OBSTETRICS

Intrapartum electronic fetal heart rate monitoring to predict acidemia at birth with the use of deep learning

Jennifer A. McCoy, MD, MSCE; Lisa D. Levine, MD, MSCE; Guangya Wan, MS; Corey Chivers, PhD; Joseph Teel, MD; William G. La Cava, PhD

BACKGROUND: Electronic fetal monitoring is used in most US hospital births but has significant limitations in achieving its intended goal of preventing intrapartum hypoxic-ischemic injury. Novel deep learning techniques can improve complex data processing and pattern recognition in medicine.

OBJECTIVE: This study aimed to apply deep learning approaches to develop and validate a model to predict fetal acidemia from electronic fetal monitoring data.

STUDY DESIGN: The database was created using intrapartum electronic fetal monitoring data from 2006 to 2020 from a large, multisite academic health system. Data were divided into training and testing sets with equal distribution of acidemic cases. Several different deep learning architectures were explored. The primary outcome was umbilical artery acidemia, which was investigated at 4 clinically meaningful thresholds: 7.20, 7.15, 7.10, and 7.05, along with base excess. The receiver operating characteristic curves were generated with the area under the receiver operating characteristic assessed to determine the performance of the models. External validation was performed using a publicly available Czech database of electronic fetal monitoring data.

RESULTS: A total of 124,777 electronic fetal monitoring files were available, of which 77,132 had <30% missingness in the last 60 minutes of the electronic fetal monitoring tracing. Of these, 21,041 were matched to a corresponding umbilical cord gas result, of which 10,182 were time-stamped within 30 minutes of the last electronic fetal monitoring reading and composed the final dataset. The prevalence rates of the outcomes in the data were 20.9% with a pH of <7.2, 9.1% with a pH of <7.15, 3.3% with a pH of <7.10, and 1.3% with a pH of <7.05. The best performing model achieved an area under the receiver operating characteristic of 0.85 at a pH threshold of <7.05. When predicting the joint outcome of both pH of <7.05 and base excess of less than −10 meq/L, an area under the receiver operating characteristic of 0.89 was achieved. When predicting both pH of <7.20 and base excess of less than −10 meq/L, an area under the receiver operating characteristic of 0.87 was achieved. At a pH of <7.15 and a positive predictive value of 30%, the model achieved a sensitivity of 90% and a specificity of 48%.

CONCLUSION: The application of deep learning methods to intrapartum electronic fetal monitoring analysis achieves promising performance in predicting fetal acidemia. This technology could help improve the accuracy and consistency of electronic fetal monitoring interpretation.

Key words: artificial intelligence, deep learning, electronic fetal monitoring, fetal acidemia, intrapartum, labor, machine learning, obstetrics

Introduction

Electronic fetal monitoring (EFM) is used in >85% of births in the United States,[1] to allow clinicians to detect changes in the fetal heart rate that may indicate academia, enabling them to intervene before irreversible consequences. EFM was disseminated into practice before robust assessment of its efficacy, and decades of work since have shown the limitations of EFM in achieving its intended goal of preventing intrapartum hypoxic-ischemic injury.[2−6] In addition, EFM use has been shown to be associated with a significant increase in obstetrical intervention, especially cesarean delivery.[6,7] Much work has been done to develop guidelines and standardized frameworks for EFM interpretation in the hopes of improving both neonatal outcomes and the precision of obstetrical interventions.[4,5,8] However, a large 2017 meta-analysis showed that continuous EFM was associated with a significantly increased risk of cesarean delivery and operative vaginal delivery but no reduction in perinatal death or cerebral palsy.[6] Subsequently, attempts have been made to further refine the visual features of EFM that may be associated with fetal acidemia,[9,10,11] whereas other studies have highlighted the limitations of the correlation among EFM patterns, umbilical artery pH, and neonatal outcomes and questioned the use of any further attempts to improve EFM interpretation.[12]

In response to many of these challenges, computerized interpretation of EFM has been explored since the 1980s. Unfortunately, existing software programs, which are largely designed to detect the same EFM features as clinicians, have not demonstrated clinical benefit in randomized controlled trials.[13−18] In recent years, artificial intelligence technologies have been explored as a potential avenue to improve EFM interpretation. Specifically, deep learning, a subtype of machine learning, has significant appeal. Deep learning techniques have shown great promise in facilitating complex data processing and pattern recognition in medicine,[19−22] and the application of deep learning to the problem of EFM interpretation has begun to be explored, with some early retrospective research, primarily performed using a single, small dataset, showing promising test characteristics for computer-generated

Cite this article as: McCoy JA, Levine LD, Wan G, et al. Intrapartum electronic fetal heart rate monitoring to predict acidemia at birth with the use of deep learning. Am J Obstet Gynecol 2024;XX:x.ex−x.ex.

0002-9378/$36.00

https://doi.org/10.1016/j.ajog.2024.04.022

Click Supplemental Materials and Video under article title in Contents at ajog.org

AJOG at a Glance

Why was this study conducted?

Electronic fetal monitoring (EFM) is used in most US hospital births but has substantial limitations in achieving its intended goal of preventing intrapartum hypoxic-ischemic injury. Novel deep learning techniques may help improve the accuracy and reliability of EFM interpretation.

Key findings

Deep learning models trained on a large multisite dataset of EFM tracings exhibited promising accuracy in predicting acidemia in umbilical cord gas at several different pH thresholds. A deep learning model exhibited an area under the receiver operating characteristic curve of 0.85 at a pH threshold of <7.05. At a pH of <7.15 and a minimum positive predictive value of 30%, the model achieved a sensitivity of 90% and a specificity of 48%.

