

Article

Cross-Sequence Robustness of Deep Learning for Placental MRI Segmentation: A Patient-Matched Study in Placenta Accreta Spectrum

Rawan AlSaad ^{1,*} , Sanaa Ahmed ², Nader Al-Dewik ^{3,4} , Thomas Farrell ^{5,6} and Rajat Thomas ¹ 

¹ AI Center for Precision Health, Weill Cornell Medicine-Qatar, Doha P.O. Box 24144, Qatar

² Department of Clinical Imaging, Women's Wellness and Research Center, Hamad Medical Corporation, Doha P.O. Box 3050, Qatar

³ Translational and Precision Medicine Research Lab, Women's Wellness and Research Center, Hamad Medical Corporation, Doha P.O. Box 3050, Qatar

⁴ Vice President for Health and Medical Sciences Office, Qatar University, Doha P.O. Box 2713, Qatar

⁵ Department of Obstetrics and Gynecology, Women's Wellness and Research Center, Hamad Medical Corporation, Doha P.O. Box 3050, Qatar

⁶ College of Medicine, Qatar University, Doha P.O. Box 2713, Qatar

* Correspondence: rta4003@qatar-med.cornell.edu

Abstract

Background: Deep learning-based placental MRI segmentation could enable quantitative assessment of placenta accreta spectrum (PAS), but models developed using one MRI sequence may not generalize reliably to another, potentially compromising downstream imaging measurements. **Objectives:** Our objectives were to quantify the magnitude and directionality of performance degradation during direct cross-sequence transfer between balanced turbo field echo (BTFE) and single-shot turbo spin echo (ssh-TSE), determine whether pooled multisequence training using both sequences mitigates this degradation, and assess whether cross-sequence degradation is consistently observed across three segmentation pipelines. **Methods:** We evaluated 7106 placental MRI image-mask pairs using three segmentation pipelines: U-Net, nnU-Net v2, and U-Mamba Bot. Models were trained using BTFE-only, ssh-TSE-only, or pooled multisequence data and evaluated on a patient-matched held-out cohort under sequence-specific, direct cross-sequence transfer, and pooled multisequence-training conditions. The primary outcome was patient-level aggregated Dice, supplemented by boundary and segmentation-derived size-error analyses. **Results:** In-domain mean Dice ranged from 0.888 to 0.917. Cross-sequence transfer significantly reduced performance across all pipelines, with markedly greater degradation when BTFE-trained models were applied to ssh-TSE (Dice loss 0.066–0.119) than when ssh-TSE-trained models were applied to BTFE (0.039–0.051; all $p < 0.001$). Pooled multisequence training recovered 0.044–0.118 Dice relative to direct cross-sequence transfer and achieved performance within -0.0020 to $+0.0058$ Dice of the corresponding sequence-specific models. Boundary-error analyses showed deterioration under sequence shift, while increases in mean absolute segmentation-derived size error on the ssh-TSE target were more pronounced for nnU-Net and U-Mamba Bot, with less conclusive findings for U-Net and inconsistent changes on BTFE. **Conclusions:** Placental MRI segmentation in PAS was strongly sequence-dependent, with substantial and directionally asymmetric degradation when models were directly transferred to an MRI sequence not represented during training across all three evaluated segmentation pipelines, although the magnitude of degradation varied among pipelines. Pooled multisequence training on BTFE and ssh-TSE largely mitigated this degradation and restored performance toward sequence-specific levels. These



Academic Editors: He Zhao and Lin Gu

Received: 19 September 2026

Revised: 17 October 2026

Accepted: 19 October 2026

Published: 23 October 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 \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

findings support sequence-aware development and validation of placental segmentation models across heterogeneous MRI protocols.

Keywords: MRI; medical imaging; placenta; pregnancy; women's health; artificial intelligence; deep learning; segmentation

1. Introduction

The placenta accreta spectrum (PAS) comprises disorders of abnormal placental adherence and invasion into the uterine wall. It is associated with substantial maternal morbidity, particularly severe hemorrhage, transfusion, hysterectomy, and injury to adjacent pelvic organs [1,2]. Antenatal recognition is important for multidisciplinary delivery planning and appropriate referral to experienced centers. Magnetic resonance imaging (MRI) enables detailed visualization of placental morphology and the uteroplacental interface because of its wide field of view, multiplanar capability, and high soft-tissue contrast [1,3].

Accurate placental segmentation is essential for quantitative MRI assessment of placental volume, shape, location, and other imaging-derived features. Although expert manual delineation remains the gold standard, it is labor-intensive and subject to inter- and intra-observer variability [4], motivating the development of automated segmentation approaches. However, automated placental segmentation remains challenging because of fetal and maternal motion, indistinct interfaces with adjacent tissues such as the myometrium, and substantial variation in placental size, shape, and position across patients and gestation [5,6].

Although deep learning approaches have achieved promising performance for automated placental segmentation on prenatal MRI, with reported mean Dice scores of approximately 0.80–0.84 [7–9], most studies have evaluated models within the imaging domains represented during model development. Generalization across distinct MRI sequences remains less well characterized. Balanced turbo field echo (BTFE) and single-shot turbo spin echo (ssh-TSE), for example, differ in image contrast and tissue-boundary appearance, such that a model trained on one sequence may not maintain its performance when applied to a different sequence acquired from the same patient [3].

Several studies have examined robustness in placental MRI segmentation [5]. Zhong et al. investigated contrast-invariant segmentation across T2*-weighted echo times [10], Wu et al. evaluated BiGC U-Net across public and institutional MRI cohorts [11], and Yang et al. developed an uncertainty-guided multisequence framework combining T2-weighted and bSSFP MRI [12]. More broadly, acquisition-related domain shift has been demonstrated in other anatomical applications, including cross-modality abdominal segmentation in the CHAOS challenge [13], multi-center and multi-vendor cardiac MRI segmentation in M&Ms [14], and cross-sequence cardiac MRI segmentation in MS-CMRSeg [15]. Together, these studies highlight imaging contrast, scanner vendor, and acquisition protocol as important sources of segmentation variability. However, patient-matched bidirectional generalization between clinically relevant placental MRI sequences remains poorly characterized, motivating our evaluation of BTFE and ssh-TSE while minimizing confounding from differences in patient populations.

Accordingly, this study aimed to determine how strongly MRI sequence influences placental segmentation performance in PAS using patient-matched BTFE and ssh-TSE acquisitions. Specifically, we sought to: (1) quantify the magnitude and directionality of bidirectional cross-sequence performance degradation; (2) determine whether pooled multisequence training mitigates this degradation and restores performance toward sequence-

specific levels; and (3) assess whether cross-sequence performance degradation was consistently observed across three complementary segmentation pipelines: U-Net, nnU-Net, and U-Mamba Bot. A targeted size-matched, patient-exclusive sensitivity analysis was additionally performed to assess whether the benefit of multisequence training could be explained solely by increased training-set size or paired sequence contributions from the same patients.

