Self-Supervised Representation Learning of Fetal Heart Rate Outperforms a Foundation Model in Identifying Fetal Acidosis

Johann Vargas-Calixto[1], Masoud Nateghi[1], Jay Ward[2], Jim Robertson[2], Yuri Sebastião[3], Jackie Patterson[3], Jeffrey Stringer[3], Reza Sameni[1,4], Gari D. Clifford[1,4]

[1]Emory University, Atlanta, GA, USA, [2] MindChild Medical, Inc., North Andover, MA, USA, [3]Division of Global Women's Health, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA, [4]Georgia Institute of Technology, Atlanta, GA, USA

Abstract

***Background**: Label scarcity in fetal monitoring is a severe limitation to training deep learning models for acidosis detection models; blood gas testing is not universal practice. Self-supervised learning (SSL) can leverage unlabeled recordings to learn signal representations without outcome labels.* ***Methods**: We used the SimMTM SSL framework to pretrain an encoder using unlabeled fetal heart rate (FHR) recordings. The frozen encoder extracted 256-dimensional features from the last hour before delivery, which were classified by a BiLSTM. We also trained the same architecture without encoder pretraining. The end-to-end model was compared to the pretrained model and to Google's TimesFM 2.5, a 200M-parameter time series foundation model. We evaluated the models using a stratified 10×5-fold cross-validation on 1,217 labeled patients (59 acidosis cases). External validation was performed on the public CTU-UHB dataset ($N = 552$, 29 acidosis cases).* ***Results**: The SSL-pretrained model achieved AUROC of $\mu = 0.704 \pm \sigma = 0.054$, significantly outperforming the end-to-end baseline (0.677 ± 0.069; $p = 0.006$, rank-biserial $r = 0.473$). TimesFM did not improve over the baseline (0.684 ± 0.066; $p = 0.608$). On CTU-UHB, the SSL model achieved the highest AUROC (0.743), followed by end-to-end (0.637) and TimesFM (0.600).* ***Conclusions**: Domain-specific SSL pretraining on FHR data significantly improves fetal acidosis detection over supervised training alone, while a general-purpose foundation model provides no benefit. The superiority of SSL was maintained during external testing. Thus, it is necessary to leverage unlabeled FHR data to address label scarcity in fetal monitoring.*

1. Introduction

Severe intrapartum fetal acidosis, is a rare but serious complication of labor associated with neonatal encephalopathy [1]. Fetal heart rate (FHR) monitoring is the main surveillance tool during labor. Yet, its interpretation is subjective, leading to high false-alarm rates, unnecessary Caesarean deliveries, and clinician alarm fatigue [2].

Deep learning has shown promise for automated FHR analysis, but development of better performing models is limited by labeled data scarcity. Arterial blood gas measurements are not universal; previous research reported lack of confirmed labels in over 70% of cases [3, 4]. Self-supervised learning (SSL) could be used for representation learning from FHR data to help leverage unlabeled recordings instead of discarding them. Here, we adapt the simple masked time-series modeling (SimMTM) framework [5] to the FHR domain. The core idea is that an encoder trained to reconstruct masked FHR segments and distinguish them from other segments will learn features that capture clinically relevant morphology, such as accelerations, decelerations, and baseline variability, without requiring outcome labels.

We evaluate whether pretrained encoders can improve downstream acidosis detection compared to training from scratch. To this aim, we compare two pretrained encoders: (1) a domain-specific encoder pretrained using the simMTM approach, and (2) Google TimesFM 2.5, a large general-purpose time series foundation model [6].