What does this add to what is known?

The application of deep learning methods to intrapartum EFM analysis achieves promising performance in predicting fetal acidemia. This technology could help improve the accuracy and consistency of EFM interpretation.

algorithms.[19,23–29] As opposed to analyzing visually identifiable EFM patterns, similar to those that clinicians visually interpret, deep learning relies on an entirely data-driven approach.

Our overarching hypothesis is that novel, data-driven deep learning approaches can detect meaningful data patterns in EFM, beyond those features that clinicians or other software programs recognize, that could help improve the predictive accuracy of EFM. Accordingly, we sought to develop and externally validate an algorithm that can use EFM data to predict fetal acidemia using deep learning data analysis techniques.

Materials and Methods

Database creation

The EFM database was created by accessing the intrapartum EFM data files stored for all deliveries in a large, multisite academic health system (University of Pennsylvania Health System) from January 1, 2006, to December 31, 2020. The EFM files were matched to a maternal medical record number and a corresponding umbilical artery laboratory result to ensure that each EFM strip corresponded to a distinct patient. We included only those tracings with EFM data available for at least the last 60 minutes before delivery and an arterial umbilical cord gas result, with a laboratory order time stamp within 30 minutes of the end of the fetal heart rate tracing. For this initial analysis, patients were excluded if there was >30% missingness in the EFM data in the last 60 minutes. Patients with 2 or more umbilical cord gas laboratory results with the same time stamp were excluded as a proxy for multiple pregnancies. Limited additional clinical and demographic data were abstracted from the medical records system and matched using the patients' medical records. A random sample of 200 medical records was selected for manual review to confirm the accuracy of the data abstraction process. The primary outcome was fetal acidemia, as determined by umbilical artery pH, which we investigated as a binary outcome at 4 clinically meaningful thresholds: 7.05, 7.10, 7.15, and 7.20. To more specifically capture fetal metabolic acidemia, we performed additional analyses with a joint outcome of umbilical artery pH and base excess.

Data preprocessing

Data preprocessing was performed to remove outliers, noise, and artifact from the EFM data. Extremes of a fetal heart rate of <50 or >200 bpm were excluded. Thresholds of beat-to-beat variation were established, with a >25 beat change in 1 second considered to be likely artifact and removed. The 4-Hz raw signal was smoothed and down-sampled to 0.25 Hz, mirroring previous work.[25,26]

Deep learning analysis

To train and evaluate deep learning models, we divided the data into 75% training and 25% test sets. The training set was subdivided during model fitting into 80/20 train and validation sets, with validation used to select the final model, for each architecture and overall. We used the Adam optimizer[30] with an adaptive learning rate[31] and trained for 100 epochs. Of note, 6 deep learning architectures were tested, including convolutional neural networks (CNNs), fully connected CNNs (FCNs),[32] long short-term memories (LSTMs), multiscale CNNs (CNN-MS),[33] a variant of residual networks called InceptionTime,[34] and Transformers.[35] The CNN and LSTM architectures were chosen to match those developed in previous work.[25] CNN-MS, FCN, and InceptionTime were chosen for their state-of-the-art performance on cross-domain time series classification benchmarks.[32,34] Transformers were included because of their breakout performance in recent years on natural language and sequence learning tasks.[35] Each architecture was trained 10 times with different random seeds and shuffles of the training and validation data.

Evaluating model performance

Receiver operating characteristic (ROC) curves were generated for the models at different pH thresholds. The median area under the receiver operating characteristic (AUROC) curve and 95% confidence intervals (CIs) were assessed. The number of parameters in each model was assessed to determine its efficiency. Internal AUROC curve validation was performed at each pH threshold to identify the best final model. For the final model, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) were assessed at classification thresholds targeting a sensitivity or specificity of 80% or a set minimum PPV, determined based on the

assessment of each model's area under the precision-recall curve.

External validation of model

The model with the best internal validation AUROC curve was chosen for external validation. External validation was performed using a publicly available database from the Czech Republic, known as CTU-UHB.[24] This database consists of 552 patients at >37 weeks of gestation with fetal monitoring data paired with cord gas pH results. They excluded multiple pregnancies or patients with known congenital anomalies or fetal growth restriction and enriched the dataset for cesarean deliveries and fetal acidemia, resulting in an elevated prevalence of 7.2% for a pH of <7.05. We externally validated our model with and without fine-tuning to account for this difference in prevalence. In the former scenario, the model was evaluated directly on the 552 cases; in the latter scenario, the final model was trained for an additional 100 epochs on a randomly chosen 50% set of patients from CTU-UHB and then evaluated on the remaining cases.

