Computers in Biology and Medicine 184 (2025) 109448

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

DeepCTG® 2.0: Development and validation of a deep learning model to detect neonatal acidemia from cardiotocography during labor

Imane Ben M'Barek [a,b,*], Grégoire Jauvion [c], Jade Merrer [d,b], Martin Koskas [b,e], Olivier Sibony [b,f], Pierre – François Ceccaldi [a,b], Erwan Le Pennec [g], Julien Stirnemann [b,h]

[a] *Department of Gynecology Obstetrics, Assistance Publique des Hôpitaux de Paris -Beaujon, Clichy, 92100, France*
[b] *Université de Paris Cité, 75006, Paris, France*
[c] *Genos Care, Paris, France*
[d] *Unité d'Épidémiologie Clinique, INSERM CIC1426, Hôpital Robert Debré, APHP Paris, France*
[e] *Department of Gynecology and Obstetrics, Assistance Publique des Hôpitaux de Paris Hôpital Bichat, 75018 Paris, France*
[f] *Department of Obstetrics and Maternal-Fetal Medicine, Assistance Publique des Hôpitaux de Paris Hôpital Robert Debré, 75019 Paris, France*
[g] *CMAP, IP Paris, École polytechnique, CNRS, 91128 Palaiseau Cédex, France*
[h] *Department of Obstetrics and Maternal-Fetal Medicine, Assistance Publique des Hôpitaux de Paris Hôpital Necker-Enfants Malades, 75015 Paris, France*

ARTICLE INFO

Keywords:
Cardiotocography
Fetal heart rate
Neonatal acidemia
Computerized cardiotocography
Neonatal morbidity
Deep learning
Convolutional neural network

ABSTRACT

Cardiotocography (CTG) is the main tool available to detect neonatal acidemia during delivery. Presently, obstetricians and midwives primarily rely on visual interpretation, leading to a significant intra-observer variability. In this paper, we build and evaluate a convolutional neural network to detect neonatal acidemia from the CTG signals during delivery on a multicenter database with 27662 cases in five centers, including 3457 and 464 cases of moderate and severe neonatal acidemia respectively (defined by a fetal pH at birth between 7.05 and 7.20, and lower than 7.05 respectively). To use all the available records, the convolutional layers are pretrained on a task which consists in predicting several features known to be associated with neonatal acidemia from the raw CTG signals. In a cross-center evaluation, the AUC varies from 0.74 to 0.83 between the centers for the detection of severe acidemia, showing the ability of deep learning models to generalize from one dataset to the other and paving the way for more accurate models trained on larger databases. The model can still be significantly improved, by adding clinical variables to account for risk factors of acidemia that may not appear in the CTG signals. Further research will also be led to integrate the model in a tool that could assist humans in the interpretation of CTG.

1. Introduction

Neonatal acidemia is one of the main complications that may arise during delivery and is associated with various fetal impairments such as cerebral palsy, hypoxic-ischemic encephalopathy or even stillbirth [1], thus there is a high potential impact of improving the detection and management of neonatal acidemia in clinical practice.

Cardiotocography (CTG) is defined as the recording of fetal heart rate (FHR) and uterine contractions (UC) during pregnancy using an electronic fetal monitor and is the main tool available to monitor the fetal well-being. In developed countries, CTG is generally monitored continuously during delivery. Presently, obstetricians and midwives primarily rely on visual interpretation based on established guidelines [2]. Although those guidelines are constantly challenged [3,4], CTG interpretation methods have remained largely unchanged since CTG was introduced in the 1960s. CTG interpretation is known to be subject to a significant inter-observer and intra-observer variability [5–7] and the effectiveness of continuous CTG monitoring is still debated [8]. Intrapartum fetal blood sampling or ST waveform analysis of fetal electrocardiogram has been suggested as a second-line test in case of abnormal fetal heart rate patterns [9,10]. In addition to their questionable contribution to reducing poor neonatal outcomes [11,12], these invasive methods are not without risks for the fetus, can be difficult to perform due to a long learning curve and require available medical staff [13,14].

Improving non-invasive CTG monitoring is a key challenge in obstetrics to increase the sensitivity while reducing the rate of false

[*] Corresponding author. . Department of gynecology obstetrics, Assistance Publique des Hôpitaux de Paris -Beaujon, Clichy, 92100, France.
E-mail address: imane.benmbarek@aphp.fr (I. Ben M'Barek).

https://doi.org/10.1016/j.compbiomed.2024.109448
Received 27 August 2024; Received in revised form 14 November 2024; Accepted 14 November 2024

positives associated with unnecessary interventions. Building and deploying models to help practitioners with CTG interpretation is a promising solution to this challenge [15]. The first developed models were based on a quantitative adaptation of the guidelines proposed by the International Federation of Gynecology and Obstetrics (FIGO), like the Omniview-SisPorto and the OxSys systems [16–18].

The availability of clinical datasets [19,20], some of which are published in open-access, as well as the ability to digitize printed CTG traces [21] led to a surge of research on computerized CTG analysis using machine learning techniques [22–26]. In the last years, there have been several publications using deep learning models. Most of them are based on convolutional neural networks [27–29], and some use more complex architectures [30].

However, no model has been shown yet to bring a clear benefit over visual interpretation in a randomized controlled trial [31,32], which proves that despite their promises, such models still need to be rethought and refined. One of the main difficulties associated with building a model is that most available clinical datasets only contain a few hundreds or thousands of births. As neonatal acidemia is a rare event and may result in heterogeneous and patient-specific impact in the CTG patterns, training robust algorithms requires larger datasets. In the present paper, we introduce DeepCTG® 2.0, a model based on a CNN predicting neonatal acidemia during delivery using CTG signals developed and evaluated on a large multicenter clinical database.

2. Material and methods

2.1. Datasets

The study was performed on a multicenter retrospective database including cases from three teaching hospitals of Assistance Publique des Hôpitaux de Paris (APHP), and two open-source datasets (Table I).

The three hospitals were Beaujon-APHP (Clichy, France), Robert-Debré-APHP (Paris, France) and Bichat-APHP (Paris, France), with the following inclusion periods: March 2006–February 2018 in Robert-Debré, January 2013–September 2022 in Bichat, and January 2020–December 2022 in Beaujon.

Two open-source datasets were included as well: the CTU-UHB dataset [33], and the SPaM dataset introduced as part of the Workshop on Signal Processing and Monitoring in Labor [34]. The CTU-UHB dataset contained 552 cases collected at the University Hospital of Brno, with CTG signals, maternofetal data, and fetal outcome. The SPaM dataset contained 300 cases collected from three participating centers (Lyon, Brno and Oxford). Each center provided 100 cases: 80 cases with pH within 7.25–7.30 and 20 with pH $\leq$ 7.05. For every case, the CTG signals and binary outcome are available.

All cases for which the CTG signals, maternofetal data and fetal outcomes were available were included. The CTG signals include the FHR signal and the UC signal, with a 4Hz frequency. Cases with premature birth (gestational age <37 weeks) were excluded, leading to a total number of 27662 cases included. Neonatal acidemia was defined according to arterial pH at birth. Three groups of outcomes were defined: normal, moderate acidemia and severe acidemia, corresponding to pH > 7.20, 7.05<pH $\leq$ 7.20 and pH $\leq$ 7.05 respectively.

2.2. Preprocessing of CTG signals

The signals were first averaged down to a 1Hz frequency. This had no impact on the performance of the model and enabled to decrease the size of the datasets and accelerate the training of the models. This choice is consistent with past studies [28].

