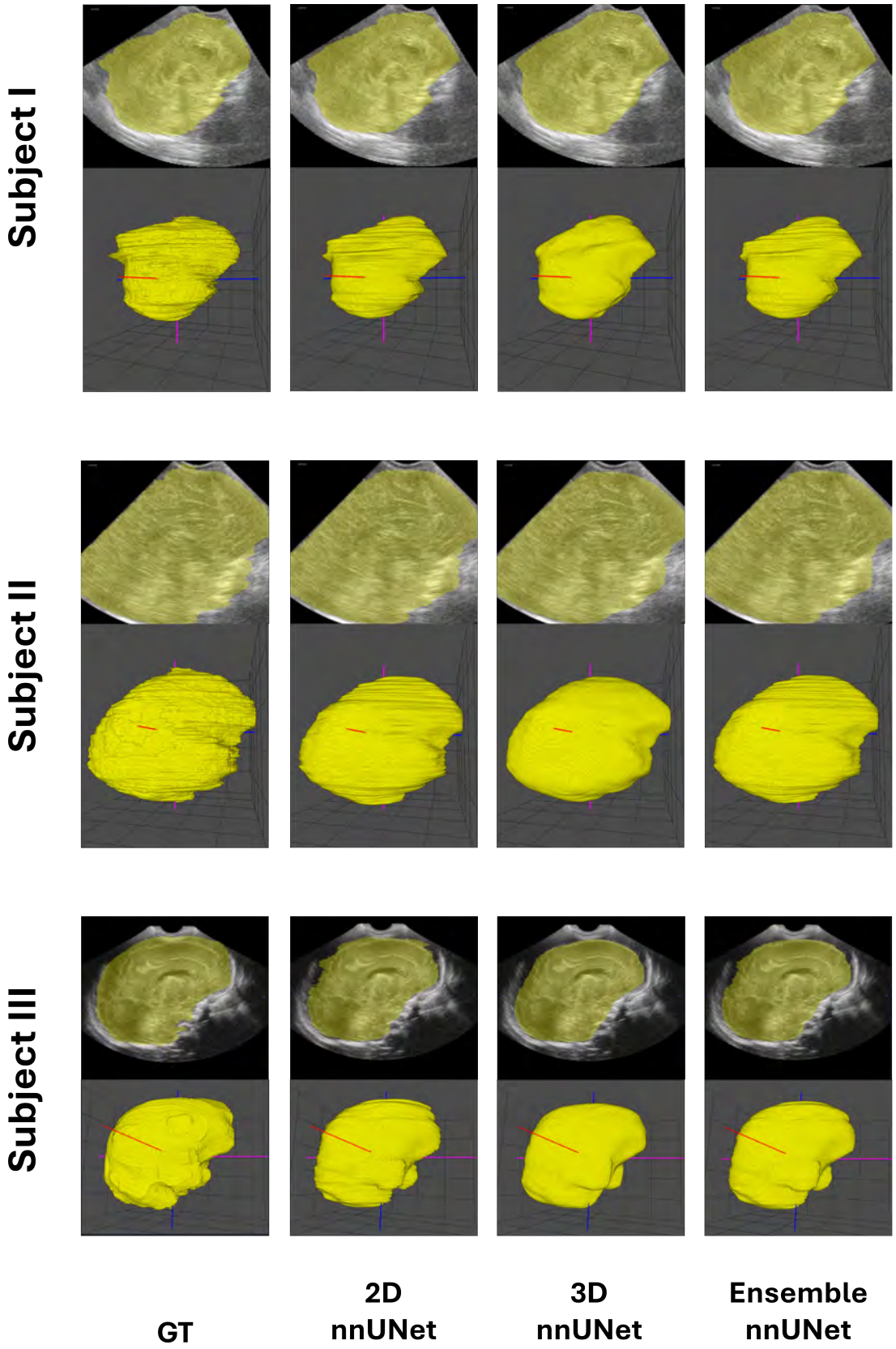


Figure 3. Qualitative comparison of neonatal brain segmentation using nnUNet models. Each row corresponds to a different patient. From left to right: Ground Truth (GT), 2D, 3D, and ensemble nnUNet models. Segmentations (GT and model outputs) are displayed in yellow as sagittal slice overlays and 3D renderings in standard anatomical orientation. Colored axes in the 3D renderings denote anatomical directions (red = left–right, blue = anterior–posterior, magenta = inferior–superior).

To further investigate model decision-making, representative explainability maps and fold-wise uncertainty estimates are shown in Figure 4. Across all models, feature-attention, Grad-CAM, saliency, and LRP-inspired relevance maps consistently identify the brain parenchyma as the primary driver of the segmentation decisions, with limited response to surrounding background structures. Notable differences are nevertheless observed between architectures. The handcrafted 2D and Contextual UNet models exhibit broader and more diffuse activation patterns, whereas the nnUNet models produce more compact and anatomically focused responses that align closely with the segmented brain region. Consistent with these observations, the handcrafted models also display more widespread fold-wise uncertainty throughout the brain, while the 2D and 3D nnUNet models exhibit lower uncertainty that is largely concentrated around the brain contour. Together, these qualitative findings are consistent with the superior quantitative performance of the nnUNet models (Table 5), suggesting that they learn more consistent and spatially coherent feature representations while maintaining stable predictions across cross-validation folds.

4.2. Computational Efficiency and Clinical Deployment

Reproducibility and computational efficiency are critical for clinical translation in NICU settings, particularly in resource-limited settings [22]. To complement accuracy analysis, we evaluated training cost, inference latency, memory usage, and computational complexity of all models (Table 6).

To better simulate the behavior of a realistic clinical production-grade software, and for a fair comparison with inference times reported in other existing studies, we perform batch inference where the model is loaded once when the application starts (or first volume is processed) and is reused for new volumes (i.e., persistent model in memory, no repeated loading). Moreover, the inference times reported include inference, binarization, and output saving. For a fair comparison, postprocessing time is reported separately for the handcrafted UNet models since the nnUNet models performed no postprocessing in our case. For handcrafted models, volume-level inference was derived from per-slice predictions, while nnUNet timings were approximated from batch inference outputs. Detailed timing procedures and times for raw inference (no output saving) and volume-by-volume inference (repeated model loading) are reported in Appendix D. Inference time and GPU memory are reported for a single trained model, reflecting the intended deployment scenario. The 5-fold cross-validation protocol was used exclusively for model development and segmentation performance evaluation.

Handcrafted 2D and Contextual UNets were the most memory-efficient models, whereas nnUNet variants required substantially more GPU memory and FLOPs (Table 6), reflecting their increased architectural complexity. These differences were also reflected in volume-level inference time. The 2D and Contextual UNets achieved total volume inference times of 0.785 s and 0.708 s, respectively, increasing to 2.965 s and 2.858 s when postprocessing was included, which qualifies as real-time 3D segmentation under our definition (given in Section 2). The 2D nnUNet required 3.759 s per volume without postprocessing, remaining within the real-time range, while the 3D nnUNet achieved a volume-level inference time of 5.695 s, corresponding to near-real-time performance. Although neurodevelopmental outcome prediction relies on longitudinal imaging, low-latency segmentation remains clinically valuable at the point of acquisition, enabling immediate quality control, scan guidance, rapid extraction of volumetric measures that can be stored for later longitudinal assessment, and timely bedside decision support when abnormalities are suspected.

To assess feasibility in low-resource settings, we evaluated inference on a CPU-only conventional machine representative of typical NICU environments; an overview of relevant clinical hardware configurations is provided in Appendix D. The evaluation was

conducted on an Intel® Core™ i7-8565U CPU @ 1.80 GHz (4 cores/8 threads) with 8 GB RAM running Windows 11. As reported in Table 7, handcrafted models remained practical, with volume-level inference times of 7.86 s for the 2D UNet and 8.90 s for the Contextual UNet, whereas the 2D and 3D nnUNet required substantially longer processing times (≈121 s and ≈244 s per volume, respectively) and higher memory usage, limiting their suitability for such settings.