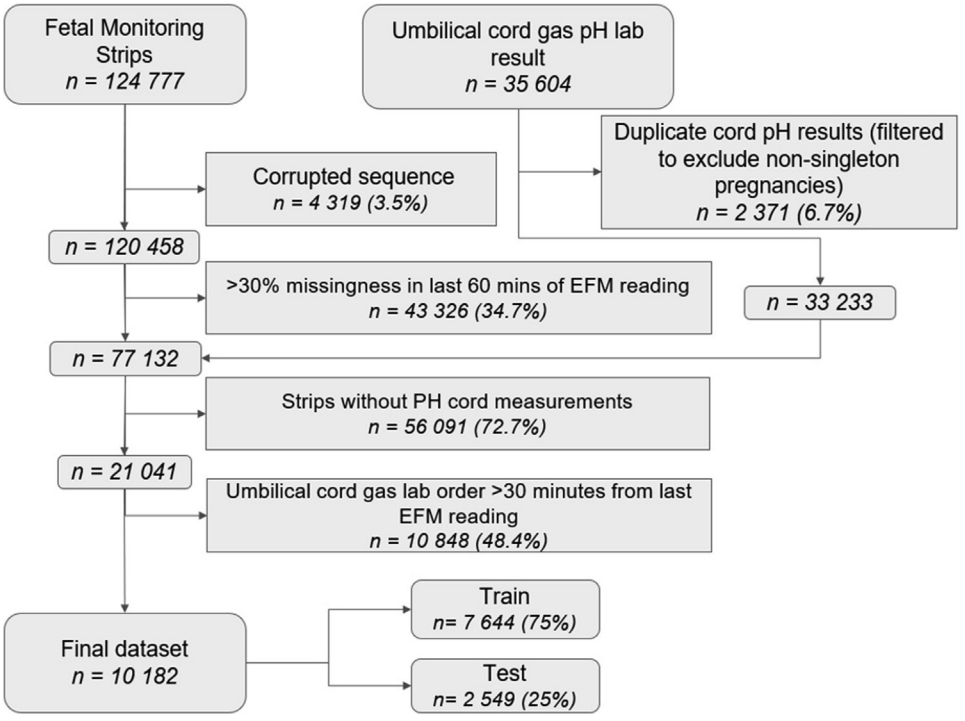
The code developed to preprocess the datasets, train the models, and evaluate performance is available at https://github.com/cavalab/ai-efm.

Results

The database creation is shown in Figure 1. A total of 124,777 fetal monitoring files were available, of which 77,132 had <30% missingness in the last 60 minutes of the EFM tracing. Of these, 21,041 were matched to a corresponding umbilical cord gas result, 10,182 of which were time-stamped within 30 minutes of the last EFM reading. This composed the final dataset for analysis. Of this final dataset of 10,182 tracings, 7644 (75%) were assigned to the training set, and 2549 (25%) were assigned to the testing set. Figure 2 shows the distribution of umbilical artery pH in the dataset, with most values >7.20, as expected. The prevalence rates of the outcomes of acidemia on umbilical artery pH were 20.9% with a pH of <7.2, 9.1% with a pH of <7.15, 3.3% with a pH of <7.10, and 1.3% with a pH of <7.05. The median maternal age was 28.0 years (interquartile range [IQR], 23.0–32.0), and the median gestational age at delivery was 39.3 weeks (IQR, 38.3–40.1). Patients were 62.8% Black, 19.3% White, 7.3% Asian, 4.4% Hispanic or Latino, 0.7% East Indian, and 5.6% other or unknown.

The test set AUROC curve values for each of the tested deep learning approaches are shown in the Supplemental Table (Supplemental Figure). The results of the best-performing model type (as determined by AUROC curve validation), known as InceptionTime, are shown in Table 1 along with the performance of the model on test data in predicting the pH outcome and the joint outcome of pH and base excess. The accompanying ROC curves for the models are shown in Figure 3. At a pH threshold of 7.05, the model achieved an

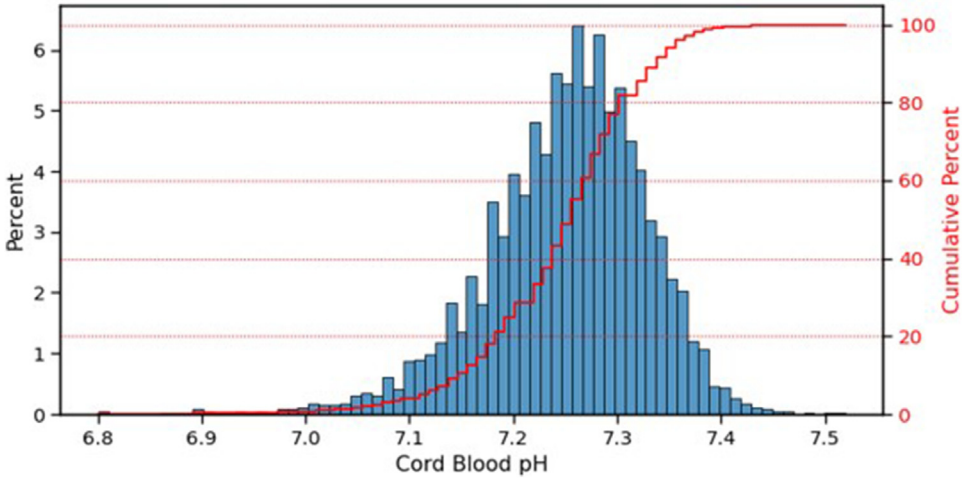
FIGURE 1
Flowchart of database creation



EFM, electronic fetal monitoring.

McCoy. Deep learning to predict fetal acidemia. Am J Obstet Gynecol 2024.

FIGURE 2
Distribution of umbilical cord gas pH across the dataset (n = 10,182)



The *red line* and axis show the cumulative distribution.

McCoy. Deep learning to predict fetal acidemia. Am J Obstet Gynecol 2024.

AUROC of 0.85. When predicting the joint outcome of both pH of <7.05 and base excess of less than −10 meq/L, an AUROC curve of 0.89 was achieved. At a pH threshold of 7.20, the model achieved an AUROC curve of 0.75. When predicting both pH of <7.20 and base excess of less than −10 meq/L, an AUROC curve of 0.87 was achieved. Next, we assessed the performance of the models when setting sensitivity, specificity, and PPV targets. The PPV set points ranged from 10% to 50%, increasing as the prevalence of the outcome increased (Table 2). At a pH of <7.15 and a PPV of 30%, the model achieved a sensitivity of 90% and a specificity of 48%.

