Development and Evaluation of Deep Learning Models for Cardiotocography Interpretation

Nicole Chiou[†]
Stanford University
nicchiou@stanford.edu

Nichole Young-Lin
Google Research
nyounglin@google.com

Christopher Kelly
Google Research
kellych@google.com

Julie Cattiau
Google Research
juliecattiau@google.com

Tiya Tiyasirichokchai
Google Research
tiyat@google.com

Abdoulaye Diack
Google Research
diack@google.com

Sanmi Koyejo
Google DeepMind
sanmik@google.com

Katherine Heller
Google Research
kheller@google.com

Mercy Asiedu[*]
Google Research
masiedu@google.com

Abstract

The inherent variability in the visual interpretation of cardiotocograms (CTGs) by obstetric clinical experts, both intra- and inter-observer, presents a substantial challenge in obstetric care. In response, we investigate automated CTG interpretation as a potential solution to enhance the early detection of fetal hypoxia during labor, which has the potential to reduce unnecessary operative interventions and improve overall maternal and neonatal care. This study employs deep learning techniques to reduce the subjectivity associated with visual CTG interpretation. Our results demonstrate that using objective umbilical cord blood pH outcome measurements, rather than clinician-defined Apgar scores, yields more consistent and robust model performance. Additionally, through a series of ablation studies, we explore the impact of temporal distribution shifts on the performance of these deep learning models. We examine tradeoffs between performance and fairness, specifically evaluating performance across demographic and clinical subgroups. Finally, we discuss the practical implications of our findings for the real-world deployment of such systems, emphasizing their potential utility in medical settings with limited resources.

1 Introduction

Intrapartum cardiotocography (CTG) is a screening technique used to monitor fetal well-being by recording the fetal heart rate (FHR) along with the maternal uterine contractions (UC) during labor. Although CTG is routinely used in medical practice, subjectivity (Bernardes et al., 1997; Palomäki et al., 2006) and intra-observer variability (Schiermeier et al., 2011) hinder the effectiveness of visual CTG interpretation. These issues are exacerbated in low-resource facilities where access to skilled interpreters is limited (Blencowe et al., 2016; Lawn et al., 2016). While machine learning approaches based on tabulated features offer promise, they often discard valuable temporal and contextual information through feature extraction (Ayres-De-Campos & Bernardes, 2004; Hoodbhoy et al., 2019; Pradhan et al., 2021; Spilka et al., 2013; 2014).

Deep learning emerges as a potential solution by analyzing the physiological time series data from CTG recordings. However, existing studies rely on proxy labels like umbilical artery blood pH or the 1-minute Apgar score, which introduce their own limitations (Asfaw et al., 2023; Daydulo et al.,

[*] Corresponding author

[†] Work done as a Student Researcher at Google

2022; Ogasawara et al., 2021; Park et al., 2022; Mendis et al., 2023; Petrozziello et al., 2019; Spairani et al., 2022; Zhou et al., 2023). Although pH serves as an objective measure of fetal hypoxia in high-resource settings, the Apgar score is a subjective clinical assessment. Apgar scores are the primary delivery outcome descriptor in low-resource settings due to their simplicity, cost-effectiveness, and the potential financial burden of umbilical blood analysis (Allanson et al., 2017; Thorp & Rushing, 1999). Additionally, continuous CTG monitoring, common in high-resource facilities, may not be feasible in low- and middle-income countries (LMICs) due to system limitations or resource constraints (Enabudoso, 2021; Mugyenyi et al., 2017; Ryu et al., 2021). Thus, to enable applications in low-resource clinical use cases, machine learning solutions must aim to accurately detect fetal compromise during intermittent periods before delivery (Enabudoso, 2021; Mendis et al., 2023).

In this work, we highlight the feasibility of using deep learning methods to reduce the subjectivity of predicting fetal hypoxia from visual CTG interpretation. We conduct ablation studies to analyze the effect of (a) the choice of objective (pH) vs subjective (Apgar) ground truth labels and (b) the evaluation of simulated low-resource environment signals on predictive performance. We propose data augmentation and statistical evaluation methods to overcome challenges with this limited dataset. Finally, we discuss the implications of training deep learning models for deployment in low-resource settings from a global health perspective.