2. Methods

2.1. Study Design and Dataset

This study was designed as a patient-matched evaluation of cross-sequence generalization in automated placental MRI segmentation (Figure 1) using the publicly available placenta accreta spectrum disorders dataset [16]. The repository included 133 unique subject IDs: 130 represented in both BTFE and ssh-TSE, one available only in BTFE, and two available only in ssh-TSE. The dataset comprised 3411 BTFE images from 131 subjects and 3695 ssh-TSE images from 132 subjects. MRI images were provided as JPEG files with corresponding PNG segmentation masks. Original image matrices varied across the dataset: BTFE ranged from 320×320 to 512×512 pixels, most commonly 512×512 (1803/3411; 52.9%) or 320×320 (1332/3411; 39.0%), whereas ssh-TSE ranged from 384×384 to 672×672 pixels and was most commonly 512×512 (2114/3695; 57.2%).

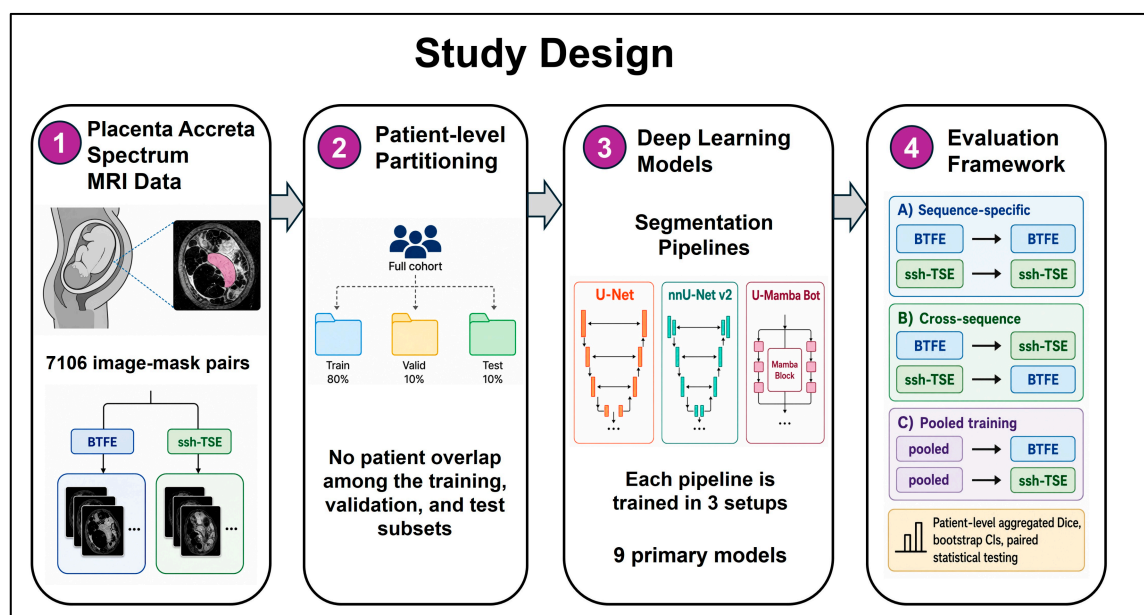


Figure 1. Overview of the study design and experimental framework. The dataset was partitioned at the patient level into training, validation, and test sets, with the same patient assignments maintained across BTFE and ssh-TSE sequences. Three segmentation pipelines (U-Net, nnU-Net, and U-Mamba Bot) were trained using BTFE-only, ssh-TSE-only, or pooled multisequence data and evaluated under sequence-specific, bidirectional cross-sequence, and pooled-training conditions.

All cases were reported as PAS, with no non-PAS controls. Patient-level metadata such as gestational age and PAS subtype were unavailable, and the repository did not provide details regarding diagnostic confirmation, scanner manufacturer/model, field strength, slice thickness or spacing, or the reference-mask annotation procedure. Although the repository did not specify whether slices without visible placenta were retained, all 716 slices in the held-out test set (345 BTFE and 371 ssh-TSE) had non-empty placental reference masks. The present evaluation therefore characterizes segmentation performance on placenta-containing slices and does not assess placenta-absent slices or complete unfiltered

MRI stacks. Because all subjects had PAS and no PAS-specific diagnostic, subtype, severity, or clinical outcome endpoint was evaluated, the study addresses segmentation robustness within a PAS cohort rather than PAS detection or severity assessment.

BTfE and ssh-TSE were separate acquisitions, and the repository did not provide registration transforms, spatial coordinates, or verified slice correspondences; the two sequences therefore could not be assumed to represent spatially identical anatomy. Direct cross-sequence reference-mask overlap was consequently not calculated.

2.2. Patient-Level Data Partitioning

Data were partitioned at the patient level, with approximately 80%, 10%, and 10% of patients allocated to the training, validation, and test sets, respectively. The partition was generated using a base random seed of 2026, with 2000 candidate patient-level assignments evaluated to balance patient proportions and sequence-specific image counts across the three subsets. The same patient assignments were maintained across both MRI sequences, and the three sets were mutually exclusive. This paired design enabled within-patient comparison of segmentation performance across both MRI sequences, reducing between-patient variability in the primary cross-sequence comparisons.

The BTfE partition comprised 2725 images from 105 patients for training, 341 images from 13 patients for validation, and 345 images from 13 patients for testing. The ssh-TSE partition comprised 2952 images from 106 patients for training, 372 images from 13 patients for validation, and 371 images from 13 patients for testing. The three subjects represented in only one sequence were confined to the training set; therefore, both the validation and test cohorts were fully paired across sequences. The common test cohort comprised the same 13 patients in each sequence, yielding 716 test images, with the patient rather than the slice treated as the independent unit of statistical analysis. For pooled multisequence training, the validation set combined both BTfE and ssh-TSE images from these same 13 validation patients, yielding 713 validation images in total.

2.3. Cross-Sequence Experimental Framework

Three segmentation pipelines were evaluated: conventional U-Net, self-configuring nnU-Net, and U-Mamba Bot. For each pipeline, three models were trained using BTfE only, ssh-TSE only, or pooled BTfE and ssh-TSE data, yielding nine primary models. All evaluations were performed using the same patient-matched test cohort across sequences. Validation data were restricted to the corresponding source training distribution, ensuring that no target-sequence test data were used for model selection, hyperparameter optimization, or other aspects of model development.

This experimental design was structured to address three questions within a common patient-matched framework: (1) what is the magnitude and directionality of performance degradation when a segmentation model is transferred between BTfE and ssh-TSE; (2) whether cross-sequence degradation is consistent across different segmentation pipelines; and (3) whether training on both MRI sequences mitigates the performance loss observed during direct cross-sequence transfer.

To examine whether the benefit of pooled multisequence training could be explained by the larger number of training images or by the same patients contributing both MRI sequences, we performed a targeted sensitivity analysis using the U-Net pipeline. Each training patient was assigned to contribute only one MRI sequence (BTfE or ssh-TSE), while maintaining representation of both sequences and matching the total training-set size to the midpoint of the two sequence-specific training sets. The resulting dataset contained 2839 images from 107 patients, comprising 1419 BTfE images from 54 patients and 1420 ssh-TSE images from 53 patients. The same patient-exclusive strategy was applied

to validation, yielding 357 images (178 BTFE and 179 ssh-TSE) from 13 patients. Patient-to-sequence assignment was based only on sequence availability and slice counts and did not use image content, segmentation performance, or test-set information. The model otherwise used the same U-Net architecture, preprocessing, augmentation, optimization, and checkpoint-selection procedure as the primary analysis and was evaluated on the unchanged patient-matched test cohort.