Finally, we externally validated our model using the publicly available CTU-UHB database.[24] On this dataset, our model exhibited similar performance, predicting a pH of <7.05 and an AUROC curve of 0.72 without additional training. With additional training and fine-tuning on a subset of CTU-UHB data, the model's performance increased (AUROC curve of 0.76).

Discussion

Principal findings

We demonstrated that, in a large, multicenter clinical database, deep learning methods applied to intrapartum EFM analysis achieve promising performance in predicting fetal acidemia. By offering a

TABLE 1
Performance of final models for predictions of umbilical artery pH and combined prediction of pH and base excess

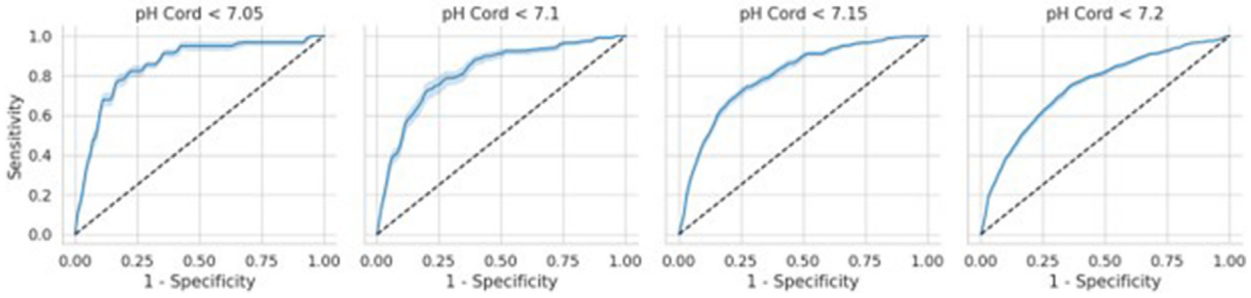
Outcome predicted	AUROC	AUPRC	Outcome rate
pH<7.05	0.85 (0.83–0.87)	0.12 (0.09–0.16)	0.013 (0.012–0.015)
pH<7.05 and base excess of less than −10 meq/L	0.89 (0.87–0.92)	0.13 (0.08–0.18)	0.008 (0.007–0.009)
pH<7.10	0.83 (0.81–0.84)	0.17 (0.15–0.20)	0.033 (0.030–0.035)
pH<7.10 and base excess of less than −10 meq/L	0.88 (0.87–0.9)	0.13 (0.10–0.16)	0.013 (0.011–0.014)
pH<7.15	0.79 (0.78–0.8)	0.28 (0.27–0.30)	0.090 (0.087–0.094)
pH<7.15 and base excess of less than −10 meq/L	0.87 (0.85–0.88)	0.15 (0.12–0.17)	0.018 (0.016–0.019)
pH<7.20	0.75 (0.74–0.75)	0.44 (0.42–0.45)	0.210 (0.205–0.215)
pH<7.20 and base excess of less than −10 meq/L	0.87 (0.86–0.89)	0.17 (0.14–0.20)	0.020 (0.018–0.022)

Along with AUROC and AUPRC predictions on patients with specified levels of pH and base excess, the outcome prevalence in the test set is provided. Data are presented as median (interquartile range) over 1000 bootstrap resamples of the test data.

AUPRC, area under the precision-recall curve; *AUROC*, area under the receiver operating characteristic.

McCoy. Deep learning to predict fetal acidemia. Am J Obstet Gynecol 2024.

FIGURE 3
Receiver operating characteristic curves for the final deep learning models



From left to right, the pH threshold for classifying fetal acidemia increases from 7.05 to 7.10, 7.15, and 7.20. The *shaded area* indicates the 95% confidence interval.

McCoy. Deep learning to predict fetal acidemia. Am J Obstet Gynecol 2024.

data-driven approach, this technology has the potential to augment the clinical interpretation of EFM. Future work will seek to assess whether this tool can help improve the accuracy and consistency of EFM interpretation.

Results in the context of what is known

Of note, 2 authors have recently summarized the existing work applying machine learning or deep learning techniques to EFM analysis.[36,37] Much of the existing work has used the small, publicly available single-center database from the Czech Republic of 552 cases of highly curated fetal monitoring data known as CTU-UHB.[24] Ogasawara et al[27] used a small, case-control dataset of 384 patients and experimented with several different deep learning models to predict the outcomes of the 1-minute Apgar score and pH of <7.20. They achieved AUROC curves ranging from 0.62 to 0.73.[27] Fergus et al[38] used the 552 tracings in the CTU-UHB database to develop CNN models, achieving an AUROC curve of 0.86. In addition, Zhao et al,[28] using CTU-UHB, experimented with CNN models, achieving AUROC curves as high as 0.98. The robustness and clinical applicability of each of these studies are significantly limited by the small, highly curated nature of the dataset used. Petrozziello et al[25] used a database of 35,429 deliveries at >36 weeks of gestation, with a 4.5% prevalence of their primary outcome of either a pH of <7.05 or a "severe compromise." They analyzed the last hour of EFM (at 0.25 Hz) and developed a CNN deep learning model that achieved a sensitivity of 42% and an AUROC curve of 0.68.[25] In our work, at a pH of <7.05, our final model exhibited a substantially higher AUROC curve of 0.85, outperforming the previous best CNN and LSTM architectures from Petrozziello et al.[25] Our work is a notable advancement of these previous efforts in that we created and used a large, clinically realistic dataset, and developed model architectures that demonstrate better performance.