2. Methods

2.1. Study Population

Intrapartum FHR recordings were obtained using transabdominal fetal electrocardiography (FECG) from a multicenter retrospective cohort from Zambia, Ghana, and India as part of the Limiting adverse birth outcomes in resource-limited setting (LABOR) project, conducted by the University of North Carolina (UNC) at Chapel Hill [7]. The study was approved by Emory IRB STUDY00000128 XIRB LABOR Study for UNC Parent Study 19-0765 (Z 31902 - Limiting Adverse Birth Outcomes in Resource-Limited Settings - The LABOR Study) [7]. Patients were assigned to two mutually exclusive cohorts: an

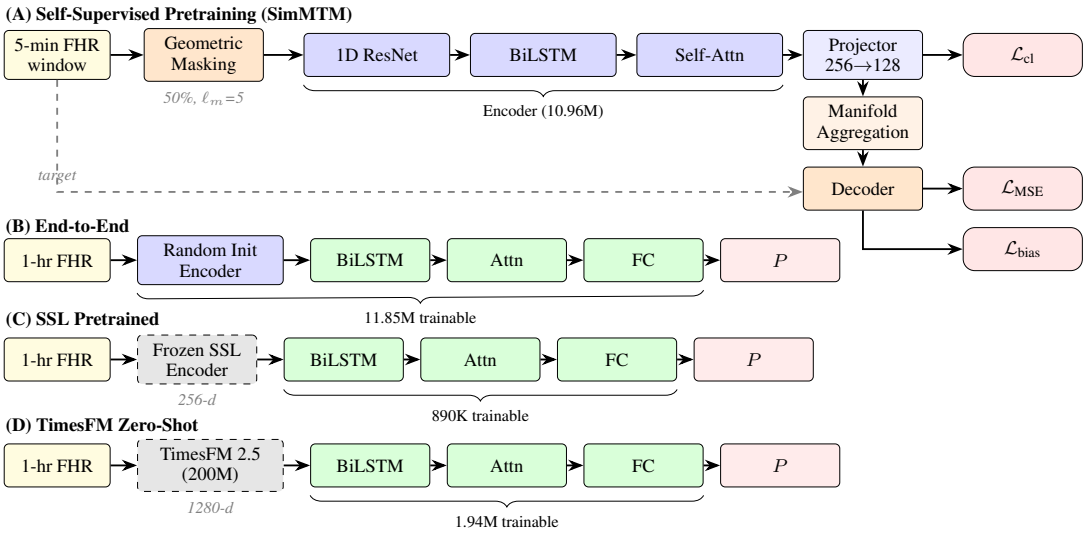

Figure 1: Framework overview. (A) SimMTM pretraining on unlabeled FHR data. Downstream classification (B) starting from random initialization, (C) using a frozen SSL encoder, and (D) using the TimesFM 2.5 encoder.

unlabeled cohort ($N_{\text{pretrain}} = 7{,}850$) comprising patients without outcome labels, and a labeled cohort ($N = 1{,}217$; 59 acidosis cases; 4.8% prevalence) comprising patients with confirmed labor outcome labels. In this dataset, acidosis was defined by a cord lactate level $\geq 8\,\text{mmol/L}$.

2.2. Signal Preprocessing

Raw transabdominal FECG collected using MERIDIAN M110R (MindChild Medical Inc., North Andover, MA) was processed to extract FECG R-peak locations and the FHR signal. The FHR was resampled to 2 Hz via block averaging to obtain a uniform sampling rate. A signal quality index (SQI) was thresholded at 0.15 to generate a binary validity mask. Gaps $\leq$15 s were linearly interpolated using 5-sample median anchors; gaps $>$15 s were zeroed to encode sensor detachment. Signals were Z-score normalized ($\mu = 140$ BPM, $\sigma = 15$ BPM) with fixed physiological constants to preserve baselines. The final 60 minutes before the end of the recordings were extracted, left-zero-padded if shorter, and segmented into overlapping 5-minute windows with a 50% stride, yielding $\sim$23 windows per patient. Windows with $<$50% valid samples were flagged via a validity mask propagated to the attention mechanism of the classifier.

2.3. SimMTM Self-Supervised Pretraining

We adapted SimMTM [5] to pretrain a domain-specific encoder on the unlabeled cohort. Fig. 1A shows the architecture, comprising a 1D ResNet backbone (8 residual blocks, kernel size 7, channels 64→512), a single-layer bidirectional LSTM (hidden= 256), and Bahdanau self-attention, projecting each 5-minute window to a 256-dimensional latent vector ($\sim$ 11M parameters).