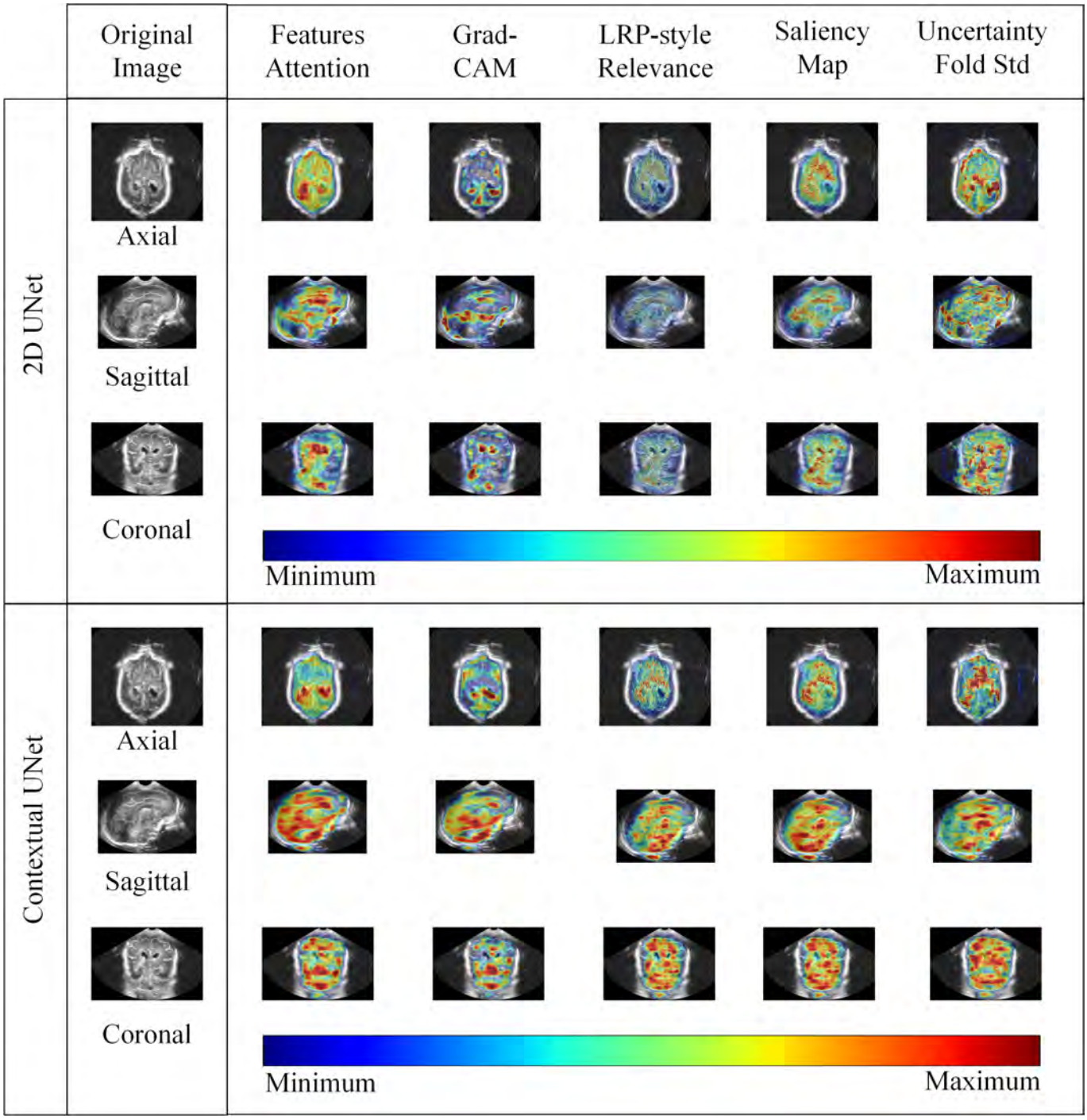


Figure 4. *Cont.*

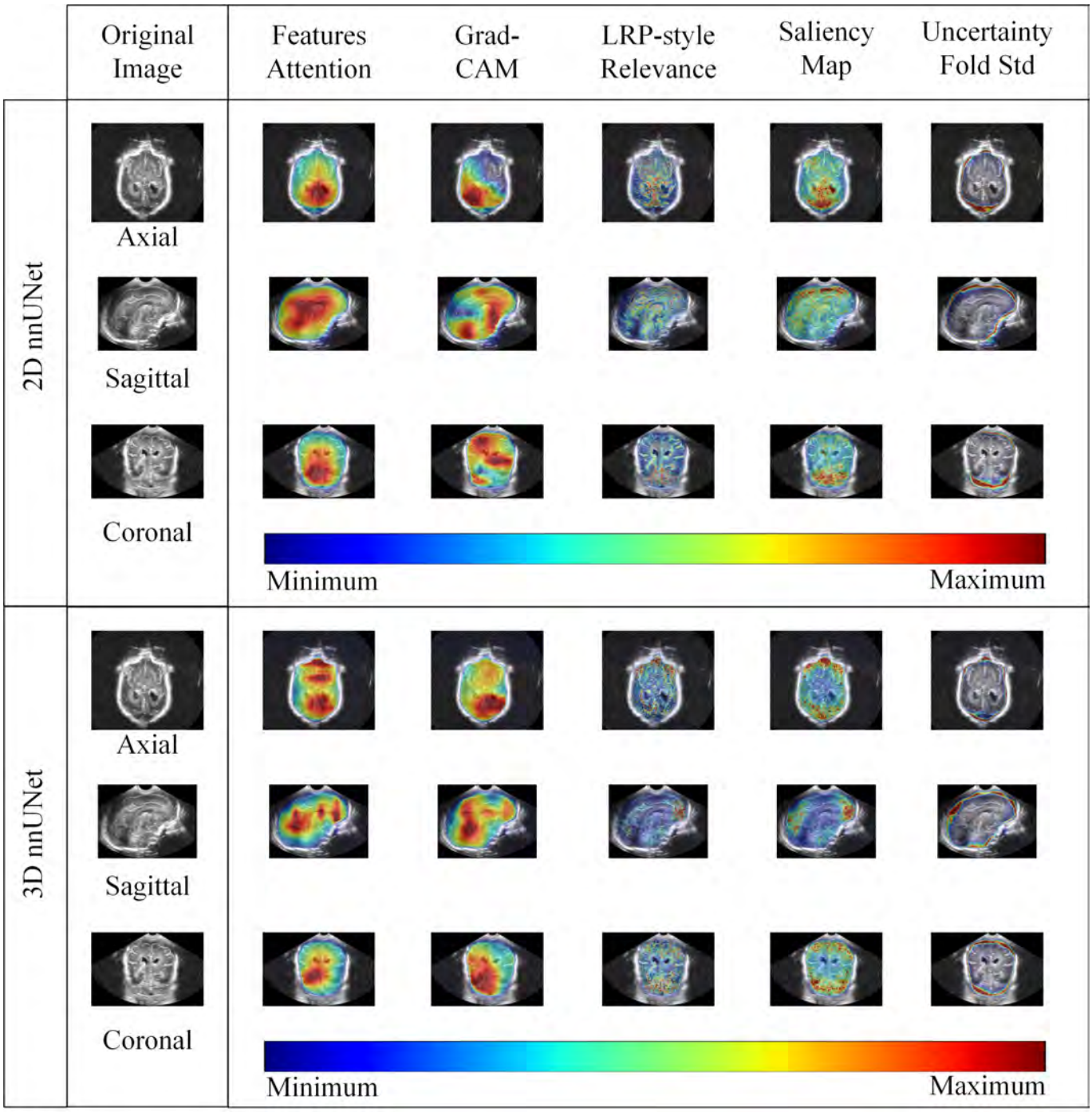


Figure 4. Representative explainability and uncertainty visualizations for the 2D UNet, Contextual UNet, 2D nnUNet, and 3D nnUNet models, obtained from the same representative 3D cUS scan. Axial, sagittal, and coronal slices are shown for each model. Columns display the original image, feature-attention maps, Grad-CAM, LRP-inspired relevance maps, saliency maps, and fold-wise uncertainty maps (standard deviation across the five cross-validation folds). Warmer colors indicate higher relative contribution or uncertainty within each visualization.

Overall, these results highlight a trade-off between accuracy and deployability. While the nnUNet ensemble achieves the highest accuracy overall (Dice = 0.97), its substantially higher computational cost limits its practicality in resource-constrained NICU environments. In contrast, the Contextual UNet provides the most favorable balance for such settings, maintaining high segmentation accuracy (Dice = 0.93) while enabling real-time

3D inference and delivering 14×, 27×, and 41× faster processing on CPU-only hardware compared with the 2D nnUNet, 3D nnUNet, and their ensemble configuration, respectively.

Table 6. Training and inference characteristics of the evaluated models, including FLOPs, peak GPU memory, and slice- and volume-level inference times under batched GPU execution. CV = 5-fold cross-validation; SF = single fold. CV training time represents the total training time for the complete 5-fold cross-validation procedure, whereas all SF training and inference metrics correspond to a single trained model. Values for the 2D UNet are averaged across the three anatomical models.

	2D UNet	Contextual UNet	2D nnUNet	3D nnUNet
Training Details				
Training images (2D/3D)	Axial: 57,663 Coronal: 68,029 Sagittal: 75,472	57,663	57,663	327
Batch size	64	64	57	2
CV training time (hours)	6.31	6.36	84.40	104.44
SF training time (hours)	1.84	2.00	17.85	20.71
Epochs to converge	≈35–40	≈35–40	250	250
Model Efficiency				
FLOPs (GMac)	1.44	1.46	34.07	1405.45
SF GPU Inference				
Peak memory (MB)	20.15	20.65	80.92	118.86
2D slice inference time (s)	0.004	0.004	0.022	–
3D volume inference time (s)	0.785	0.708	3.759	5.695
Postprocessing (not GPU optimized)				
Postprocessing time (s)	2.18	2.15	–	–

Table 7. CPU-only inference performance of the evaluated models under the single-model deployment scenario, reporting slice- and volume-level inference times and peak CPU memory usage. Values for the 2D UNet are averaged across the three anatomical models.

Model	2D UNet	Contextual UNet	2D nnUNet	3D nnUNet
2D slice inference time (s)	0.040	0.052	0.701	–
3D volume inference time (s)	7.859	8.899	121.047	244.089
Peak inference CPU memory	352.73	515.41	501.25	581.84

When modestly higher computational resources are available, the 2D nnUNet represents a viable alternative, offering improved accuracy at still acceptable inference latency. Higher-capacity nnUNet configurations are best suited for GPU-enabled platforms, underscoring the importance of aligning model choice with the intended clinical deployment scenario. Ultimately, the ideal scenario for real-time application and seamless integration into NICU workflows involves embedding the chosen model directly within the US machine's proprietary hardware, optimizing for on-device processing and immediate feedback during image acquisition [22].

To illustrate practical deployment, we additionally implemented these models in an interactive web application designed for bedside use, enabling clinicians to upload volumes, select a model, review segmentation overlays, and access volumetric metrics through a simple browser interface (see Appendix B).

4.3. Contextualization with Existing DL Methods for Neonatal cUS Segmentation

Only limited number of studies have applied DL for neonatal cUS segmentation, hence, the field remains relatively underexplored compared with other imaging modalities such as MRI. As outlined in Section 2, none of these few cUS studies perform TB segmentation and they primarily focus on segmenting localized regions such as cerebral ventricles or sulcal structures, which are intrinsically different and often more difficult tasks than TB segmentation. This makes direct, quantitative comparison impossible. As a result, reported DSC values and inference times across studies are not directly comparable, as they depend heavily on the structure of interest, input image resolution, preprocessing pipelines, and hardware differences.

Nevertheless, to contextualize our work within the broader field of neonatal cUS research, we summarize prior DL approaches applied to neonatal cUS, regardless of target structure, and grouped by 2D and 3D approaches in Table 8. We report their segmentation accuracy (DSC), input image sizes, hardware, and inference times. This comparison is intended purely as contextual background, not as a head-to-head benchmarking.

Regarding segmentation accuracy, previous works reported DSC scores between 0.72 and 0.908, with most methods clustering around the 0.80–0.89 range, highlighting the challenges of this modality [12,13,17,39–46]. In our study, our simplest model (the axial 2D UNet) achieved 0.92 ± 0.04 DSC for TB segmentation, and our best model (the nnUNet ensemble) reached 0.97 ± 0.02.

Table 8. Contextualization within existing methods. For our proposed 2D and 3D nnUNet models, the inference time includes resampling to the model input size and back to the original image size; therefore, inference image size is variable, where: D × H × W = 126–238 × 123–343 × 145–309 (median: 172 × 231 × 201).