Concerning previous literature assessing the performance of clinician interpretation of EFM to predict acidemia, many efforts have been made over recent decades to standardize and

TABLE 2
Performance of final models

pH threshold	AUROC	AUPRC	Operating target	Sensitivity	Specificity	PPV	NPV
7.05	0.85	0.13	Sensitivity = 0.8	0.79	0.78	0.05	1.00
			Specificity = 0.8	0.76	0.80	0.04	1.00
			PPV = 0.1	0.97	0.26	0.10	1.00
7.1	0.81	0.16	Sensitivity = 0.8	0.81	0.65	0.11	0.99
			Specificity = 0.8	0.71	0.80	0.07	0.99
			PPV = 0.2	0.93	0.33	0.20	0.99
7.15	0.79	0.29	Sensitivity = 0.8	0.80	0.62	0.26	0.97
			Specificity = 0.8	0.65	0.80	0.17	0.96
			PPV = 0.3	0.90	0.48	0.30	0.98
7.2	0.75	0.44	Sensitivity = 0.8	0.80	0.54	0.43	0.91
			Specificity = 0.8	0.56	0.80	0.32	0.87
			PPV = 0.5	0.88	0.39	0.50	0.92

Sensitivity, specificity, and PPV are reported at classification thresholds that target a sensitivity or specificity of 80% or a minimum PPV based on assessment of each model's AUPRC.

AUPRC, area under the precision-recall curve; *AUROC*, area under the receiver operating characteristic; *NPV*, negative predictive value; *PPV*, positive predictive value.

McCoy. Deep learning to predict fetal acidemia. Am J Obstet Gynecol 2024.

optimize visual interpretation of EFM.[4,5,8] Cahill et al[9] conducted a prospective cohort study to assess the performance of both standard *Eunice Kennedy Shriver* National Institute of Child Health and Human Development criteria and more complex visually interpreted features of EFM, in the 120 minutes before delivery in predicting cord blood acidemia of <7.10, with a maximum AUROC curve of 0.77, sensitivities ranging from 63% to 73%, specificities of 44% to 78%, and PPVs of 2% to 4%. In contrast, our final model exhibited better discriminative performance at this pH threshold (AUROC curve of 0.81), with a PPV of 11% at 80% sensitivity and 65% specificity. Another study assessing the performance of visual interpretation of the last 30 minutes of EFM in predicting acidemia demonstrated high specificity of certain visual features but very low sensitivities and PPVs ranging from 0.0% to 3.4%.[10] In studies of EFM interpretation in clinical practice, a large Cochrane database meta-analysis found that continuous EFM during labor did not significantly reduce the risk of cord blood acidosis at delivery (risk ratio, 0.92; 95% CI, 0.27−3.11).[6] Other studies have shown that the sensitivity of EFM interpretation for predicting acidemia in actual clinical practice ranges from 30% to 45% and that fewer than 15% of neonates delivered via cesarean delivery for the indication of nonreassuring fetal heart tones are found to have acidemia at delivery.[39] Compared with these previous assessments, the performance of our models suggests a meaningful improvement in sensitivity and PPV while maintaining adequate specificity.

Research implications

This work is an essential step in the research needed to understand how to best harness the power of artificial intelligence to improve the interpretation of EFM. These initial results provide proof of concept that a purely data-driven model can achieve promising predictive performance. The database and the models developed in this work can now be used to compare the performance of the deep learning model with expert clinician interpretation, which is a work that is underway. In addition, future research will seek to identify whether there are particular types of tracings or subsets of patients in whom the model's performance may be most useful to enhance clinical decision-making and to begin to test the performance of the model prospectively.

In addition to advancing important clinical research on improving EFM interpretation, our work contributes to the practice of machine learning for time series model development. Our results suggest that deep learning approaches that model the local frequency content of signals at various time scales (eg, CNN-MS and InceptionTime) perform well for EFM analysis and acidemia prediction. Both architectures allow the models to learn nonlinear features that span local (ie, a few seconds) and global scales (ie, several minutes). Moreover, the results suggest that CNN-based architectures are preferable over autoregressive approaches, such as LSTMs and Transformers, not to mention faster to train. Transformers exhibited poor performance on this task, contributing to a growing body of work suggesting fundamental limitations that need to be addressed with this architecture if its promise is to be realized for time series classification.[40]

Clinical implications

Although umbilical artery acidemia as an outcome measure has limitations, it is a clinically relevant outcome and affects important neonatal care decisions. Umbilical artery acidemia as measured by pH has been consistently shown to be associated with neonatal morbidity,[41,42] with a large 2010 meta-analysis finding an odds ratio for hypoxic-ischemic encephalopathy (HIE) of 13.8.[42] Accordingly, umbilical artery acidemia is used to determine the need for neonatal therapeutic hypothermia treatment when there is concern for HIE.[43] Although we hope to expand these models to predict other clinical outcomes, including Apgar score and HIE, in future work, our results demonstrate significant promise. A future decision support model could be integrated into clinical care to help bolster the capacity of clinicians to accurately identify fetal acidemia to help better focus obstetrical intervention on true cases of acidemia. Such a model could also help improve interrater and intrarater reliability in EFM interpretation and reduce inconsistencies and cognitive biases in clinical management.[6] Furthermore, a model that is appropriately and continuously adjusted to remove biases could play a pivotal role in mitigating healthcare disparities and health outcome inequities. Although much further work is needed to move this technology toward clinical application and to rigorously evaluate its potential clinical effect, it holds substantial promise to help improve intrapartum care for laboring patients.