2.4. Image Preprocessing and Augmentation

For the conventional U-Net pipeline, images and masks were converted to single-channel grayscale and resized to 256×256 pixels using bilinear interpolation for MRI images and nearest-neighbor interpolation for segmentation masks. Image intensities were robustly normalized by clipping to the 1st–99th percentile range and rescaling to [0,1]. Masks were binarized after resizing. Training-time augmentation included horizontal mirroring with probability 0.5, rotation within $\pm 10^\circ$ with probability 0.7, contrast scaling between 0.85 and 1.15 with probability 0.5, and gamma transformation between 0.85 and 1.15 with probability 0.3. Vertical flipping was not applied.

For nnU-Net and U-Mamba Bot, the original grayscale images and binary masks were converted losslessly to PNG and passed through the respective official 2D planning and preprocessing pipelines. Framework-specific configurations were applied before training. This preserved the self-configuring behavior of these pipelines rather than imposing the preprocessing and spatial configuration of one model on all architectures. Accordingly, comparisons among models represent complete segmentation pipelines, incorporating differences in architecture, preprocessing, spatial configuration, and training framework, rather than an architecture-only ablation under standardized conditions.

2.5. Model Development and Training

The conventional 2D U-Net [17] comprised four encoder levels, a bottleneck, and a symmetric decoder with transposed-convolution upsampling and skip connections. Convolutional blocks consisted of 3×3 convolutions with batch normalization, ReLU activation, and dropout. The initial feature width was 32 channels and dropout was 0.20. U-Net optimization used the sum of binary cross-entropy and soft Dice losses with AdamW, an initial learning rate of 1×10^{-4} , weight decay of 1×10^{-5} , and cosine-annealing learning-rate scheduling. Training used a batch size of 8, mixed-precision computation, and gradient-norm clipping at 5.0. Models were trained for a maximum of 100 epochs with early stopping after 12 epochs without improvement in validation loss, and the checkpoint with the lowest validation loss was retained. Training stopped at epochs 46, 40, and 43 for the BTFE-only, ssh-TSE-only, and pooled multisequence U-Net models, respectively, with the best checkpoints selected at epochs 34, 28, and 31.

nnU-Net v2 [18] was implemented using the official 2D framework. A custom split file was supplied so that the framework used the predefined patient-level training and validation partitions described in Section 2.2 rather than generating its default internal five-fold split. Fold 0 was used for training and inference. The 100-epoch configuration was selected to maintain a consistent maximum number of training epochs across the three evaluated pipelines, rather than using the default 1000-epoch nnU-Net schedule.

U-Mamba Bot [19] was implemented using the official 2D U-Mamba framework. A custom trainer limited training to 100 epochs while retaining the original training configuration and polynomial learning-rate schedule, rescaled across the 100-epoch period. The same predefined patient-level split and fold 0 were used.

One model was trained for each pipeline and training source, yielding nine primary models. No repeated training runs or model ensembles were used. Training and validation

histories for all six nnU-Net and U-Mamba Bot models are shown in Supplementary Figure S2. Validation loss decreased rapidly during early training and was largely stable toward the end of the 100-epoch schedule. For both nnU-Net and U-Mamba Bot, inference used the final checkpoint after 100 epochs, without validation-based checkpoint selection. All experiments were conducted on an NVIDIA A100 80-GB GPU.

2.6. Segmentation Outcomes

The primary outcome was patient-level aggregated Dice, calculated by combining all evaluated slices for each patient and computing Dice across the corresponding predicted and reference pixels. Confidence intervals for the primary Dice analyses were calculated using BCa bootstrap with 10,000 patient-level resamples. Confidence intervals for the RVD and size-matched sensitivity analyses used percentile bootstrap with 2000 patient-level resamples.

Cross-sequence degradation was defined as the paired difference between sequence-specific in-domain Dice and Dice after direct transfer from the alternate sequence. The effect of pooled training was quantified relative to cross-sequence performance using paired patient-level differences. These prespecified within-pipeline comparisons were assessed using two-sided Wilcoxon signed-rank tests. Effect estimates are reported as mean paired differences with 95% BCa confidence intervals, together with the number of patients showing change in the expected direction.

As a secondary measure of segmentation-derived size error, relative volume difference (RVD) was calculated between predicted and reference segmentations based on foreground pixel counts. Absolute RVD was used to quantify the magnitude of this relative size difference irrespective of over- or under-segmentation. RVD was calculated from foreground pixel counts aggregated across all slices per patient, with absolute RVD subsequently averaged across patients. Because physical voxel spacing and slice geometry were unavailable, RVD was interpreted as a pixel-count-based proxy for relative size error rather than a measure of actual placental volume.

To further characterize segmentation error beyond regional overlap, the 95th-percentile Hausdorff distance (HD95) was calculated between predicted and reference segmentation boundaries. Dice and HD95 were computed on a common 256×256 evaluation grid for all three pipelines. U-Net predictions were generated directly on this grid, whereas foreground probability maps exported by nnU-Net and U-Mamba Bot were resized to 256×256 using bilinear interpolation before thresholding; reference masks were resized to the same grid using nearest-neighbor interpolation. Because physical pixel spacing was unavailable for the public dataset, HD95 was reported in pixels. HD95 was infinite when one mask was empty and the other was non-empty. As no evaluated test slice had an empty reference mask, all non-finite HD95 values represented empty predictions (complete segmentation misses). HD95 summary statistics were calculated from finite values, while the number of empty predictions was reported separately for each evaluation condition.

3. Results

3.1. Overall Model Performance

Across sequence-specific testing, mean patient-level Dice ranged from 0.888 to 0.917 across the three segmentation pipelines (Table 1). On BTFE, nnU-Net achieved the highest mean Dice (0.917), followed by U-Mamba Bot (0.915) and U-Net (0.908). On ssh-TSE, U-Mamba Bot achieved the highest mean Dice (0.903), followed by nnU-Net (0.899) and U-Net (0.888). In contrast, performance decreased when models were transferred to a sequence not represented during training, particularly for BTFE→ssh-TSE, where mean Dice ranged from 0.784 to 0.822. Pooled multisequence training yielded mean Dice val-

ues within 0.006 of the corresponding sequence-specific models across all pipeline–target combinations (Table 1).

Table 1. Patient-level segmentation performance across training and testing conditions. Values are mean patient-level aggregated Dice scores with 95% bias-corrected and accelerated (BCa) bootstrap confidence intervals, estimated from 10,000 patient-level resamples. Performance is reported for sequence-specific evaluation, direct cross-sequence transfer, and pooled multisequence training across U-Net, nnU-Net, and U-Mamba Bot. Each condition includes the same 13 held-out patients for the corresponding target sequence.