The model inputs 60-min CTG segments which were built with the following methodology:

Table 1
Description of the datasets.

	Multicenter APHP dataset			Public datasets		
	Hôpital Beaujon	Hôpital Robert-Debré	Hôpital Bichat	CTU-UHB	SPaM	
Number of cases	383	19194	7242	549	294	**27662**
Fetal outcome: share of cases per pH at birth						
pH $\leq$ 7.05	14 %	1 %	2 %	8 %	20 %	**464**
7.05 < pH $\leq$ 7.20	19 %	8 %	22 %	27 %	0 %	**3457**
pH > 7.20	66 %	91 %	75 %	65 %	80 %	**23741**
Delivery mode: share of cases per delivery mode						
Vaginal	68 %	70 %	72 %	92 %	100 %	
Operative	21 %	21 %	16 %			
Cesarean during labor	10 %	8 %	12 %	8 %		
CTG signals characteristics						
Share of cases with more than 60 min of signal	94.0 %	92.7 %	93.8 %	84.9 %	100.0 %	
Mean share of FHR missing points	7.9 %	9.9 %	8.6 %	15.7 %	7.6 %	
Mean share of UC missing points	23.5 %	8.0 %	16.5 %	21.3 %	7.9 %	

APHP assistance publique des hôpitaux de Paris, CTG cardiotocography, FHR fetal heart rate, UC uterine contractions.

- Missing segments of data lasting less than 10 min were interpolated using linear interpolation. Missing segments of over 10 min were not filled.
- For every case, the latest segment without any missing data (after interpolation) and lasting 10 min at least was selected. If such a segment did not exist, or if it started more than 90 min before delivery, the case was excluded. This led to the exclusion of 74 cases, representing less than 0.3 % of the total number of cases.
- When this segment lasted more than 60 min, we selected the last 60 min. When it lasted less than 60 min, we padded the signals with zeroes on the left. Hence, given our 1Hz sampling, the size of the CTG segments that were analyzed is 3600.

Then, FHR and UC values were normalized by subtracting their mean and dividing by their standard deviation computed over the whole dataset.

2.3. Deep learning model

The classification model inputs a 60-min CTG segment and outputs the probabilities for each of the three possible outcomes (normal, moderate acidemia and severe acidemia).

The model is a CNN with the following architecture:

- The CTG segments of shape (3600, 2) were processed with four convolutional blocks formed with a one-dimensional convolutional layer and a max-pooling layer, parametrized by the number of convolutional kernels and their size. We used 32, 32, 64 and 64 for the number of kernels, and 6, 6, 5, and 5 for their size. The output shape of the processed segments is (4, 64), which is flattened to a vector of shape 256.

- Then, two fully-connected layers with ReLU activation and 256 units were applied successively.
- Finally, a fully-connected layer with Softmax activation and three output units was applied to output the probabilities for the three classes.

The following alternative architectures have been considered to process the CTG segments (the fully-connected layers are left unchanged):

- Convolutional blocks with different numbers of kernels: either two times less kernels (16, 16, 32 and 32) or two times more kernels.
- Stacked Long Short-Term Memory (LSTM) layers: this architecture processes the CTG segments with LSTM layers, a recurrent neural network (RNN) variant that excels in capturing long-range dependencies in sequential data. Stacking several of them allows to learn complex temporal features. We have evaluated this architecture with 64 hidden units, and 1 or 3 stacked layers.
- Hybrid architecture combining convolutional blocks and a transformer: in this architecture, the CTG segments are processed first with two convolutional blocks (both of size 6 with 64 kernels) and then with a transformer, formed with several self-attention layers as described in Vaswani et al. [35]. The convolutional blocks extract local features from the input segments, while the transformer layers handle long-range dependencies via self-attention mechanisms. The application of convolutional blocks early in the processing has two main advantages: it reduces the temporal dimension (decreasing the computational power required by the transformer), and at the same time increases the features dimension (increasing the number of parameters in the transformer and hence its expressive power). We have evaluated this architecture with two sets of parameters for the transformer: 2 self-attention layers, 4 attention heads, 128 units in the feedforward layer, and 4 self-attention layers, 8 attention heads, 256 units in the feedforward layer.

Those different architectures have been evaluated on the Robert Debré dataset (the training was performed on cases from other datasets).

2.4. Training and validation of the model

The model was evaluated separately on the five centers included in the study. For each center, we built a training, a validation and a test dataset the following way:

- Only cases from the four other centers were included in the training and validation datasets. This ensured that the performance was representative of the performance that could be reached when using the model in a new hospital. 80 % of those cases were selected to be included in the training dataset, and the remaining 20 % were selected to be included in the validation dataset. Then, the training and validation datasets were formed with 10000 cases sampled with replacement from those cases: 6000 normal cases, 3000 cases of moderate acidemia and 1000 cases of severe acidemia. This choice enabled to overweight the cases with acidemia (i.e. pH < 7.05) given its low prevalence.
- The test dataset was formed with the cases in the center being evaluated.

The model was trained with the cross-entropy loss commonly used for classification tasks. We used early stopping with a patience parameter equal to three: training was stopped when the cross-entropy loss on the validation dataset did not improve during three consecutive epochs.

We pretrained the convolutional layers on a task which consisted in predicting from the raw CTG segments the value of several features extracted from the segments. The features extracted from the FHR segment are the minimum and maximum value of the baseline, the areas covered by accelerations and decelerations, and the short-term and long-term variabilities of the signal. The features extracted from the UC segment are the number of contractions and the total duration of the contractions. Those features are extracted following the methodology used in the DeepCTG® 1.0 model [29] and fully described in the corresponding paper. Pretraining the convolutional layers on this alternative task presented two advantages: this enabled to use all parts of CTG signals (and not only the segment just before delivery), increasing the size of the dataset by a factor between 5 and 10, and the definition of the task included features known to be linked to neonatal acidemia (Fig. 1). We report the performance of the model with no pretraining and with this pretraining methodology.

2.5. Performance assessment and statistical methods

The receiver operating characteristic (ROC) curve of the model was built on every dataset for two binary classification tasks: the detection of cases of severe acidemia (by merging the two other classes) and the detection of cases of moderate or severe acidemia (by merging those two classes). The area under those ROC curves (AUC) was evaluated, with a 90 % confidence interval estimated by bootstrapping the evaluation datasets with 100 bootstraps (built by drawing cases with replacement). Considering the width of the confidence intervals, AUCs were rounded to two decimal places only. For each dataset the performance reached by another model, DeepCTG® 1.0, was evaluated. This model is based on a logistic regression fed with features extracted from the CTG signals [29].

