MEASURE: Multi-Task Slice Selection and Regression for Neonatal Brain Biometry

Jiyang Lee[1*], Woori Bae[1*], Dabin Kim[2], Jong-Min Lee[1†], and Seh Hyun Kim[2,3†]

[1] Department of Artificial Intelligence, Hanyang University, Seoul, South Korea
[2] Division of Neonatology, Department of Pediatrics, Seoul National University Children's Hospital, Seoul, South Korea
[3] Department of Pediatrics, Seoul National University College of Medicine, Seoul, South Korea
ljm@hanyang.ac.kr, neobrain@snu.ac.kr

Abstract. Quantitative brain biometry from MRI is essential for evaluating brain growth in preterm-born neonates, yet manual measurement is time-consuming and suffers from inter-rater variability. Existing automation approaches require dense voxel-wise annotations that are costly to obtain and sensitive to domain shifts. We present MEASURE (**M**easurement **E**stimation via **A**natomical **S**lice **U**nderstanding and **RE**gression), the first annotation-efficient deep learning framework for Kidokoro-protocol neonatal brain biometry. Our two-stage approach mirrors the clinical workflow: a ViT-based multi-task slice selector identifies optimal measurement planes, followed by a multi-task regression network with measurement-specific spatial attention that jointly predicts all biometrics without voxel-level supervision. Experiments on 470 developing Human Connectome Project subjects demonstrate that explicit plane selection consistently outperforms end-to-end 3D volumetric baselines, and our proposed predictor outperforms multi-task regression approaches on simulated thick-slice data.

Keywords: Neonatal Brain MRI · Brain Biometry · Multi-task Learning

1 Introduction

Preterm birth affects approximately 15 million infants annually, with very preterm infants (<32 weeks) at substantial risk for neurodevelopmental impairment [1]. MRI at term-equivalent age is essential for characterizing brain development and injury [2, 6], where quantitative biometry—measuring anatomical structures on standardized planes provides objective markers that correlate with long-term outcomes [3, 5]. Common biometric measurements include transcerebellar diameter (TCD), biparietal width (BPW), interhemispheric distance (IHD), deep

[*] Equal contribution
[†] Corresponding author

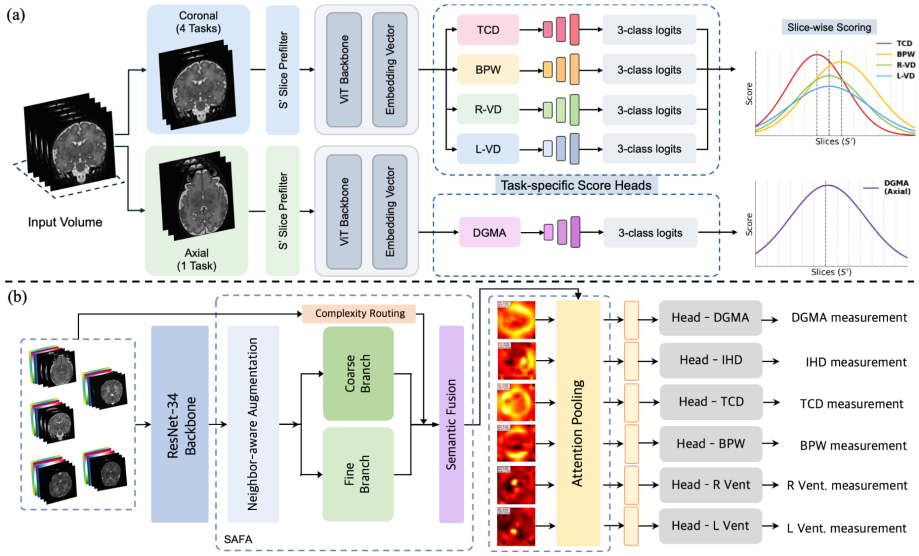

Fig. 1. Overview of MEASURE. (a) ViT-based multi-task slice selector model structure. (b) Multi-task regression model with proposed SAFA, attention pooling.

gray matter area (DGMA), and left/right ventricular diameters (L-VD, R-VD). However, manual measurement requires expertise and takes 15–30 minutes per scan with considerable inter-rater variability (intraclass correlation coefficients of 0.70–0.90 [2]).