2 Methods

2.1 Dataset description

The CTU-UHB Intrapartum Cardiotocography Database is an open-source collection of 552 CTG recordings (Chudáček et al., 2014; Goldberger et al., 2000 (June 13). Each CTG records the fetal heart rate (FHR) and corresponding uterine contractions (UC) for up to 90 minutes before delivery. The data are associated with fetal outcomes, along with fetal and maternal metadata. We defined three outcome label categories:

$$Y_i = \begin{cases} \mathbb{1}\{P_i < 7.20\} & \text{, pH} \\ \mathbb{1}\{A_i < 7\} & \text{, Apgar} \\ \mathbb{1}\{(P_i < 7.20) \cup (A_i < 7)\} & \text{, LOR} \end{cases}$$

where $\mathbb{1}\{\cdot\}$ is the binary indicator function, P_i is the umbilical arterial cord blood pH, A_i is the 1-minute Apgar score, and Y_i is the assigned ground truth label (abnormal = 1, normal = 0) for the ith CTG recording. The pH classification task and the LOR classification task, defined as the logical inclusive "OR" of the abnormal pH and Apgar criteria, use a cut-off threshold of 7.20. This threshold was chosen for relevance to clinical cut-offs, appropriate class balance, and to enable comparison to prior work (Ogasawara et al., 2021).

2.2 Preprocessing and data splitting

The preprocessing pipeline for CTG signals included removing repeated zero signals at the beginning and end of the recorded signal, data quality assessment, imputation of missing values, signal smoothing, data augmentation, cropping to 30-minute signal length, and downsampling. In addition to using the last 30 minutes of the signal for training, we also investigated pre-training on cropped signals excluding the last 30 minutes and then fine-tuning model parameters on the last 30 minutes. Furthermore, we evaluated trained models on the following varied 30-minute segments: (1) the signal at the 30-60 minute mark and (2) randomly cropped 30-minute signals sampled over the entire recording.

We performed stratified data splitting to ensure similar distributions over the predicted target variable across training, validation, and test splits. We assigned 10% of the record identifiers to a held-out test set before assigning the remaining 90% to 10-fold cross-validation (cv) splits by record identifiers, ensuring a representative distribution of labels in each split.

2.3 Neural network model architecture and training

We utilized the CTG-net neural network model proposed by Ogasawara et al. (2021) for our base model for its ability to learn the relationship between FHR and UC. The CTG-net architecture takes signals of 1800 time points long (30 minutes downsampled at 1 Hz) as input. A high-level depiction of the training pipeline is shown in Figure 1. Further details regarding the architecture can be found in Ogasawara et al. (2021).

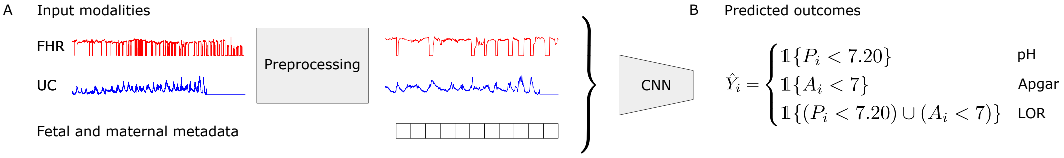
A
Input modalities
FHR
UC
Fetal and maternal metadata
Preprocessing
CNN
B
Predicted outcomes
$\hat{Y}_i = \begin{cases} \mathbb{1}\{P_i < 7.20\} \\ \mathbb{1}\{A_i < 7\} \\ \mathbb{1}\{(P_i < 7.20) \cup (A_i < 7)\} \end{cases}$
pH
Apgar
LOR

Figure 1: (A) Possible input modalities for the deep learning model, which outputs (B) a predicted outcome depending on the classification task.

We also ran the following experiments: (1) training with FHR or UC as input using a 1D CNN model variation and (2) adding metadata features as a vector to the input. Model hyperparameters were optimized separately for two-channel (FHR and UC) versus 1-channel (FHR) input models. Out of 500 random hyperparameter configurations, the best hyperparameters were selected based on the highest validation AUROC averaged over 10 cv folds. See Appendix A for more details.

All neural networks were trained on an NVIDIA V100 GPU in TensorFlow using the Keras API (Abadi et al., 2015; Chollet, 2015). We used Adam as the optimizer, initialized model weights and optimizer state with a fixed random seed, and trained each model for 300 epochs (Kingma & Ba, 2017). We perform channel-specific maximum absolute value scaling on time series and metadata normalization on tabular features before input into the neural network.