The impact of the following parameters on the performance of the model was evaluated:

- Signal quality: each one of the three APHP datasets was broken down according to signal quality into three equally sized subsets, based on the proportion of missing values in the FHR segment fed to the model. Those subsets are named “High quality”, “Medium quality” and “Low quality”. The ROC curve of the model was evaluated on each subset to evaluate how the CTG signal quality impacts the performance of the model.
- Sample size and pretraining: models were trained on subsets of the training datasets, built by randomly sampling without replacement 10 %, 50 % or 100 % of the cases. On every subset, two models were trained, with and without pretraining the convolutional layers. This enabled to evaluate how both the size of the training datasets and the pretraining of the convolutional layers impact the performance of the model. For the sake of clarity, this evaluation was done on the Robert Debré dataset only (the training was performed on cases from other datasets).
- Input features: on the Robert Debré dataset, we evaluated the model by using as inputs the FHR signal only, the UC signal only or both signals.
- Obstetric risk factors: the model was trained and evaluated on a low-risk subgroup excluding the following cases: breech presentation, gestational age >42 weeks, maternal body mass index (BMI) > 30 kg/m^2, gestational diabetes, pre-eclampsia, birthweight <10th percentile and >90th percentile, or suspicion of intrauterine infection during labor. This subgroup comprised 16424 cases, including 237 cases (1.4 %) of severe acidemia and 1934 cases (11.8 %) of moderate acidemia. Those exclusion rules are based on variables that are known to influence the interpretation of CTG during labor. A further analysis was conducted on the false negatives in this subgroup, defined as cases for which a normal outcome was predicted while the true outcome was a severe acidemia. By analyzing jointly the FHR and the clinical variables, those cases were carefully classified into five categories: (1) low signal quality (cases with a high share of FHR missing data), (2) bradycardia during pushing (a clear bradycardia was noticed during pushing, but the model

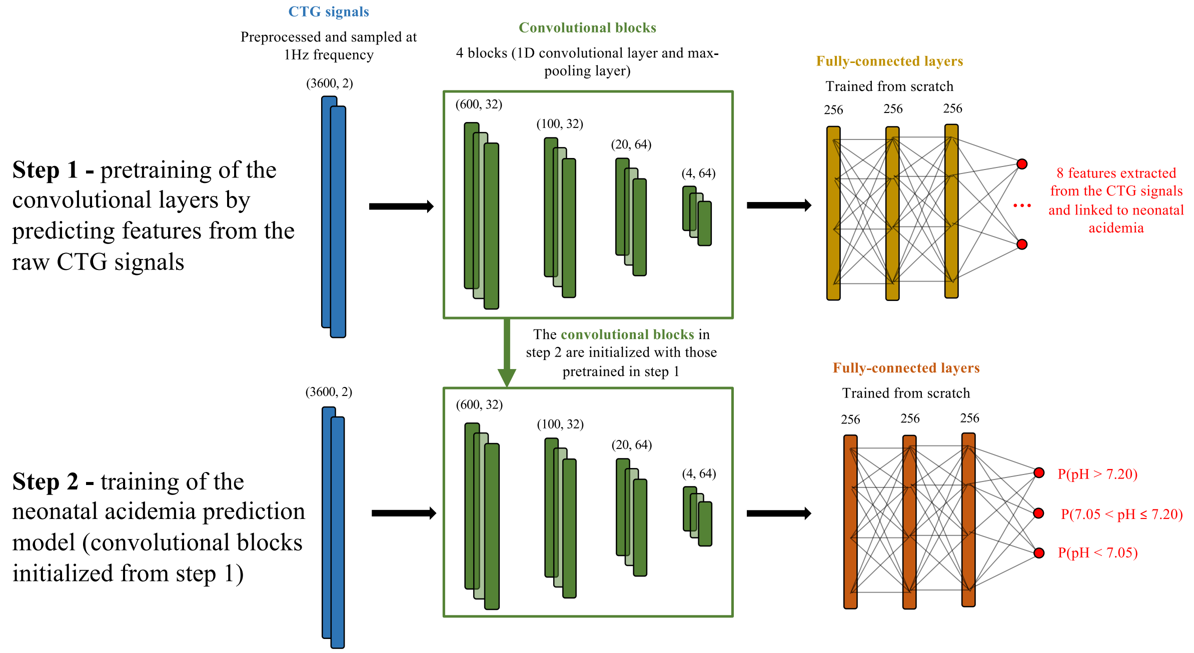
CTG signals
Preprocessed and sampled at 1Hz frequency
(3600, 2)
Convolutional blocks
4 blocks (1D convolutional layer and max-pooling layer)
(600, 32)
(100, 32)
(20, 64)
(4, 64)
Fully-connected layers
Trained from scratch
256
256
256
8 features extracted from the CTG signals and linked to neonatal acidemia
Step 1 - pretraining of the convolutional layers by predicting features from the raw CTG signals
The convolutional blocks in step 2 are initialized with those pretrained in step 1
(3600, 2)
(600, 32)
(100, 32)
(20, 64)
(4, 64)
Fully-connected layers
Trained from scratch
256
256
256
P(pH > 7.20)
P(7.05 < pH ≤ 7.20)
P(pH < 7.05)
Step 2 - training of the neonatal acidemia prediction model (convolutional blocks initialized from step 1)

Fig. 1. Architecture of the model.

misinterpreted it), (3) other FHR anomalies (those anomalies could be detected visually but the model did not detect them), (4) evolution post CTG (cases where the neonatal acidemia appeared after the CTG recording finished and hence could not have been detected by the model), (5) unexplained (neonatal acidemia could not be detected visually by analyzing the case, hence it was hard to explain why the model was wrong).

The model was also evaluated on a low-risk subgroup built using rules based on clinical variables that are known to influence the interpretation of CTG.

2.6. Evaluation of clinical practice

For every case in the three APHP datasets, it is known whether an anomaly in the CTG signals was detected by the practitioners during delivery. The sensitivity and specificity of the detection of moderate and severe acidemia by practitioners were estimated based on this information, enabling to position clinical practice on the ROC curves.

This work had the approval of Robert Debré hospital's Ethical committee (IRB 00006477).

3. Results

3.1. Description of the datasets

The database contains 27662 cases, including 464 cases (1.7 %) of severe acidemia (pH $\leq$ 7.05) and 3457 cases (12.5 %) of moderate acidemia (7.05<pH $\leq$ 7.20). The rate of cesarean delivery per center ranges from 8 % to 12 %. In all datasets, the CTG signals lasted more than 60 min in a large majority of cases (from 84.9 % for the CTU-UHB dataset to 100 % for the SPaM dataset). The proportion of missing data in the FHR signal is 15.7 % in the CTU-UHB dataset and below 10 % in the other datasets.

The low-risk subgroup comprises 16424 cases, including 237 cases (1.4 %) of severe acidemia and 1934 cases (11.8 %) of moderate acidemia (Table II).

Table 2
Description of the low-risk subgroup.

	Multicenter APHP dataset			Public datasets		
	Hôpital Beaujon	Hôpital Robert-Debré	Hôpital Bichat	CTU-UHB	SPaM	
Number of cases	151	11099	4524	356	294	**16424**
Fetal outcome: share of cases per pH at birth						
pH $\leq$ 7.05	7 %	1 %	2 %	7 %	20 %	**237**
7.05 < pH $\leq$ 7.20	21 %	7 %	22 %	28 %	0 %	**1934**
pH > 7.20	72 %	92 %	76 %	65 %	80 %	**14253**

3.2. Training and validation of the model

The model has been trained and evaluated on all datasets. Fig. 2 shows the loss and AUC evaluated on the train and validation datasets at every epoch during training, highlighting the effectiveness of our training methodology to prevent overfitting. The model performs better to detect severe acidemia than moderate acidemia, with a difference in AUC between 0.03 and 0.06 depending on the datasets. The AUC vary significantly between the datasets, from 0.70 to 0.83 in the detection of moderate and severe acidemia, and from 0.74 to 0.83 in the detection of severe acidemia (Fig. 3, Table III). The worse performance is reached on the CTU-UHB dataset, while the best performance occurs with the SPaM dataset. The model performs better than DeepCTG® 1.0 (i.e. a model based on a logistic regression), with an increase in AUC of 0.05. For the SPaM dataset, the detection of moderate and severe acidemia and the detection of severe acidemia only are identical tasks because the dataset did not contain any case of moderate acidemia. Because of the relatively low number of cases of acidemia in the datasets, the confidence intervals are wide, especially for the detection of severe acidemia.