For neonatal brain injury assessment using the Kidokoro system [2], the only prior automation for it's brain biometry quantitative measurement is the pipeline by van Eijk et al. [7], which depends on brain segmentation requiring manual quality control. Nearly all deep learning approaches also rely on dense annotation: segmentation-based methods derive measurements from delineated boundaries [8, 17], while landmark-based methods detect keypoints through heatmap regression [9, 18]. The recent FeTA Challenge [12, 13] highlighted how segmentation-reliant methods can struggle to surpass simple age-based regression.

We propose MEASURE, a framework that predicts neonatal brain biometry for Kidokoro protocol [2] directly from MRI without dense annotations. Our two-stage approach mirrors the clinical workflow: a ViT-based multi-task slice selector identifies measurement-relevant planes, then a multi-task regression network jointly predicts measurements for each biometric. Our contributions are threefold: (1) this the first annotation-efficient deep learning framework for Kidokoro-protocol neonatal biometry quantitative measurement, eliminating dense supervision requirements; (2) a principled two-stage design mirroring clinical workflow; and (3) comprehensive evaluation demonstrating that explicit plane selection outperforms end-to-end volumetric approaches.

2 Method

The MEASURE framework operates in two stages: a multi-task slice selector identifies measurement-relevant planes, then a multi-task predictor jointly returns biometry measurements.

2.1 Multi-Task Slice Selector

The selector predicts slice-wise suitability scores and returns the Top-K most suitable slices per task, labeled according to the Kidokoro protocol [2]. As shown in Fig. 1(a), we use a multi-task selector for the coronal view (four heads for five measurements, as BPW and IHD share a target slice) and a single-task selector for the axial view (DGMA).

Each slice is resized to 224×224 with aspect-ratio–preserving padding. From S=29 candidates per view, we retain the top S' slices (S'=22, ≈70%) ranked by nonzero pixel fraction, preserving anatomical ordering. Each retained slice s is represented as a 2.5D input [4] by stacking (s−1, s, s+1) along the channel dimension, yielding $\mathbf{x} \in \mathbb{R}^{S' \times 3 \times H \times W}$.

From the manual target slice index c_g for each selector head g, we derive per-slice labels: strong positive (class 2) for $s = c_g$, weak positive (class 1) for its immediate neighbors ($|s - c_g| = 1$), and negative (class 0) otherwise. This 3-class scheme models annotation ambiguity: adjacent slices are near-optimal measurement planes, so penalizing them as hard negatives would be counterproductive when the predictor consumes the Top-K candidates. A ViT backbone [14] extracts per-slice embeddings $\mathbf{f}_s \in \mathbb{R}^D$, and task-specific 3-class heads yield suitability scores $a_{s,g} = p_{s,g}(2)$ (the strong-positive probability), where $p_{s,g}(c) = \mathrm{softmax}(\mathbf{z}_{s,g})_c$ for $c \in \{0, 1, 2\}$.

2.2 Selector Objective

The selector is optimized with class-weighted cross-entropy for slice discrimination:

$$\mathcal{L}_{\mathrm{CE}}^{w} = -\frac{1}{S'G} \sum_{s=1}^{S'} \sum_{g=1}^{G} \alpha_{y_{s,g}} \log p_{s,g}(y_{s,g}) \tag{1}$$

where $\alpha_0 < \alpha_1 < \alpha_2$, and G is the number of selector heads in the corresponding view. We further use a set-level distribution loss that concentrates probability mass within the ground-truth neighborhood:

$$\mathcal{L}_{\mathrm{set}} = -\frac{1}{G} \sum_{g=1}^{G} \sum_{s=1}^{S'} q_{s,g} \log \pi_{s,g} \tag{2}$$

where $\pi_{s,g} = \mathrm{softmax}_s(\tilde{s}_{s,g})$ with $\tilde{s}_{s,g} = \sum_c c \cdot p_{s,g}(c)$, and $q_{s,g}$ is a triangular target $(1, 2, 1)$ centered at c_g, normalized. The final loss is $\mathcal{L}_{\mathrm{sel}} = \beta_1 \mathcal{L}_{\mathrm{CE}}^{w} + \beta_2 \mathcal{L}_{\mathrm{set}}$.

