Multimodal Convolutional Neural Networks to detect fetal compromise during labor and delivery

Alessio Petrozziello, Christopher Redman, Aris Papageorghiou, Ivan Jordanov, Antoniya Georgieva

***Abstract* — The gold standard to assess whether a baby is at risk of oxygen deprivation during childbirth, is monitoring continuously the fetal heart rate with cardiotocography (CTG). The aim is to identify babies that could benefit from an emergency operative delivery (e.g., Cesarean section), in order to prevent death or permanent brain injury. The long, dynamic and complex CTG patterns are poorly understood and known to have high false positive and false negative rates. Visual interpretation by clinicians is challenging and reliable accurate fetal monitoring in labor remains an enormous unmet medical need.**

In this work, we applied deep learning methods to achieve data-driven automated CTG evaluation. Multimodal Convolutional Neural Network (MCNN) and Stacked MCNN models were used to analyze the largest available database of routinely collected CTG and linked clinical data (comprising more than 35000 births). We also assessed in detail the impact of the signal quality on the MCNN performance.

On a large hold-out testing set from Oxford (n = 4429 births), MCNN improved the prediction of cord acidemia at birth when compared with *Clinical Practice* and previous computerized approaches. On two external datasets, MCNN demonstrated better performance compared to current feature extraction-based methods.

Our group is the first to apply deep learning for the analysis of CTG. We conclude that MCNN hold potential for the prediction of cord acidemia at birth and further work is warranted. Despite the advances, our deep learning models are currently not suitable for the detection of severe fetal injury in the absence of cord acidemia – a heterogeneous, small, and poorly understood group. We suggest that the most promising way forward are hybrid approaches to CTG interpretation in labor, in which different diagnostic models can estimate the risk for different types of fetal compromise, incorporating clinical knowledge with data-driven analyses.

***Index Terms*— Clinical decision making, Deep learning, Convolutional Neural Networks, Fetal heart rate, Sensitivity, Specificity.**

I. Introduction

During labor, materno-fetal respiratory exchange is transiently compromised by uterine contractions leading to reduced oxygen supply to the fetus. The fetus responds by adjusting its cardiac output, redistributing blood to prioritize the heart and brain, and adapting metabolically. The failure of oxygen delivery can cause fetal brain injury or even death. Such events are usually associated with changes in the fetal heart rate (FHR). Because of this, it is recommended that the fetal heart rate is monitored during labor to detect FHR abnormalities, which may in turn reduce adverse outcomes related to oxygen deprivation (hypoxia) [1]. Most women in high income countries will have continuous monitoring using a cardiotocogram (CTG, Figure 1), which continuously displays the FHR alongside uterine contractions.

In practice, the CTG is examined visually in real time, to identify those babies that may benefit from emergency delivery (Cesarean or instrumental vaginal birth). The CTG signals are complex and reflect periodic changes in the fetal sleep state, responses to the stresses of uterine contractions, responses to maternal position, anesthesia, pregnancy complications, infection, and stage of labor, in addition to patterns that reflect severe oxygen deprivation.

There has been little progress in monitoring the health of babies in labor over the past several decades [2] [3]. In the UK alone, during labor at term, about 100 healthy babies die and about 1100 sustain brain injury each year [4] [5] [6]. Globally, of the approximately 2.6 million stillbirths that occurred in 2015, most of those that occurred during childbirth are considered to be preventable with CTG monitoring and appropriate intervention [7]. It must also be noted that due to the high false positive rate, performing CTG is also associated with harm due to unnecessary interventions. Therefore, the challenge is how CTG monitoring in labor can be improved to maximize sensitivity, while reducing the false positive rate.

Current clinical knowledge on how to interpret the CTG stems from basic animal research [8]; but more recently computerized versions of expert clinical interpretation have been developed [9] [10]. Their aim is to improve consistency of the interpretation by substituting the subjective assessment (with its intrinsically poor inter-observer agreement) with objective pattern recognition.

Submission date: XX/XX/XXXX

This independent research is supported by the National Institute for Health Research (NIHR), Dr Georgieva, CDF-2016-09-004. The views expressed here are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health.

A. Georgieva, C. Redman and A. Papageorghiou are with the Nuffield Department of Women's & Reproductive Health, University of Oxford, UK. (corresponding author phone: +44(0)1865857854; fax: +44(0)1865769141; e-mail: Antoniya.georgieva@wrh.ox.ac.uk; aris.papageorghiou@wrh.ox.ac.uk, christopher.redman@wrh.ox.ac.uk).

A. Petrozziello and I. Jordanov are with the School of Computing, University of Portsmouth, UK. (alessio.petrozziello@port.ac.uk, ivan.jordanov@port.ac.uk).