The evaluation of clinical practice gave different results on the three APHP datasets: for the detection of severe acidemia, the sensitivity ranged from 42 % to 67 % and the specificity ranged from 12 % to 20 %. The performance of the model is better for the Bichat dataset, similar for

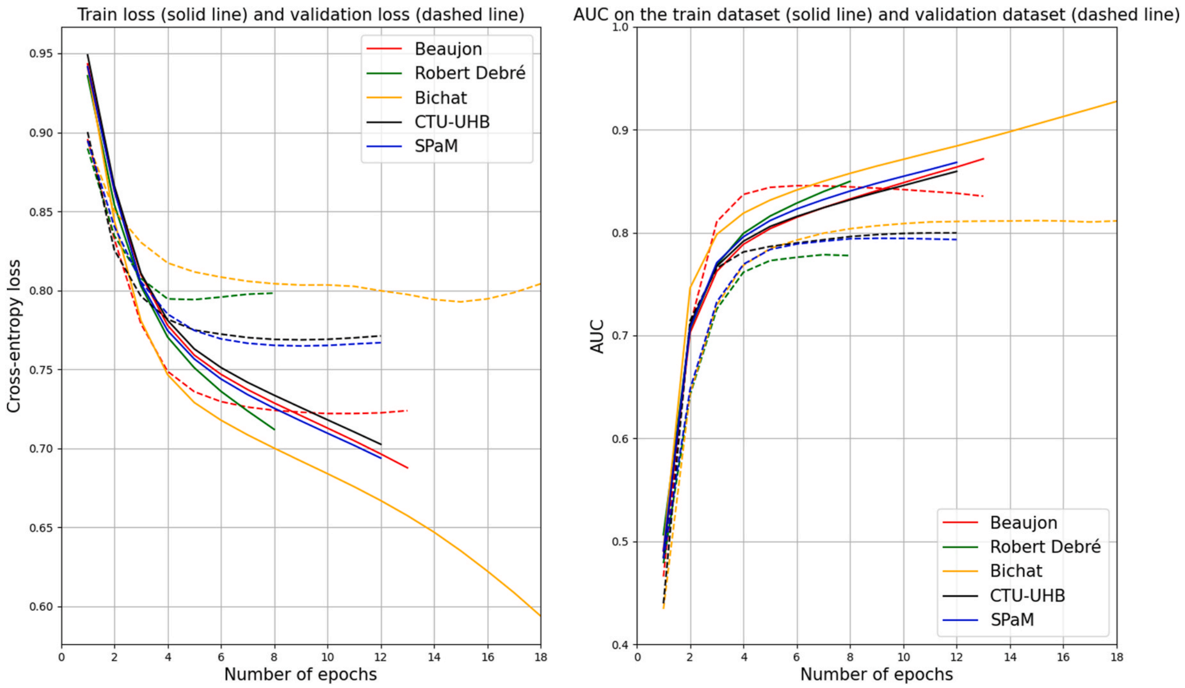

Fig. 2. Loss and AUC on train and validation datasets during training.

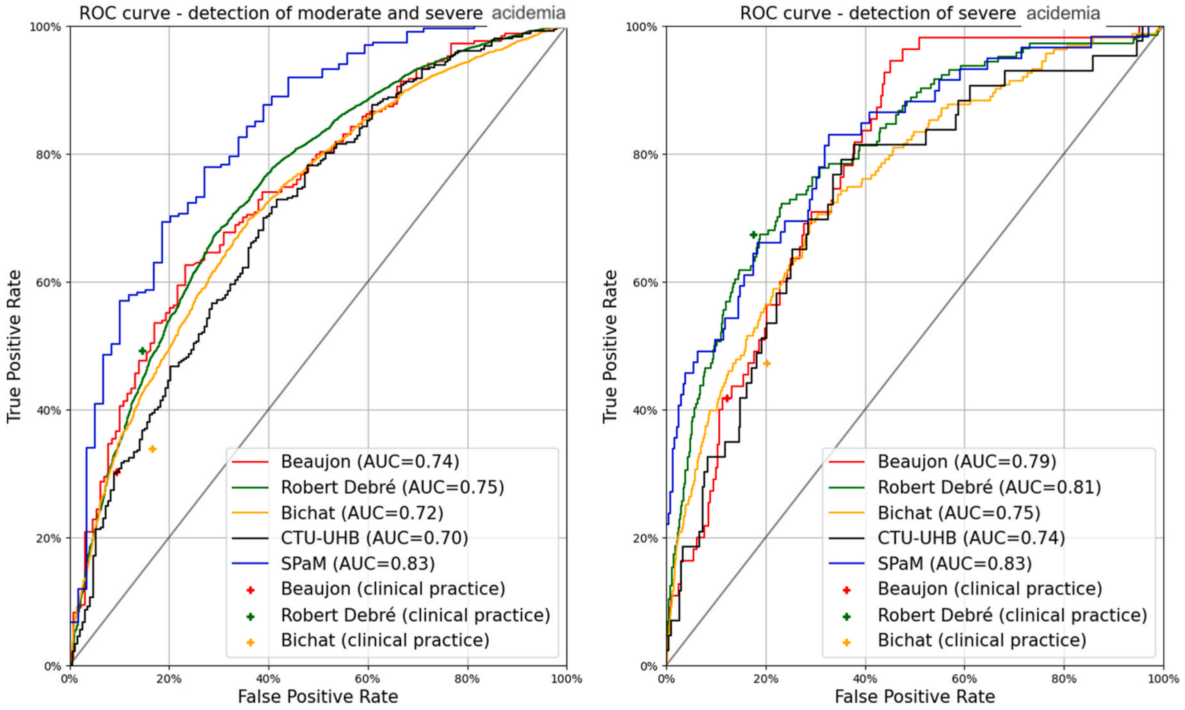

Fig. 3. ROC curve of the model on each center to detect moderate and severe acidemia.

the Beaujon dataset, and worse for the Robert Debré dataset.

3.3. *Comparison of different deep learning architectures*

The models compared in that section were not pretrained (every architecture was trained from scratch). The number of parameters in the part of the model processing the CTG segments (excluding the fully-connected layers) is reported as a measure of the complexity of the model. The CNN with smaller kernels (four times less parameters) have a lower AUC (0.70 compared to 0.72), and using larger kernels (four times more parameters) does not increase the AUC, suggesting that the CNN architecture that was chosen throughout the paper has the right complexity regarding the problem and the size of the datasets. The architectures based on LSTM layers have a significantly lower AUC. The Transformer architecture gives a slightly lower AUC (Table IV).

Table 3
AUC of the model per dataset with 90 % confidence interval.