2.3 Multi-Task Biometry Predictor

As shown in Fig. 1(b), the model receives three slices (center, 2 neighbors) with three coordinate channels [21]. A pretrained ResNet-34 [22] extracts feature maps $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$.

Slice-and-Structure Aware Feature Augmentation (SAFA). SAFA addresses two challenges in 2.5D biometry prediction: (1) limited through-plane context from single-slice features, and (2) spatially varying complexity across brain regions.

Through-plane augmentation: For the 3 neighboring slices in input (s_{-1}, s_0, s_{+1}), we compute inter-slice difference maps $\Delta^+ = s_{+1} - s_0$ and $\Delta^- = s_0 - s_{-1}$. A lightweight CNN ψ encodes these differences, and the output is fused with backbone features $\mathbf{F}$ via learned projection:

$$\mathbf{F}' = \mathbf{F} + \gamma \cdot \mathrm{Conv}_{1\times1}(\psi([\Delta^+; \Delta^-])) \tag{3}$$

where γ is a learnable scale initialized to 0.

Spatially adaptive routing: The features $\mathbf{F}'$ are processed by parallel fine and coarse branches. The fine branch applies dilated depthwise convolutions [26, 27] at the $\mathbf{F}'$ spatial size to capture high-frequency structural detail. Coarse branch uses a strided convolution followed by bilinear upsampling, capturing global context. Their outputs $\mathbf{F}_{\text{fine}}, \mathbf{F}_{\text{coarse}} \in \mathbb{R}^{C\times H\times W}$ are blended pixel-wise via a soft routing mask:

$$\mathbf{F}_{\text{out}} = \mathbf{r} \odot \mathbf{F}_{\text{fine}} + (\mathbf{1} - \mathbf{r}) \odot \mathbf{F}_{\text{coarse}} \tag{4}$$

where $\mathbf{r} = \sigma((\mathbf{c} - \tau) \cdot t) \in [0, 1]^{H\times W}$. Here $\mathbf{c}$ is a Sobel-derived edge complexity map computed from the center input slice, τ is a fixed threshold, and t is a learnable temperature controlling routing sharpness, encouraging $\mathbf{r}$ toward binary decisions at high values. Specifically, complexity map is normalized per-sample to [0, 1], τ is fixed as the midpoint of [0,1].

Measurement-Specific Attention Pooling. Motivated by spatial attention mechanisms [31], each of the N measurements learns its own spatial attention map over the routed feature map $\mathbf{F}_{\text{out}} \in \mathbb{R}^{C\times H\times W}$, rather than sharing a single global average pooling operation:

$$\mathbf{a}_i = \mathrm{softmax}(\phi_i(\mathbf{F}_{\text{out}}) + \lambda\, \mathbf{b}_i), \quad \mathbf{m}_i = \sum_{h,w} \mathbf{a}_i(h, w) \cdot \mathbf{F}_{\text{out}}(h, w), \quad i = 1, \ldots, N_{meas.} \tag{5}$$

where $\phi_i : \mathbb{R}^{C\times H\times W} \to \mathbb{R}^{H\times W}$ is a learnable attention head and $\mathbf{b}_i \in \mathbb{R}^{H\times W}$ a learnable spatial prior (shown in Fig. 1(b)), yielding measurement-specific feature vectors $\mathbf{m}_i \in \mathbb{R}^C$ for individual regression heads.

Anatomical Attention Regularization (AAR). Three terms regularize the attention maps: *symmetry* (applied only to L-VD and R-VD, whose attention maps should be horizontal mirror images), *smoothness* (spatial coherence, applied to all six maps), and *distinctiveness* (penalizing cosine similarity between

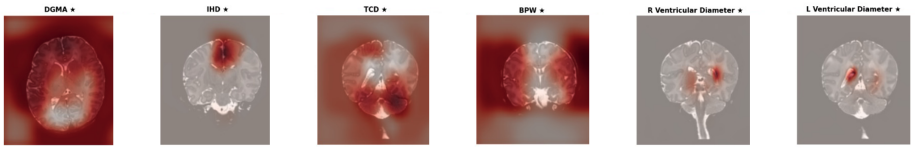

Fig. 2. Visualization of 6 measurement pooling attention maps for a subject.