Strengths and limitations

This study contributes to our understanding of state-of-the-art deep learning approaches for time series classification and their potential for improving intrapartum EFM analysis. In particular, we have established the performances of several contemporary model architectures concerning several clinically important thresholds for acidemia that had not been systematically explored in previous studies. The multicentered nature and size of our study help strengthen its generalizability to other obstetrical populations. Furthermore, most clinical sites represented in our database employ a policy of universal umbilical cord gas collection, as opposed to selective collection at provider discretion, ensuring that our database is more representative of a general laboring population.

This study did not explore the sensitivity of the results to additional experimental design variables, including sampling rate, laboratory order delay threshold, missingness threshold, prediction horizon, and training window. Future studies could illuminate the effect of these choices on the efficacy of models more broadly. In addition, we did not consider other formulations of the

prediction task, such as in regression settings or ordinal classification of the pH outcome variable. It is possible that other formulations would result in more clinically impactful models. Furthermore, there are limitations to both the EFM data and umbilical cord pH as an outcome that are important to acknowledge. Umbilical cord pH is an imperfect measure of intrapartum hypoxia, and there are limitations to its correlation with other short- and long-term neonatal outcomes.[9,39,44] Moreover, umbilical cord pH may not capture other nonhypoxic causes of fetal compromise. We also acknowledge the risk of potential confounding in the EFM data by possible recording of maternal heart rate at times by the EFM and the fact that models built using the US standard EFM tracing speed of 3 cm/minute may not be generalizable to other practice settings that use different tracing speeds.

Conclusions

Through exploration of several novel deep learning approaches, we have created, and both internally and externally validated, a deep learning model that demonstrated promising performance in predicting fetal acidemia. Further development of this technology has the potential to improve the accuracy and consistency of EFM interpretation. ■

References

1. Martin JA, Hamilton BE, Sutton PD, Ventura SJ, Menacker F, Munson ML. Births: final data for 2002. Natl Vital Stat Rep 2003;52:1–113.
2. Nelson KB, Dambrosia JM, Ting TY, Grether JK. Uncertain value of electronic fetal monitoring in predicting cerebral palsy. N Engl J Med 1996;334:613–8.
3. McCusker J, Harris DR, Hosmer DW. Association of electronic fetal monitoring during labor with cesarean section rate and with neonatal morbidity and mortality. Am J Public Health 1988;78:1170–4.
4. Macones GA, Hankins GD, Spong CY, Hauth J, 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. Obstet Gynecol 2008;112:661–6.
5. ACOG practice bulletin no. 106: intrapartum fetal heart rate monitoring: nomenclature, interpretation, and general management principles. Obstet Gynecol 2009;114:192–202.
6. Alfirevic Z, Devane D, Gyte GM, Cuthbert A, Devane D. Continuous cardiotocography (CTG) as a form of electronic fetal monitoring (EFM) for fetal assessment during labour. Cochrane Database Syst Rev 2017;2:CD006066.
7. Ananth CV, Chauhan SP, Chen H-Y, D'Alton ME, Vintzileos AM. Electronic fetal monitoring in the United States: temporal trends and adverse perinatal outcomes. Obstet Gynecol 2013;121:927–33.
8. Ayres-de-Campos D, Spong CY, Chandraharan E. FIGO Intrapartum Fetal Monitoring Expert Consensus Panel. FIGO consensus guidelines on intrapartum fetal monitoring: cardiotocography. Int J Gynaecol Obstet 2015;131:13–24.
9. Cahill AG, Tuuli MG, Stout MJ, López JD, Macones GA. A prospective cohort study of fetal heart rate monitoring: deceleration area is predictive of fetal acidemia. Am J Obstet Gynecol 2018;218:523.e1–12.
10. Cahill AG, Roehl KA, Odibo AO, Macones GA. Association and prediction of neonatal acidemia. Am J Obstet Gynecol 2012;207:206.e1–8.
11. Martí Gamboa S, Giménez OR, Mancho JP, Moros ML, Sada JR, Mateo SC. Diagnostic accuracy of the FIGO and the 5-tier fetal heart rate classification systems in the detection of neonatal acidemia. Am J Perinatol 2017;34:508–14.
12. Johnson GJ, Salmanian B, Denning SG, Belfort MA, Sundgren NC, Clark SL. Relationship between umbilical cord gas values and neonatal outcomes: implications for electronic fetal heart rate monitoring. Obstet Gynecol 2021;138:366–73.
13. INFANT Collaborative Group. Computerised interpretation of fetal heart rate during labour (INFANT): a randomised controlled trial. Lancet 2017;389:1719–29.
14. Steer PJ, Kovar I, McKenzie C, Griffin M, Linsell L. Computerised analysis of intrapartum fetal heart rate patterns and adverse outcomes in the INFANT trial. BJOG 2019;126:1354–61.
15. Wilson E, Dunn L, Beckmann M, Kumar S. Measuring the impact of cardiotocograph decision support software on neonatal outcomes: a propensity score-matched observational study. Aust N Z J Obstet Gynaecol 2021;61:876–81.
16. Ayres-de-Campos D, Sousa P, Costa A, Bernardes J. Omniview-SisPorto 3.5 – a central fetal monitoring station with online alerts based on computerized cardiotocogram+ST event analysis. J Perinat Med 2008;36:260–4.
17. Nunes I, Ayres-de-Campos D. Computer analysis of foetal monitoring signals. Best Pract Res Clin Obstet Gynaecol 2016;30:68–78.
18. Nunes I, Ayres-de-Campos D, Ugwumadu A, et al. Central fetal monitoring with and without computer analysis: a randomized controlled trial. Obstet Gynecol 2017;129:83–90.
19. Petrozziello A, Redman CWG, Papageorghiou AT, Jordanov I, Georgieva A. Multimodal convolutional neural networks to detect fetal compromise during labor and delivery. IEEE Access 2019;7:112026–36.
20. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436–44.
21. Nie D, Trullo R, Lian J, et al. Medical image synthesis with deep convolutional adversarial networks. IEEE Trans Biomed Eng 2018;65:2720–30.
22. Tomašev N, Glorot X, Rae JW, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. Nature 2019;572:116–9.
23. Balayla J, Shrem G. Use of artificial intelligence (AI) in the interpretation of intrapartum fetal heart rate (FHR) tracings: a systematic review and meta-analysis. Arch Gynecol Obstet 2019;300:7–14.
24. Chudáček V, Spilka J, Burša M, et al. Open access intrapartum CTG database. BMC Pregnancy Childbirth 2014;14:16.
25. Petrozziello A, Jordanov I, Aris Papageorghiou T, Christopher Redman WG, Georgieva A. Deep learning for continuous electronic fetal monitoring in labor. Annu Int Conf IEEE Eng Med Biol Soc 2018;2018:5866–9.
26. Spilka J, Frecon J, Leonarduzzi R, Pustelnik N, Abry P, Doret M. Sparse support vector machine for intrapartum fetal heart rate classification. IEEE J Biomed Health Inform 2017;21:664–71.
27. Ogasawara J, Ikenoue S, Yamamoto H, et al. Deep neural network-based classification of cardiotocograms outperformed conventional algorithms. Sci Rep 2021;11:13367.
28. Zhao Z, Deng Y, Zhang Y, Zhang Y, Zhang X, Shao L. DeepFHR: intelligent prediction of fetal acidemia using fetal heart rate signals based on convolutional neural network. BMC Med Inform Decis Mak 2019;19:286.
29. Ben M'Barek I, Jauvion G, Vitrou J, Holmström E, Koskas M, Ceccaldi PF. DeepCTG® 1.0: an interpretable model to detect fetal hypoxia from cardiotocography data during labor and delivery. Front Pediatr 2023;11:1190441.
30. Kingma DP, Ba J. Adam: a method for stochastic optimization. arXiv. 2014. Available at: arXiv:1412.6980. Accessed May 9, 2024.
31. Orabona F, Tommasi T. Training deep networks without learning rates through coin betting. arXiv. 2017. Available at: arXiv:1705.07795. Accessed May 9, 2024.
32. Wang Z, Yan W, Oates T. Time series classification from scratch with deep neural networks: a strong baseline. arXiv. 2016. Available at: arXiv:1611.06455. Accessed May 9, 2024.
33. Cui Z, Chen W, Chen Y. Multi-scale convolutional neural networks for time series

