Real-Time Prediction of Impending Severe Fetal Distress Using Deep Learning on Fetal Heart Rate Monitoring

Donghyeon Seo*[1], Hyungyu Park*[2], Ji Hoi Kim[2], Kwangsoo Kim[3,4,5], Sohee Kim[2], Chan-Wook Park[2], Joong Shin Park[2], Un Yung Choi[2], Yu Gyeom Yoon[6], Tae Kwan Lee[6], Hyung-Chul Lee[7,8], Garam Lee[9†], and Seung Mi Lee[2,6,10†]

[1] Interdisciplinary Program of Medical Informatics, College of Medicine, Seoul National University
[2] Department of Obstetrics and Gynecology, Seoul National University
[3] Department of Transdisciplinary Medicine, Seoul National University Hospital
[4] Center for Data Science, Healthcare AI Research Institute, Seoul National University Hospital
[5] Department of Medicine, College of Medicine, Seoul National University
[6] Interdisciplinary Program in Artificial Intelligence, Seoul National University
[7] Department of Anesthesiology and Pain Medicine, Seoul National University
[8] Department of Anesthesiology and Pain Medicine, Seoul National University Hospital
[9] Biomedical Research Institute, Seoul National University Hospital
[10] Department of Obstetrics and Gynecology & Healthcare AI Research Institute, Seoul National University Hospital
lbsm@snu.ac.kr

Abstract. Antepartum fetal heart rate (FHR) monitoring is crucial for assessing fetal well-being. Severe Fetal Distress, such as "Category III" FHR status, indicates potential fetal acidemia and demands timely clinical decisions. Early prediction of severe fetal distress before onset remains a critical unmet need, offering an opportunity for preventative interventions. This study aimed to develop a deep learning model capable of predicting impending severe fetal distress 5 minutes before onset. The dataset was derived from multicenter cohort (N=18,302 patients from 12 hospitals) consisting of a development (N=16,460) and an internal validation (N=1,842) cohort, along with an independent external validation cohort (N=510 patients from SNUH, yielding 183,016 continuous monitoring windows). A deep learning model was developed to predict severe fetal distress 5 minutes before onset using FHR and uterine contraction (UC) signals. The severe fetal distress occurred in 2.68% (441/16,460) of the development cohort and 2.12% (39/1,842) of the internal validation cohort. In the external validation cohort, 46 patients (9.0%, 46/510) experienced severe fetal distress in 52 abnormal windows. The developed

[10] *These authors contributed equally to this work.
† Corresponding authors. Correspondence to: Seung Mi Lee, lbsm@snu.ac.kr.

model predicted the severe fetal distress 5 minutes before onset, achieving an AUROC of 0.855 (95% CI, 0.799-0.902) in the internal validation cohort. The model also demonstrated excellent performance in the external validation cohort, simulating continuous real-world monitoring, with an AUROC of 0.917 (95% CI, 0.882–0.946). A deep learning model demonstrated high accuracy in predicting impending severe fetal distress in both internal and external cohorts. This tool shows promise as an early warning system to support clinical decision-making and facilitate preventative interventions to mitigate fetal compromise.

Keywords: Deep Learning · Fetal Heart Rate Monitoring · Early Prediction · Severe Fetal Distress.

1 Introduction

Antepartum fetal heart rate (FHR) monitoring is a fundamental component of assessing fetal well-being in modern obstetrics. Specific fetal heart rate patterns are highly indicative of potential fetal acidemia and hypoxia, necessitating rapid clinical decision-making. In particular, high-risk patterns, such as the Category III FHR status [1, 2], require immediate evaluation and prompt delivery. However, the current clinical paradigm relies on reacting to these patterns after they occur. If these high-risk states could be predicted before their onset, it would provide a crucial window of opportunity for preventative interventions, significantly improving neonatal outcomes [3].

Despite this clinical need, predicting Category III status remains challenging. True Category III events are relatively rare, and in real-world practice, there is a lack of objective tools that can provide early warnings just before a definitive Category III pattern emerges [4]. Furthermore, manual interpretation of FHR monitoring is burdened by high inter- and intra-observer variability [5], and existing automated systems often suffer from inappropriate warning timings—either alerting too late or generating excessive false alarms [6].