Training → Test	U-Net	nnU-Net	U-Mamba Bot
BTFE → BTFE	0.908 (0.895–0.922)	0.917 (0.901–0.929)	0.915 (0.900–0.929)
BTFE → ssh-TSE	0.822 (0.795–0.844)	0.791 (0.727–0.835)	0.784 (0.720–0.832)
ssh-TSE → ssh-TSE	0.888 (0.869–0.904)	0.899 (0.878–0.916)	0.903 (0.882–0.919)
ssh-TSE → BTFE	0.870 (0.843–0.886)	0.866 (0.838–0.886)	0.867 (0.839–0.887)
Pooled → BTFE	0.914 (0.903–0.925)	0.915 (0.902–0.928)	0.916 (0.902–0.929)
Pooled → ssh-TSE	0.894 (0.878–0.908)	0.901 (0.879–0.917)	0.902 (0.882–0.919)

3.2. Bidirectional Cross-Sequence Degradation

Cross-sequence transfer resulted in significant performance degradation for all three segmentation pipelines (Table 2). When ssh-TSE-trained models were applied to BTFE, mean Dice decreased by 0.039–0.051 relative to BTFE in-domain performance. The degradation was more pronounced in the opposite direction: transferring BTFE-trained models to ssh-TSE reduced Dice by 0.066 for U-Net, 0.108 for nnU-Net, and 0.119 for U-Mamba Bot (Supplementary Figure S1). Cross-sequence degradation was observed in 12/13 patients for U-Net during ssh-TSE→BTFE transfer and in all 13 patients for the remaining five pipeline–target combinations (all $p < 0.001$). Mean paired Dice reductions ranged from 0.039 to 0.119, with 95% confidence intervals reported in Table 2. The directional asymmetry is illustrated in Figure 2, with substantially greater Dice loss for BTFE→ssh-TSE than for ssh-TSE→BTFE across all three segmentation pipelines.

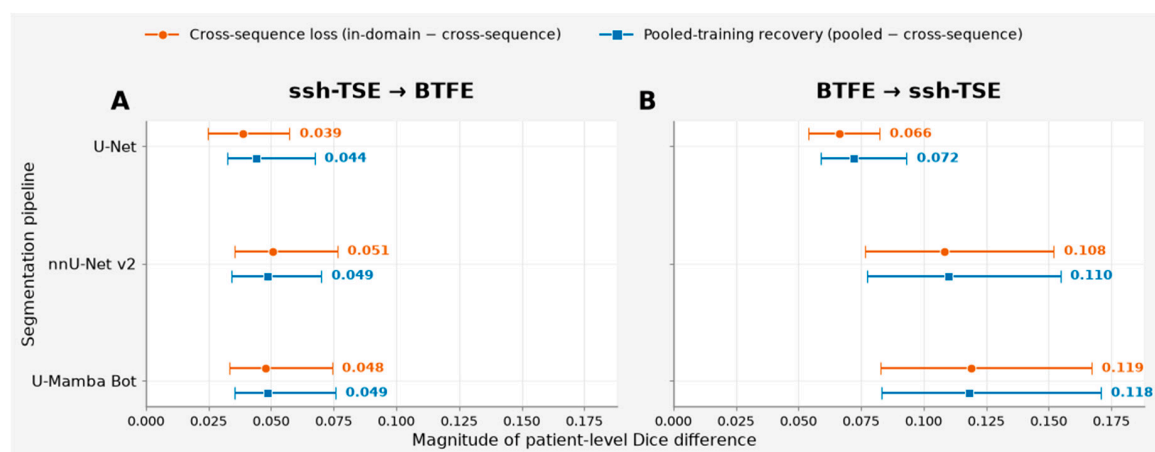


Figure 2. Bidirectional cross-sequence Dice loss and recovery with pooled multisequence training. Mean paired patient-level Dice differences are shown for U-Net, nnU-Net v2, and U-Mamba Bot for (A) ssh-TSE → BTFE transfer and (B) BTFE → ssh-TSE transfer. Orange circles represent cross-sequence Dice loss, defined as sequence-specific in-domain Dice minus cross-sequence Dice. Blue squares represent recovery with pooled multisequence training, defined as pooled-training Dice minus cross-sequence Dice. Whiskers indicate 95% bias-corrected and accelerated (BCa) bootstrap confidence intervals based on 10,000 patient-level resamples. Numeric labels indicate the corresponding mean paired Dice differences.

Table 2. Cross-sequence performance degradation and recovery with pooled multisequence training. Cross-sequence Dice loss was calculated as sequence-specific in-domain Dice minus Dice after direct transfer from the alternate sequence. Improvement with pooled training was calculated as pooled-training Dice minus cross-sequence Dice. The pooled–sequence-specific difference was calculated as pooled-training Dice minus the corresponding sequence-specific Dice. Values are mean paired patient-level differences with 95% bias-corrected and accelerated (BCa) bootstrap confidence intervals based on 10,000 patient-level resamples. *p* values are from two-sided Wilcoxon signed-rank tests. Positive pooled–sequence-specific values indicate higher Dice with pooled training. The pooled-versus-sequence-specific comparisons are interpreted descriptively and do not constitute tests of statistical equivalence.

Target Sequence	Pipeline	Cross-Sequence Dice Loss (95% BCa CI)	Patients with Degradation, n/13	<i>p</i>	Improvement with Pooled Training (95% BCa CI)	Patients Improved with Pooled Training, n/13	<i>p</i>	Pooled–Sequence-Specific Dice Difference (95% BCa CI)
BTFE	U-Net	0.039 (0.025–0.058)	12/13	<0.001	+0.044 (0.033–0.067)	13/13	<0.001	+0.0054 (0.0010–0.0110)
BTFE	nnU-Net	0.051 (0.035–0.076)	13/13	<0.001	+0.049 (0.034–0.069)	13/13	<0.001	−0.0020 (−0.0071–0.0041)
BTFE	U-Mamba Bot	0.048 (0.033–0.074)	13/13	<0.001	+0.049 (0.035–0.075)	13/13	<0.001	+0.0008 (−0.0075–0.0044)
ssh-TSE	U-Net	0.066 (0.054–0.083)	13/13	<0.001	+0.072 (0.059–0.094)	13/13	<0.001	+0.0058 (−0.0018–0.0128)
ssh-TSE	nnU-Net	0.108 (0.078–0.154)	13/13	<0.001	+0.110 (0.077–0.155)	13/13	<0.001	+0.0017 (−0.0012–0.0075)
ssh-TSE	U-Mamba Bot	0.119 (0.084–0.170)	13/13	<0.001	+0.118 (0.083–0.172)	13/13	<0.001	−0.0009 (−0.0053–0.0036)

The impact of cross-sequence degradation on segmentation-derived size estimates was also assessed using absolute relative volume difference (RVD; Supplementary Table S2). On ssh-TSE, where Dice degradation was greatest, mean absolute RVD increased from 4.8% to 9.1% for U-Net, from 5.4% to 17.5% for nnU-Net, and from 4.1% to 19.2% for U-Mamba Bot under cross-sequence transfer. The corresponding paired increases were 4.3 percentage points (95% CI, −0.2 to 8.5), 12.1 percentage points (95% CI, 6.5–18.7), and 15.1 percentage points (95% CI, 9.5–21.4), respectively; thus, the confidence interval included zero for U-Net but excluded zero for nnU-Net and U-Mamba Bot. Pooled multisequence training reduced mean absolute RVD to 4.6–4.8%. On BTFE, changes in absolute RVD were not directionally consistent: mean absolute RVD increased from 5.1% to 9.8% for U-Net and from 5.0% to 6.5% for U-Mamba Bot, but decreased from 5.0% to 4.2% for nnU-Net. These findings indicate that worsening of segmentation-derived relative size error under sequence shift was most evident for the ssh-TSE target rather than uniformly across both transfer directions.