2.4 Evaluation

The performance of the neural network models was evaluated using the area under the receiver operating characteristic curve (AUROC). We also reported the sensitivity at a fixed specificity threshold of 90% for comparison with clinical performance. A two-tailed Welch's t-test was used to compare the average AUROC computed over bootstrapped samples for the various approaches. We also evaluated performance disparities across subgroups for both pH and Apgar prediction tasks. Subgroup variables included demographic and clinical attributes as well as signal quality descriptors for each of the FHR and UC channels. A comprehensive description of the cut-off thresholds and formulas used to define binary subgroup variables is found in Appendix B.

3 Results

Performance by prediction outcome. Our baseline method for predicting LOR that takes both FHR and UC signals achieves comparable AUROC (0.68 ± 0.07) as prior work (0.68 ± 0.03) (Ogasawara et al., 2021). We achieved a lower (0.27 ± 0.18) sensitivity at 90% specificity than clinician performance (0.45, 95% CI: 0.23-0.68), however, this was not significant (Singh et al., 2022). Baseline Apgar prediction was slightly higher (0.69 ± 0.12) and baseline pH was slightly lower (0.62 ± 0.09) compared to baseline LOR, but differences were not statistically significant. A summary of model performance can be found in Table 1.

Table 1: Average model performance, trained and evaluated on the last 30 minutes of CTG (standard error shown in parentheses).

Inputs	AUROC		Sensitivity at 90% Specificity	
	pH	Apgar	pH	Apgar
UC	0.55 (0.09)	0.62 (0.11)	0.13 (0.09)	0.00 (0.00)
FHR	0.54 (0.09)	0.39 (0.09)	0.17 (0.09)	0.00 (0.00)
FHR + metadata	0.55 (0.08)	0.38 (0.16)	0.06 (0.06)	0.26 (0.20)
FHR + UC	0.62 (0.09)	**0.69 (0.12)**	0.27 (0.11)	**0.32 (0.20)**
FHR + UC + metadata	**0.69 (0.08)**	0.57 (0.11)	**0.44 (0.13)**	0.05 (0.12)