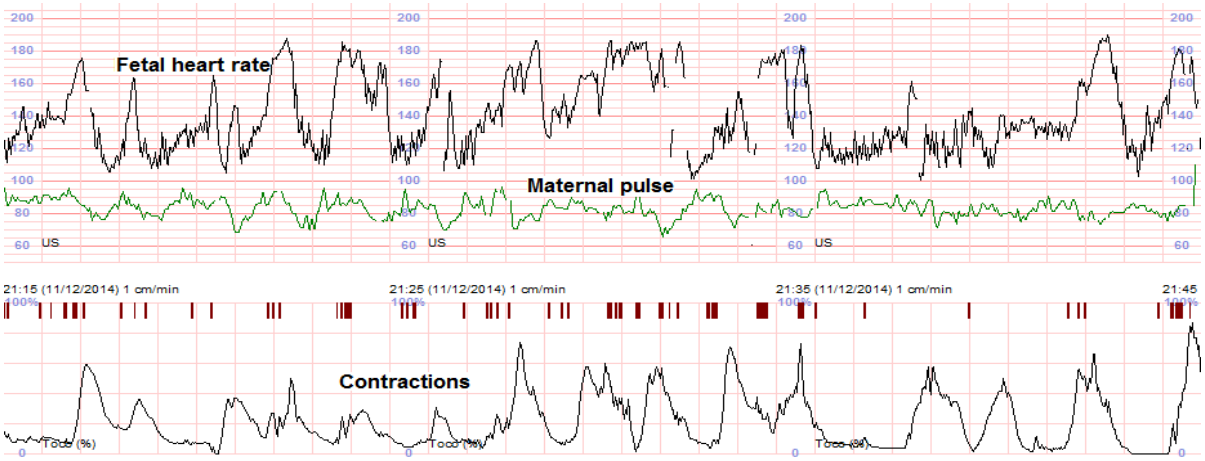


Figure 1 Cardiotocogram (CTG) in labor (a 30min snippet).

In randomized clinical trials two such systems, in their current form, showed no benefit over standard visual CTG interpretation. [9] [10] They were designed to replicate "expert opinion", which in itself is limited. For this and other reasons [11], the negative results do not prove that computerized analysis has nothing to offer, but that it needs to be rethought. For example, the latest data-driven methods can use more sophisticated signal processing to extract features that are associated with adverse outcome. Modern classifiers have been used, such as Bayesian Support Vector Machines [12] [13], classic Artificial Neural Networks [14], and 'sparse learning' approaches [15]. These methods and models allow to go beyond what is classically observed in the CTG by eye [11]

One of the difficulties associated with such approaches is that most clinical datasets contain only a few hundred or a few thousands births at best [16] [17]; as fetal compromise is rare, and as the signals associated with adverse outcome are heterogeneous and patient-specific, small datasets mean that training robust algorithms is very challenging.

The investigation presented here arises from our prior work with the Oxford digital cohort, which is unique in its detail and, to our knowledge, is over ten times larger than any other CTG database. We have already developed a basic prototype diagnostic system (OxSys 1.5) that objectively quantifies the CTG in the context of clinical risk factors; and relates these to perinatal outcome [18]. OxSys 1.5 compares favourably to clinical assessment on retrospective data with a higher *Sensitivity* for fetal compromise (37.6% vs. 32.2%, $p<0.05$) and higher *Specificity* (85.5% vs. 83.6%, $p<0.001$). It is a relatively simple system that employs only two FHR features and two clinical risk factors [18]. The main CTG feature used by OxSys 1.5 is the decelerative capacity (DC) of the phase rectified signal averaging algorithm – a combined measure of the frequency, depth, and slope of any dips in the fetal heart rate [18] [19]. However, the size of our database confers scope for substantial improvement of OxSys.

Deep Learning methods have been successful in various real-world applications by 'learning' the most relevant, unbiased information from large datasets [20] [21] [22]. Hence, our aim was to apply Deep Learning to interrogate our CTG archive and establish optimal ways to classify the CTG into 'high' and 'low' risk. We recently presented our initial simulations and experiments of applying Long Short Term Memory (LSTM) and Convolutional Neural Networks (CNN) to CTG assessment [23]. We demonstrated that CNN compared favourably to LSTM. The LSTM is generally more suitable for forecasting patterns rather than classifying them, and there were also vanishing gradient problems during back-propagation when learning on long CTG records. On the other hand, CNN worked effectively with prolonged data through the use of moving filters and max-pooling. Despite the fact that traditionally CNN are applied successfully for image recognition [24], CNN has also shown promising results for time series, such as in our preliminary work, as well as the analysis of neonatal EEG to detect seizures [25]. The principle and implementation are the same as with standard CNNs for image classification, but for time-series, 1-dimensional convolutions are used instead of two-dimensional matrices.