Boundary-based and slice-level analyses further supported the cross-sequence findings (Supplementary Table S3). In-domain and pooled models showed substantially lower HD95 than cross-sequence models. For nnU-Net, median HD95 increased from 6.02 pixels under ssh-TSE in-domain evaluation to 19.30 pixels following BTFE → ssh-TSE transfer. The degradation was not restricted to isolated slices: the proportion of slices with Dice below 0.75 increased from 5.1% under ssh-TSE in-domain evaluation to 34.2% following BTFE → ssh-TSE transfer. Complete segmentation misses, defined as empty predictions for non-empty reference masks, were uncommon but occurred more frequently in some cross-sequence conditions: 1, 2, and 2 slices for nnU-Net, U-Mamba Bot, and U-Net, respectively, during BTFE → ssh-TSE transfer, and 0, 3, and 5 slices during ssh-TSE → BTFE transfer. By comparison, sequence-specific evaluation produced no empty predictions except for one U-Net BTFE slice, and pooled multisequence evaluation produced none (Supplementary Table S3).

3.3. Pooled Multisequence Training

Training jointly on BTFE and ssh-TSE consistently mitigated the performance loss observed under direct cross-sequence transfer. On BTFE, pooled training improved mean Dice relative to ssh-TSE→BTFE transfer by 0.044–0.049 across the three pipelines, whereas improvements on ssh-TSE relative to BTFE → ssh-TSE transfer ranged from 0.072 to 0.118 (Table 2). The 95% BCa confidence intervals excluded zero for all six paired effects, and pooled training improved Dice in all 13 held-out patients for each pipeline–target combination (all $p < 0.001$).

Direct comparison with the corresponding sequence-specific models showed only small differences in performance. Mean paired differences between pooled and sequence-specific Dice ranged from −0.0020 to +0.0058 across the six pipeline–target combinations. For U-Net, the pooled-minus-sequence-specific differences were +0.0054 (95% BCa CI, 0.0010–0.0110) on BTFE and +0.0058 (−0.0018 to 0.0128) on ssh-TSE. For nnU-Net, the corresponding differences were −0.0020 (−0.0071 to 0.0041) and +0.0017 (−0.0012 to 0.0075), respectively; for U-Mamba Bot, they were +0.0008 (−0.0075 to 0.0044) and −0.0009 (−0.0053 to 0.0036). Although the 95% BCa confidence intervals for four of the six comparisons were entirely contained within ± 0.01 Dice, no equivalence margin was prespecified, and ± 0.01 should therefore not be interpreted as a formal equivalence threshold. These comparisons are consequently interpreted descriptively and should not be considered evidence of statistical equivalence, non-inferiority, or superiority.

3.4. Size-Matched Patient-Exclusive Sensitivity Analysis

To assess whether the advantage of multisequence training persisted after controlling for training-set size and paired sequence contributions from the same patients, we trained an additional patient-exclusive, size-matched U-Net using 2839 training images, with approximately equal contributions from BTFE (1419 images) and ssh-TSE (1420 images). The model achieved mean patient-level Dice scores of 0.905 (95% CI, 0.890–0.919) on BTFE and 0.880 (95% CI, 0.865–0.895) on ssh-TSE (Supplementary Table S1). Compared with direct cross-sequence transfer, Dice increased by 0.0352 (95% CI, 0.0218–0.0527) on BTFE and by 0.0581 (95% CI, 0.0413–0.0762) on ssh-TSE, with higher Dice in all 13 test patients for both target sequences.

Performance remained numerically close to the corresponding sequence-specific U-Net, with paired differences of -0.0034 (95% CI, -0.0094 to 0.0018) for BTFE and -0.0083 (95% CI, -0.0163 to -0.0001) for ssh-TSE. The size-matched model performed modestly below the original full pooled U-Net, with differences of -0.0088 (95% CI, -0.0137 to -0.0043) and -0.0141 (95% CI, -0.0219 to -0.0064), respectively. These findings indicate that the benefit of multisequence exposure persisted when training-set size was matched and each patient contributed only one MRI sequence, although the larger full pooled dataset provided an additional modest performance gain. Because this sensitivity analysis was performed using U-Net only and a single training run, it should be interpreted as supportive evidence for this experimental setting rather than as a pipeline-general result.

3.5. Qualitative Assessment of Cross-Sequence Segmentation and Pooled Recovery

Qualitative assessment of three patient-matched cases was consistent with the overall pattern of cross-sequence degradation and recovery observed quantitatively (Figure 3). The patients were selected based on their patient-level cross-sequence performance patterns in the held-out test cohort: Patient example 1 showed pronounced BTFE $\rightarrow$ ssh-TSE degradation with relatively preserved performance in the reverse direction; Patient example 2 represented a non-extreme cross-sequence case; and Patient example 3 was selected as a case with comparatively greater patient-level degradation in the ssh-TSE $\rightarrow$ BTFE direction. For each patient and target sequence, Figure 3 shows a representative slice with the corresponding slice-level Dice, allowing direct comparison between the displayed contours and the reported segmentation performance. Because patient selection was based on aggregated patient-level behavior, the performance of an individual displayed slice does not necessarily reproduce the relative magnitude of degradation observed across all slices from that patient.

Patient example 1 showed the clearest visual effect of sequence shift. For the ssh-TSE target, the BTFE-trained model exhibited substantial boundary disagreement, with a slice-level Dice of 0.566. Pooled multisequence training produced closer correspondence with the reference contour and increased Dice to 0.807. In contrast, segmentation of the displayed BTFE slice was already highly accurate after ssh-TSE $\rightarrow$ BTFE transfer (Dice 0.944) and changed only marginally with pooled training (Dice 0.947), illustrating the marked directional asymmetry in this example.

Patient example 2 demonstrated a more moderate cross-sequence pattern. On the ssh-TSE target, slice-level Dice increased from 0.680 with BTFE $\rightarrow$ ssh-TSE transfer to 0.879 with pooled training, with visibly improved delineation of the placental boundary. On the corresponding BTFE slice, Dice increased from 0.871 to 0.936, again showing improved agreement with the reference contour following multisequence training.