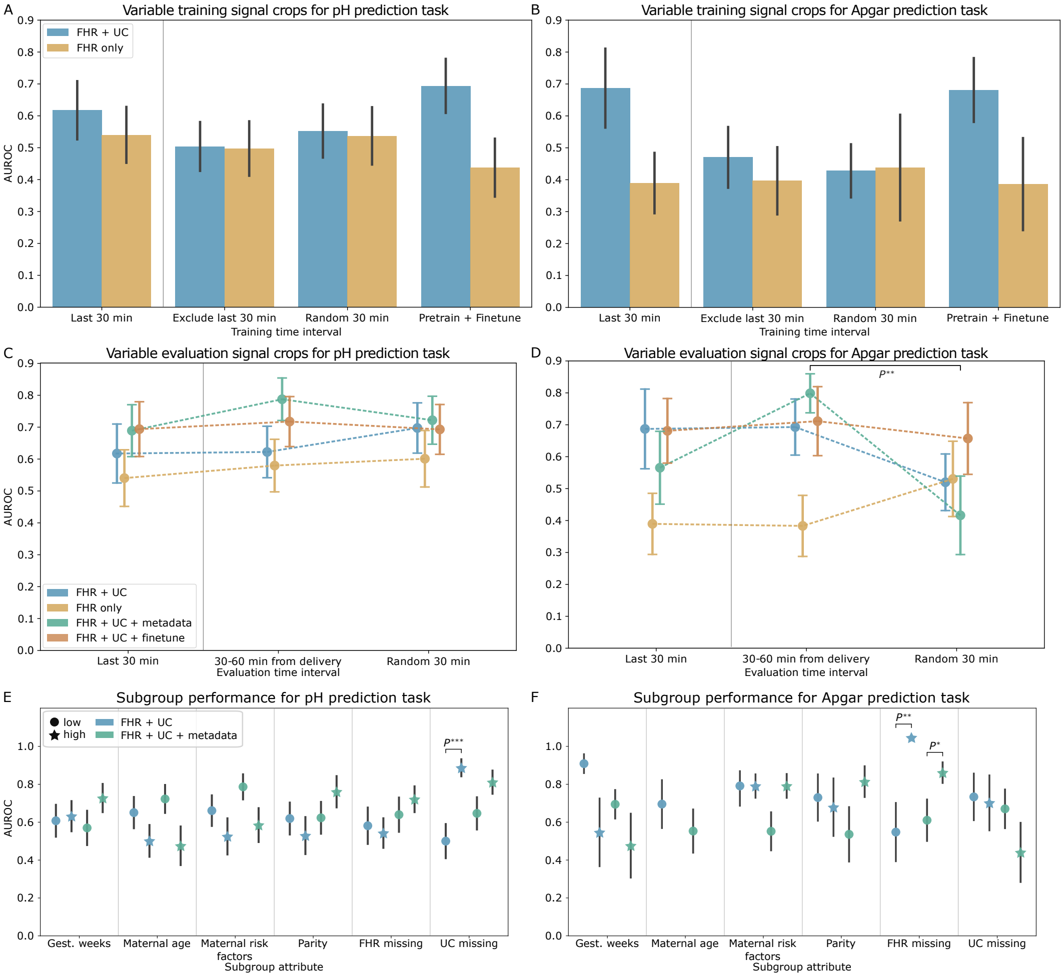

Figure 2: P^{*}, P^{**}, and P^{***} represent a p-value < 0.05, 0.01, and 0.001 respectively. Error bars depict the standard error. (A-D) AUROC for models trained on (A-B) and evaluated on (C-D) CTG signal from different time intervals. Markers to the left of the vertical line indicate the reference model. (E-F) Subgroup AUROC performance. Results for the high maternal age subgroup are omitted for Apgar prediction and further details can be found in Appendix C.

Comparing FHR only, UC only, and FHR + UC. The FHR + UC model achieved the highest AUROC performance for both pH and Apgar classification tasks, followed by UC only and then FHR only models. Excluding either of the channels also resulted in a significant reduction in sensitivity at 90% specificity for both tasks.

Performance with maternal and fetal metadata. Adding metadata to the FHR + UC model increased the performance for the pH prediction task by 0.07 points to 0.69 ± 0.08, though the results were not significant. Adding metadata for the Apgar prediction task degraded the FHR + UC model performance (0.57 ± 0.11).

Evaluation of temporal distribution shifts during training. Figures 2A and 2B show model performance when trained on different time points and evaluated on the last 30 minutes of the held-out test set. No significant differences in AUROC were observed for both pH and Apgar prediction tasks. However, Apgar prediction performance had higher variability across the different trained models compared to pH prediction performance, which was more stable. Pre-training on windowed signals before the last 30 minutes, then fine-tuning on the last 30 minutes achieved the highest AUROC for predicting pH using FHR + UC (0.69 ± 0.09) followed by the model trained on the last 30 minutes alone (0.62 ± 0.09). For the Apgar prediction task, pre-training and fine-tuning the model (0.68 ± 0.10) achieved a similar performance as training on the last 30 minutes alone (0.69 ± 0.12).

Evaluation of temporal distribution shifts during testing. Figures 2C and 2D show model performance when trained on the last 30 minutes and evaluated on different time points of the held-out set. No significant differences in model performance were observed when testing on various signal time points for the pH classification task. Apgar prediction performance generally had higher variability across different time points, demonstrating reduced robustness to temporal distribution shifts. The FHR + UC + metadata model yielded slightly higher performance when evaluated on the 30-60 minutes before delivery, compared to the last 30 minutes and randomly sampled 30 minutes for both pH and Apgar tasks. However, this difference was only significant when comparing performance on the random 30 minutes and the 30-60 minute interval for the Apgar prediction task.

Subgroup evaluation. Figures 2E and 2F show the AUROC performance of the subgroup analysis for pH and Apgar prediction, respectively. We found significant differences in baseline performance between subgroups with low and high UC signal missingness with pH evaluation and for FHR missingness subgroups with Apgar prediction. With metadata, the performance disparities observed with pH prediction were mitigated. However, including metadata increased the AUROC performance disparities for demographic and clinical-related subgroups on this task, although none of these differences were statistically significant.

4 Discussion

Our study demonstrates deep learning's potential for predicting fetal hypoxia from CTG tracings, emphasizing the importance of rigorous evaluations by choice of label, time interval, and subgroup performances. The FHR + UC baseline model achieved comparable performance to prior work and clinical practice (Ogasawara et al., 2021; Singh et al., 2022), highlighting the importance of both FHR and UC channels for accurate pH prediction. We found objective pH labels yielded more consistent performance than subjective Apgar scores, suggesting future work focus on quantitative measures like umbilical cord blood pH (Allanson et al., 2017; Thorp & Rushing, 1999).