	DSC	Inference Image Size	Used Resources	Time (s)
2D Segmentation				
Wang et al. (2018) [40]	0.908	512 × 512	GPU: Nvidia Titan-X, CPU: Intel Xeon (3.00 GHz), RAM: 8GB	0.022
Valanarasu et al. (2020) [41]	0.8901	512 × 512	–	0.01
Tabrizi et al. (2020) [42]	0.86	128 × 128	GPU: NVIDIA GeForce GTX TITAN X	17.4
Martin et al. (2021) [43]	0.81	320 × 320	(1) GPU: NVidia Tesla V100, RAM: 32 G (2) GPU: Nvidia GTX 1080, RAM: 8 GB (3) GPU: Nvidia Quadro M1000 M, RAM: 2 GB	3.5 9.6 70.4
Valanarasu et al. (2022) [44]	0.8943	128 × 128	Two GPUs: NVIDIA-RTX 2080 Ti	–
Regalado et al. (2024) [46]	0.777	550 × 868	GPU: NVIDIA GeForce GTX 1080	–
3D Segmentation				
Martin et al. (2018) [45]	0.816	256 × 256 × 256	GPU: NVIDIA GeForce GTX 1080, RAM: 8 GB	5
Gontard et al. (2021) [39]	0.8	397 × 200 × 200	GPU: Nvidia Geforce 920MX, CPU: i3, RAM: 12 MB	<60
Martin et al. (2021) [43]	0.82	64 × 320 × 320	(1) GPU: NVidia Tesla V100, RAM: 32 GB (2) GPU: Nvidia GTX 1080, RAM: 8 GB (3) GPU: Nvidia Quadro M1000 M, RAM: 2 GB	10.2 18.0 –

Table 8. *Cont.*

	DSC	Inference Image Size	Used Resources	Time (s)
3D Segmentation				
Szentimrey et al. (2022) [12]	0.72	128 × 128 × 128	GPU: Nvidia V100-SXM2, CPU: Intel Xeon Gold 6148 Skylake (2.4 GHz), RAM:16GB	5.01
Szentimrey et al. (2024) [13]	0.757	128 × 128 × 128	GPU: Nvidia V100-SXM2, CPU: Intel Skylake (2.4 GHz), RAM: 32GB	2
2D UNet	**0.92**	D × 256 × 256 ($D \in [145, 309]$)	(1) Two GPUs: NVIDIA RTX A5000 (24 GB each), CPU: Intel Xeon Gold 5220R, RAM: 125 GB. (2) CPU: Intel Core i7-8565U (1.80 GHz), RAM: 8GB	0.785 7.859
Contextual UNet	**0.93**	D × 256 × 256 ($D \in [145, 309]$)	(1) Two GPUs: NVIDIA RTX A5000 (24 GB each), CPU: Intel Xeon Gold 5220R, RAM: 125 GB. (2) CPU: Intel Core i7-8565U (1.80 GHz), RAM: 8GB	0.708 8.899
2D nnUNet	**0.96**	D × H × W	(1) Two GPUs: NVIDIA RTX A5000 (24 GB each), CPU: Intel Xeon Gold 5220R, RAM: 125 GB (2) CPU: Intel Core i7-8565U (1.80 GHz), RAM: 8GB	3.759 121.047
3D nnUNet	**0.96**	D × H × W	(1) Two GPUs: NVIDIA RTX A5000 (24 GB each), CPU: Intel Xeon Gold 5220R, RAM: 125 GB (2) CPU: Intel Core i7-8565U (1.80 GHz), RAM: 8GB	5.695 244.089

Bold rows indicate the proposed methods.

In terms of computational efficiency, previous studies have primarily relied on dedicated GPU hardware, with 2D inference times ranging from 0.01 to 70.4 s and 3D inference times from 2 to 18 s [12,39–45]. Only a few works have explored low-resource scenarios. In contrast, our proposed models achieved competitive or faster inference across both high-end and CPU-only environments. Notably, the 2D UNet and Contextual UNet models achieved sub-second GPU inference of 0.79 and 0.71 s per 3D volume, respectively, and CPU inference of 7.9 and 8.9 s per 3D volume, indicating strong suitability for real-world clinical settings, including resource-constrained NICUs. Furthermore, our 2D and 3D nnUNet models delivered both top-tier accuracy and efficient GPU inference times (3.76 and 5.70 s), falling within the lower range of volume-level inference time reported in prior work, despite greater model complexity.

Although our work targets TB rather than structure-specific segmentation as in the methods summarized in Table 8, our proposed models achieve sub-second to few-second inference times on GPUs and remain usable on CPU-only systems, indicating feasibility for real-time or near-real-time clinical deployment. For instance, the Contextual UNet model, given its low computational demands, holds promise for broader clinical use and could be integrated into bedside US systems for real-time segmentation, where fast and lightweight processing is essential.

4.4. Clinical Relevance: Brain Volume Growth and Neurodevelopmental Outcomes

To evaluate the clinical value of automatically computed TBVs, we examined whether longitudinal brain growth differed between infants with favorable and adverse neurodevelopmental outcomes. This analysis used the filtered cohort described in Section 3.1, comprising 109 infants with 659 3D cUS volumes, all of whom had at least two valid volumes and did not present PHVD.

Linear Mixed-Effects Models (LMMs) were used to estimate TBV growth trajectories and to assess group differences via a PMA × neurodevelopmental outcome interaction term. LMMs appropriately account for the non-independence of longitudinal volumes by incorporating subject-specific random intercepts and random slopes for PMA, allowing

each infant to have an individualized baseline and growth rate. Model assumptions were evaluated using residual diagnostic plots, which confirmed approximately normal residual distributions and largely homoscedastic variance. Thus, the LMM was appropriate, and no non-parametric alternative was required. Final sample sizes (n = 82 favorable, n = 27 adverse) were sufficient to ensure stable slope estimates.

TBVs were obtained following automatic brain segmentation using our trained 3D nnUNet model (selected for its superior performance), and PMA at scan time served as the longitudinal axis. As previously described in Section 3.1, neurodevelopmental outcomes were classified at two years corrected age using the Bayley-III scale, where an adverse outcome corresponded to any domain score below 85.

Figure 5 shows the predicted TBV trajectories for both outcome groups. Infants with favorable outcomes exhibited a slightly steeper growth trajectory. The estimated TBV growth rate was 15.40 cm^3/week (95% CI: 14.40–16.39) in the adverse outcome group and 16.94 cm^3/week (95% CI: 14.83–19.06) in the favorable group, with a significant PMA × outcome interaction (p = 0.007). Domain-specific analyses also showed significant differences in growth rates for language ($p < 0.0001$), motor (p = 0.003), and cognitive (p = 0.001). Full slope estimates and p-values are provided in Table 9.

Table 9. Longitudinal TBV growth rates for favorable vs. adverse outcome groups across Bayley domains. Slopes are in cm^3/week.

Metric	Group	Slope	95% CI Lower	95% CI Upper	p-Value
Overall	Adverse	15.40	14.40	16.39	**0.007**
	Favorable	16.94	14.83	19.06	
Linguistic	Adverse	14.69	13.63	15.75	**<0.0001**
	Favorable	17.08	14.85	19.31	
Motor	Adverse	15.05	13.90	16.21	**0.003**
	Favorable	16.94	14.52	19.35	
Cognitive	Adverse	14.40	13.01	15.79	**0.001**
	Favorable	16.83	13.99	19.67	

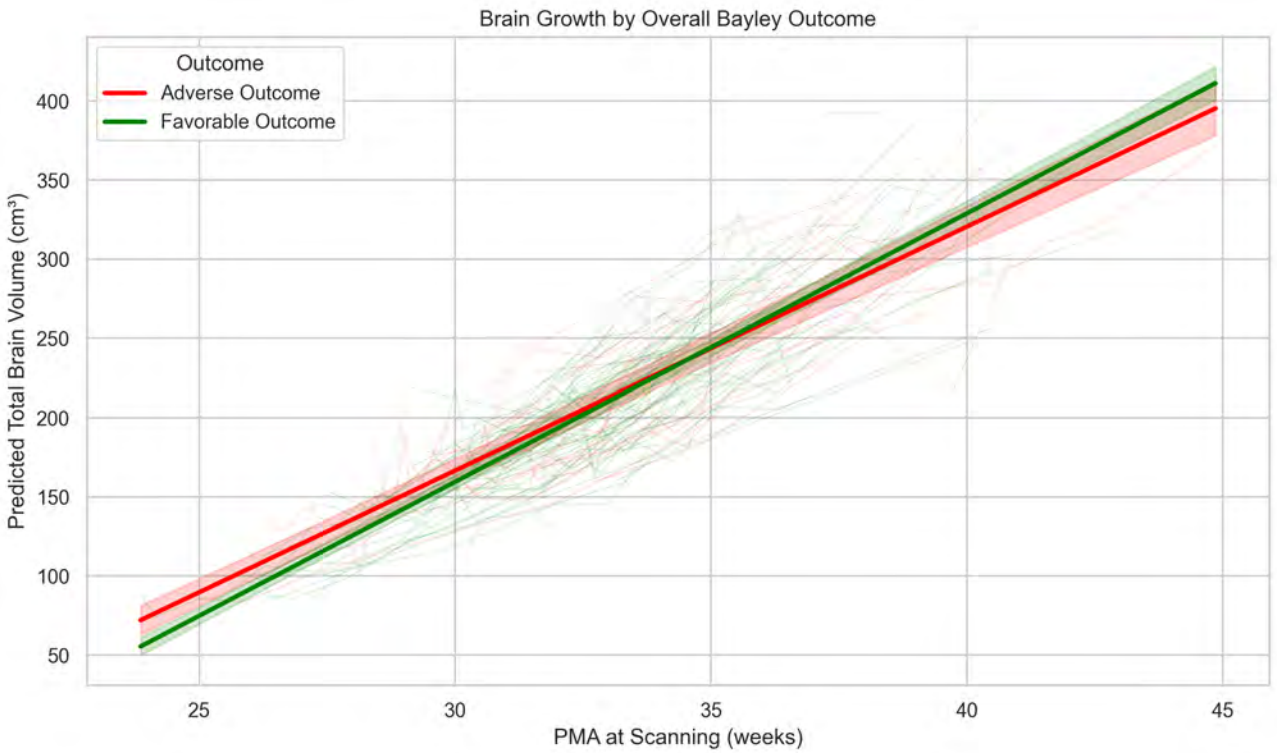
Bold p-values indicate statistically significant differences ($p < 0.05$).

These findings are consistent with prior MRI and US-based studies linking impaired early brain growth to later neurodevelopmental outcomes in preterm infants [15,16,51–55]. Our findings support the value of early longitudinal volumetric cUS analysis as a practical and noninvasive bedside tool for early neurodevelopmental risk stratification and demonstrate the potential of fully automated tools for faster and more precise quantitative assessment.