During pretraining, 50% of each window was randomly masked, with an average length ℓ_m=5 samples and five masked copies per signal. Three objectives were jointly minimized: (1) $\mathcal{L}_{\text{cl}}$, a contrastive loss based on KL divergence over cosine similarities ($\tau = 0.2$); (2) $\mathcal{L}_{\text{MSE}}$, a variance-normalized MSE computed on valid samples only; and (3) $\mathcal{L}_{\text{bias}}$, a bias loss penalizing baseline shift errors independently from morphology. Training used Adam (lr $= 10^{-4}$, batch size 64) for 1,000 epochs.

2.4. TimesFM Zero-shot Evaluation

We compared our domain-specific encoder to TimesFM 2.5 [6], a 200M-parameter decoder-only causal transformer pretrained on billions of time points from non-physiological domains. Each 5-minute window was left-padded to 608 samples (19 patches of 32), processed through 20 frozen transformer layers, and the last-patch embedding (1280-dim) was extracted. We did not apply a projection layer for zero-shot evaluation.

2.5. Classification Model

Both pretrained encoders (SSL and TimesFM) were used to generate latent window sequences for each participant. These sequences were fed to a 2-layer bidirectional LSTM (hidden 128), Bahdanau self-attention with a valid-mask mechanism which forced zero attention on invalid windows, layer normalization, and a fully connected classification head. Weighted binary cross-entropy loss was used ($w_+ = 10$) to address class imbalance. Training used early stopping on the validation AUROC (patience = 100), for a maximum of 1,000 epochs. For the end-to-end baseline, an encoder with the same architecture pretrained with SimMTM was randomly initialized and

Table 1: Internal cross-validation results. Mean ± SD. Sens@$X\%$: sensitivity at $X\%$ FPR.

Model	AUROC	AUPRC	Sens@10%	Sens@20%
SSL Pretrained	**0.704** ±0.054	0.139 ±0.043	**0.229** ±0.116	**0.410** ±0.129
TimesFM 2.5	0.684 ±0.066	0.127 ±0.051	0.198 ±0.119	0.360 ±0.138
End-to-End	0.677 ±0.069	**0.140** ±0.057	0.215 ±0.119	0.383 ±0.130

trained jointly with the classifier ($\sim$ 12M parameters).

2.6. Cross-Validation and Evaluation

Three conditions were compared using 10×5 repeated stratified K-fold cross-validation (20% internal validation split): (1) SSL Pretrained (frozen encoder, Fig. 1B); (2) TimesFM zero-shot (frozen encoder, Fig. 1C); and (3) end-to-end classification (Fig. 1D). All models were trained and tested on the same partitions. Friedman omnibus tests assessed global differences, followed by treatment-vs-control Wilcoxon signed-rank tests and Holm-Bonferroni correction, where the end-to-end approach was the control. The effect sizes were quantified by matched-pairs rank-biserial correlation. Finally, we performed a zero-shot cross-domain and cross-population external validation on the CTU-UHB dataset ($N = 552$) [8]. This database contains FHR collected mostly with Doppler ultrasound, at 4 Hz, and using different biomarkers for acidosis. The classifiers were retrained on the full internal labeled cohort and zero-shot evaluated once on the CTU-UHB database, resampled to 2 Hz and normalized. For evaluation, we defined the target acidosis class as a pH < 7.0 or base deficit > 10.0 ($N_{\text{acidosis}} = 29$).

3. Results

Table 1 summarizes the internal CV results. SSL pretraining yielded the highest AUROC of all conditions (0.704 ± 0.054). The improvement was significant as compared to the end-to-end baseline (0.704 vs. 0.677, $\Delta = +0.028$, $p = 0.006$, $r = 0.473$, medium effect). TimesFM 2.5 did not differ from the end-to-end baseline (0.684 vs. 0.677, $p = 0.608$, $r = 0.084$, negligible effect).

The SSL advantage was consistent across operating points: for a 20% false-positive rate, the SSL model detected 41.0% of acidotic neonates compared to 38.3% (end-to-end) and 36.0% (TimesFM). However, these differences were not statistically significant. AUPRC was comparable across all conditions ($\sim$0.13–0.14), showing the difficulty of achieving high precision at 4.8% prevalence. Fig. 2 shows the mean ROC and precision-recall curves across 50 folds. The SSL-pretrained model ROC shows separation the baseline across the full operating range. Table 2 shows the results in the external CTU-UHB dataset, where the SSL-pretrained model was best.