	DeepCTG® 2.0 (convolutional neural network)		DeepCTG® 1.0 (logistic regression)	
Dataset	AUC to detect moderate and severe acidemia	AUC to detect severe acidemia	AUC to detect moderate and severe acidemia	AUC to detect severe acidemia
Beaujon	0.74 (0.70–0.79)	0.79 (0.74–0.83)	0.71 (0.66–0.76)	0.76 (0.72–0.82)
Robert Debré	0.75 (0.73–0.76)	0.81 (0.78–0.84)	0.72 (0.70–0.73)	0.77 (0.74–0.81)
Bichat	0.72 (0.71–0.73)	0.75 (0.72–0.79)	0.72 (0.70–0.73)	0.73 (0.70–0.76)
CTU-UHB	0.70 (0.67–0.74)	0.74 (0.67–0.81)	0.67 (0.63–0.71)	0.71 (0.64–0.79)
SPaM	0.83 (0.77–0.88)		0.78 (0.74–0.84)	

3.4. *Validation of the model on a low-risk subgroup*

The AUC scores range from 0.65 to 0.85 for detecting moderate and severe acidemia and from 0.68 to 0.85 for detecting severe acidemia alone. The worse performance is obtained on the CTU-UHB dataset. The performance on the low-risk subgroup is similar when the model is trained on the whole dataset or on the low-risk subgroup only (Table V). There are 58 false negatives (cases where the model predicted a normal outcome while the true outcome was a severe acidemia), which are described in Table VI.

3.5. *Impact of missing data*

For the three APHP datasets, the performance of the model significantly increases with the quality of the signal (Fig. 4). The difference in AUC between the "Low quality" and "High quality" cases is 0.06 for the Bichat and Robert Debré datasets and 0.11 for the Beaujon dataset. Clinical practice is also impacted by the signal quality: the sensitivity between the "Low quality" and "High quality" cases ranges from 47 % to 52 % for the Robert Debré dataset, and from 31 % to 38 % for the Bichat dataset, with similar specificities.

3.6. *Impact of pretraining and of the size of the datasets*

The AUC increases with the number of cases used for training, from 0.69 to 0.82 to detect severe acidemia when using 10 % and 100 % of the available cases. The AUC significantly increases when the convolutional layers is pretrained, from 0.72 to 0.75 to detect moderate and severe acidemia, and from 0.78 to 0.82 to detect severe acidemia (Table VII).

3.7. *Impact of the input features fed to the model*

The highest AUC is reached when both FHR and UC signals are used as inputs (Table VIII). As expected, the FHR signal is the most important feature, and a model based on UC signal only as a poor performance (AUC is 0.61 for the detection of moderate and severe acidemia).

4. Discussion

We introduced DeepCTG® 2.0, a convolutional neural network to detect neonatal acidemia from CTG signals. The model was trained and

Table 5
AUC of the model on the low-risk subgroup.

	DeepCTG® 2.0 (convolutional neural network)			DeepCTG® 2.0 (convolutional neural network)	
Dataset	AUC to detect moderate and severe acidemia	AUC to detect severe acidemia	Dataset	AUC to detect moderate and severe acidemia	AUC to detect severe acidemia
Beaujon	0.70 (0.61–0.78)	0.77 (0.66–0.90)	**Beaujon**	0.74 (0.66–0.81)	0.77 (0.60–0.87)
Robert Debré	0.75 (0.74–0.76)	0.82 (0.78–0.87)	**Robert Debré**	0.74 (0.72–0.75)	0.80 (0.75–0.85)
Bichat	0.74 (0.72–0.75)	0.75 (0.71–0.80)	**Bichat**	0.73 (0.71–0.74)	0.76 (0.70–0.80)
CTU-UHB	0.65 (0.60–0.69)	0.68 (0.59–0.78)	**CTU-UHB**	0.66 (0.60–0.72)	0.69 (0.59–0.80)
SPaM	0.85 (0.79–0.89)		**SPaM**	0.84 (0.79–0.87)	
Model trained on whole database/ evaluated on low-risk subgroup			*Model trained on low-risk subgroup/ evaluated on low-risk subgroup*		

Table 6
Categorization of the false negatives in the low-risk subgroup.

Category	Number of cases (%)
Low signal quality (missing data)	14 (24)
Bradycardia during pushing	11 (19)
Other aFHR	12 (21)
Evolution post CTG*	10 (17)
Unexplained	11(19)
Total	**58 (100)**

aFHR abnormal fetal heart rate.
*ombilical cord prolaps, delay between the end of the record and the birth, severe bradycardia.

Table 4
AUC of the model with different deep learning architectures.

Architecture of the model	Number of parameters to process CTG segments (in thousands)	AUC to detect moderate and severe acidemia	AUC to detect severe acidemia		
CNN Number of kernels: 32, 32, 64, 64	37	0.72 (0.71–0.73)	0.78 (0.74–0.81)	0.76 (0.75–0.77)	0.75 (0.72–0.79)
CNN Number of kernels: 16, 16, 32, 32	10	0.70 (0.69–0.72)	0.77 (0.73–0.80)	0.75 (0.74–0.77)	0.74 (0.70–0.77)
CNN Number of kernels: 64, 64, 128, 128	149	0.72 (0.71–0.73)	0.78 (0.74–0.81)	0.76 (0.75–0.77)	0.75 (0.72–0.79)
LSTM (1 layer) 64 hidden units	17	0.67 (0.66–0.68)	0.73 (0.69–0.76)	0.54 (0.53–0.56)	0.60 (0.56–0.66)
LSTM (3 stacked layers) 64 hidden units	84	0.69 (0.68–0.70)	0.75 (0.70–0.77)	0.54 (0.53–0.56)	0.60 (0.56–0.66)
Transformer 2 attention layers, 4 heads, 128 feedforward units	45	0.71 (0.70–0.73)	0.78 (0.73–0.80)	0.74 (0.72–0.75)	0.75 (0.71–0.78)
Transformer 4 attention layers, 8 heads, 256 feedforward units	112	0.69 (0.68–0.71)	0.76 (0.71–0.78)	0.72 (0.71–0.74)	0.73 (0.69–0.76)

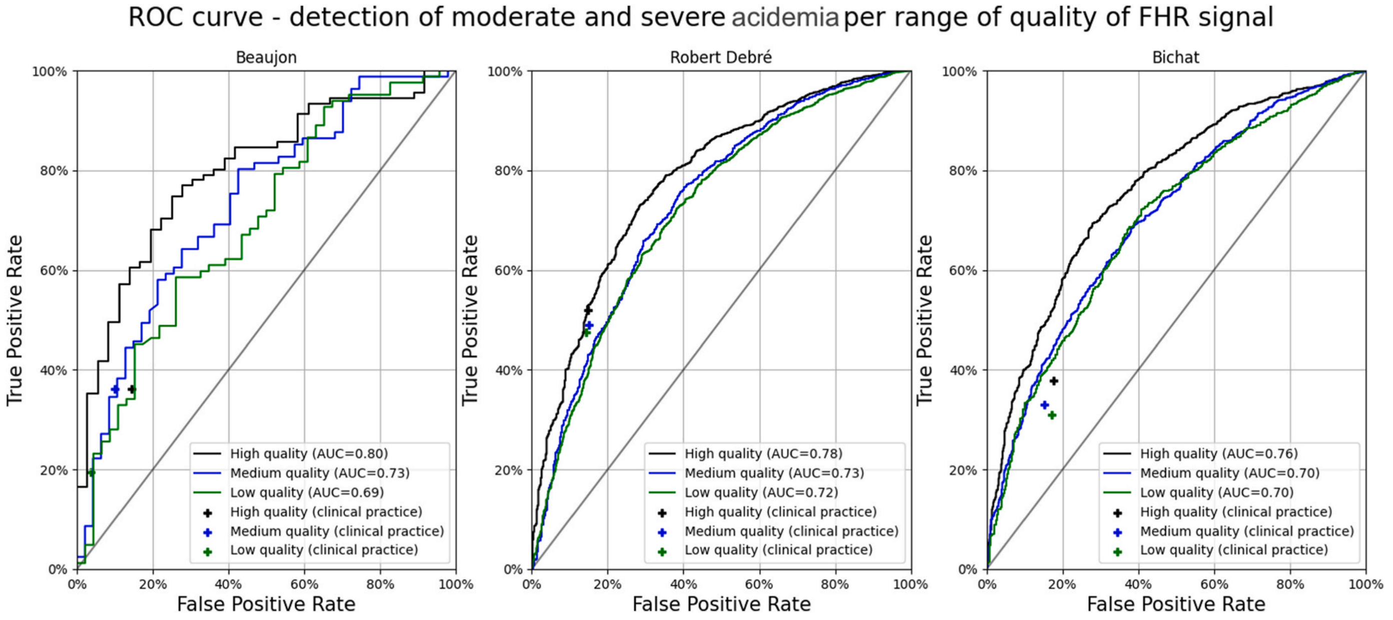

Fig. 4. Impact of signal quality on the ROC curve.