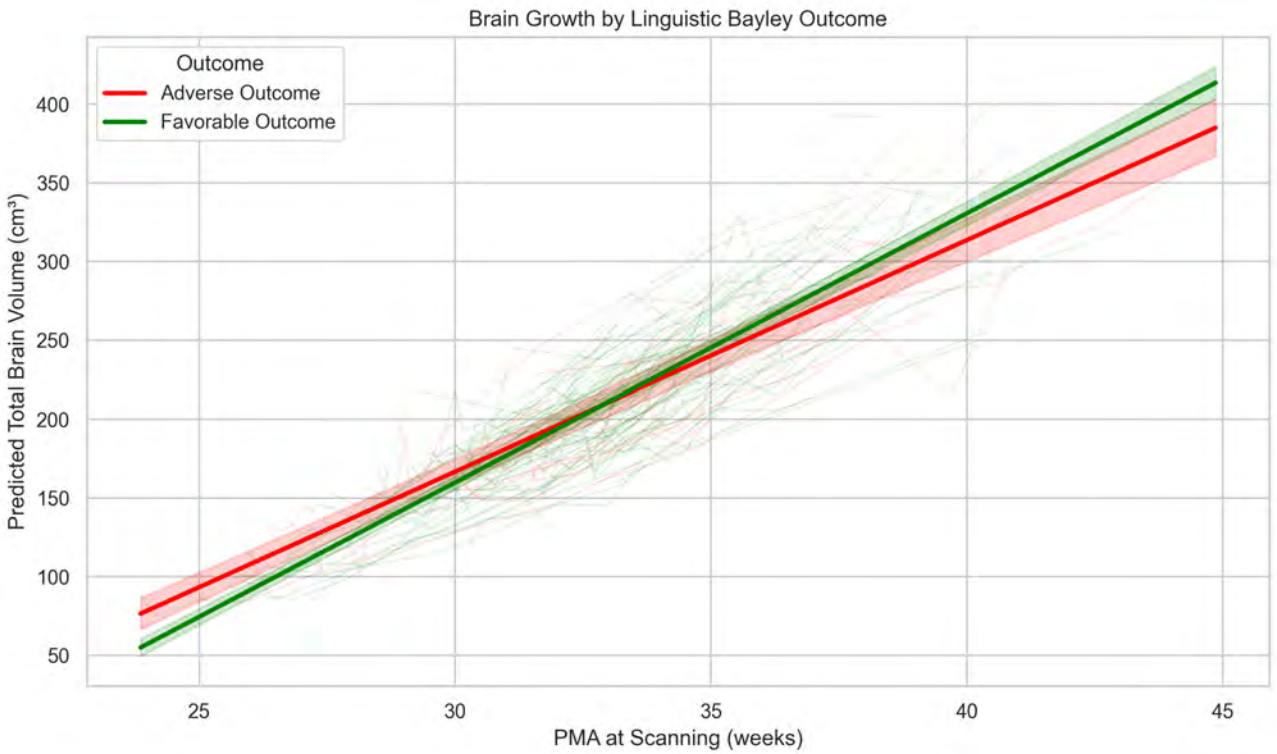
4.5. Limitations and Future Work

This analysis has several limitations. While the longitudinal mixed-effects analysis was sufficiently powered to model TBV trajectories, the modest sample size, particularly the small number of infants with adverse outcomes, may limit statistical power for subgroup analyses and segmentation model generalization. Additionally, the TBV provides a broad estimate of brain development, whereas regional brain volumes (e.g., the thalamus, cerebellum, and ventricles) may be more closely associated with specific neurodevelopmental outcomes [15,16,52–55]. Regional segmentation was beyond the scope of the present study because suitable annotations for these structures were not available. Developing such datasets would require extensive expert annotation by specialized neonatologists, or substantial training of annotators, particularly given the low tissue contrast, anatomical variability, and frequent abnormalities encountered in preterm neonatal cUS. Incorporating

structure-specific volumes could substantially improve predictive accuracy and therefore represents an important direction for future work.

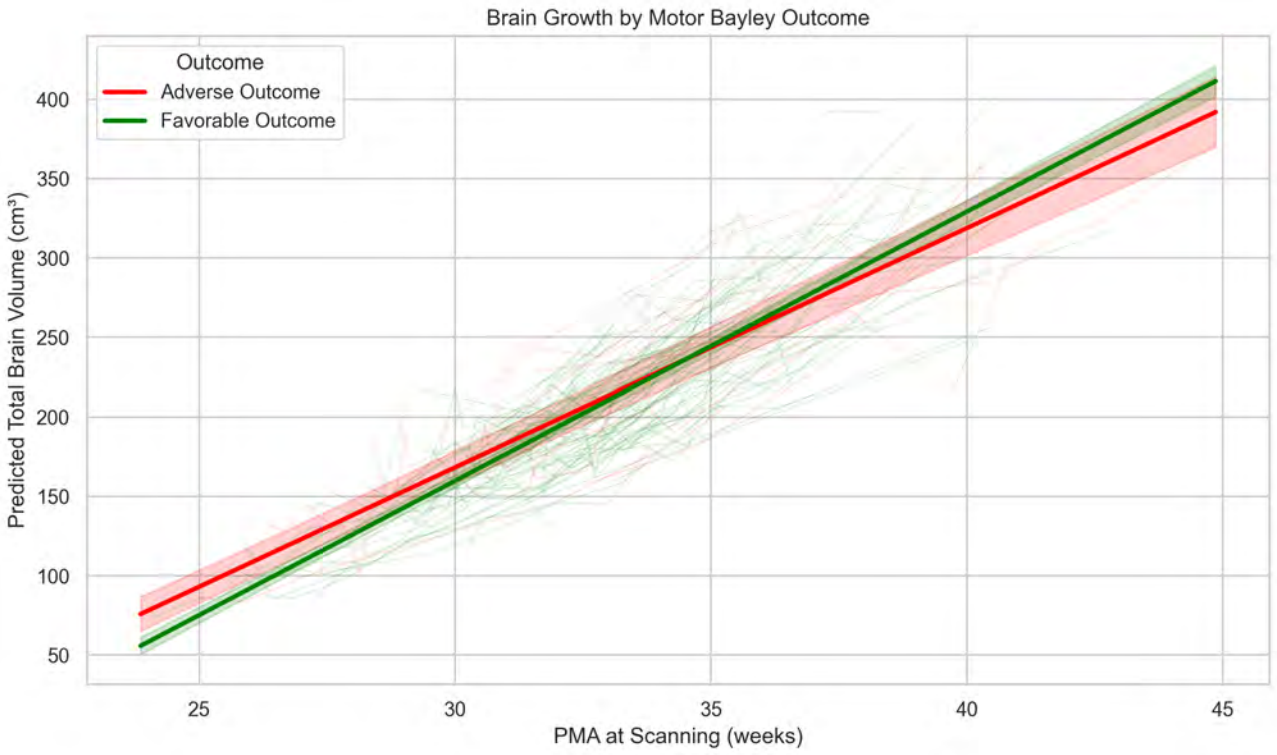


(**a**) Overall Bayley Outcome

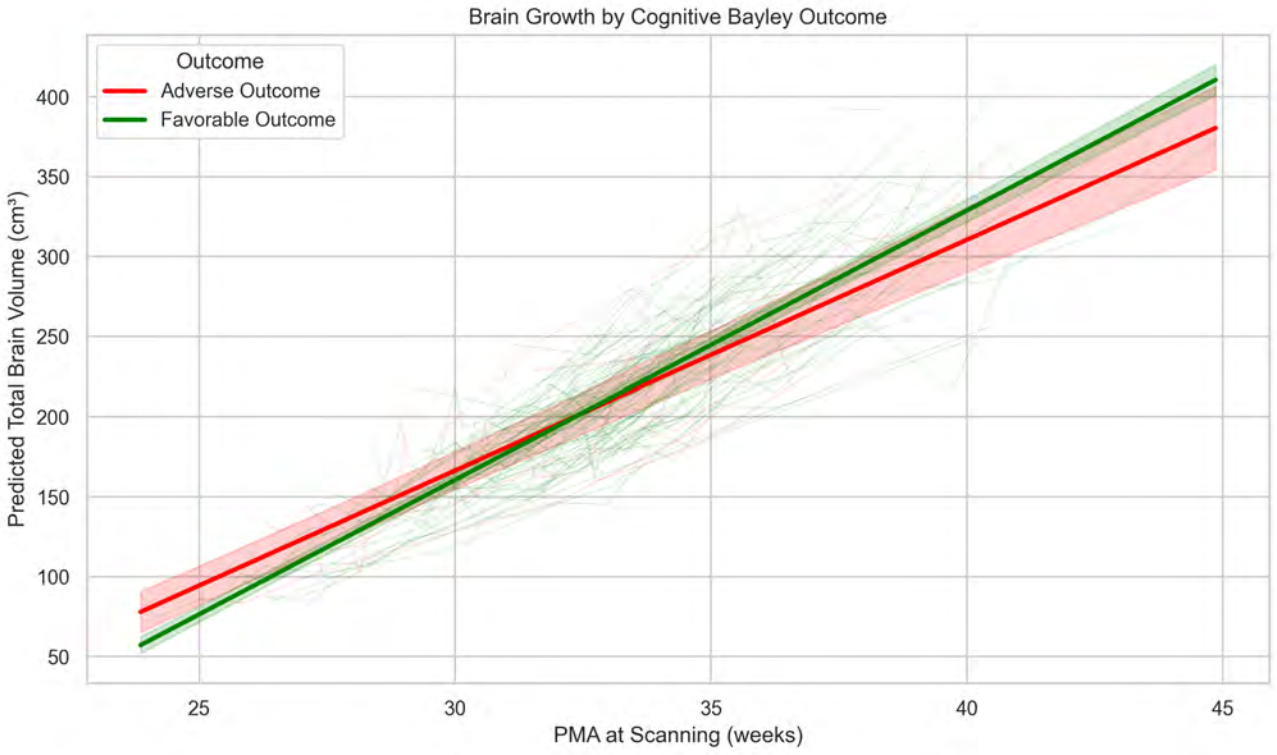


(**b**) Language Score

Figure 5. *Cont.*



(**c**) Motor Score



(**d**) Cognitive Score

Figure 5. Predicted TBV growth trajectories across postmenstrual age (PMA), stratified by neurodevelopmental outcome (adverse = red, favorable = green). Shaded regions show the 95% confidence intervals of the LMM-predicted fixed-effect curves. Panels illustrate growth differences for (**a**) overall Bayley outcome, (**b**) language, (**c**) motor, and (**d**) cognitive scores.

Further validation is warranted to confirm the robustness and generalization of our findings. Future work should include assessment of intra- and inter-observer variability of the GT annotations, and testing on external multi-center datasets acquired with different US systems. The GT segmentations were generated by two biomedical engineers following the same anatomical protocol under the supervision of an experienced neonatologist. While an older subset of the dataset had already been manually annotated on a slice-by-slice basis, a key-slice interpolation strategy was adopted for the newly annotated scans to accelerate the annotation process during dataset expansion while maintaining consistency with the established protocol. Given the substantial time required for manual volumetric annotation and the limited availability of expert clinical time, formal quantitative assessment of inter- and intra-observer variability was beyond the scope of this study. These steps are essential to ensure consistent performance across diverse imaging conditions and annotation variability.

Although deployment was evaluated on standard clinical hardware, optimization strategies to further reduce model complexity and inference latency were not explored in this study. Future work will focus on real-time optimization through approaches such as model pruning, quantization, and lightweight architectures (e.g., MobileUNet or efficient transformers), with the goal of reducing computational and memory requirements.

Future work will investigate foundation models and self-supervised learning strategies to reduce reliance on large manually annotated datasets and improve generalization across diverse neonatal US acquisitions. Another promising direction is multi-modal learning, integrating longitudinal 3D cUS with complementary information such as structural MRI and relevant clinical variables (e.g., gestational age, birth weight, and perinatal factors) to further improve early neurodevelopmental outcome prediction. In addition, although the explainability visualizations presented in this study provide valuable qualitative insight into model behavior, more rigorous and quantitatively validated interpretability frameworks are needed. This includes evaluating the consistency of explanations across independent datasets and clinical settings, integrating explainability with calibrated uncertainty estimation, and assessing the relationship between explanation quality and model reliability. Such developments may further improve the transparency, robustness, and clinical trustworthiness of AI-assisted neonatal neuroimaging systems.

Centralized server-based inference could also offer an alternative, although this raises challenges related to patient privacy, data governance, and secure transmission, especially in settings lacking robust IT infrastructure. Addressing these challenges will require national-level investment, standardized imaging protocols, and strategic support systems. Moreover, the absence of unified platforms for healthcare AI highlights the importance of developing end-to-end, standardized, and interoperable solutions tailored to clinical workflows in neonatal care.