To address these gaps, this study aimed to define severe fetal distress—a composite of clinically significant precursor patterns—and develop a deep learning model capable of predicting its onset 5 minutes in advance. The primary contributions of this study are threefold: (1) the utilization of a large-scale, multicenter development cohort with rigorous internal and external validation to ensure generalizability; (2) the proposal of a 5-minute prediction window, which provides an operational and actionable timeframe for clinicians; and (3) the integration of both FHR and uterine contraction (UC) signals to achieve superior predictive performance.

2 Methods

2.1 Study Design and Cohorts

This was a retrospective, multicenter study. The dataset was partitioned into three distinct cohorts: a development cohort (N=16,460 patients; 16,460 win-

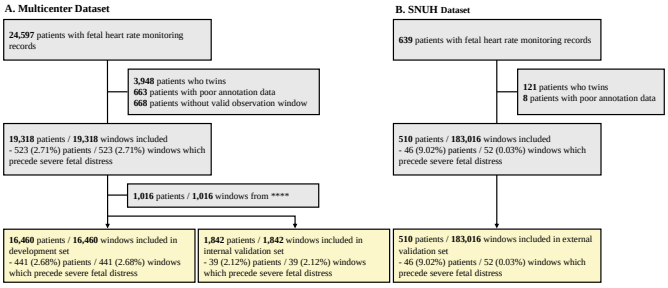

Fig. 1. Study Cohort Flow Chart

dows), an internal validation cohort (N=1,842 patients; 1,842 windows), and an independent external validation cohort (N=510 patients; 183,016 windows) (Fig. 1). The development and internal validation datasets were sourced from the NIA (AI-Hub) fetal monitoring dataset, while the external validation dataset was obtained from Seoul National University Hospital (SNUH). This retrospective study was approved by the Institutional Review Board of Seoul National University Hospital (IRB No. H-2212-137-1389).

The study included FHR monitoring data. We excluded data from multiple gestations (twins), records with poor annotation quality confirmed by clinicians, and records from which the defined observation windows could not be extracted.

We defined the target outcome as "severe fetal distress". A 5-minute window was labeled as severe fetal distress if it exhibited at least one of the following characteristics: prolonged deceleration, repetitive late deceleration, repetitive variable deceleration, bradycardia, or absent variability (Table 1). All datasets were annotated by an obstetrics and gynecology resident to ensure clinical reliability.

2.2 Data Preprocessing and Windowing Strategy

The primary task was formulated to predict the severe fetal distress 5 minutes before onset (Fig. 2.A). For the development and internal validation sets, the multicenter dataset was randomly partitioned using a stratified 9:1 split. This ensured that the ratio of abnormal to normal cases remained consistent across both

Table 1. Label Characteristics

Class	FHR & UC Pattern	Development & Internal Validation (N=18,302)	External Validation (N=183,016)
1	Severe Fetal Distress (Abnormal)	480	52
	- Prolonged Deceleration	35	43
	- Repetitive Late Deceleration	10	9
	- Repetitive Variable Deceleration	449	1
	- Absent Variability	0	0
	- Bradycardia	2	0
0	Normal	17,822	182,964