Furthermore, training models on the last 30 minutes of signal, the time interval most closely correlated with the delivery outcome, yielded the best performance. Pre-training the model on signal data excluding this critical interval, followed by fine-tuning on the last 30 minutes, improved performance. The robustness of our pH classification model to out-of-distribution time points was demonstrated by consistent performance across randomly sampled intervals within 90 minutes of delivery, simulating the intermittent CTG monitoring setting standard in LMICs. Although in-distribution training and testing on the last 30 minutes was hypothesized to yield the highest performance, evaluating the 30-60 minute interval performed slightly better for some experiments. We speculate that the high proportion of missing signal within the last 30 minutes of recording decreased the amount of discriminative information, thus leading to worse classification performance. While clinical experience supports that the CTG signal recorded closest to delivery corresponds the most with the delivery outcome, real-world complications during the end stages of labor (e.g., sensor displacement or clinical intervention) may contribute to reduced signal quality and have practical performance implications.

Subgroup analyses revealed performance disparities across demographic, clinical, and signal quality subgroups for the baseline model. While incorporating fetal and maternal metadata attributes during training enhanced pH classification performance, we found that demographic and clinical subgroup disparities were exacerbated.

This study had several limitations that constrain the generalizability of our findings. First, we used CTGs from 552 patients at a single hospital in Prague, Czech Republic. To enhance the robustness of our findings, future investigations should involve a larger and more diverse dataset sourced from maternity centers worldwide, encompassing varied clinical contexts, demographics, and outcomes. Secondly, the absence of automated CTG digitization infrastructure in many resource-limited settings necessitates the simulation of intermittent CTG use cases from facilities with digitized recordings (Sbrollini et al., 2017; Parer, 1983). Additionally, our study did not include a comparison of algorithmic performance against clinicians viewing the same dataset, prompting future research to explore different human and algorithmic use combinations. Finally, further work is needed to understand how such prediction algorithms can be optimally integrated into clinical workflows to improve neonatal outcomes.

5 Conclusion

We develop an end-to-end deep learning approach to interpret CTGs and propose a framework to evaluate these models. Our major findings indicate that utilizing objective pH measurements, as opposed to clinician-defined Apgar scores, results in more consistent, robust performance under temporal distribution shifts. This is especially important when transferring models to settings that only have intermittent CTG measurements. The model and evaluation framework we propose can be applied more generally to paired time-series datasets especially where sample size is limited.

Acknowledgments

The authors thank Stephen Pfohl and Chirag Nagpal for valuable discussions and feedback.

References

Martín Abadi, Ashish Agarwal, Paul Barham, et al. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL https://www.tensorflow.org/. Software available from tensorflow.org.

E. R. Allanson, T. Waqar, C. R.H. White, Tunçalp, and J. E. Dickinson. Umbilical lactate as a measure of acidosis and predictor of neonatal risk: a systematic review. *BJOG: An International Journal of Obstetrics & Gynecology*, 124:584–594, 3 2017. ISSN 1471-0528. doi: 10.1111/1471-0528.14306.

Daniel Asfaw, Ivan Jordanov, Lawrence Impey, Ana Namburete, Raymond Lee, and Antoniya Georgieva. Multimodal deep learning for predicting adverse birth outcomes based on early labour data. *Bioengineering*, 10, 6 2023. doi: 10.3390/bioengineering10060730.

Diogo Ayres-De-Campos and João Bernardes. Comparison of fetal heart rate baseline estimation by sisporto® 2.01 and a consensus of clinicians. *European Journal of Obstetrics and Gynecology and Reproductive Biology*, 117:174–178, 12 2004. ISSN 03012115. doi: 10.1016/j.ejogrb.2004.03.013.

J. Bernardes, A. Costa-Pereira, D. Ayres-De-Campos, H. P. Van Geijn, and L. Pereira-Leite. Evaluation of interobserver agreement of cardiotocograms. *International Journal of Gynecology and Obstetrics*, 57:33–37, 1997. ISSN 00207292. doi: 10.1016/S0020-7292(97)02846-4.