Finally, although the accompanying web interface provides a practical mechanism for clinicians to visualize segmentations and retrieve brain-volume metrics, the application has not yet undergone formal validation in clinical settings. Future work will focus on clinician-centered evaluation to assess usability, interpretability, and integration within routine neonatal imaging workflows, as well as to determine its practical utility in real-world clinical environments.

5. Conclusions

This study addresses the underexplored challenge of automated brain analysis from 3D cUS in preterm neonates, aiming to make 3D cUS more accessible and clinically useful through scalable, reproducible, and clinically feasible analysis, particularly in resource-limited settings. We developed and evaluated DL-based methods for fully automated

TB segmentation from 3D cUS, emphasizing not only segmentation accuracy but also computational efficiency and clinical deployability.

Specifically, we compared plane-specific 2D UNets, a Contextual 2D UNet, and self-configuring nnUNet variants, evaluating segmentation accuracy, computational efficiency, and feasibility on both GPU and CPU-only hardware, and post hoc explainability through feature-attention maps, Grad-CAM, LRP-inspired relevance maps, saliency maps, and fold-wise uncertainty visualization. Importantly, although the nnUNet ensemble achieved the highest segmentation accuracy, the Contextual UNet emerged as the most clinically deployable model, maintaining high accuracy while enabling real-time inference on standard CPU hardware commonly available in NICU environments.

By combining segmentation accuracy, clinical relevance, deployment feasibility, and model interpretability, this work lays a practical foundation for integrating DL-based TB segmentation into routine NICU imaging pipelines, supporting early quantitative assessment of brain development in preterm infants.

Author Contributions: Conceptualization, R.K., J.P., I.B.-F., S.P.L.-L. and L.C.G.; methodology, R.K., J.P. and L.C.G.; software, R.K., J.P., S.T.H.S. and S.A.H.S.; validation, R.K., S.T.H.S. and S.A.H.S.; formal analysis, R.K.; investigation, R.K.; resources, J.P., I.B.-F., S.P.L.-L. and L.C.G.; data curation, R.K., J.P., I.B.-F. and S.P.L.-L.; writing—original draft preparation, R.K., S.T.H.S. and S.A.H.S.; writing—review and editing, all authors; visualization, R.K.; supervision, J.P., I.B.-F., S.P.L.-L., M.A.D. and L.C.G.; project administration, I.B.-F., S.P.L.-L., M.A.D. and L.C.G.; funding acquisition, I.B.-F., S.P.L.-L., M.A.D. and L.C.G. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the PARENT project, funded by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Actions (MSCA–ITN–ETN 2020), grant agreement No. 956394 (https://parenth2020.com/ (accessed on 29 May 2026)). This work was also supported by the GALATEA project, funded by the European Union's Horizon Europe research and innovation programme under the Marie Skłodowska-Curie Actions, grant agreement No. 101183057 (https://galateahe.eu/ (accessed on 29 May 2026)), and by the project PID2023-149783OB-I00, funded by MCIN/AEI/10.13039/501100011033 and by FEDER, UE.

Institutional Review Board Statement: The data used in this study were obtained from a longitudinal cohort of preterm infants recruited at Puerta del Mar University Hospital, Cádiz, Spain, since May 2018. The study was conducted in accordance with the Declaration of Helsinki, and the protocol was approved by the Biomedical Research Ethics Committee of Cádiz, Andalusia, Spain (protocol code PI0052/2017, and protocol code ITI-0019-2019, approved on 2 February 2020, and protocol code PID2019-110484RB-I00, approved on 30 July 2021). All clinical and imaging data were anonymized prior to analysis.

Informed Consent Statement: Informed consent for participation was obtained from the parents/legal guardians of all subjects involved in the study.

Data Availability Statement: The complete implementation presented in this study, including model training code, inference code, trained models, sample 3D cUS volumes, and the web application, is openly available at https://github.com/RoaKhaled/Total-Brain-Segmentation-from-cUS.git (accessed on 29 May 2026). The ultrasound imaging dataset presented in this study is not readily available due to ethical and privacy restrictions associated with the use of clinical data from human participants.

Conflicts of Interest: The author Syed Adil Hussain Shah was employed by the company GPI SpA. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

NDDs	Neurodevelopmental Disorders
cUS	Cranial Ultrasound
DL	Deep Learning
TB	Total Brain
TBV	Total Brain Volume
NICU	Neonatal Intensive Care Unit
ML	Machine Learning
SVM	Support Vector Machine
RF	Random Forests
DSC	Dice Similarity Coefficient
PMA	Postmenstrual Age
BSID-III	Bayley Scales of Infant and Toddler Development (Third Edition)
GA	Gestational Age
SD	Standard Deviation
PHVD	Post-Hemorrhagic Ventricular Dilatation
IQR	Interquartile Range
GT	Ground Truth
BCE	Binary Cross-Entropy
CE	Cross-Entropy
SGD	Stochastic Gradient Descent
RAS	Right–Anterior–Superior
VDE	Volumetric Difference Error
VSI	Volumetric Similarity Index
ASD	Average Surface Distance
FLOPs	Floating-Point Operations
LMMs	Linear Mixed-Effects Models

Appendix A. 3D cUS Acquisition and Export Details

3D cUS volumes were acquired using two GE Healthcare systems: Voluson i and Voluson S8. In both systems, transfontanellar volumetric scans were obtained using the manufacturer's 4D acquisition mode, which performs an automated mechanical sweep to generate a single 3D volume. Acquisitions were performed using a center frequency of 6.5 MHz and a fixed 90° scan angle from the third coronal plane, with an anterior-to-posterior beam sweep over 40–50 s.

The Voluson i used the SVNA5-8B transducer (5–8 MHz). Although no longer manufactured, this micro-convex transducer was part of the Voluson i platform and was suitable for neonatal transfontanellar imaging. The Voluson S8 used the RIC5-9A-RS transducer (4–10 MHz), which despite being categorized as an intracavitary transducer, provides excellent depth penetration and resolution for neonatal cUS due to its micro-convex surface, small footprint and compatibility with small acoustic windows. The approximate footprint of both probes ranges from 20 × 23 mm to 26 × 28 mm.

Raw volumes were saved and analyzed offline using 4D-View software (Version 17.0, 2001–2019 GE Healthcare Austria GmbH & Co OG, 2001–2019) and then exported from the 4D-view format (.vol) to the standardized .nrrd format, for compatibility with medical imaging toolkits. The resulting voxel spacing of the exported 3D volumes ranged from 0.366 mm to 0.528 mm across all three spatial dimensions. Because the Voluson S8 is a more recent system with enhanced imaging capabilities, the final dataset contains variable image quality across volumes.

Appendix B. Clinician-Facing Web Application

We developed a lightweight clinician-facing web application to demonstrate how the proposed total brain segmentation models can be integrated into a bedside research workflow for neonatal cranial ultrasound (cUS). The application allows users to load a 3D cUS volume, select a segmentation model, and visually inspect the resulting segmentation together with derived volumetric measures. All processing is performed locally on the user's machine, using available CPU or GPU resources, and no imaging data are transmitted to external servers, supporting privacy-preserving and offline use.

The application is implemented using a Flask-based Python backend for model inference and data handling, coupled with a lightweight HTML/JavaScript front-end for visualization and interaction. Pretrained Contextual UNet, 2D nnUNet and 3D nnUNet models are loaded locally and applied to the input volume. For each case, the application generates a total brain segmentation volume (NIfTI format), synchronized axial, coronal, and sagittal views with segmentation overlays, and an automatically computed total brain volume (TBV) based on voxel spacing metadata.

The full implementation, including pretrained models and the web application code, is publicly available to support reproducibility (the link to the publicly accessible repository has been temporarily withheld to preserve anonymity during peer review). A representative view of the web interface is shown in Figure A1.

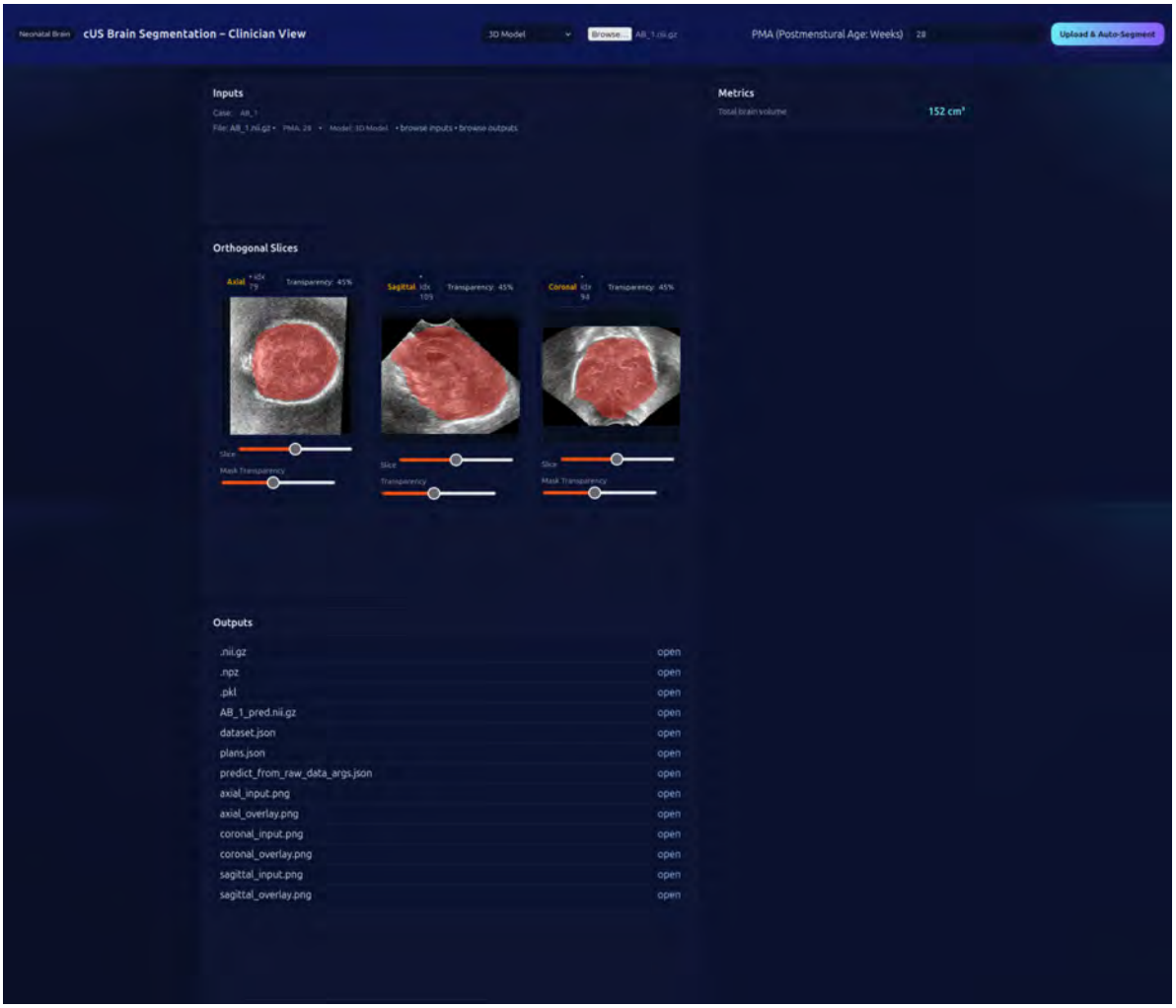


Figure A1. Clinician-facing web application illustrating model selection, orthogonal slice visualization with segmentation overlays, and derived total brain volume metrics.