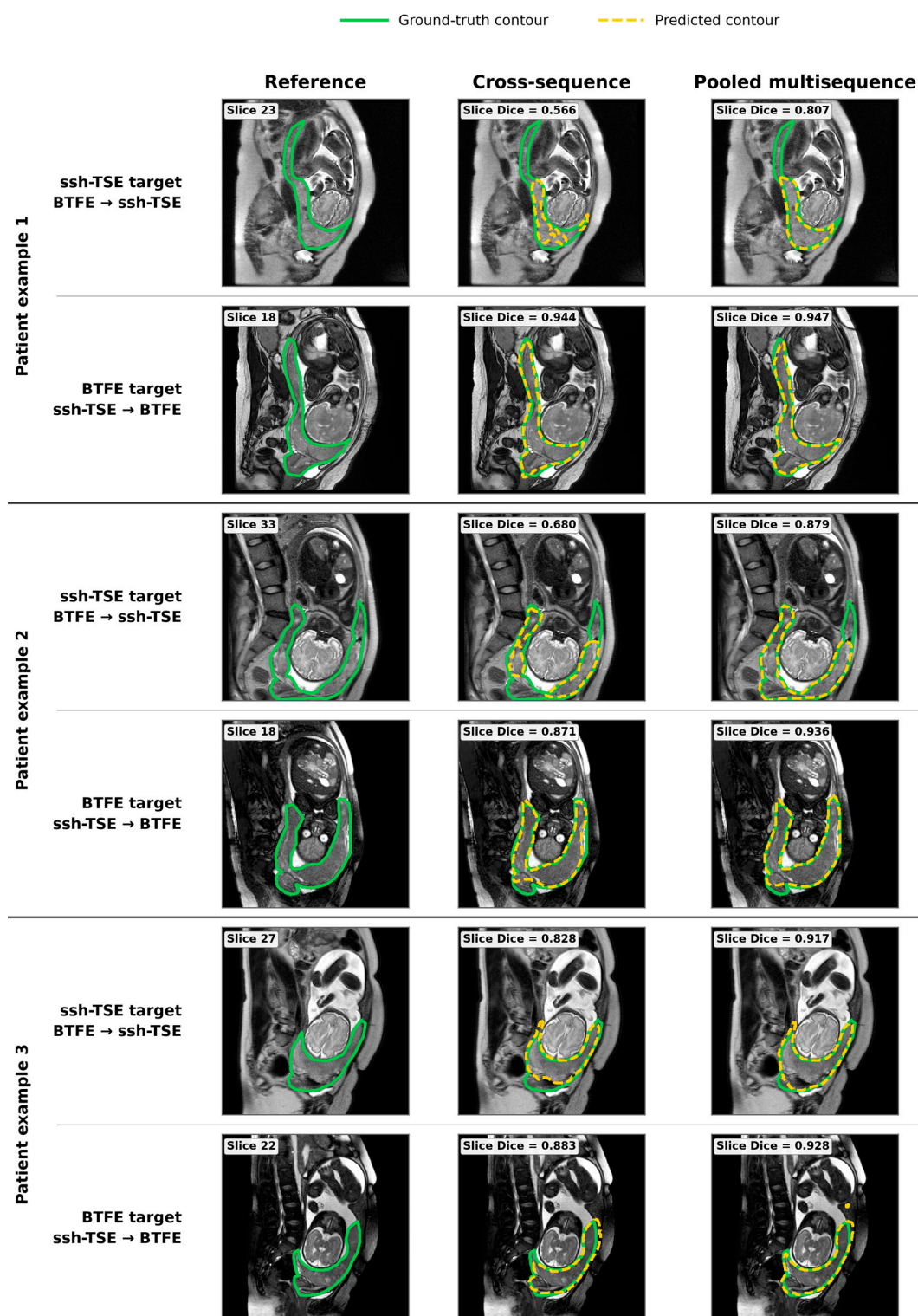


Figure 3. Patient-matched qualitative examples of cross-sequence placental segmentation and recovery with pooled multisequence training. Three held-out patients are shown, each with ssh-TSE and BTFE as the target sequence. The patients were selected based on distinct patient-level cross-sequence performance patterns observed in the test cohort. For each target sequence, the reference placental contour is compared with the prediction obtained from a model trained only on the alternate MRI sequence and with the prediction from the pooled multisequence model. Ground-truth contours are shown in green and predicted contours in yellow. Dice values shown in the figure are slice-level scores calculated for the specific images displayed and therefore may differ from the patient-level aggregated Dice values used to select the qualitative cases. nnU-Net is shown as the representative segmentation pipeline.

For Patient example 3, cross-sequence segmentation remained comparatively strong on both displayed slices but improved further with pooled training. Slice-level Dice increased from 0.828 to 0.917 on the ssh-TSE target and from 0.883 to 0.928 on the BTFE target. Together, the three examples illustrate substantial within- and between-patient heterogeneity in cross-sequence segmentation performance, while consistently showing that pooled multisequence training either improved or maintained boundary agreement on the illustrated slices. These qualitative findings complement the cohort-level analyses demonstrating that sequence shift can affect placental segmentation across both transfer directions, with greater average degradation for BTFE $\rightarrow$ ssh-TSE. Although the selected examples illustrate improvements with pooled training, residual segmentation errors remained, with 2.9–5.4% of pooled-training slices across pipelines and target sequences achieving Dice below 0.75 (Supplementary Table S3).

4. Discussion

4.1. Main Findings

This study demonstrates that placental MRI segmentation in PAS is materially sequence-dependent even when the underlying patients are identical. Cross-sequence transfer significantly reduced Dice for all three pipelines, with a consistently larger penalty when BTFE-trained models were applied to ssh-TSE than in the reverse direction. Training on pooled multisequence data mitigated this degradation and achieved performance close to that of sequence-specific models. However, because pooled training increased both sequence diversity and the total amount of training data, the present design cannot determine the relative contribution of these two factors.

Although the magnitude of cross-sequence degradation differed numerically among pipelines, these differences should not be interpreted as evidence that one pipeline was more robust than another. Each model configuration was trained once, and the confidence intervals for the pipeline-specific effects overlapped. In addition, U-Net used fixed 256×256 resizing and percentile-based intensity normalization, whereas nnU-Net and U-Mamba Bot retained their framework-specific planning and preprocessing. Because the study evaluated each framework as a complete segmentation pipeline, it cannot separate the effects of architecture, preprocessing, spatial configuration, and training strategy.

The sensitivity analysis further suggests that the improvement associated with multi-sequence training was not solely attributable to the larger number of training images or to paired acquisition of both sequences from the same patients. When training-set size was reduced to a level comparable with the sequence-specific datasets and each patient contributed only one sequence, U-Net performance remained substantially higher than under direct cross-sequence transfer and remained close to the corresponding sequence-specific performance. The modest additional advantage of the full pooled model nevertheless suggests that greater data volume and/or paired multisequence exposure may provide further benefit.

The qualitative examples further illustrated visible boundary discrepancies associated with cross-sequence transfer, complementing the quantitative findings.