Table 7
AUC of the model with and without pretraining, and as a function of the number of cases included in the training dataset.

	With pretraining		Without pretraining	
Share of cases from the training datasets	AUC to detect moderate and severe acidemia	AUC to detect severe acidemia	AUC to detect moderate and severe acidemia	AUC to detect severe acidemia
10 %	0.68 (0.66–0.69)	0.69 (0.65–0.73)	0.66 (0.65–0.67)	0.67 (0.64–0.70)
50 %	0.74 (0.73–0.75)	0.81 (0.78–0.84)	0.72 (0.70–0.72)	0.76 (0.72–0.80)
100 %	0.75 (0.74–0.76)	0.82 (0.79–0.84)	0.72 (0.71–0.73)	0.78 (0.74–0.81)

Table 8
AUC of the model when fed with FHR signal only, UC signal only, or both signals.

Input signals	AUC to detect moderate and severe acidemia	AUC to detect severe acidemia
FHR only	0.73 (0.71–0.74)	0.79 (0.76–0.82)
UC only	0.61 (0.60–0.63)	0.69 (0.66–0.72)
FHR and UC	0.75 (0.73–0.76)	0.81 (0.78–0.84)

validated on a large multicenter database. In a cross-center validation, the AUC varied from 0.74 to 0.83 between the centers for the detection of severe acidemia, performing better than a simpler logistic regression model. The pretraining of the model on a task which consisted in predicting features known to be associated with neonatal acidemia significantly improved the performance.

The performance of the model significantly varied between the datasets: the highest AUC was achieved on the SPaM dataset, mainly because this dataset did not contain any case of moderate acidemia. The lowest AUC was obtained on the CTU-UHB dataset, probably because of the low quality of the CTG signals in this dataset. Those results should be interpreted cautiously given the large width of the confidence intervals, coming from the relatively low number of pathological cases in the datasets.

The comparison with other published models is challenging as the models are evaluated on different clinical databases or using different outcomes. Several published studies were based on the CTU-UHB dataset [24,36] and reported an AUC between 0.72 and 0.74 [37–40] for the detection of acidemia with pH thresholds of 7.05 or 7.10. Petrozziello et al. reported the highest AUC (0.82), however it was obtained on a subset of the dataset excluding the cases of moderate acidemia [41].

Comparing the model with clinical practice gave different conclusions across datasets, with a significant variability in the sensitivity and specificity in the three centers. The evaluation of clinical practice in Petrozziello et al. [41] also gave different results, with a lower sensitivity around 31 % to detect severe acidemia. Further studies should be led to characterize the cases for which the model performs better than clinical practice and conversely.

The CNN architecture used throughout the paper compared favorably to other kinds of neural networks based on LSTMs (a type of recurrent neural networks) or transformers. Using a larger CNN did not bring any increase in performance at the cost of a higher complexity of the model. It was shown that pretraining the convolutional layers on a task that enabled to use all the available CTG signals (and not only the 60 min before delivery) significantly increased the performance of the model. Further work will be conducted on pretraining, especially to pretrain the layers on a larger dataset containing possibly other timeseries than CTG signals, as suggested by Zerveas et al. [42]. The use of a foundational model for time series like TimeGPT [43] could also be considered.

The specification of the outcome is debated in the literature [15,36,44]. Most studies about computerized CTG use an outcome based on the fetal pH at birth [45], with thresholds between 7.05 and 7.20 [36], and some use the Apgar at one or 5 min [46]. The pH is a poor proxy for fetal compromise and unfavorable neonatal outcomes. However, as it is the most readily available, we will use it to develop a detection algorithm that could be adapted to any better proxy agreed by professionals and patient [34,44]. It could be refined to a composite outcome that includes both clinical and biological variables [44,47]. While most previously published models use two classes (e.g. normal and severe acidemia), we chose to model three (i.e. normal, moderate and severe acidemia). This choice prevented the model from relying on a single threshold and allowed it to trigger various alerts for practitioners.

Although we hypothesized that the model should perform better in a low-risk subgroup, since our model does not yet incorporate clinical covariables and risk-factors for acidemia [37,48], this was not the case, possibly because of the reduction in the sample size used to train the model. A careful analysis of the false negatives in this subgroup highlighted specific areas that could require improvement. While poor signal quality was a significant cause, the model misinterpreted bradycardia in a significant proportion of these false negatives, possibly because of the similarity of this pattern with the normal maternal heart rate [11]. Building specific algorithms to detect maternal heart rate and process the FHR accordingly would help in those cases [49].

Several other areas for improvement of the model will be addressed in future developments. First, we will add clinical variables and risk-factors to the CTG as an input to the model. This has been done in several studies [25,50,51]. Second, a strong correlation was highlighted between the quality of the CTG signals and the performance of the model [41,49], suggesting that advanced techniques for missing data imputation could improve the model, although a previous study by Asfaw et al. [52] shows a modest improvement to the classification model. Finally, to be useful for practitioners as a tool assisting them in the interpretation of CTG, the model should produce interpretable indicators to support the prediction of fetal acidemia (i.e. the proportion of deceleration, the absence of acceleration, bradycardia, or even the proportion of missing data).

5. Conclusion

We validated a deep learning model for the prediction of neonatal acidemia during labor on a large multicenter database with encouraging and robust results. Pretraining the model on a prediction task that leveraged the whole signals and based on expert knowledge of neonatal acidemia significantly increased the performance. Although we acknowledge that significant adjustments are required before such a model can be implemented in clinical practice, this work is a first step to build a tool that could assist humans in the interpretation of CTG.

CRediT authorship contribution statement

Imane Ben M'Barek: Writing – review & editing, Writing – original draft, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Grégoire Jauvion:** Writing – review & editing, Writing – original draft, Software, Methodology, Data curation, Conceptualization. **Jade Merrer:** Writing – review & editing, Project administration, Methodology, Conceptualization. **Martin Koskas:** Writing – review & editing, Data curation, Conceptualization. **Olivier Sibony:** Writing – review & editing, Data curation, Conceptualization. **Pierre – François Ceccaldi:** Writing – review & editing, Data curation. **Erwan Le Pennec:** Writing – review & editing, Validation, Supervision, Methodology, Data curation, Conceptualization. **Julien Stirnemann:** Writing – review & editing, Validation, Supervision, Methodology, Data curation, Conceptualization.

Data availability statement

The datasets presented in this article are not readily available because of legal and ethical reasons.

Funding

This research received no external funding

Declaration of competing interest

Grégoire Jauvion is CEO of Genos Care a company specialized on medical software.