Appendix C. Segmentation Performance Distributions

Box plots summarizing Dice, volumetric, and surface-based segmentation metrics for the evaluated models, including plane-specific 2D UNets and comparisons between the axial 2D UNet, Contextual UNet, and nnUNet ensemble.

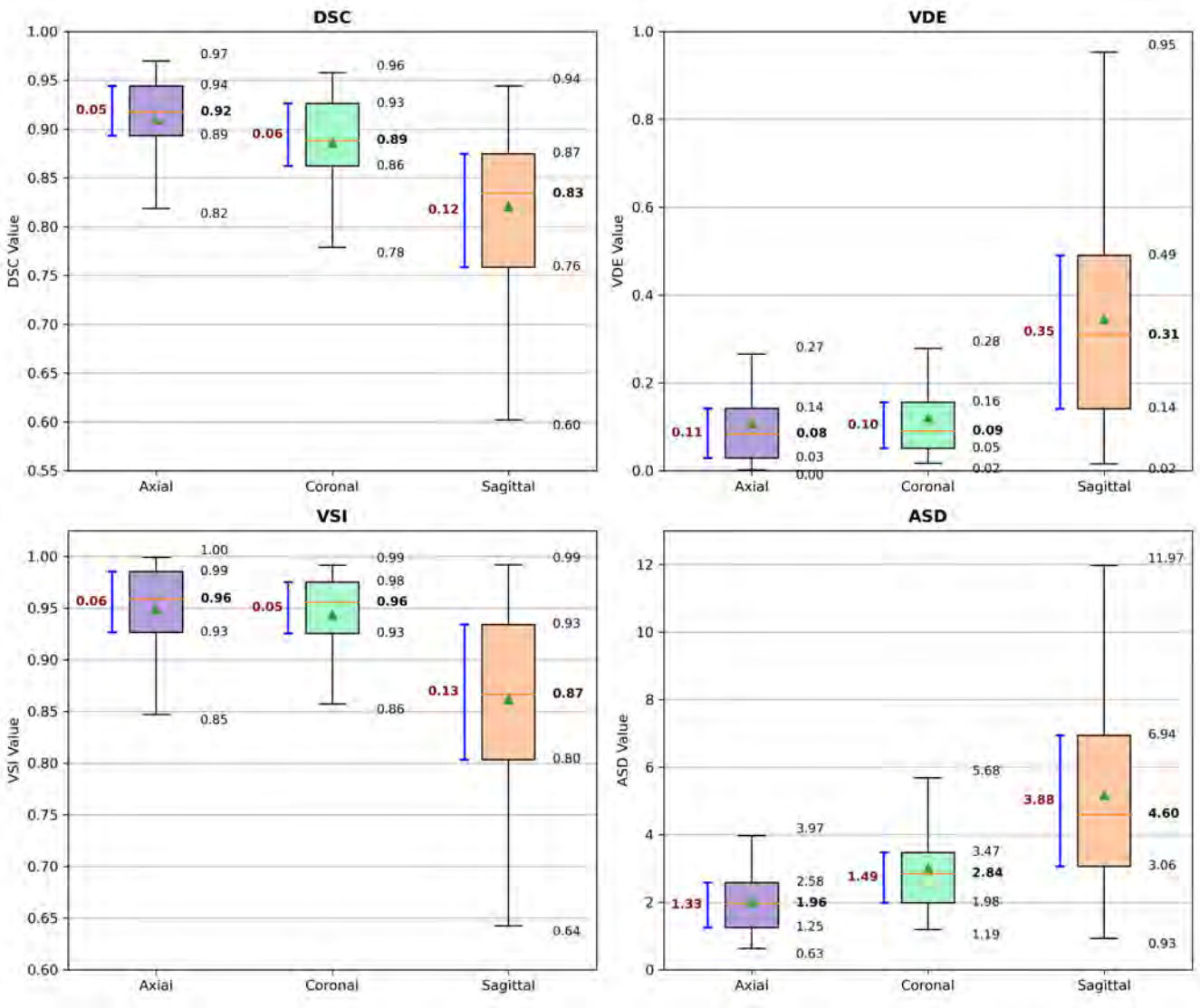


Figure A2. Comparison of segmentation performance of 2D UNet trained across different anatomical planes: axial, coronal, and sagittal, and reported using DSC, VDE, VSI, and ASD metrics. Boxplots show the median (bold), mean (green triangle), standard deviation (red), and minimum, Q1, Q3, and maximum values (black). Colors correspond to the axial (purple), coronal (green), and sagittal (orange) planes.

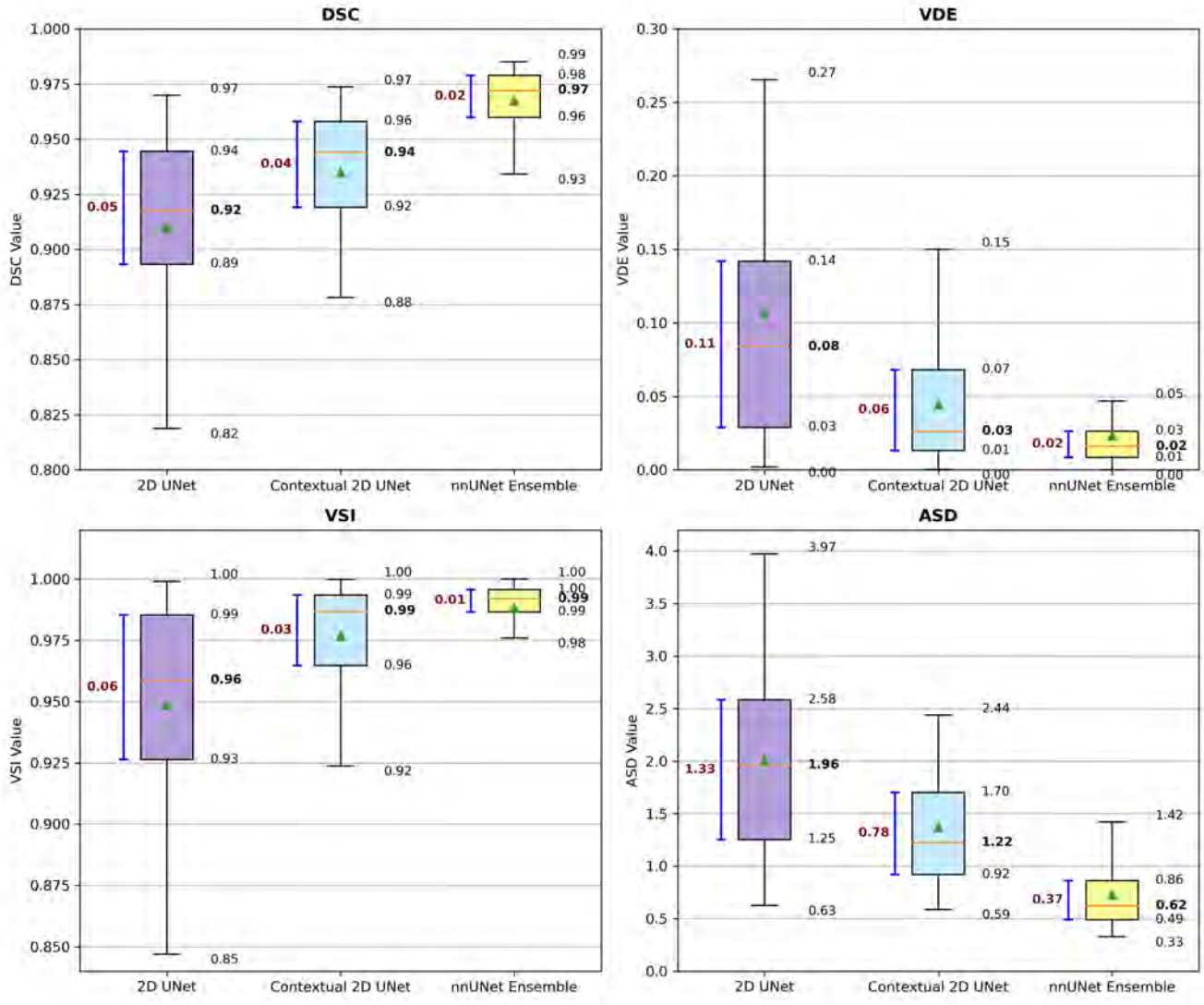


Figure A3. Comparison of segmentation performance between axial 2D UNet, Contextual UNet, and nnUNet Ensemble, reported using DSC, VDE, VSI, and ASD metrics. Boxplots show the median (bold), mean (green triangle), standard deviation (red), and minimum, Q1, Q3, and maximum values (black). Colors correspond to the axial (purple), coronal (green), and sagittal (orange) planes. Note that y-axes ranges differ between Figure A2 and this figure to preserve readability, as the distributions vary substantially across models.

Appendix D. Computational Evaluation Details and Clinical Hardware Context

Appendix D.1. Inference Time Measurement Protocol

For 2D and Contextual UNet models, volume-level inference time was computed by summing the per-slice inference times over all slices in each volume, then the volume-level times were averaged to produce the reported volume inference time. Reported postprocessing time accounts for 3D reconstruction, hole filling, largest connected component filtering, and smoothing.

For nnUNet models inference time, approximation was necessary as nnUNet operates as an end-to-end automated framework that does not expose per-slice or per-volume inference timings. Therefore, for the volume-level inference time of the 2D nnUNet, we computed an average slice-level inference time by dividing the total batch-inference time by the total number of slices. This value was then multiplied by the slice count of each volume to obtain its volume inference time, after which the volume-level times were averaged to

produce the reported volume inference time. For the volume-level inference time of the 3D nnUNet, the total batch-inference time was divided by the number of volumes.

Appendix D.2. Volume-by-Volume Inference with Repeated Model Loading

In addition to persistent-model inference, we evaluated volume-by-volume inference, where the model is reloaded from disk for each processed US volume. The corresponding inference times are reported in Table A1, complementing the main results presented in Section 4.2.

Table A1. GPU inference time of all the evaluated models with: (1) Raw Inference—Batched; (2) Raw Inference—Non-Batched; (3) Inference with Output Saving—Non-Batched. Raw inference refers to inference without saving the output. Batched inference refers to loading the model once and reusing it per volume. Non-Batched inference refers to model reloading per volume. All reported inference times correspond to deployment using a single trained model. Reported values include slice-level and volume-level inference times (in seconds). Note: nnUNet automatically saves output, therefore, the reported inference time here does not change between raw inference and inference with output saving.