Hannah Blencowe, Simon Cousens, Fiorella Bianchi Jassir, Lale Say, Doris Chou, Colin Mathers, Dan Hogan, Suhail Shiekh, Zeshan U. Qureshi, Danzhen You, and Joy E. Lawn. National, regional, and worldwide estimates of stillbirth rates in 2015, with trends from 2000: A systematic analysis. *The Lancet Global Health*, 4:e98–e108, 2 2016. ISSN 2214109X. doi: 10.1016/S2214-109X(15)00275-2.

Patricia A. Cavazos-Rehg, Melissa J. Krauss, Edward L. Spitznagel, et al. Maternal age and risk of labor and delivery complications. *Maternal and Child Health Journal*, 19:1202, 6 2015. ISSN 15736628. doi: 10.1007/S10995-014-1624-7.

François Chollet. Keras. https://github.com/fchollet/keras, 2015.

Václav Chudáček, Jiří Spilka, Miroslav Burša, et al. Open access intrapartum ctg database. *BMC Pregnancy and Childbirth*, 14:16, 2014.

Yared Daniel Daydulo, Bheema Lingaiah Thamineni, Hanumesh Kumar Dasari, and Genet Tadese Aboye. Deep learning based fetal distress detection from time frequency representation of cardiotocogram signal using morse wavelet: research study. *BMC Medical Informatics and Decision Making*, 22, 12 2022. ISSN 14726947. doi: 10.1186/s12911-022-02068-1.

Ehigha Enabudoso. Electronic fetal monitoring. *Contemporary Obstetrics and Gynecology for Developing Countries: Second Edition*, pp. 159–173, 8 2021. doi: 10.1007/978-3-030-75385-6_15/COVER.

A. L. Goldberger, L. A. N. Amaral, L. Glass, et al. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000 (June 13). Circulation Electronic Pages: http://circ.ahajournals.org/content/101/23/e215.full PMID:1085218; doi: 10.1161/01.CIR.101.23.e215.

Zahra Hoodbhoy, Mohammad Noman, Ayesha Shafique, Ali Nasim, Devyani Chowdhury, and Babar Hasan. Use of machine learning algorithms for prediction of fetal risk using cardiotocographic data. *International Journal of Applied and Basic Medical Research*, 9:226, 2019. ISSN 2229-516X. doi: 10.4103/ijabmr.ijabmr_370_18.

A. M. Jukic, D. D. Baird, C. R. Weinberg, D. R. Mcconnaughey, and A. J. Wilcox. Length of human pregnancy and contributors to its natural variation. *Human Reproduction (Oxford, England)*, 28: 2848, 2013. ISSN 14602350. doi: 10.1093/HUMREP/DET297.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

Joy E. Lawn, Hannah Blencowe, Peter Waiswa, et al. Stillbirths: Rates, risk factors, and acceleration towards 2030. *The Lancet*, 387:587–603, 2 2016. ISSN 1474547X. doi: 10.1016/S0140-6736(15) 00837-5.

Lisa D. Levine, Adi Hirshberg, and Sindhu K. Srinivas. Term induction of labor and risk of cesarean delivery by parity. *The Journal of Maternal-Fetal & Neonatal Medicine*, 27:1232, 2014. ISSN 14764954. doi: 10.3109/14767058.2013.864274.

Lochana Mendis, Marimuthu Palaniswami, Fiona Brownfoot, and Emerson Keenan. Computerised cardiotocography analysis for the automated detection of fetal compromise during labour: A review. *Bioengineering*, 10, 9 2023. ISSN 23065354. doi: 10.3390/BIOENGINEERING10091007.

Godfrey R. Mugyenyi, Esther C. Atukunda, Joseph Ngonzi, Adeline Boatin, Blair J. Wylie, and Jessica E. Haberer. Functionality and acceptability of a wireless fetal heart rate monitoring device in term pregnant women in rural southwestern uganda. *BMC Pregnancy and Childbirth*, 17:1–11, 6 2017. ISSN 14712393. doi: 10.1186/S12884-017-1361-1/TABLES/2.

Jun Ogasawara, Satoru Ikenoue, Hiroko Yamamoto, et al. Deep neural network-based classification of cardiotocograms outperformed conventional algorithms. *Scientific Reports*, 11:1–9, 6 2021. ISSN 2045-2322. doi: 10.1038/s41598-021-92805-9.