In this paper, we focused further on the use of CNNs to detect indications of fetal hypoxia by: (1) introducing multimodal CNN (MCNN) for CTG interpretation allowing the network to easily scale in the number and type of input signal; (2) including an array containing information about the CTG signal loss and investigating in detail its impact on the models' performance; (3) developing *Stacked* MCNN to analyze separately and link sequentially the CTG, before and after the onset of active pushing (the second dynamic stage of labor, when the baby is delivered over a relatively short time), usually less than one hour; (4) validating the performance of our models using external multicenter datasets.

II. Data and Methods

We investigated and implemented machine learning algorithms that are using a total of 35429 births, a subset of the Oxford archive (UK, Figure 2). These were split into 85% training and 15% testing sets. The testing subset was identified by a random selection of 15% of cases within each outcome group, ensuring similar rates of compromise in training and testing. The algorithms were then tested on external datasets from hospitals in Lyon (France) and Brno (Czech Republic).

A. Oxford Data

The Oxford archive [18] comprises data from all women and their babies undergoing monitoring during labor at the John Radcliffe Hospital, Oxford, UK, between 1993 and 2011 that met the following inclusion criteria (n = 58748 births):

- Delivery at 36 weeks gestation or more;
- CTG in labor comprising fetal heart rate and contractions (Figure 1), longer than 15minutes, ending within three hours of birth (98% of the Oxford CTGs end within the hour; 92% within the 30min; and 86% within the 10min preceding birth).

In this particular study, in order to define precise outcome groups of interest, we have selected those who also had:

- Validated cord blood gas analysis immediately after birth as an indicator of fetal blood oxygenation. In practice, the acidity of the blood, measured by pH, is the available index of an increased risk for long term compromise of the baby [26]. Cord gases were analyzed at the discretion of the clinician – in about 65% of all continuously monitored births in our unit.

Excluded were babies with breech presentation and congenital abnormalities. The inclusion/exclusion criteria resulted in 35429 births with CTG in labor and clinical details of the labor outcome (Figure 2). The births were grouped in five exclusive groups according to the outcome of labor, defined using the clinical presentation as well as the values of cord arterial pH at birth:

- Severe compromise (a composite outcome of: stillbirth; neonatal death; neonatal encephalopathy; intubation or cardiac massage followed by admission to neonatal intensive care for≥48 hours) and cord pH<7.05 (acidemia);
- Severe compromise and cord pH≥7.05 (no acidemia);
- Moderate compromise: arterial cord pH below 7.05;
- Intermediate: arterial cord pH≥7.05 and <7.15;
- Normal: arterial cord pH≥7.15.

In the cases with severe compromise but no acidemia, the role of oxygen deprivation during labor is debated. How the compromise occurs and whether it is *visible* in the CTG is not well established. Also, the *Intermediate* group comprises a 'middle ground' that takes into account the fact that poor outcome is part of a spectrum that evolves during labor, and represents stress, which is not necessarily abnormal. *Intermediate* cases as well as those with compromise but no acidemia are typically excluded from consideration in CTG research (see also the Discussion below). Nevertheless, for completeness, we also report here the main result for an *Additional Testing Set* of 885 CTGs. The *Main Testing Set* (hold-out) then includes 4429 CTGs. This dataset matches the selection criteria of the external datasets, allows comparison to be made.