	2D UNet	Contextual UNet	2D nnUNet	3D nnUNet
Raw Inference—Batched				
2D slice inference time (s)	0.001	0.001	0.022	–
3D volume inference time (s)	0.253	0.225	3.759	5.695
Raw Inference—Non-Batched				
2D slice inference time (s)	0.001	0.001	0.076	–
3D volume inference time (s)	0.253	0.223	13.268	15.984
Inference with Output Saving—Non-Batched				
2D slice inference time (s)	0.004	0.004	0.076	–
3D volume inference time (s)	0.782	0.662	13.268	15.984

Appendix D.3. Clinical Hardware Tiers in NICU Settings

In typical clinical settings, especially within NICUs, the computing infrastructure varies significantly depending on the specific task. Therefore, several hardware tiers can be encountered:

1. Standard clinical workstations used for image review via PACS or EHR systems typically rely on mid-range or older-generation Intel Core i5/i7 CPUs (e.g., 6th–10th generation) with 8–16 GB of RAM and integrated GPUs for display purposes [22,59,60]. Dedicated high-performance GPUs are uncommon in these settings [22].
2. Dedicated US systems often incorporate embedded processing units optimized for real-time image acquisition and 2D analysis. Modern high-end US systems increasingly integrate specialized embedded processors (DSPs, FPGAs, embedded CPUs) and, in some cases, embedded GPUs (e.g., NVIDIA Jetson modules) to support 3D/4D rendering and on-device AI acceleration [22,61].
3. Centralized hospital servers may be available for computationally intensive tasks, providing access to enterprise-grade GPUs (e.g., NVIDIA Tesla or A100/H100) for large-scale AI inference through PACS-integrated or container-based infrastructures [62–64]. However, network latency and queuing can limit real-time responsiveness, which is particularly critical for point-of-care NICU applications.

In many NICU settings, bedside image review and analysis are performed on lightweight laptops or mobile workstations due to mobility, cost, and data privacy constraints [22], motivating our CPU-only evaluation described in Section 4.2.

References

1. Volpe, J.J. Primary Neuronal Dysmaturation in Preterm Brain: Important and Likely Modifiable. *J. Neonatal-Perinat. Med.* **2020**, *14*, 1–6. [CrossRef] [PubMed]
2. Sarda, S.P.; Sarri, G.; Siffel, C. Global Prevalence of Long-Term Neurodevelopmental Impairment Following Extremely Preterm Birth: A Systematic Literature Review. *J. Int. Med. Res.* **2021**, *49*, 3000605211028026. [CrossRef] [PubMed]
3. Dan, A.M.; Vasilescu, D.I.; Dragomir, I.; Vasilescu, S.L.; Voicu, D.; Cîrstoiu, M.M. Cranial Ultrasonography—Standards in Diagnosis of Intraventricular Hemorrhage and Ventricular Dilatation in Premature Neonates. *Children* **2025**, *12*, 768. [CrossRef] [PubMed]
4. Van, P.V.; Beck, J.; Meunier, H.; Venot, P.; Caby, G.M.; Bednarek, N.; Loron, G. Assessment of Brain Two-Dimensional Metrics in Infants Born Preterm at Term Equivalent Age: Correlation of Ultrasound Scans with Magnetic Resonance Imaging. *Front. Pediatr.* **2022**, *10*, 961556. [CrossRef] [PubMed]
5. Dudink, J.; Steggerda, S.J.; Horsch, S. State-of-the-Art Neonatal Cerebral Ultrasound: Technique and Reporting. *Pediatr. Res.* **2020**, *87*, 3–12. [CrossRef] [PubMed]
6. Arulnathan, E.; Saxena, S. Principles of a controlled imaging environment for neonatal brain MRI: Strategies for optimal image quality and safety. *Neuroradiol. J.* **2024**, *38*, 394–400. [CrossRef] [PubMed]
7. Liu, L. Application of Brain Ultrasound in Premature Infants with Brain Injury. *Front. Neurol.* **2023**, *14*, 1095280. [CrossRef] [PubMed]
8. Blanchard, M.; Martin, M.; Quétin, P.; Delachartre, P. Quantitative Comparison of 3D Freehand Ultrasound and MRI Images of the Neonatal Brain. In Proceedings of the 2020 IEEE International Ultrasonics Symposium (IUS), Las Vegas, NV, USA, 7–11 September 2020; pp. 1–4. [CrossRef]
9. Kebaya, L.; Stubbs, K.; Lo, M.; Al-Saoud, S.; Karat, B.; St Lawrence, K.; De Ribaupierre, S.; Duerden, E. Three-dimensional cranial ultrasound and functional near-infrared spectroscopy for bedside monitoring of intraventricular hemorrhage in preterm neonates. *Sci. Rep.* **2023**, *13*, 3730. [CrossRef] [PubMed]
10. Kim, Y.J.; Choi, Y.H.; Cho, H.; Lee, S.M.; Park, J.E.; Cheon, J.; Kim, W.S.; Kim, I.O. Comparison Between 3-Dimensional Cranial Ultrasonography and Conventional 2-Dimensional Cranial Ultrasonography in Neonates: Impact on Reinterpretation. *Ultrasonography* **2017**, *37*, 63–70. [CrossRef] [PubMed]
11. Benavente-Fernández, I.; Ruiz-González, E.; Lubián-Gutiérrez, M.; Lubián-Fernández, S.P.; Fontela, Y.C.; Roca-Cornejo, C.; Olmo-Duran, P.; Lubián-López, S.P. Ultrasonographic Estimation of Total Brain Volume: 3D Reliability and 2D Estimation. Enabling Routine Estimation During NICU Admission in the Preterm Infant. *Front. Pediatr.* **2021**, *9*, 708396. [CrossRef] [PubMed]
12. Szentimrey, Z.; de Ribaupierre, S.; Fenster, A.; Ukwatta, E. Automated 3D U-Net based Segmentation of Neonatal Cerebral Ventricles from 3D Ultrasound Images. *Med. Phys.* **2022**, *49*, 1034–1046. [CrossRef] [PubMed]
13. Szentimrey, Z.; Al-Hayali, A.; de Ribaupierre, S.; Fenster, A.; Ukwatta, E. Semi-supervised learning framework with shape encoding for neonatal ventricular segmentation from 3D ultrasound. *Med. Phys.* **2024**, *51*, 6134–6148. [CrossRef] [PubMed]
14. van Gils, R.H.; Helder, O.; Kornelisse, R.F.; Reiss, I.; Dankelman, J. 3D scanner measuring preterm infants' head circumference and cranial volume: Validation in a simulated care setting. *Front. Med. Eng.* **2024**, *2*, 1463793. [CrossRef]
15. Ruiz-González, E.; Benavente-Fernández, I.; Lubián-Gutiérrez, M.; Segado-Arenas, A.; Zafra-Rodríguez, P.; Méndez-Abad, P.; Lubián-López, S.P. Ultrasonographic Evaluation of the Early Brain Growth Pattern in Very Low Birth Weight Infants. *Pediatr. Res.* **2023**, *94*, 296–303. [CrossRef] [PubMed]
16. González, E.; López, S.; Natalia, J.L.; Segado-Arenas, A.; Lubián-Gutiérrez, M.; Almagro, Y.; Zafra, P.; Méndez-Abad, P.; Benavente-Fernández, I. Relationship of Early Brain Growth Pattern Measured by Ultrasound with Neurological Outcome at Two Years of Age in Very Low Birth Weight Infants. *Eur. J. Pediatr.* **2023**, *182*, 5119–5129. [CrossRef] [PubMed]
17. Khaled, R.; Ortega-Leon, A.; Pizarro, J.; Benavente-Fernández, I.; Lubián López, S.P.; Gontard, L.C. Automatic Neonatal Cranial Ultrasound Segmentation Using Deep Learning: A Review. In *Proceedings of the Neurodevelopmental Impairments in Preterm Children: DETERMINED (2022)*; CEUR-WS.org: Ljubljana, Slovenia, 2023; Volume 3363, pp. 49–64. Available online: https://ceur-ws.org/Vol-3363/paper05.pdf (accessed on 29 May 2026).
18. Ciceri, T.; Squarcina, L.; Giubergia, A.; Bertoldo, A.; Brambilla, P.; Peruzzo, D. Review on deep learning fetal brain segmentation from Magnetic Resonance images. *Artif. Intell. Med.* **2023**, *143*, 102608. [CrossRef] [PubMed]
19. Mahalakshmi, P.; Rajagopal, A. Lightweight and Real-Time Deep Learning Models for Fetal Head Segmentation: A Review of Techniques and Applications. In Proceedings of the 2025 International Conference on NexGen Networks and Cybernetics (IC2NC), Erode, India, 1–3 December 2025; pp. 929–933. [CrossRef]