Outi Palomäki, Tiina Luukkaala, Riikka Luoto, and Risto Tuimala. Intrapartum cardiotocography – the dilemma of interpretational variation. *Journal of Perinatal Medicine*, 34:298–302, 8 2006. ISSN 0300-5577. doi: 10.1515/JPM.2006.057.

J. T. Parer. *Handbook of fetal heart rate monitoring*. Saunders, 1983. URL `https://cir.nii.ac.jp/crid/1130000794602978560`.

Tae Jun Park, Hye Jin Chang, Byung Jin Choi, et al. Machine learning model for classifying the results of fetal cardiotocography conducted in high-risk pregnancies. *Yonsei Medical Journal*, 63: 692–700, 7 2022. ISSN 05135796. doi: 10.3349/ymj.2022.63.7.692.

Alessio Petrozziello, Christopher W.G. Redman, Aris T. Papageorghiou, Ivan Jordanov, and Antoniya Georgieva. Multimodal convolutional neural networks to detect fetal compromise during labor and delivery. *IEEE Access*, 7:112026–112036, 2019. ISSN 21693536. doi: 10.1109/ACCESS.2019.2933368.

Astik Kumar Pradhan, Jitendra Kumar Rout, Aurobinda Bharat Maharana, Bunil Kumar Balabantaray, and Niranjan Kumar Ray. A machine learning approach for the prediction of fetal health using ctg. *International Conference on Information Technology*, pp. 239–244, 3 2021. doi: 10.1109/ocit53463.2021.00056.

Dennis Ryu, Dong Hyun Kim, Joan T Price, et al. Comprehensive pregnancy monitoring with a network of wireless, soft, and flexible sensors in high-and low-resource health settings. *PNAS*, 118, 2021. doi: 10.1073/pnas.2100466118/-/DCSupplemental.

Agnese Sbrollini, Angela Agostinelli, Luca Burattini, et al. Ctg analyzer: A graphical user interface for cardiotocography. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, 1:2606–2609, 9 2017. ISSN 1557170X. doi: 10.1109/EMBC.2017.8037391.

S. Schiermeier, G. Westhof, A. Leven, H. Hatzmann, and J. Reinhard. Intra- and interobserver variability of intrapartum cardiotocography: a multicenter study comparing the figo classification with computer analysis software. *Gynecologic and Obstetric Investigation*, 72:169–173, 10 2011. ISSN 1423-002X. doi: 10.1159/000327133.

Suraj Singh, Rakesh Kumar, Anand Agarwal, Amita Tyagi, and Surender Bisht. Intrapartum cardiotocographic monitoring and its correlation with neonatal outcome. *Journal of Family Medicine and Primary Care*, 11:7398, 2022. ISSN 2249-4863. doi: 10.4103/jfmpc.jfmpc_1525_22.

Edoardo Spairani, Beniamino Daniele, Maria Gabriella Signorini, and Giovanni Magenes. A deep learning mixed-data type approach for the classification of fhr signals. *Frontiers in Bioengineering and Biotechnology*, 10:887549, 8 2022. ISSN 22964185. doi: 10.3389/FBIOE.2022.887549/BIBTEX.

Jiří Spilka, George Georgoulas, Petros Karvelis, et al. Automatic evaluation of fhr recordings from ctu-uhb ctg database. *Lecture Notes in Computer Science (LNCS)*, 8060:47–61, 2013. ISSN 03029743. doi: 10.1007/978-3-642-40093-3_4/COVER.

Jiří Spilka, George Georgoulas, Petros Karvelis, Václav Chudáček, Chrysostomos D. Stylios, and Lenka Lhotská. Discriminating normal from “abnormal” pregnancy cases using an automated fhr evaluation method. *Lecture Notes in Computer Science (LNCS)*, 8445:521–531, 2014. ISSN 16113349. doi: 10.1007/978-3-319-07064-3_45/COVER.

J. A. Thorp and R. S. Rushing. Umbilical cord blood gas analysis. *Obstetrics and Gynecology Clinics of North America*, 26:695–709, 12 1999. ISSN 0889-8545. doi: 10.1016/S0889-8545(05)70107-8.

Zhixin Zhou, Zhidong Zhao, Xianfei Zhang, Xiaohong Zhang, Pengfei Jiao, and Xuanyu Ye. Identifying fetal status with fetal heart rate: Deep learning approach based on long convolution. *Computers in Biology and Medicine*, 159:106970, 6 2023. ISSN 0010-4825. doi: 10.1016/J.COMPBIOMED.2023.106970.