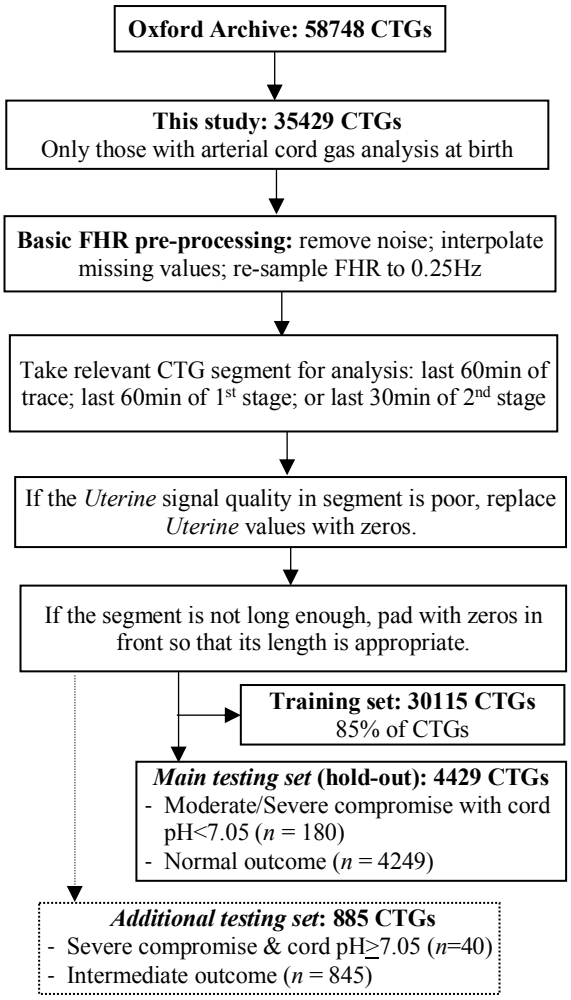


Figure 2 Data preparation before training and testing. First and second stage refer to the different stages of labor – prior to and after the onset of pushing.

The CTG data was originally available at 4Hz for the fetal heart rate and 2Hz for the uterine signal (as default output from the monitors). Basic pre-processing was applied as described in [23]: abrupt increases/decreases were removed and missing values were linearly interpolated. The signals were then averaged down (i.e. smoothed) to 0.25Hz as a standard sampling rate for most OxSys algorithms and computerized antepartum or intrapartum CTG analysis [23]. The original 4Hz sampling rate is too frequent given that the average fetal heart rate beats less often than 3 times a second (<180bpm). To allow computationally reasonable timeframe, we settled on 0.25Hz for this particular study, but future work could examine different sampling rates. If the missing values were at the beginning of trace, they were coded as zeroes. As a result, one hour of data corresponded to 900 heart rate and 900 contraction signal samples. Thirty minutes corresponded to 450 signal samples each.

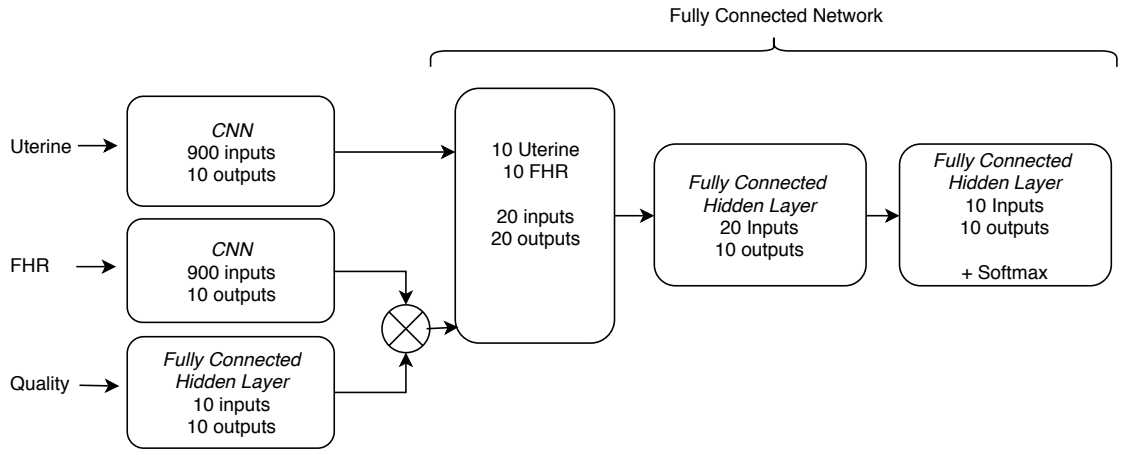


Figure 3 Multimodal Convolutional Network topology. Uterine – time series showing contractions; FHR – fetal heart rate time series; Quality – 10-dimentional array showing the amount of signal loss in the FHR for a sliding 15min window; MCNN – multimodal convolutional neural network. Both convolutional layers take 900 data points as input and output 10 values.

B. External datasets from Lyon and Brno

We also tested the methods on two external datasets:

1) The Signal Processing and Monitoring (SPaM) in Labor Workshop 2017 database

The SPaM dataset is available at the Workshop webpage[1] where full details of its characteristics are given). It comprises monitoring data of 300 women in labor, collected from the three participating centres (Lyon, Brno and Oxford). Each centre provided 100 cases: 80 with normal pH and 20 with pH<7.05, i.e., case sets were selected specifically to have a higher than usual rate of cases with fetal compromise. We tested only with the 200 SPaM cases from Lyon and Brno to ensure that it was truly independent data.