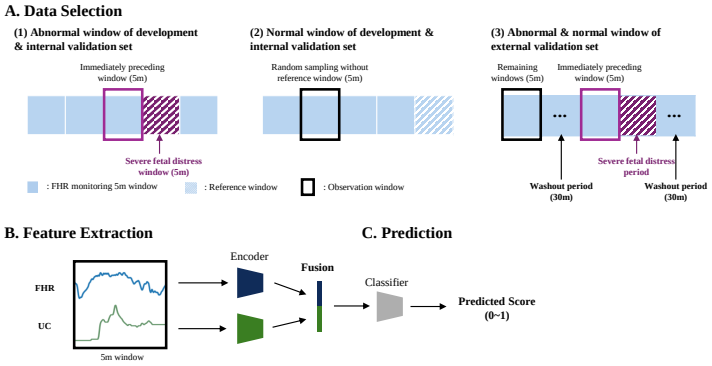

Fig. 2. Overall Framework of Model Development

the training (development) and internal validation cohorts. Data were segmented into 5-minute windows. For abnormal cases (severe fetal distress), the 5-minute window immediately preceding the event was extracted as the observation window. For normal controls, one 5-minute window was randomly selected from the record, excluding the final 5 minutes. This 1-patient-to-1-window matching prevented patient-specific bias during model training. For the external validation set, the 5-minute segment immediately preceding the onset of severe fetal distress was extracted as abnormal cases. For normal controls, the remaining continuous record was segmented into 5-minute windows, excluding a washout period (30 minutes prior to the abnormal observation window and 30 minutes after the offset of severe fetal distress). Data normalization was performed by calculating the mean and standard deviation for each sample and channel independently (z-score normalization).

2.3 Model Architecture and Training

We designed a multi-stream 1D ResNet architecture [11, 12] to effectively process the heterogeneous input signals (Fig. 2.B). Instead of stacking inputs into a single channel, each input variable (FHR, UC) was processed through its own dedicated ResNet encoder [13]. Each ResNet encoder consisted of three stages with one residual block per stage and channel widths of 8, 16, and 32, respectively. This parallel structure allowed the model to extract distinct feature vectors optimized for each signal type independently. These extracted feature vectors were then concatenated to form a comprehensive representation of the fetal status. Finally, the fused vector was passed through a dense layer classifier to generate a final probability score (0 to 1), indicating the risk of an impending severe fetal distress(Fig. 2.C).

Given the significant prevalence of normal control samples in our dataset, standard training may lead the model to be biased toward the majority class. To counteract this, we adopted Focal Loss (α_t=0.85, γ=2.0) [14], which dynamically

scales the Cross Entropy loss based on the prediction confidence. This ensures that the optimization process is driven by hard-to-predict event cases rather than the overwhelming number of simple negative samples. The Focal Loss (FL) is defined as:

$$FL(p_t) = -\alpha_t(1-p_t)^\gamma \log(p_t) \tag{1}$$

where p_t is the model's estimated probability for the target class, γ is the focusing parameter that controls the rate at which easy examples are down-weighted, and α_t is a balancing factor for the rare class. The model was trained using the Adam optimizer with a weight decay of 1e-4. The initial learning rate was set to 0.01, modulated by a ReduceLROnPlateau scheduler (factor=0.5, patience=5). Training was conducted with a batch size of 256 for up to 100 epochs, with an early stopping patience of 15 epochs.

The model was implemented in Python 3.9.19 and PyTorch 2.5.1, with training and evaluation performed on two NVIDIA RTX A6000 GPUs.

2.4 Evaluation Metrics and Statistical Analysis

The primary evaluation metric was the Area Under the Receiver Operating Characteristic curve (AUROC). The optimal classification threshold was determined using the Youden Index from the internal validation cohort and was subsequently applied directly to the external validation cohort. Secondary metrics included accuracy, sensitivity, specificity. The 95% confidence intervals (CIs) for all metrics were calculated using bootstrapping with 5,000 iterations.

3 Results

3.1 Event Rate and Outcome Incidence

The general characteristics of the development and internal validation cohorts are summarized in Table 2. In the development cohort, severe fetal distress was identified in 441 windows (2.68%). In the internal validation cohort, they occurred in 39 windows (2.12%). In the external validation cohort, 52 abnormal windows were detected out of 183,016 total windows. The detailed subcategories of these events are presented in Table 1.