all map pairs):

$$\mathcal{L}_{\text{AAR}} = \|\mathbf{a}_{\text{L}} - \text{flip}_h(\mathbf{a}_{\text{R}})\|^2 + \frac{1}{6}\sum_{i=1}^{6}(\|\nabla_x \mathbf{a}_i\|_1 + \|\nabla_y \mathbf{a}_i\|_1) + \sum_{i\neq j}\max(0, \langle \hat{\mathbf{a}}_i, \hat{\mathbf{a}}_j\rangle - \tau_d) \tag{6}$$

where $\mathbf{a}_{\text{L}}, \mathbf{a}_{\text{R}} \in \mathbb{R}^{H\times W}$ denote the attention maps for L-VD and R-VD respectively, flip_h denotes flipping, $\hat{\mathbf{a}}_i = \mathbf{a}_i^{\text{flat}}/\|\mathbf{a}_i^{\text{flat}}\|_2$ is the flattened and ℓ_2-normalized attention map, and τ_d is a margin controlling the distinctiveness penalty (distinct from the routing threshold τ). As shown in Fig. 2, the learned spatial attention weighting maps are anatomically accurate even without voxel-level supervision.

2.4 Predictor Objective

For each measurement i, heads predict a value $\hat{y}_i$ and log-variance $\log\sigma_i^2$. The loss combines four terms. Concordance Correlation Coefficient (CCC) loss [25] directly optimizes clinical agreement:

$$\mathcal{L}_{\text{CCC}} = 1 - \frac{2\sigma_{\hat{y}}\sigma_y\rho_{\hat{y}y}}{\sigma_{\hat{y}}^2 + \sigma_y^2 + (\mu_{\hat{y}} - \mu_y)^2} \tag{7}$$

Negative log-likelihood [30] loss:

$$\mathcal{L}_{\text{NLL}} = \frac{1}{2}\log\sigma^2 + \frac{(y-\hat{y})^2}{2\sigma^2} \tag{8}$$

A bilateral symmetry constraint penalizes left/right ventricular prediction differences beyond a learnable margin: $\mathcal{L}_{\text{sym}} = \text{ReLU}(|\hat{y}_L - \hat{y}_R| - m)$. The final objective combines these with MSE loss and AAR regularization:

$$\mathcal{L}_{\text{pred}} = \alpha_1\mathcal{L}_{\text{CCC}} + \alpha_2 w(e)\mathcal{L}_{\text{NLL}} + \alpha_3\mathcal{L}_{\text{MSE}} + \alpha_4\mathcal{L}_{\text{sym}} + \alpha_5 w'(e)\mathcal{L}_{\text{AAR}} \tag{9}$$

where $w(e) = \min(1, e/20)$ warms up the NLL term. For the AAR regularization, $w'(e) = \min(1, \max(0, (e-10)/20))$ implements a delayed start: or can cause the model to lock onto incorrect anatomical regions if starts too early.

3 Experiments

3.1 Dataset and Preprocessing

We use the developing Human Connectome Project (dHCP) [11], a publicly available neonatal brain MRI dataset acquired at 3T with T2-weighted sequences

Table 1. Biometry prediction performance and ablation study. **1st**, 2nd, 3rd. ↑higher is better; ↓lower is better.