Acknowledgements

The authors want to thank the clinical research and innovation department and the information systems department of APHP for their support.

References

[1] C. Acun, S. Karnati, S. Padiyar, S. Puthuraya, H. Aly, M. Mohamed, Trends of neonatal hypoxic-ischemic encephalopathy prevalence and associated risk factors in the United States, 2010 to 2018, Am. J. Obstet. Gynecol. (22) (2022 Jun 8) S0002–S9378, 00443-4.

[2] FIGO consensus guidelines on intrapartum fetal monitoring: cardiotocography - Ayres-de-Campos, in: International Journal of Gynecology & Obstetrics, Wiley Online Library, 2015 [Internet]. [cited 2021 Jun 30]. Available from: https://obgyn.onlinelibrary.wiley.com/doi/10.1016/j.ijgo.2015.06.020.

[3] Zaima A. Intrapartum Fetal Monitoring Guideline. :33.

[4] D. Ayres-de-Campos, J. Bernardes, F.I.G.O. Subcommittee, Twenty-five years after the FIGO guidelines for the use of fetal monitoring: time for a simplified approach? Int. J. Gynaecol. Obstet. 110 (1) (2010 Jul) 1–6.

[5] S.C. Blackwell, W.A. Grobman, L. Antoniewicz, M. Hutchinson, C. Gyamfi Bannerman, Interobserver and intraobserver reliability of the NICHD 3-tier fetal heart rate interpretation system, Am. J. Obstet. Gynecol. 205 (4) (2011 Oct) 378.e1–378.e5.

[6] L. Hruban, J. Spilka, V. Chudáček, P. Janků, M. Huptych, M. Burša, et al., Agreement on intrapartum cardiotocogram recordings between expert obstetricians, J. Eval. Clin. Pract. 21 (4) (2015 Aug) 694–702.

[7] M'Barek I. Ben, M. Ben, B. Barek, G. Jauvion, E. Holmström, A. Agman, J. Merrer, et al., Large-scale analysis of interobserver agreement and reliability in cardiotocography interpretation during labor using an online tool, BMC Pregnancy Childbirth 24 (1) (2024 Feb 14) 136.

[8] Z. Alfirevic, G.M. Gyte, A. Cuthbert, D. Devane, Continuous cardiotocography (CTG) as a form of electronic fetal monitoring (EFM) for fetal assessment during labour, Cochrane Database Syst. Rev. 2017 (2) (2017 Feb 3) CD006066.

[9] E. Chandraharan, Should national guidelines continue to recommend fetal scalp blood sampling during labor? J. Matern. Fetal Neonatal Med. 29 (22) (2016 Nov) 3682–3685.

[10] B.H. Al Wattar, A. Lakhiani, A. Sacco, A. Siddharth, A. Bain, A. Calvia, et al., Evaluating the value of intrapartum fetal scalp blood sampling to predict adverse neonatal outcomes: a UK multicentre observational study, Eur. J. Obstet. Gynecol. Reprod. Biol. 240 (2019 Sep) 62–67.

[11] M. Tarvonen, J. Markkanen, V. Tuppurainen, R. Jernman, V. Stefanovic, S. Andersson, Intrapartum cardiotocography with simultaneous maternal heart rate registration improves neonatal outcome, Am. J. Obstet. Gynecol. 230 (4) (2024 Apr 1) 379.e1–379.e12.

[12] J.P. Neilson, Fetal Electrocardiogram (ECG) for Fetal Monitoring during Labour, Neilson, JP, Cochrane Library, 2015 [cited 2024 Oct 6]; Available from: https://www.cochranelibrary.com/cdsr/doi/10.1002/14651858.CD000116.pub5/full.

[13] H. Sabir, H. Stannigel, A. Schwarz, T. Hoehn, Perinatal hemorrhagic shock after fetal scalp blood sampling, Obstet. Gynecol. 115 (2 Pt 2) (2010 Feb) 419–420.

[14] T.P. Schaap, K.A. Moormann, J.H. Becker, M.E.M.H. Westerhuis, A. Evers, H.A. A. Brouwers, et al., Cerebrospinal fluid leakage, an uncommon complication of fetal blood sampling: a case report and review of the literature, Obstet. Gynecol. Surv. 66 (1) (2011 Jan) 42–46.

[15] Ben M'Barek I, Jauvion G, Ceccaldi PF. Computerized cardiotocography analysis during labor – A state-of-the-art review. Acta Obstet. Gynecol. Scand. [Internet]. [cited 2022 Dec 21];n/a(n/a). Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/aogs.14498.

[16] D. Ayres-de-Campos, P. Sousa, A. Costa, J. Bernardes, Omniview-SisPorto 3.5 - a central fetal monitoring station with online alerts based on computerized cardiotocogram+ST event analysis, J. Perinat. Med. 36 (3) (2008) 260–264.

[17] A. Georgieva, C.W.G. Redman, A.T. Papageorghiou, Computerized data-driven interpretation of the intrapartum cardiotocogram: a cohort study, Acta Obstet. Gynecol. Scand. (2017 Jul 7).

[18] D. Ayres-de-Campos, M. Rei, I. Nunes, P. Sousa, J. Bernardes, SisPorto 4.0 - computer analysis following the 2015 FIGO Guidelines for intrapartum fetal monitoring, J. Matern. Fetal Neonatal Med. 30 (1) (2017 Jan) 62–67.

[19] Open access intrapartum CTG database | BMC Pregnancy and Childbirth | Full Text [Internet]. [cited 2021 Dec 30]. Available from: https://bmcpregnancychildbirth.biomedcentral.com/articles/10.1186/1471-2393-14-16.

[20] A. Houzé de l'Aulnoit, A. Parent, S. Boudet, B. Rogoz, R. Demailly, R. Beuscart, et al., Development of a comprehensive database for research on foetal acidosis, Eur. J. Obstet. Gynecol. Reprod. Biol. 274 (2022 Jul 1) 40–47.

[21] Z. Cömert, A. Şengür, Y. Akbulut, Ü. Budak, A.F. Kocamaz, V. Bajaj, Efficient approach for digitization of the cardiotocography signals, Phys. Stat. Mech. Appl. 537 (2020 Jan 1) 122725.

[22] M.A. Gatellier, J.D. Jonckheere, L. Storme, V. Houfflin-Debarge, L. Ghesquière, C. Garabedian, Fetal heart rate variability analysis for neonatal acidosis prediction, J. Clin. Monit. Comput. (2020).

[23] P. Abry, J. Spilka, R. Leonarduzzi, V. Chudáček, N. Pustelnik, M. Doret, Sparse learning for Intrapartum fetal heart rate analysis, Biomedical Physics & Engineering Express (2018 May 4).

[24] I. Ben M'Barek, G. Jauvion, J. Vitrou, E. Holmström, M. Koskas, P.F. Ceccaldi, DeepCTG® 1.0: an interpretable model to detect fetal hypoxia from cardiotocography data during labor and delivery, Frontiers in Pediatrics [Internet] (2023) [cited 2023 Sep 25];11. Available from: https://www.frontiersin.org/articles/10.3389/fped.2023.1190441.

[25] A. Houzé de l'Aulnoit, M. Génin, S. Boudet, R. Demailly, C. Ternynck, G. Babykina, et al., Use of automated fetal heart rate analysis to identify risk factors for umbilical cord acidosis at birth, Comput. Biol. Med. 115 (2019 Dec) 103525.