4.2. Research and Clinical Implications

An important implication of these findings is that MRI sequence should be considered explicitly as a source of domain variation in placental segmentation studies, rather than as a secondary acquisition detail. Prior work has addressed contrast invariance within T2*-weighted imaging [10], cross-dataset robustness [11], and multisequence feature fusion [12], but our patient-matched design shows that substantial degradation can arise even when patient identity is held constant. The marked directional asymmetry is also im-

portant; robustness cannot be inferred from a single transfer direction, because transferring models from BTFE to ssh-TSE and from ssh-TSE to BTFE represent distinct generalization challenges. Future benchmarks should therefore report bidirectional transfer, sequence-stratified performance, and patient-level paired analyses where matched acquisitions are available.

Cross-sequence degradation was observed across all three evaluated segmentation pipelines, although its magnitude differed among pipelines. Because the three pipelines retained their native preprocessing and framework-specific training configurations, these differences should be interpreted at the level of complete end-to-end segmentation pipelines rather than as isolated architecture effects. The consistent degradation observed across pipelines nevertheless indicates that MRI acquisition sequence is an important source of deployment variability that should be evaluated explicitly during model development and validation. Future work should therefore examine source-only domain-generalization strategies, including stronger intensity perturbation and intensity harmonization, as well as size-matched multisequence training and sequence-aware adaptation, using prospectively specified transformation parameters and validation procedures.

Clinically, this matters because segmentation is typically an upstream component rather than an endpoint. Errors at the placental boundary can propagate into volume, shape, radiomic, or spatial-interface measurements and may therefore alter downstream quantitative PAS assessments without producing an obvious model failure. The additional boundary and size-error analyses further illustrate this concern, although the effects differed by metric and transfer direction. Boundary error increased under cross-sequence evaluation, with nnU-Net median HD95 increasing from 6.02 pixels under ssh-TSE in-domain evaluation to 19.30 pixels following BTFE → ssh-TSE transfer. Segmentation-derived relative size error showed a less uniform pattern: on the ssh-TSE target, increases in mean absolute RVD were more pronounced for nnU-Net and U-Mamba Bot, while the increase for U-Net was less conclusive. Changes on BTFE were smaller and not directionally consistent. This distinction indicates that degradation in segmentation overlap does not necessarily translate proportionally into relative size error. The degradation was also distributed across a substantial proportion of slices rather than being driven solely by isolated failures, with 34.2% of BTFE → ssh-TSE nnU-Net slices showing Dice <0.75 compared with 5.1% under ssh-TSE in-domain evaluation.

This may be relevant to future quantitative MRI and radiomics applications in PAS, because segmentation instability could propagate into downstream feature extraction and potentially affect subsequent predictive modeling. If placental segmentation is incorporated into future clinical pipelines, sequence-aware safeguards may be warranted, for example by using acquisition metadata to verify that an examination falls within a validated protocol, select an appropriate model, or trigger quality-control review when it does not. The present study did not directly assess effects on PAS diagnosis, severity classification, or clinical decision-making, which should be investigated in future work.

More broadly, sequence-robust placental segmentation could provide a foundation for future multimodal approaches that integrate reproducible quantitative MRI features with clinical and molecular biomarkers. Reliable delineation across MRI sequences would support consistent extraction of placental volume, morphology, radiomic features, and characteristics of the uteroplacental interface, which could then be combined with established clinical risk factors and emerging biological markers. PAS-associated transcriptomic signatures have recently been identified in maternal plasma, urine, and peripheral blood [20], providing a complementary source of information that may reflect underlying disease biology not captured by imaging alone. Integrating these structural imaging phenotypes with circulating molecular signatures could therefore support more comprehensive PAS

characterization and, ultimately, more individualized risk assessment. Such multimodal integration was not evaluated in the present study and remains an important direction for future research.

4.3. Study Limitations

Although the patient-matched design and BCa confidence intervals quantify sampling uncertainty efficiently, they do not compensate for the limited number of independent patients. In addition, each model configuration was trained once; therefore, seed-to-seed variability was not quantified, particularly under cross-sequence transfer. The reported confidence intervals capture uncertainty across held-out patients but not uncertainty arising from model retraining, and small numerical differences among pipelines should not be interpreted as evidence of relative robustness. Comparisons also reflect complete segmentation pipelines rather than architecture-only effects because each framework retained its intended preprocessing, spatial configuration, and training strategy. Finally, pooled training increased both sequence diversity and sample size, preventing complete separation of these effects.

Several limitations arise from the public dataset. Physical voxel spacing, PAS subtype, and detailed acquisition and annotation metadata were unavailable. Because BTFE and ssh-TSE were separate acquisitions, cross-sequence differences should be interpreted as acquisition-domain effects that may reflect contrast, spatial characteristics, motion, and annotation conventions rather than contrast alone; the absence of registration and spatial metadata also prevented direct assessment of cross-sequence mask agreement. All evaluated test slices contained non-empty placental reference masks, so performance on placenta-absent slices and complete unfiltered MRI stacks remains unknown, and comparisons with prior studies using different or incompletely reported slice-selection protocols should be made cautiously. The MRI images were JPEG exports with unknown intensity windowing and varying matrix sizes. Therefore, the observed directional asymmetry cannot be attributed solely to sequence contrast, as resizing and other acquisition-related factors may contribute. Validation in larger independent, multisite cohorts with heterogeneous acquisition protocols is required.

5. Conclusions

This patient-matched study demonstrates that placental MRI segmentation in PAS is strongly influenced by MRI sequence, with significant and directionally asymmetric degradation between BTFE and ssh-TSE across all three evaluated segmentation pipelines. These findings support the development and evaluation of placental segmentation models across heterogeneous MRI sequences rather than within a single acquisition domain. Future work should prioritize external multisite validation, sequence-aware modeling, and evaluation of how segmentation robustness affects downstream quantitative MRI assessment in PAS, including the integration of segmentation-derived imaging phenotypes with clinical and molecular biomarkers for biologically informed PAS characterization and risk stratification.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/bioengineering13111247/s1>, Table S1: Size-matched patient-exclusive multisequence U-Net sensitivity analysis; Table S2: Absolute relative volume difference across sequence-specific, cross-sequence, and pooled multisequence conditions; Table S3: Additional slice-level segmentation performance and boundary-error analysis across sequence-specific, cross-sequence, and pooled multisequence conditions; Figure S1: Patient-level paired changes in Dice under cross-sequence transfer; Figure S2: Training and validation loss trajectories for nnU-Net and U-Mamba Bot.

Author Contributions: Conceptualization: R.A. and R.T.; Methodology: R.A., S.A. and R.T.; Investigation: R.A. and R.T.; Software: R.A.; Visualization: R.A.; Validation: S.A., N.A.-D. and T.F.; Project Administration: R.A.; Writing—Original Draft: R.A. and N.A.-D.; Writing—Review and Editing: R.T., S.A., N.A.-D. and T.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable. This study involved secondary analysis of publicly available, de-identified data and did not involve the collection of new data from human participants. Therefore, institutional review board (IRB) approval and informed consent were not required.