Method	Type	ICC ↑						Pearson r ↑						MAE ↓					
		TCD	BPW	R-VD	L-VD	DGMA	IHD	TCD	BPW	R-VD	L-VD	DGMA	IHD	TCD	BPW	R-VD	L-VD	DGMA	IHD
Comparison Methods																			
PMA	–	.697	.722	.052	.222	.367	.032	.730	.735	.060	.334	.463	.149	1.80	4.88	1.10	0.91	0.85	0.86
SUNETRV2	3D	.649	.787	.104	.094	.758	.101	.690	.807	.192	.172	.781	.229	2.00	2.68	0.93	1.08	0.55	0.80
BrainAge	3D	.780	.903	.459	.552	.821	.416	.805	.909	.505	.588	.835	.445	1.64	1.84	0.82	0.92	0.49	0.83
ViT3D	3D	.675	.812	.175	.259	.731	.161	.710	.825	.246	.358	.750	.272	1.97	2.63	0.94	1.02	0.59	0.80
MECDS‡	2D	.816	.907	.622	.673	.814	.418	.853	.915	.666	.712	.832	.479	1.43	1.88	0.69	0.77	0.48	0.77
SePL‡	2D	.800	.909	.629	.660	.824	.352	.842	.924	.673	.719	.848	.438	1.50	1.76	0.69	0.78	0.47	0.76
MEASURE (Ours)	2D	**.847***	**.924***	**.674***	**.722***	**.837**	**.485***	**.880***	**.930**	**.711***	**.759***	**.855**	**.564***	**1.33***	**1.62***	**0.66**	**0.73**	**0.46**	**0.71***
Ablation Study (MEASURE variants)																			
Base	–	.734	.904	.447	.558	.806	.367	.763	.912	.511	.627	.821	.424	1.73	1.88	0.82	0.88	0.51	0.81
w/o Attn Pool & AAR	–	.828	.919	.596	.673	.833	**.502**	.840	.920	.635	.702	.841	.525	1.52	1.81	0.73	0.80	0.48	0.76
w/o SAFA	–	.857°	.920	.699°	.748°	.829	.459	.870	.924	.718°	.770°	.836	.500	1.34	1.77	**0.64**°	0.71°	0.48	0.76
Full Model	–	.847	**.924**	.674	.722	.837	.486	**.880**	**.930**	.711	.759	**.855**	**.564**	**1.33**	**1.62**†	0.66	0.73	0.46	**0.71**†

‡ Input slices chosen by our selector. * Sig. better than 2nd-best comparison method (p<0.05). † Sig. better than w/o SAFA (p<0.01). ° Not statistically significant.

at isotropic resolution. We select 470 subjects scanned at term-equivalent age (36–42 weeks PMA), consistent with the Kidokoro scoring system [2]. Expert measurements for six biometrics were obtained by a single experienced clinician following the Kidokoro protocol measurement procedure [2]. Intra-rater reliability, assessed via re-measurement at a >3-month interval on a randomly sampled 97-subject subset. Routine clinical neonatal MRI typically uses thick-slice acquisitions due to time constraints and motion sensitivity [29]. To evaluate under thick-slice conditions, we simulate clinical acquisitions by first applying a Gaussian slice profile along the through-plane axis, downsampling to 5 mm slice thickness and 0.8 mm in-plane resolution via bilinear interpolation, and adding Rician noise [28] to both axial and coronal views. Expert measurements for six biometrics are obtained following the Kidokoro protocol. All experiments use 5-fold cross-validation with subject-level splits, reporting Intraclass Correlation Coefficient (ICC), Pearson r, and Mean Absolute Error (MAE).

3.2 Neonatal Biometry Prediction

We compare against three baseline categories: PMA regression as the lower bound [12]; 3D models (SwinUNETR-V2 [10], 3D ViT [15], BrainAge [16]) to assess whether explicit plane selection outperforms end-to-end volumetric regression; and recent multi-task regression methods adapted from related domains (MECDS [19], SePL [20]), as no DL prior exists for Kidokoro neonatal biometry. For 2D methods, inputs are slices chosen by our selector. van Eijk et al. [7] is the only prior automated Kidokoro pipeline, but is not open-sourced and was validated on a private dataset, precluding direct comparison.

Table 1 shows MEASURE ranks first across all six biometrics and all three metrics, with statistically significant improvements on the majority of comparisons. 3D models consistently underperform 2D slice-based methods, validating our two-stage design: explicit plane selection concentrates features on measurement-relevant anatomy, while 3D networks must simultaneously local-

Table 2. Bland–Altman analysis [23]: model vs. human intra-rater variability (>3-month interval). LoA Ratio <1 indicates model is more consistent than human re-measurement.