classification. arXiv. 2016. Available at: arXiv: 1603.06995. Accessed May 9, 2024.

34. Ismail Fawaz H, Lucas B, Forestier G, et al. InceptionTime: finding AlexNet for time series classification. Data Min Knowl Disc 2020;34: 1936–62.

35. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. arXiv. 2017. Available at: arXiv:1706.03762. Accessed May 9, 2024.

36. O'Sullivan ME, Considine EC, O'Riordan M, Marnane WP, Rennie JM, Boylan GB. Challenges of developing robust AI for intrapartum fetal heart rate monitoring. Front Artif Intell 2021;4:765210.

37. Ben M'Barek I, Jauvion G, Ceccaldi PF. Computerized cardiotocography analysis during labor — A state-of-the-art review. Acta Obstet Gynecol Scand 2023;102:130–7.

38. Fergus P, Chalmers C, Montanez CC, Reilly D, Lisboa P, Pineles B. Modelling segmented cardiotocography time-series signals using one-dimensional convolutional neural networks for the early detection of abnormal birth outcomes. IEEE Trans Emerg Top Comp Intell 2021;5:882–92.

39. Clark SL, Hamilton EF, Garite TJ, Timmins A, Warrick PA, Smith S. The limits of electronic fetal heart rate monitoring in the prevention of neonatal metabolic acidemia. Am J Obstet Gynecol 2017;216:163.e1–6.

40. Zeng A, Chen M, Zhang L, Xu Q. Are Transformers Effective for Time Series Forecasting? Proc AAAI Conf Artif Intell 2023;37: 11121–8.

41. Bligard KH, Cameo T, McCallum KN, et al. The association of fetal acidemia with adverse neonatal outcomes at time of scheduled cesarean delivery. Am J Obstet Gynecol 2022;227: 265.e1–8.

42. Malin GL, Morris RK, Khan KS. Strength of association between umbilical cord pH and perinatal and long term outcomes: systematic review and meta-analysis. BMJ 2010;340: c1471.

43. Shankaran S, Laptook AR, Ehrenkranz RA, et al. Whole-body hypothermia for neonates with hypoxic–ischemic encephalopathy. N Engl J Med 2005;353:1574–84.

44. Zullo F, Di Mascio D, Raghuraman N, et al. Three-tiered fetal heart rate interpretation system and adverse neonatal and maternal outcomes: a systematic review and meta-analysis. Am J Obstet Gynecol 2023;229: 377–87.

Author and article information

From the Maternal Fetal Medicine Research Program, Department of Obstetrics and Gynecology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA (Drs McCoy and Levine); School of Data Science, University of Virginia, Charlottesville, VA (Mr Wan); Proscia Inc, Philadelphia, PA (Dr Chivers); Department of Family Medicine and Community Health, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA (Dr Teel); and Computational Health Informatics Program, Department of Pediatrics, Boston Children's Hospital, Harvard Medical School, Boston, MA (Dr La Cava).