20. Suha, S.A.; Shahriyar, R. Deep learning for fetal brain imaging: A systematic review and framework towards privacy-preserving neurodevelopmental informatics. *Neurosci. Inform.* **2025**, *5*, 100241. [CrossRef]
21. Fang, Y.; Wang, S.; Zhang, Y.; Yuan, C.; Song, L.; Zhao, P.; Su, F.; Liu, J.; Wu, L. A review of deep learning-based segmentation and registration of infant brain MRI. *Biomed. Signal Process. Control* **2026**, *112*, 108676. [CrossRef]
22. Xu, Y.; Khan, T.; Song, Y.; Meijering, E. Edge Deep Learning in Computer Vision and Medical Diagnostics: A Comprehensive Survey. *Artif. Intell. Rev.* **2025**, *58*, 93. [CrossRef]
23. Ahmed, M.; Spooner, B.; Isherwood, J.; Lane, M.; Orrock, E.; Dennison, A. A Systematic Review of the Barriers to the Implementation of Artificial Intelligence in Healthcare. *Cureus* **2023**, *15*, e46454. [CrossRef] [PubMed]
24. Schouten, J.; Kalden, M.; Twist, E.; Reiss, I.; Gommers, D.; van Genderen, M.; Taal, H. From Bytes to Bedside: A Systematic Review on the Use and Readiness of Artificial Intelligence in the Neonatal and Pediatric Intensive Care Unit. *Intensive Care Med.* **2024**, *50*, 1767–1777. [CrossRef] [PubMed]
25. Kim, S.; Fischetti, C.; Guy, M.; Hsu, E.; Fox, J.; Young, S.D. Artificial Intelligence (AI) Applications for Point of Care Ultrasound (POCUS) in Low-Resource Settings: A Scoping Review. *Diagnostics* **2024**, *14*, 1669. [CrossRef] [PubMed]
26. Devi, C.N.; Chandrasekharan, A.; Sundararaman, V.K.; Alex, Z.C. Neonatal Brain MRI Segmentation: A review. *Comput. Biol. Med.* **2015**, *64*, 163–178. [CrossRef] [PubMed]
27. Jaware, T.H.; Khanchandani, K.B.; Zurani, A. An Accurate Automated Local Similarity Factor-Based Neural Tree Approach toward Tissue Segmentation of Newborn Brain MRI. *Am. J. Perinatol.* **2018**, *36*, 1157–1170. [CrossRef] [PubMed]
28. Makropoulos, A.; Gousias, I.S.; Ledig, C.; Aljabar, P.; Serag, A.; Hajnal, J.V.; Edwards, A.D.; Counsell, S.J.; Rueckert, D. Automatic Whole Brain MRI Segmentation of the Developing Neonatal Brain. *IEEE Trans. Med. Imaging* **2014**, *33*, 1818–1831. [CrossRef] [PubMed]
29. Gui, L.; Lisowski, R.; Faundez, T.; Hüppi, P.S.; Lazeyras, F.; Kocher, M. Morphology-Driven Automatic Segmentation of MR Images of the Neonatal Brain. *Med. Image Anal.* **2012**, *16*, 1565–1579. [CrossRef] [PubMed]
30. Richter, L.; Fetit, A.E. Accurate Segmentation of Neonatal Brain MRI with Deep Learning. *Front. Neuroinform.* **2022**, *16*, 1006532. [CrossRef] [PubMed]
31. Moeskops, P.; Benders, M.J.; Chiţ, S.M.; Kersbergen, K.J.; Groenendaal, F.; de Vries, L.S.; Viergever, M.A.; Išgum, I. Automatic Segmentation of MR Brain Images of Preterm Infants Using Supervised Classification. *NeuroImage* **2015**, *118*, 628–641. [CrossRef] [PubMed]
32. Saladi, S.; Karuna, Y.; Koppu, S.; Reddy, G.R.; Mohan, S.; Mallik, S.; Qin, H. Segmentation and Analysis Emphasizing Neonatal MRI Brain Images Using Machine Learning Techniques. *Mathematics* **2023**, *11*, 285. [CrossRef]
33. Ding, Y.; Acosta, R.P.; Enguix, V.; Suffren, S.; Ortmann, J.; Luck, D.; Dolz, J.; Lodygensky, G.A. Using Deep Convolutional Neural Networks for Neonatal Brain Image Segmentation. *Front. Neurosci.* **2020**, *14*, 207. [CrossRef] [PubMed]
34. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*; Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F., Eds.; Springer: Cham, Switzerland, 2015; pp. 234–241. [CrossRef]
35. Zhou, S.K.; Greenspan, H.; Davatzikos, C.; Duncan, J.S.; van Ginneken, B.; Madabhushi, A.; Prince, J.L.; Rueckert, D.; Summers, R.M. A Review of Deep Learning in Medical Imaging: Imaging Traits, Technology Trends, Case Studies with Progress Highlights, and Future Promises. *Proc. IEEE* **2021**, *109*, 820–838. [CrossRef] [PubMed]
36. Li, M.; Jiang, Y.; Zhang, Y.; Haisheng, Z. Medical Image Analysis Using Deep Learning Algorithms. *Front. Public Health* **2023**, *11*, 1273253. [CrossRef] [PubMed]
37. Shen, D.D.; Bao, S.L.; Wang, Y.; Chen, Y.C.; Zhang, Y.C.; Li, X.C.; Ding, Y.C.; Jia, Z.Z. An Automatic and Accurate Deep Learning-Based Neuroimaging Pipeline for the Neonatal Brain. *Pediatr. Radiol.* **2023**, *53*, 1685–1697. [CrossRef] [PubMed]
38. Isensee, F.; Jaeger, P.F.; Kohl, S.A.A.; Petersen, J.; Maier-Hein, K.H. nnU-Net: A Self-Configuring Method for Deep Learning-Based Biomedical Image Segmentation. *Nat. Methods* **2021**, *18*, 203–211. [CrossRef] [PubMed]
39. Gontard, L.C.; Pizarro, J.; Sanz-Peña, B.; Lubián López, S.P.; Benavente-Fernández, I. Automatic Segmentation of Ventricular Volume by 3D Ultrasonography in Post Haemorrhagic Ventricular Dilatation Among Preterm Infants. *Sci. Rep.* **2021**, *11*, 567. [CrossRef] [PubMed]
40. Wang, P.; Cuccolo, N.G.; Tyagi, R.; Hacihaliloglu, I.; Patel, V.M. Automatic Real-Time CNN-based Neonatal Brain Ventricles Segmentation. In Proceedings of the 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), Washington, DC, USA, 4–7 April 2018; pp. 716–719. [CrossRef]
41. Valanarasu, J.M.J.; Yasarla, R.; Wang, P.; Hacihaliloglu, I.; Patel, V.M. Learning to Segment Brain Anatomy from 2D Ultrasound with Less Data. *arXiv* **2019**, arXiv:1912.08364.
42. Tabrizi, P.R.; Mansoor, A.; Obeid, R.; Cerrolaza, J.J.; Perez, D.A.; Zember, J.; Penn, A.; Linguraru, M.G. Ultrasound-Based Phenotyping of Lateral Ventricles to Predict Hydrocephalus Outcome in Premature Neonates. *IEEE Trans. Biomed. Eng.* **2020**, *67*, 3026–3034. [CrossRef] [PubMed]

43. Martin, M.; Sciolla, B.; Sdika, M.; Quétin, P.; Delachartre, P. Automatic Segmentation and Location Learning of Neonatal Cerebral Ventricles in 3D Ultrasound Data Combining CNN and CPPN. *Comput. Biol. Med.* **2021**, *131*, 104268. [CrossRef] [PubMed]
44. Valanarasu, J.M.J.; Sindagi, V.A.; Hacihaliloglu, I.; Patel, V.M. KiU-Net: Overcomplete Convolutional Architectures for Biomedical Image and Volumetric Segmentation. *IEEE Trans. Med. Imaging* **2022**, *41*, 965–976. [CrossRef] [PubMed]
45. Martin, M.; Sciolla, B.; Sdika, M.; Wang, X.; Quetin, P.; Delachartre, P. Automatic Segmentation of the Cerebral Ventricle in Neonates Using Deep Learning with 3D Reconstructed Freehand Ultrasound Imaging. In Proceedings of the 2018 IEEE International Ultrasonics Symposium (IUS), Kobe, Japan, 22–25 October 2018; pp. 1–4. [CrossRef]
46. Regalado, M.; Carreras, N.; Mata, C.; Oliver, A.; Lladó, X.; Agut, T. Automatic Segmentation of Sylvian Fissure in Brain Ultrasound Images of Pre-Term Infants Using Deep Learning Models. *Ultrasound Med. Biol.* **2025**, *51*, 543–550. [CrossRef] [PubMed]
47. Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Torr, P.H.; et al. Rethinking Semantic Segmentation from a Sequence-to-sequence Perspective with Transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 6881–6890. [CrossRef]
48. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. TransUnet: Transformers Make Strong Encoders for Medical Image Segmentation. *arXiv* **2021**, arXiv:2102.04306. [CrossRef]
49. Hatamizadeh, A.; Xu, Z.; Yang, D.; Li, W.; Roth, H.; Xu, D. UNetFormer: A Unified Vision Transformer Model and Pre-Training Framework for 3D Medical Image Segmentation. *arXiv* **2022**, arXiv:2204.00631. [CrossRef]
50. Neha, F.; Bhati, D.; Shukla, D.K.; Dalvi, S.M.; Mantzou, N.; Shubbar, S. An analytics-driven review of U-Net for medical image segmentation. *Healthc. Anal.* **2025**, *8*, 100416. [CrossRef]
51. Mayrink, M.L.d.; Villela, L.D.; Méio, M.D.B.B.; Soares, F.V.M.; de Abranches, A.D.; Nehab, S.R.G.; Reis, A.B.R.; de Paula Barros, L.B.; de Rodrigues, M.C.C.; Junior, S.C.G.; et al. The trajectory of head circumference and neurodevelopment in very preterm newborns during the first two years of life: A cohort study. *J. Pediatr.* **2024**, *100*, 483–490. [CrossRef] [PubMed]
52. Trimarco, E.; Jafrasteh, B.; Jiménez-Luque, N.; Marín Almagro, Y.; Román Ruiz, M.; Lubián Gutiérrez, M.; Ruiz González, E.; Segado Arenas, A.; Lubián-López, S.P.; Benavente-Fernández, I. Thalamic volume in very preterm infants: Associations with severe brain injury and neurodevelopmental outcome at two years. *Front. Neurol.* **2024**, *15*, 1427273. [CrossRef] [PubMed]
53. Sheng, M.; Guo, T.; Mabbott, C.; Chau, V.; Synnes, A.; Vries, L.; Grunau, R.; Miller, S. Ventricular volume in infants born very preterm: Relationship with brain maturation and neurodevelopment at 4.5 years. *J. Pediatr.* **2022**, *248*, 51–58.e2. [CrossRef] [PubMed]
54. Pagnozzi, A.M.; van Eijk, L.; Pannek, K.; Boyd, R.N.; Saha, S.; George, J.; Bora, S.; Bradford, D.; Fahey, M.; Ditchfield, M.; et al. Early brain morphometrics from neonatal MRI predict motor and cognitive outcomes at 2-years corrected age in very preterm infants. *NeuroImage* **2023**, *267*, 119815. [CrossRef] [PubMed]
55. Pagnozzi, A.M.; Pannek, K.; Boyd, R.N.; van Eijk, L.; George, J.M.; Bora, S.; Bradford, D.; Fahey, M.; Ditchfield, M.; Malhotra, A.; et al. Brain MRI before and at term equivalent age predicts motor and cognitive outcomes in very preterm infants. *NeuroImage Rep.* **2025**, *5*, 100262. [CrossRef] [PubMed]
56. Bayley, N. *Bayley Scales of Infant and Toddler Development*, 3rd ed.; The Psychological Corporation: San Antonio, TX, USA, 2006.
57. Johnson, S.; Moore, T.; Marlow, N. Using the Bayley-III to assess neurodevelopmental delay: Which cut-off should be used? *Pediatr. Res.* **2014**, *75*, 670–674. [CrossRef] [PubMed]
58. Jafrasteh, B.; Lubián López, S.P.; Benavente-Fernández, I. MELAGE: A Purely Python Based Neuroimaging Software (Neonatal). *arXiv* **2023**, arXiv.2309.07175. [CrossRef]
59. Sevocity. System Requirements—Sevocity EHR Service. Available online: https://www.sevocity.com/ehr-service/system-requirements/ (accessed on 29 May 2025).
60. Nexus SISINF. CHILI PACS Hardware and Operating System Requirements. 2017. Available online: https://www.nexus-spain.es/departamentales/ris-pacs?file=files/nexus-sisinf/Downloads/hw-sles-requirements_170518_en.pdf&cid=28107 (accessed on 29 May 2025).
61. Bench International. The Future of Medical Devices: From Hardware to AI-Powered Software. Available online: https://benchinternational.com/the-future-of-medical-devices-from-hardware-to-ai-powered-software/ (accessed on 29 May 2025).
62. Liquid Web. GPU Servers for AI and Big Data Analytics in Healthcare. Available online: https://www.liquidweb.com/gpu/use-cases/healthcare/ (accessed on 29 May 2025).
63. Super Micro Computer, Inc. Healthcare & IoT Edge Solutions. Available online: https://www.supermicro.com/en/solutions/iot-edge/healthcare (accessed on 29 May 2025).
64. AceCloud Blog. Saving Lives with GPUs: The Future of Healthcare Imaging and Diagnosis. 2024. Available online: https://acecloud.ai/blog/future-of-healthcare-imaging-and-diagnosis (accessed on 29 May 2025).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.