A HYPERPARAMETER TUNING

To optimize the neural network model hyperparameters, we performed an architecture search over the number of temporal, depthwise, separable filters, the kernel width for the separable convolution, the number of hidden layers, and the hidden layer dimension with fixed model training hyperparameters (loss function: binary cross-entropy, learning rate: $3e{-}4$, batch size: 128, dropout: 0.2). The range of values considered and the sampling function used to conduct the model architecture and training hyperparameter search is depicted in Table 2. The resulting optimized model architecture parameters and fixed hyperparameter values used to conduct the architecture search are shown in Table 3.

Table 2: Model architecture and training hyperparameter search sampling parameters.

Hyperparameter	Sampling function	Value range
# temporal filters	Uniform	[4, 8]
# depthwise filters	Uniform	[4, 8]
# separable filters	Uniform	[4, 8]
Kernel width	Uniform	[3, 9]
# hidden layers	Uniform	[0, 2]
Hidden layer dimension	$\lfloor 2^{x/2} \rfloor$ for $x \in [12, 20)$	[64, 724]
Loss function	Uniform	{BCE, Focal BCE}
Learning rate	Log-Uniform	[$1e{-}5$, $1e{-}1$]
Batch size	Uniform	[50, 300]
Dropout rate	Uniform	{0.0, 0.1, 0.2, 0.3, 0.4, 0.5}

Table 3: Selected model architecture hyperparameters and default model training hyperparameters used during the architecture search.

Hyperparameter	Default value	
	2-channel	1-channel
# temporal filters	6	7
# depthwise filters	8	6
# separable filters	5	4
Kernel width	5	3
# hidden layers	2	2
Hidden layer dimension	128	181
Loss function	BCE	
Learning rate	$3e{-}4$	
Batch size	128	
Dropout rate	0.2	

We generated confidence intervals for the estimated metrics by performing 1000 iterations of bootstrap resampling of the 10% held-out test set, corresponding to a bootstrap size of 56. As mentioned in Section 2, we selected the best hyperparameter settings using the highest validation AUROC averaged over cross-validation folds. We then selected the best-trained model overall by choosing the model trained on the cross-validation fold yielding the highest validation AUROC. The same procedure was used to select the model from the pre-training phase to use for downstream fine-tuning.

B SUBGROUP DEFINITION CRITERIA

Subgroups were split on a binary subgroup variable and held-out CTG recordings were assigned to two disjoint sets according to a cut-off threshold value, shown in Table 4. The thresholds were chosen according to clinical understanding (Cavazos-Rehg et al., 2015; Jukic et al., 2013; Levine et al., 2014). For the subgroup analysis, the low and high groups consisted of recordings below and above the threshold respectively.

Table 4: Binary demographic, clinical, and signal quality subgroup variable cut-off thresholds.

Subgroup variable	Low group	High group
Gestational age	≤ 40 weeks	> 40 weeks
Maternal age	< 35 years	≥ 35 years
Maternal risk	Diabetes $\cap$ Hypertension $\cap$ Pre-eclampsia $\cap$ Liq. praecox $\cap$ Pyrexia $\cap$ Meconium	Diabetes $\cup$ Hypertension $\cup$ Pre-eclampsia $\cup$ Liq. praecox $\cup$ Pyrexia $\cup$ Meconium
Parity	Nulliparous (0)	Multiparous (≥ 1)
FHR signal missingness	$\leq 20\%$ missing	$> 20\%$ missing
UC signal missingness	$\leq 20\%$ missing	$> 20\%$ missing

C Invalid performance metrics

The maternal age subgroup performance metrics for the Apgar prediction task were not comparable because only the normal class was present in the test set for the high maternal age subgroup. This yielded invalid AUROC and sensitivity metrics. For this subgroup, none of the 10 records in the held-out test set had an associated abnormal Apgar score (mode: 9, median: 9, min: 8). However, both the FHR + UC and FHR + UC + metadata models predicted all normal Apgar scores for this subgroup, achieving perfect accuracy. We speculate that the model defaults to predicting normal Apgar scores due to the class imbalance in the dataset. This may lead to an over-optimistic estimate of performance on the test set since the high maternal age subgroup has an under-representation of abnormal Apgar scores. Therefore, the model's ability to identify abnormal cases for this subgroup remains yet to be evaluated.