Received Feb. 5, 2024; revised April 17, 2024; accepted April 18, 2024.

The authors report no conflict of interest.

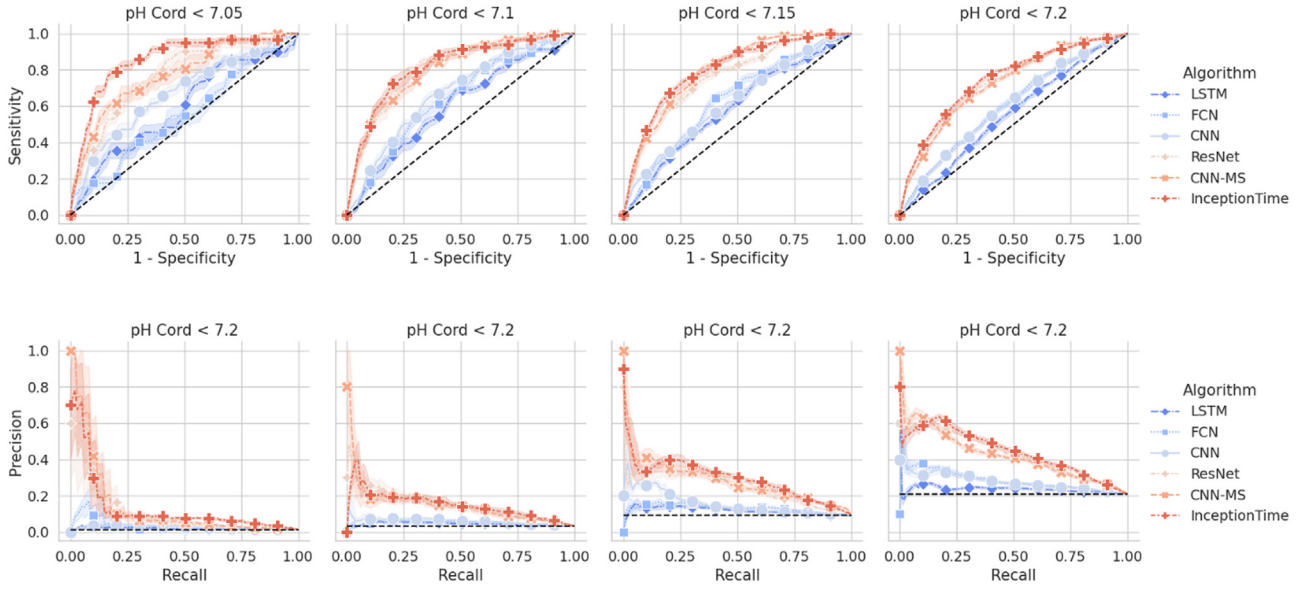
This study received funding from the Women's Reproductive Health Research (grant numbers: 5 K12 HD 1265-22 and T32-HD007440).

The funding source had no role in the study design; collection, analysis, and interpretation of data; writing of the report; or decision to submit the article for publication.

This study was presented at the 44th annual pregnancy meeting of the Society for Maternal-Fetal Medicine, National Harbor, MD, February 10—14, 2024.

Corresponding author: Jennifer A. McCoy, MD, MSCE. Jennifer.mccoy@pennmedicine.upenn.edu

SUPPLEMENTAL FIGURE
Performance of different deep learning models



Receiver operator characteristic curves (top panel) and precision-recall curves (bottom panel). From left to right, the pH threshold for classifying fetal distress increases from 7.05 to 7.10, 7.15, and 7.20.

CNN, convolutional neural network; *CNN-MS*, multiscale convolutional neural network; *FCN*, fully connected network; *LSTM*, long short-term memory; *ResNet*, residual neural network.

McCoy. Deep learning to predict fetal acidemia. Am J Obstet Gynecol 2024.

SUPPLEMENTAL TABLE
Median test set AUROC curve performance for various deep learning models and pH thresholds

		AUROC at pH threshold			
Algorithm	No. of parameters	7.05	7.1	7.15	7.2
InceptionTime	492,000	0.85 (0.83–0.87)[a]	0.81 (0.79–0.83)[a]	0.79 (0.78–0.80)[a]	0.75 (0.74–0.75)[a]
CNN-MS	44,000	0.78 (0.75–0.80)	0.80 (0.79–0.82)	0.79 (0.78–0.80)[a]	0.72 (0.72–0.73)
ResNet	6,000,000	0.78 (0.75–0.81)	0.81 (0.79–0.82)[a]	0.76 (0.75–0.77)	0.73 (0.72–0.74)
CNN	267,000	0.64 (0.60–0.68)	0.67 (0.66–0.69)	0.62 (0.61–0.63)	0.60 (0.59–0.61)
FCN	405,000	0.52 (0.49–0.56)	0.63 (0.61–0.65)	0.63 (0.62–0.64)	0.61 (0.60–0.62)
LSTM	12,000	0.62 (0.58–0.65)	0.61 (0.59–0.63)	0.59 (0.58–0.61)	0.56 (0.55–0.57)
Transformer	161,000	0.66 (0.56–0.75)	0.54 (0.47–0.60)	0.54 (0.50–0.58)	0.55 (0.52–0.58)

Each algorithm was trained with 10 random seed initializations, and the best performance on an internal validation set was used to choose the final model. Statistics are calculated via 1000 bootstrap resamples of test set performance.

AUROC, area under receiver operating characteristic.

[a] Indicates the best median AUROC scores among the algorithms at each pH threshold.

McCoy. Deep learning to predict fetal acidemia. Am J Obstet Gynecol 2024.