Data Availability Statement: The data used in this study are publicly available through the Mendeley Data repository: Placenta accreta spectrum disorders (PASDs), Version 1, DOI: 10.17632/284gwmf5bh.1. The dataset is available under a Creative Commons Attribution 4.0 International (CC BY 4.0) license. Code Availability Statement: The code used for model training, evaluation, and statistical analysis is publicly available on GitHub at: https://github.com/rawan-alsaad/placenta_MRI (accessed on 16 October 2026).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. American College of Obstetricians and Gynecologists and the Society for Maternal–Fetal Medicine. Obstetric Care Consensus No. 7: Placenta Accreta Spectrum. *Obstet. Gynecol.* **2018**, *132*, e259–e275. [CrossRef] [PubMed]
2. Silver, R.M.; Branch, D.W. Placenta Accreta Spectrum. *N. Engl. J. Med.* **2018**, *378*, 1529–1536. [CrossRef] [PubMed]
3. Himoto, Y.; Fujimoto, K.; Chigusa, Y.; Yoshida, A.; Minamiguchi, S.; Kido, A.; Nakamoto, Y. Toward Clinical Implementation of Magnetic Resonance Imaging for Placental Function. *Magn. Reson. Med. Sci.* **2025**, *24*, 343–353. [CrossRef] [PubMed]
4. Shahedi, M.; Spong, C.Y.; Dormer, J.D.; Do, Q.N.; Xi, Y.; Lewis, M.A.; Herrera, C.; Madhuranthakam, A.J.; Twickler, D.M.; Fei, B. Deep learning-based segmentation of the placenta and uterus on MR images. *J. Med. Imaging* **2021**, *8*, 054001. [CrossRef] [PubMed]
5. Jittou, A.; Fazazy, K.E.; Riffi, J. Placenta segmentation redefined: Review of deep learning integration of magnetic resonance imaging and ultrasound imaging. *Vis. Comput. Ind. Biomed. Art.* **2025**, *8*, 17. [CrossRef] [PubMed]
6. Wang, G.; Zuluaga, M.A.; Pratt, R.; Aertsen, M.; Doel, T.; Klusmann, M.; David, A.L.; Deprest, J.; Vercauteren, T.; Ourselin, S. Slic-Seg: A minimally interactive segmentation of the placenta from sparse and motion-corrupted fetal MRI in multiple views. *Med. Image Anal.* **2016**, *34*, 137–147. [CrossRef] [PubMed]
7. Liu, Y.; Zabihollahy, F.; Yan, R.; Lee, B.; Janzen, C.; Devaskar, S.U.; Sung, K. Evaluation of Spatial Attentive Deep Learning for Automatic Placental Segmentation on Longitudinal MRI. *J. Magn. Reson. Imaging* **2023**, *57*, 1533–1540. [CrossRef] [PubMed]
8. Shahedi, M.; Dormer, J.D.; Do, Q.N.; Xi, Y.; Lewis, M.A.; Herrera, C.L.; Spong, C.Y.; Madhuranthakam, A.J.; Twickler, D.M.; Fei, B. Automatic Segmentation of Uterine Cavity and Placenta on MR Images Using Deep Learning. *Proc. SPIE Int. Soc. Opt. Eng.* **2022**, *12036*, 1203611. [CrossRef] [PubMed]
9. Huang, J.; Do, Q.N.; Shahedi, M.; Xi, Y.; Lewis, M.A.; Herrera, C.L.; Owen, D.; Spong, C.Y.; Madhuranthakam, A.J.; Twickler, D.M.; et al. Deep learning based automatic segmentation of the placenta and uterine cavity on prenatal MR images. *Proc. SPIE Int. Soc. Opt. Eng.* **2023**, *12465*, 124650N. [CrossRef] [PubMed]
10. Zhong, X.; Liu, R.; Nichols, E.S.; Zhang, X.; Laine, A.F.; Duerden, E.G.; Wang, Y. Contrast-Invariant Self-supervised Segmentation for Quantitative Placental MRI. In *Perinatal, Preterm and Paediatric Image Analysis*; Springer: Cham, Switzerland, 2026; Volume 16118, pp. 141–152. [CrossRef] [PubMed]
11. Wu, D.; Coldman, S.; Zhou, G.; Li, X.; Whitby, E.; Guo, L. Automatic Segmentation of Placenta from MR images Using a Novel BiGC U-Net. *IEEE J. Biomed. Health Inform.* **2026**, *30*, 8813–8823. [CrossRef] [PubMed]
12. Yang, H.; Li, P.; Li, J.; Liu, J.; Xia, W.; Qiu, L.; Zhang, D.; Fu, L. Accurate segmentation of placenta accreta spectrum: An uncertainty-guided fusion network for multi-sequence MRI. *Biomed. Signal Process. Control* **2026**, *122*, 110473. [CrossRef]
13. Kavur, A.E.; Gezer, N.S.; Baris, M.; Aslan, S.; Conze, P.H.; Groza, V.; Pham, D.D.; Chatterjee, S.; Ernst, P.; Ozkan, S.; et al. CHAOS Challenge—Combined (CT-MR) healthy abdominal organ segmentation. *Med. Image Anal.* **2021**, *69*, 101950. [CrossRef] [PubMed]
14. Campello, V.M.; Gkontra, P.; Izquierdo, C.; Martin-Isla, C.; Sojoudi, A.; Full, P.M.; Maier-Hein, K.; Zhang, Y.; He, Z.; Ma, J.; et al. Multi-Centre, Multi-Vendor and Multi-Disease Cardiac Segmentation: The M&Ms Challenge. *IEEE Trans. Med. Imaging* **2021**, *40*, 3543–3554. [CrossRef] [PubMed]

15. Zhuang, X.; Xu, J.; Luo, X.; Chen, C.; Ouyang, C.; Rueckert, D.; Campello, V.M.; Lekadir, K.; Vesal, S.; RaviKumar, N.; et al. Cardiac segmentation on late gadolinium enhancement MRI: A benchmark study from multi-sequence cardiac MR segmentation challenge. *Med. Image Anal.* **2022**, *81*, 102528. [[CrossRef](#)] [[PubMed](#)]
16. Yang, T. Placenta Accreta Spectrum Disorders (PASDs). 2023. Available online: <https://data.mendeley.com/datasets/284gwmf5bh/1> (accessed on 18 October 2026).
17. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*; Springer International Publishing: Cham, Switzerland, 2015.
18. 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](#)]
19. Ma, J.; Li, F.; Wang, B. U-Mamba: Enhancing Long-range Dependency for Biomedical Image Segmentation. *arXiv* **2024**, arXiv:2401.04722.
20. Farrell, T.; Linga, B.G.; Ibrahim, F.E.; Derouiche, I.; Murali, J.; Al-Janahi, N.; Ahmed, S.; Minisha, F.; Kumar, V.; Roshanuddin, J.; et al. Gene expression profiling of maternal plasma, urine, and blood reveals molecular signatures of placenta accreta and low-lying placenta. *Sci. Rep.* **2026**. [[CrossRef](#)]

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.