[26] W. Alsaggaf, Z. Cömert, M. Nour, K. Polat, H. Brdesee, M. Toğaçar, Predicting Fetal Hypoxia Using Common Spatial Pattern and Machine Learning from Cardiotocography Signals, 2020.

[27] J. Ogasawara, S. Ikenoue, H. Yamamoto, M. Sato, Y. Kasuga, Y. Mitsukura, et al., Deep neural network-based classification of cardiotocograms outperformed conventional algorithms, Sci. Rep. 11 (2021 Jun 28) 13367.

[28] A. Mohannad, C. Shibata, K. Miyata, T. Imamura, S. Miyamoto, H. Fukunishi, Predicting high risk birth from real large-scale cardiotocographic data using multi-input convolutional neural networks. Nonlinear Theory and its Applications, IEICE 12 (3) (2021) 399–411.

[29] A. Petrozziello, C.W.G. Redman, A.T. Papageorghiou, I. Jordanov, A. Georgieva, Multimodal convolutional neural networks to detect fetal compromise during labor and delivery, IEEE Access 7 (2019) 112026–112036.

[30] M. Liu, Y. Lu, S. Long, J. Bai, W. Lian, An attention-based CNN-BiLSTM hybrid neural network enhanced with features of discrete wavelet transformation for fetal acidosis classification, Expert Syst. Appl. 186 (2021 Dec 30) 115714.

[31] P. Brocklehurst, D.J. Field, E. Juszczak, S. Kenyon, L. Linsell, M. Newburn, et al., The INFANT trial, Lancet 390 (10089) (2017 Jul 1) 28.

[32] I. Nunes, D. Ayres-de-Campos, A. Ugwumadu, P. Amin, P. Banfield, A. Nicoll, et al., Central fetal monitoring with and without computer analysis: a randomized controlled trial, Obstet. Gynecol. 129 (1) (2017 Jan) 83–90.

[33] V. Chudáček, J. Spilka, M. Burša, P. Janků, L. Hruban, M. Huptych, et al., Open access intrapartum CTG database, BMC Pregnancy Childbirth 14 (2014 Jan 13) 16.

[34] A. Georgieva, P. Abry, V. Chudáček, P.M. Djurić, M.G. Frasch, R. Kok, et al., Computer-based intrapartum fetal monitoring and beyond: a review of the 2nd Workshop on signal processing and monitoring in labor (october 2017, oxford, UK), Acta Obstet. Gynecol. Scand. 98 (9) (2019) 1207–1217.

[35] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, et al., Attention is all you need [internet], arXiv (2023) [cited 2024 Jan 10]. Available from: http://arxiv.org/abs/1706.03762.

[36] L. Mendis, M. Palaniswami, F. Brownfoot, E. Keenan, Computerised cardiotocography analysis for the automated detection of fetal compromise during labour: a review, Bioengineering 10 (9) (2023 Sep) 1007.

[37] M.A. Gatellier, J.D. Jonckheere, L. Storme, V. Houfflin-Debarge, L. Ghesquière, C. Garabedian, Fetal heart rate variability analysis for neonatal acidosis prediction, J. Clin. Monit. Comput. (2020).

[38] P. Abry, J. Spilka, R. Leonarduzzi, V. Chudáček, N. Pustelnik, M. Doret, Sparse learning for Intrapartum fetal heart rate analysis, Biomedical Physics & Engineering Express 4 (3) (2018 May) 034002.

[39] J. Ogasawara, S. Ikenoue, H. Yamamoto, M. Sato, Y. Kasuga, Y. Mitsukura, et al., Deep neural network-based classification of cardiotocograms outperformed conventional algorithms, Sci. Rep. 11 (1) (2021 Jun 28) 13367.

[40] P. Fergus, C. Chalmers, C.C. Montanez, D. Reilly, P. Lisboa, B. Pineles, Modelling segmented cardiotocography time-series signals using one-dimensional convolutional neural networks for the early detection of abnormal birth outcomes, IEEE Transactions on Emerging Topics in Computational Intelligence 5 (6) (2021 Dec) 882–892.

[41] A. Petrozziello, C.W.G. Redman, A.T. Papageorghiou, I. Jordanov, A. Georgieva, Multimodal convolutional neural networks to detect fetal compromise during labor and delivery, IEEE Access 7 (2019) 112026–112036.

[42] G. Zerveas, S. Jayaraman, D. Patel, A. Bhamidipaty, C. Eickhoff, A Transformer-Based Framework for Multivariate Time Series Representation Learning [Internet]. arXiv, 2020 [cited 2024 Jan 10]. Available from: http://arxiv.org/abs/2010.02803.

[43] A. Garza, M. Mergenthaler-Canseco, TimeGPT-1 [Internet]. arXiv, 2023 [cited 2024 Jan 10]. Available from: http://arxiv.org/abs/2310.03589.

[44] J. Savchenko, P.G. Lindqvist, S. Brismar Wendel, Comparing apples and oranges? Variation in choice and reporting of short-term perinatal outcomes of term labor: a systematic review of Cochrane reviews, Eur. J. Obstet. Gynecol. Reprod. Biol. 276 (2022 Sep) 1–8.

[45] L.C. Gilstrap, K.J. Leveno, J. Burris, M. Lynne Williams, B.B. Little, Diagnosis of birth asphyxia on the basis of fetal pH, Apgar score, and newborn cerebral dysfunction, Am. J. Obstet. Gynecol. 161 (3) (1989 Sep 1) 825–830.

[46] A. Mohannad, C. Shibata, K. Miyata, T. Imamura, S. Miyamoto, H. Fukunishi, et al., Predicting high risk birth from real large-scale cardiotocographic data using multi-input convolutional neural networks. Nonlinear Theory and its Applications, IEICE 12 (3) (2021) 399–411.

[47] P. Olofsson, Umbilical cord pH, blood gases, and lactate at birth: normal values, interpretation, and clinical utility, Am. J. Obstet. Gynecol. 228 (5) (2023 May 1) S1222–S1240.

[48] A.M. Vintzileos, J.C. Smulian, Abnormal fetal heart rate patterns caused by pathophysiologic processes other than fetal acidemia, Am. J. Obstet. Gynecol. (2023 Mar 17) [Internet], [cited 2023 Apr 12];0(0). Available from: https://www.ajog.org/article/S0002-9378(22)00346-5/fulltext.

[49] S. Boudet, A. Houzé de l'Aulnoit, L. Peyrodie, R. Demailly, D. Houzé de l'Aulnoit, Use of deep learning to detect the maternal heart rate and false signals on fetal heart rate recordings, Biosensors 12 (9) (2022 Sep) 691.

[50] A. Georgieva, C.W.G. Redman, A.T. Papageorghiou, Computerized data-driven interpretation of the intrapartum cardiotocogram: a cohort study, Acta Obstet. Gynecol. Scand. 96 (7) (2017 Jul) 883–891.

[51] Spilka J, Leonarduzzi R. Fetal Heart Rate Classification: First vs. Second Stage of Labor.

[52] D. Asfaw, I. Jordanov, L. Impey, A. Namburete, R. Lee, A. Georgieva, Fetal heart rate classification with convolutional neural networks and the effect of gap imputation on their performance, in: G. Nicosia, V. Ojha, E. La Malfa, G. La Malfa, P. Pardalos, Fatta G. Di, et al. (Eds.), Machine Learning, Optimization, and Data Science, Springer Nature Switzerland, Cham, 2023, pp. 459–469 (Lecture Notes in Computer Science).