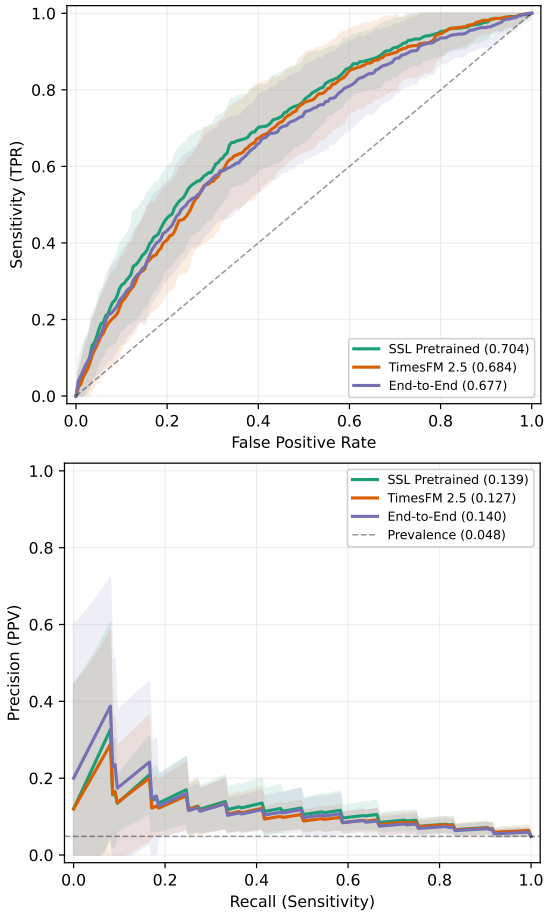

Figure 2: Mean ROC (top) and precision-recall (bottom) curves with ±1 SD shading across 50 cross-validation folds.

Table 2: External validation on CTU-UHB

Model	AUROC	AUPRC	Sens@20%
SSL Pretrained	**0.743**	**0.151**	**0.448**
TimesFM 2.5	0.600	0.087	0.310
End-to-End	0.637	0.112	0.310

4. Discussion

Our results demonstrate that SSL pretraining on unlabeled FHR produces representations that significantly improve acidosis detection, with a medium effect size, compared to an end-to-end training approach. This improvement is achieved with only $\sim$ 1M trainable parameters in the classifier, down from $\sim$ 12M for the end-to-end model.

TimesFM 2.5 showed no advantage over random initialization, despite its 200M parameters and pretraining on billions of time points across diverse contexts, including financial, meteorological, and energy time series. This is an informative negative result, suggesting that the temporal patterns in other contexts do not necessarily transfer to

fetal cardiac physiology. The domain gap appears too wide for zero-shot transfer learning. This suggests that domain-specific pretraining is preferable to general-purpose foundation models in clinical applications.

The external validation on CTU-UHB (AUROC=0.743) provides evidence that the representations learned with SSL capture generalizable physiological features rather than dataset-specific artifacts. We performed this evaluation under triple distribution shift: different acquisition modality (Doppler vs. transabdominal electrodes), different label definition (pH and base deficit composite vs. lactate), and different population (Czech Republic vs. Zambia, Ghana, and India). Our model maintained above chance discrimination under these conditions and achieved 44.8% sensitivity at a 20% false-positive rate. These results suggest that the encoder learned transferable FHR morphology during SSL pretraining. Moreover, the advantage of SSL pretraining increased on external data when compared to end-to-end training ($\Delta_{\text{AUROC}} = 0.106$), and the performance of the TimesFM model degraded more severely for external testing. In the literature, published models trained and evaluated within the Doppler domain on CTU-UHB, and that used a more lenient acidosis threshold (pH < 7.05, 40 cases), have reported AUROCs of 0.68–0.74 and $\sim$50% sensitivity at a 20% false-positive rate [4, 9, 10]. Despite the cross-domain evaluation, our model achieved an AUROC at the top of the range reported by other same-domain studies, validating our approach.