Measure	Human BLoA	Model BLoA	LoA Ratio	Coverage
BPW	−0.24 (−7.77; 7.30)	0.71 (−4.37; 5.79)	0.66 (0.45; 1.10)	97.7%
IHD	−0.32 (−2.22; 1.57)	−0.01 (−1.84; 1.82)	0.99 (0.77; 1.29)	90.8%
TCD	0.17 (−2.47; 2.80)	0.12 (−3.29; 3.53)	1.28 (1.04; 1.64)	88.5%
DGMA	−0.13 (−1.39; 1.13)	−0.11 (−1.36; 1.14)	1.17 (0.90; 1.48)	90.8%
L-VD	−0.37 (−3.04; 2.30)	0.05 (−1.92; 2.03)	0.73 (0.62; 0.87)	97.7%
R-VD	−0.30 (−2.79; 2.18)	0.09 (−1.60; 1.69)	0.68 (0.54; 0.88)	98.9%

Table 3. Ablation on selector loss components.

			Top-1					**Recall@2**				
Cross Entropy	Class-wtd CE	Set-level loss	TCD	BPW	R-VD	L-VD	DGMA	TCD	BPW	R-VD	L-VD	DGMA
✓	×	×	.933	.833	.833	.967	.933	1.00	.900	1.00	1.00	1.00
×	✓	×	**.967**	**.867**	.833	.900	**.967**	1.00	.933	1.00	1.00	1.00
×	×	✓	.900	.833	.800	.933	.900	1.00	.900	1.00	1.00	.967
×	✓	✓	.867	**.867**	**.867**	**1.00**	.933	**1.00**	**.967**	**1.00**	**1.00**	**1.00**

ize and measure. Performance varies by structure: BPW and DGMA are well-predicted across methods (ICC > 0.80), while IHD is the most challenging due to its narrow dynamic range ($\sim$2 mm); its MAE of 0.71 mm remains clinically acceptable despite low ICC. Ventricular measures show the clearest benefit from measurement-specific attention pooling, confirming that lateralized structures require spatially distinct attention. For contextual reference, van Eijk et al. reported ICCs of 0.50–0.85 on the six shared biometrics using a segmentation-dependent pipeline with manual quality control [7], broadly consistent with MEASURE's range. Table 2 further shows that for BPW, L-VD, R-VD, and IHD, model variability is narrower than human intra-rater variability (LoA ratio < 1), with R-VD and L-VD achieving ratios of 0.68 and 0.73 at >97% coverage. The ablation (Table 1, lower) reveals Attention Pooling as the most impactful component. To validate SAFA and AAR, we apply paired Wilcoxon signed-rank tests with Benjamini–Hochberg FDR correction [24] across all 12 ablation comparisons. SAFA yields FDR-significant MAE improvements on BPW and IHD (p_{FDR}=0.036 each), the two structures most sensitive to through-plane context and fine-grained spatial detail. AAR shows a trend toward BPW improvement. Cells where ablated variants numerically exceed the full model are marked $^\circ$ in Table 1, none are statistically significant after correction. The apparent ventricular ICC advantage of **w/o AAR** is expected: the symmetry loss compresses predicted L/R asymmetry, slightly reducing prediction variance and ICC for lateralized structures. We consider this a deliberate trade-off for anatomically interpretable attention maps (Fig. 2).

Table 4. Cross-domain evaluation on the FeTA 2024 dataset using MAE and ICC.

	MAE (mm) ↓					ICC ↑				
Model	**bBIP**	**sBIP**	**HV**	**LCC**	**TCD**	**bBIP**	**sBIP**	**HV**	**LCC**	**TCD**
GA only	2.91	3.50	1.27	2.84	2.55	0.937	0.916	0.817	0.809	0.895
Image only	3.69	4.34	1.52	2.98	2.52	0.852	0.835	0.751	0.749	0.862
Image + GA	2.88	3.16	1.51	2.91	2.09	0.932	0.921	0.729	0.799	0.916

3.3 Ablation Study

Selector Objective. Table 3 compares selector objectives. Since the predictor consumes top-2 candidates, we report Recall@2. The adopted objective (class-weighted CE + set-level loss) provides the most stable top-2 retrieval, achieving near-perfect Recall@2 across measurements with competitive Top-1 performance.