3.2 Model Performance and Analysis

The deep learning model demonstrated robust discriminative performance (Fig. 3). In the internal validation cohort, the model achieved an AUROC of 0.855 (95% CI, 0.799–0.902) for predicting impending severe fetal distress 5 minutes before onset. When applied to the external validation cohort—the model maintained excellent performance with an AUROC of 0.917 (95% CI, 0.882–0.946). At the Youden index–derived threshold, the model achieved a sensitivity of 0.846 and specificity of 0.774 in the internal validation cohort; when the same threshold was

Table 2. Baseline Characteristics of Development and Internal Validation Sets

Variable	Development (N=16,460)	Internal Validation (N=1,842)	P-value
Gestational age (w)	36.84 ± 2.70	36.85 ± 2.72	0.8593
Birth weight (g)	2,851.54 ± 677.01	2,854.32 ± 665.49	0.9631
APGAR 1min	7.73 ± 1.42	7.76 ± 1.33	0.9471
APGAR 5min	8.92 ± 1.06	8.94 ± 0.96	0.587
UA pH	7.30 ± 0.09	7.30 ± 0.08	0.0589
UA BE	-3.74 ± 3.32	-3.63 ± 3.44	0.23
UA pO2	35.75 ± 29.22	33.97 ± 25.36	0.5498
UA pCO2	48.44 ± 11.61	48.07 ± 11.51	0.2096
Number of 5min windows	4.32 ± 1.13	4.34 ± 1.13	0.2579
Delivery	n (%)	n (%)	0.6578
- Vaginal Delivery	6,939 (42.2)	790 (42.9)	
- Cesarean Section	9,496 (57.7)	1,048 (56.9)	
- VBAC	25 (0.2)	4 (0.2)	
Sex	n (%)	n (%)	0.1122
- Male	8,663 (52.6)	933 (50.7)	
- Female	7,797 (47.4)	909 (49.3)	
NICU admission	n (%)	n (%)	0.9578
- Yes	6,887 (41.8)	771 (41.9)	
- No	9,573 (58.2)	1,071 (58.1)	
Anomaly	n (%)	n (%)	0.5218
- Yes	2,887 (17.5)	334 (18.1)	
- No	13,573 (82.5)	1,508 (81.9)	
Institution	Mode: 'snu' (n=4,112)	Mode: 'snu' (n=458)	0.9779

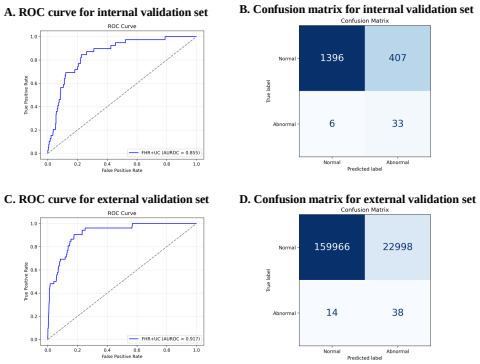

Fig. 3. Model Performance

Table 3. Model Performance Comparison across Different Input Configurations

Feature	AUROC (95% CI)
FHR	0.811 [0.746, 0.865]
UC	0.786 [0.709, 0.853]
FHR+UC	0.855 [0.799, 0.902]

applied to the external validation cohort, sensitivity was 0.731 with specificity of 0.874 (Table 4).

Through a comparative analysis of input configurations (Table 3) in the internal validation cohort, the combination of FHR, UC yielded the highest discrimination performance (AUROC 0.855), outperforming models using FHR alone (AUROC 0.811) or UC alone (AUROC 0.786). This demonstrates that UC provides essential contextual information for interpreting fetal status [7]. Consequently, this dual-channel configuration (FHR + UC) was selected as the final input and was used for subsequent validation and analysis.

We evaluated the model's predictive performance at extended lead times (Table 4). In the external validation set, the model achieved AUROCs of 0.864, 0.851, and 0.870 for predicting events 10, 15, and 20 minutes prior to onset, respectively, indicating the model's capacity to capture early morphological changes well before the critical event.

Figure 4 illustrates the temporal trajectory of the predicted risk scores for representative samples from internal validation set. The risk scores of abnormal case samples exhibited a progressive increase starting approximately 20 minutes prior to onset of the event, crossing the decision threshold well in advance of the clinical onset, which confirms the model's early warning capability.

To further interpret the model's decision-making process, we visualized the latent space distribution using UMAP and sampled representative normal and abnormal samples from both dense clusters and outlier regions for Grad-CAM analysis (Fig. 5). These qualitative visualizations suggested modality-specific saliency patterns, with focal FHR activations and more contextual UC activations, and were intended for interpretation rather than quantitative validation.