Our study also has some limitations. The labeled cohort contains only 59 acidosis cases, which results in approximately 12 positives per test fold, limiting statistical power for operating-point-specific sensitivity and AUPRC. The high fold-to-fold variability ($\sigma = 0.054$) also reflects this limitation. Finally, the external validation defined acidosis on different biomarkers, which shows the difficulty of cross-database and cross-setting comparisons; alignment of label definitions across datasets is an open challenge.

5. Conclusions

Self-supervised pretraining on unlabeled FHR significantly improved fetal acidosis classification when compared to end-to-end supervised training alone. A general-purpose foundation model was inferior to a Domain-specific SSL approach, indicating that domain-specific approaches are needed to address the label scarcity problem in fetal monitoring. Future work will evaluate different window sizes, explore minority-class augmentation in latent space, examine a wider array of foundation models, and evaluate on multi-center prospective cohorts.

Acknowledgments

The collection and analysis of data in this study were funded by the Gates Foundation (Award OPP1191684 and INV003266). JVC, RZ, and GC are partially funded by the NICHD (R01HD110480) and Google.org (AI for the Global Goals Impact Challenge Award). GDC and RS are co-inventors of the FECG machine and software used in this study, which is licensed to Mindchild Medical Inc., in which both authors have ownership interests. The terms of this arrangement have been approved by Emory University in accordance with its conflict-of-interest policies.

References

[1] Low JA, et al. Intrapartum fetal asphyxia: definition, diagnosis, and classification. Am J Obstet Gynecol 1997; 176(5):957–959.

[2] Jackson M, Holmgren CM, Esplin MS, Henry E, Varner MW. Frequency of fetal heart rate categories and short-term neonatal outcome. Obstet Gynecol 2011;118(4):803–808.

[3] Vargas-Calixto J, Wu YW, Kuzniewicz M, Cornet MC, Forquer H, Gerstley L, Hamilton E, Warrick PA, Kearney RE. Accounting for nulliparity in the prediction of hypoxic-ischemic encephalopathy using cardiotocography. In 2023 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI). IEEE, 2023; 1–4.

[4] McCoy JA, Levine LD, Wan G, Chivers C, Teel J, La Cava WG. Intrapartum electronic fetal heart rate monitoring to predict acidemia at birth with the use of deep learning. Am J Obstet Gynecol 2025;232(1):116–e1.

[5] Dong J, Wu H, Zhang H, Zhang L, Wang J, Long M. SimMTM: A simple pre-training framework for masked time-series modeling. Advances in Neural Information Processing Systems 2023;36:29996–30025.

[6] Das A, Kong W, Leber A, Mathews R, Mozer M, Rose S, Yoon J. A decoder-only foundation model for time-series forecasting. In Proceedings of the International Conference on Machine Learning (ICML). 2024; .

[7] Adu-Amankwah A, Bellad MB, Benson AM, Beyuo TK, Bhandankar M, Charanthimath U, Chisembele M, Cole SR, Dhaded SM, Enweronu-Laryea C, et al. Limiting adverse birth outcomes in resource-limited settings (LABOR): protocol of a prospective intrapartum cohort study. Gates Open Research 2022;6:115.

[8] Chudáček V, Spilka J, Burša M, Janků P, Hruban L, Huptych M, Lhotská L. Open access intrapartum CTG database. BMC pregnancy and childbirth 2014;14(1):16.

[9] Ogasawara J, Ikenoue S, Yamamoto H, Sato M, Kasuga Y, Mitsukura Y, Ikegaya Y, Yasui M, Tanaka M, Ochiai D. Deep neural network-based classification of cardiotocograms outperformed conventional algorithms. Scientific reports 2021;11(1):13367.

[10] M'Barek IB, Jauvion G, Merrer J, Koskas M, Sibony O, Ceccaldi PF, Le Pennec E, Stirnemann J. DeepCTG® 2.0: Development and validation of a deep learning model to detect neonatal acidemia from cardiotocography during labor. Computers in Biology and Medicine 2025;184:109448.

Address for correspondence:

Johann Vargas-Calixto (cavarga@emory.edu)
101 Woodruff Circle, Suite 4100, Atlanta, GA 30322