Cross-domain Evaluation on Fetal Brain Biometry. To assess architectural generalizability, we adapt MEASURE to fetal brain biometry using the FeTA 2024 Challenge dataset [12, 13] (70 subjects, 21–35 weeks GA). This evaluation tests whether our two-stage design transfers to a fundamentally different population (fetal vs. neonatal) and gestational range (21–35 vs. 36–42 weeks), rather than evaluating thick-slice robustness. The FeTA dataset provides only isotropic reconstructions, which we use directly without any simulations.

We retain core components (ViT slice selector, ResNet-34 with SAFA, measurement specific attention pooling) while replacing prediction heads for five fetal targets (bBIP, sBIP, HV, LCC, TCD). Following the challenge protocol, GA is provided as input. Table 4 shows the image-only configuration failed to converge, consistent with FeTA 2024 findings where most methods could not surpass GA-only regression [13]. We attribute this to limited training data and anatomical variability caused by domain differences. The image+GA configuration improves predictions for larger, visually distinctive structures (TCD, sBIP), while smaller structures (HV, LCC) show no improvement over GA-only. These results demonstrate evidence of architectural adaptability, and that learned image features provide complementary value for sufficiently salient structures (TCD, sBIP), consistent with FeTA 2024 challenge findings [13]. This result highlighting the inherent difficulty of fetal biometry from T2w image for further research.

4 Conclusion

We presented MEASURE, a deep learning framework for neonatal brain biometry that does not require voxel-level dense annotations. By decomposing the problem into slice selection and measurement-specific regression—mirroring clinical workflow, our approach achieves strong agreement with expert measurements while eliminating dense annotation requirements. Experiments on the dHCP dataset demonstrate consistent improvements over both 3D volumetric and 2D

slice-based baselines, and cross-domain evaluation on the FeTA 2024 dataset confirms architectural transferability. Limitations include evaluation on simulated thick-slice data from a single site. bridging the domain gap to real clinical thick-slice acquisitions with potential motion artifacts will require further validation.

Acknowledgments. This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2020-II201373, Artificial Intelligence Graduate School Program(Hanyang University)). Also by the National IT Industry Promotion Agency(NIPA), an agency under the MSIT and with the support of the Daegu Digital Innovation Promotion Agency (DIP), the organization under the Daegu Metropolitan Government.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Blencowe, H., Cousens, S., Oestergaard, M.Z., et al.: National, regional, and worldwide estimates of preterm birth rates in the year 2010 with time trends since 1990 for selected countries: a systematic analysis and implications. The Lancet **379**(9832), 2162–2172 (2012)
2. Kidokoro, H., Neil, J.J., Inder, T.E.: New MR imaging assessment tool to define brain abnormalities in very preterm infants at term. AJNR Am. J. Neuroradiol. **34**(11), 2208–2214 (2013)
3. Setänen, S., Lahti, K., Lehtonen, L., et al.: Neurological examination combined with brain MRI or cranial US improves prediction of neurological outcome in preterm infants. Early Hum. Dev. **127**, 68–72 (2018)
4. Kumar, A., Jiang, H., Imran, M., et al.: A flexible 2.5D medical image segmentation approach with in-slice and cross-slice attention. Comput. Biol. Med. **182**, 109173 (2024)
5. George, J.M., Fiori, S., Fripp, J., et al.: Validation of an MRI brain injury and growth scoring system in very preterm infants. AJNR Am. J. Neuroradiol. **38**(7), 1435–1442 (2017)
6. Woodward, L.J., Anderson, P.J., Austin, N.C., Howard, K., Inder, T.E.: Neonatal MRI to predict neurodevelopmental outcomes in preterm infants. N. Engl. J. Med. **355**(7), 685–694 (2006)
7. van Eijk, L., Pannek, K., George, J.M., et al.: Automating quantitative measures of an established conventional MRI scoring system for preterm-born infants. AJNR Am. J. Neuroradiol. **42**(10), 1870–1877 (2021)
8. She, J., Huang, H., Ye, Z., et al.: Automatic biometry of fetal brain MRIs using deep and machine learning techniques. Sci. Rep. **13**(1), 17860 (2023)
9. Avisdris, N., Yehuda, B., Ben-Zvi, O., et al.: Automatic linear measurements of the fetal brain on MRI with deep neural networks. Int. J. Comput. Assist. Radiol. Surg. **16**(9), 1481–1492 (2021)
10. He, Y., Nath, V., Yang, D., Tang, Y., et al.: SwinUNETR-V2: Stronger Swin Transformers with stagewise convolutions for 3D medical image segmentation. In: Greenspan, H., et al. (eds.) MICCAI 2023. LNCS, vol. 14223, pp. 416–426. Springer, Cham (2023)