Table 4. Performance comparison by time window(min) for Internal and External Validation sets. Values are presented as Mean [95% CI].

Window	AUROC	Accuracy	Sensitivity	Specificity
Internal Validation Set				
$[-5, 0]$	0.855 [0.799, 0.902]	0.776 [0.757, 0.795]	0.846 [0.718, 0.949]	0.774 [0.755, 0.794]
$[-10, -5]$	0.685 [0.578, 0.787]	0.767 [0.749, 0.787]	0.391 [0.217, 0.609]	0.772 [0.753, 0.792]
$[-15, -10]$	0.717 [0.583, 0.846]	0.770 [0.751, 0.790]	0.538 [0.308, 0.769]	0.772 [0.753, 0.791]
$[-20, -15]$	0.675 [0.471, 0.862]	0.771 [0.752, 0.791]	0.556 [0.222, 0.889]	0.772 [0.753, 0.792]
External Validation Set				
$[-5, 0]$	0.917 [0.882, 0.946]	0.874 [0.873, 0.876]	0.731 [0.596, 0.846]	0.874 [0.873, 0.876]
$[-10, -5]$	0.864 [0.814, 0.909]	0.874 [0.873, 0.876]	0.615 [0.481, 0.750]	0.874 [0.873, 0.876]
$[-15, -10]$	0.851 [0.804, 0.892]	0.874 [0.873, 0.876]	0.538 [0.404, 0.673]	0.874 [0.873, 0.876]
$[-20, -15]$	0.870 [0.823, 0.909]	0.874 [0.873, 0.876]	0.686 [0.549, 0.804]	0.874 [0.873, 0.876]

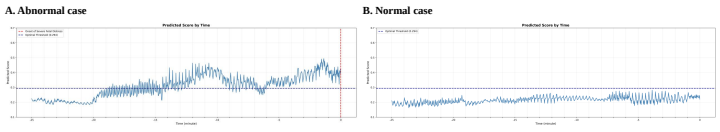

Fig. 4. Temporal Profiles of Predicted Risk Scores for Representative Cases

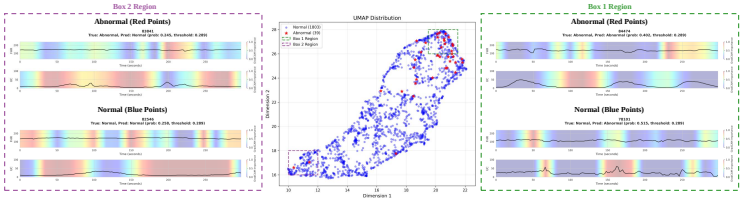

Fig. 5. UMAP Data Distribution and Model Attention Visualization by Grad-CAM

4 Discussion

4.1 Principal Findings

In this multicenter retrospective study, we developed and validated a deep learning model to predict impending severe fetal distress 5 minutes before its onset. The model demonstrated high discriminative accuracy, achieving an AUROC of 0.855 in the internal validation cohort and 0.917 in the external validation cohort. Notably, the model maintained predictive capability even at extended lead times of 10 to 20 minutes. Ultimately, 5-minute warning enables timely preventative interventions—such as maternal repositioning or oxygen administration—to effectively mitigate fetal compromise.

4.2 Comparison with Prior Work

Unlike previous studies focusing on detecting or classifying already manifested fetal heart rate patterns [8, 9], our work shifts the paradigm toward proactive prediction. This 5-minute early warning enables a preventive rather than reactive approach. Furthermore, robust external validation results demonstrate the model's strong generalizability across different cohorts [10].

4.3 Limitations and Future Work

This study has several limitations. First, the labeling of impending severe fetal distress was performed by a single resident, which may introduce subjective bias, although the criteria were strictly defined. Second, while a 5-minute observation window was chosen for clinical immediacy, it remains to be determined if this is the optimal length for capturing all relevant physiological trends. Finally, as a retrospective study, the true impact of real-time alerts on clinical workflow and patient outcomes remains unmeasured.