11. Makropoulos, A., Robinson, E.C., Schuh, A., et al.: The developing human connectome project: A minimal processing pipeline for neonatal cortical surface reconstruction. NeuroImage **173**, 88–112 (2018)
12. Payette, K., Li, H.B., de Dumast, P., et al.: Fetal brain tissue annotation and segmentation challenge results. Med. Image Anal. **88**, 102833 (2023)
13. Zalevskyi, V., Sanchez, T., Kaandorp, M., et al.: Advances in automated fetal brain MRI segmentation and biometry: Insights from the FeTA 2024 Challenge. arXiv preprint arXiv:2505.02784 (2025)
14. Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)
15. Kunanbayev, K., Shen, V., Kim, D.S.: Training ViT with limited data for Alzheimer's disease classification: an empirical study. In: Linguraru, M.G., et al. (eds.) MICCAI 2024. LNCS, vol. 15010, pp. 334–343. Springer, Cham (2024)
16. Wood, D.A., Townend, M., Guilhem, E., et al.: Optimising brain age estimation through transfer learning: A suite of pre-trained foundation models. Hum. Brain Mapp. **45**(4), e26625 (2024)
17. Luis, A., Uus, A., Matthew, J., et al.: Towards automated fetal brain biometry reporting for 3-dimensional T2-weighted 0.55–3T magnetic resonance imaging at 20–40 weeks gestational age range. Pediatr. Radiol. (2025). https://doi.org/10.1007/s00247-025-06403-2
18. Avisdris, N., Joskowicz, L., Dromey, B., et al.: BiometryNet: landmark-based fetal biometry estimation from standard ultrasound planes. In: Wang, L., et al. (eds.) MICCAI 2022. LNCS, vol. 13434, pp. 279–289. Springer, Cham (2022)
19. Jiang, H., Huang, H., Cai, J., Xu, M., Shen, Z., Fei, M., Wang, X., Zhang, L., Wang, Q.: Multi-task Screening for Cervical Diseases via Feature Routing and Asymmetric Distillation. In: Gee, J.C., et al. (eds.) MICCAI 2025. LNCS, vol. 15973, pp. 389–398. Springer, Cham (2025)
20. Ko, S., Yang, H., Bum, J., Le, D.-T., Choo, H.: Self-Propagative Multi-Task Learning for Predicting Cardiometabolic Risk Factors. In: Gee, J.C., et al. (eds.) MICCAI 2025. LNCS, vol. 15974, pp. 564–573. Springer, Cham (2025)
21. Liu, R., Lehman, J., Molino, P., et al.: An intriguing failing of convolutional neural networks and the CoordConv solution. In: NeurIPS (2018)
22. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR, pp. 770–778 (2016)
23. Bland, J.M., Altman, D.G.: Statistical methods for assessing agreement between two methods of clinical measurement. The Lancet **327**(8476), 307–310 (1986)
24. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B (Methodological) **57**(1), 289–300 (1995)
25. Lin, L.I.: A concordance correlation coefficient to evaluate reproducibility. Biometrics **45**(1), 255–268 (1989)
26. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. In: ICLR (2016)
27. Howard, A.G., Zhu, M., Chen, B., et al.: MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)
28. Gudbjartsson, H., Patz, S.: The Rician distribution of noisy MRI data. Magn. Reson. Med. **34**(6), 910–914 (1995)
29. Dubois, J., Alison, M., Counsell, S.J., et al.: MRI of the neonatal brain: a review of methodological challenges and neuroscientific advances. J. Magn. Reson. Imaging **53**(5), 1318–1343 (2021)

30. Kendall, A., Gal, Y.: What uncertainties do we need in Bayesian deep learning for computer vision? In: NeurIPS, pp. 5574–5584 (2017)
31. Woo, S., Park, J., Lee, J.-Y., Kweon, I.S.: CBAM: Convolutional block attention module. In: ECCV, pp. 3–19 (2018)