Future research should prioritize prospective clinical trials to evaluate real-time deployment, focusing on workflow impact and alarm fatigue. Additionally, incorporating FHR morphological patterns (such as accelerations, decelerations, and variability) as structured inputs and integrating explainability modules will be essential to enhance model robustness and foster clinical trust.

Acknowledgments. This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) [NO.RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University)]

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Macones, G.A., Hankins, G.D.V., Spong, C.Y., Hauth, J.C., Moore, T.: The 2008 National Institute of Child Health and Human Development workshop report on electronic fetal monitoring: update on definitions, interpretation, and research guidelines. Obstetrics & Gynecology **112**(3), 661–666 (2008). https://doi.org/10.1097/AOG.0b013e3181841395
2. American College of Obstetricians and Gynecologists: ACOG Practice Bulletin No. 106: Intrapartum fetal heart rate monitoring: nomenclature, interpretation, and general management principles. Obstetrics & Gynecology **114**(1), 192–202 (2009). https://doi.org/10.1097/AOG.0b013e3181aef106
3. Sabljić, J., Čogelja, K., Runjić, E., Markoski, B., Barbača, M., Modrić, T., Bačić, B.: Outcome of Preterm Neonates > 32 Weeks Gestation in Relation to Three-Tiered Fetal Heart Rate Categorization. Medicina **61**(7), 1171 (2025). https://doi.org/10.3390/medicina61071171
4. Vimalesan, E., Khursheed, R., Maldar, A., Dalal, A., Vyshnavi, K.: Prediction of neonatal acidemia at birth with total deceleration area on intrapartum cardiotocogram. Gynecology and Obstetrics Clinical Medicine **5**(1), e000078 (2025). https://doi.org/10.1136/gocm-2024-000078
5. Varisco, G., Steyde, G., Peri, E., Hoogendoorn, I., Signorini, M.G., van Laar, J.O.E.H., Mischi, M., van der Hout-van der Jagt, M.B.: Predicting Intrapartum Acidemia: A Review of Approaches Based on Fetal Heart Rate. Bioengineering **13**(2), 146 (2026)
6. McCoy, J.A., Levine, L.D., Wan, G., Chivers, C., Teel, J., La Cava, W.G.: Intrapartum electronic fetal heart rate monitoring to predict acidemia at birth with the use of deep learning. American Journal of Obstetrics and Gynecology **232**(1), 116.e1–116.e9 (2025)
7. Pardasani, R., Vitullo, R., Harris, S., Yapici, H.O., Beard, J.: Development of a novel artificial intelligence algorithm for interpreting fetal heart rate and uterine activity data in cardiotocography. Frontiers in Digital Health **7**, 1638424 (2025)
8. Ben M'Barek, I., Jauvion, G., Merrer, J., Koskas, M., Sibony, O., Ceccaldi, P.F., 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 **184**, 109448 (2025)
9. McCoy, J.A., La Cava, W.G.: Deep learning to predict fetal acidemia: a response. American Journal of Obstetrics and Gynecology **232**(2), e46 (2025)
10. Lin, Z., Liu, X., Wang, N., Li, R., Liu, Q., Ma, J., Wang, L., Wang, Y., Hong, S.: Deep learning with information fusion and model interpretation for long-term prenatal fetal heart rate data. npj Women's Health **2**(1), 31 (2024). https://doi.org/10.1038/s44294-024-00033-z
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778. IEEE (2016). https://doi.org/10.1109/CVPR.2016.90

12. Hannun, A.Y., Rajpurkar, P., Haghpanahi, M., Tison, G.H., Bourn, C., Turakhia, M.P., Ng, A.Y.: Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nature Medicine **25**(1), 65–69 (2019). https://doi.org/10.1038/s41591-018-0268-3
13. Baltrušaitis, T., Ahuja, C., Morency, L.P.: Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence **41**(2), 423–443 (2018). https://doi.org/10.1109/TPAMI.2018.2798607
14. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 2980–2988. IEEE, Venice (2017). https://doi.org/10.1109/ICCV.2017.324