Four established and well-documented algorithms for computerized CTG analysis were tested on this data and reported previously [11]. The four algorithms performed comparably: on the Lyon subset, the median True Positive Rate (TPR) for methods was 77.5% at median False Positive Rate (FPR) of 24%; and on the Brno subset, the median TPR was 55% at median FPR of 28.5%.

2) The Czech Technical University / University Hospital Brno (CTU-UHB)

The CTU-UHB [27] comprises 552 cases of which 40 (7%) have cord acidemia at birth below 7.05. We refer to the details provided in [27] and two published methods reporting results on the CTU-UHB database (even though the data were not strictly used as an unseen hold-out testing set in those methods) and compared our models to them: Spilka et al [28] had 40% TPR at 14% FPR, and Georgoulas et al [29] had 72%TPR at 35%FPR.

C. Development of deep learning models

To tackle the problem of imbalanced training dataset (4% compromised babies vs. 96% healthy ones), we used a weighted binary cross-entropy error: data were weighted in such a way that one misclassification from the compromised group contributed to the error as much as 24 misclassifications from the healthy group (reflecting the incidence of 1 in 24 of compromised cases in our data).

We also tested other approaches to overcome this problem, namely down-sampling and bootstrapping techniques; however, this resulted in worse generalization performance on the new data (data not shown).

For all models, we used *Bayesian* optimization with *Gaussian Process*, a popular model for parameter optimization [30], to maximize the models' TPR at 15% FPR.

D. Multimodal Convolutional Neural Networks (MCNN)

We proposed a multimodal Convolutional Neural Network (MCNN), comprising different input layers and independent learning branches (Figure 3). The MCNN allowed us to use a variety of input sources: FHR, uterine contractions, and a FHR quality score vector (comprised of ten signal quality scores).

The ten signal quality scores were each calculated on a 15 minute moving window, with a 5 minute step over 60 minute FHR. The raw 4Hz data was used to calculate the ratio of valid signal data points, out of the total number of signal points [18].

The heart rate and uterine signals were fed into two distinct 12-layer convolutional networks branches, while the FHR quality vector was used as a score multiplier of the FHR convolutional branch, giving a weight for each output.

We assessed the quality of the CTG contraction signals by an established autoregressive model [31], imposing the following restriction: longer than 40 minutes of acceptable quality, of which more than 20 minutes of excellent quality. We found in our preliminary work [23], that the classification results were improved when only the uterine signals that met this condition were used in the networks. Where they did not, the data were input into the network as zeros (Figure 3 and Figure 4). In effect, the data were tagged as missing (zero entries) and did

[1] https://www.wrh.ox.ac.uk/research/spam-in-labour

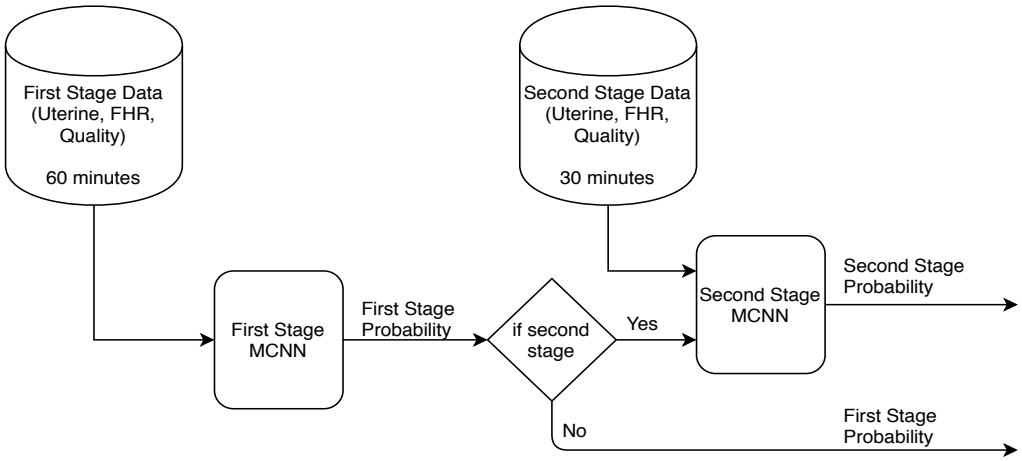


Figure 4 Stacked MCNN topology for 1st and 2nd stage classification. Uterine – time series showing contractions; FHR – fetal heart rate time series; Quality – 10-dimentional array showing the amount of signal loss in the FHR for a sliding 15min window; MCNN – multimodal convolutional neural network. First and second stage refer to the different stages of labor – prior to and after the onset of pushing.

not contribute to the data analysis in as many as 76% of the 35429 CTGs (i.e. only 24% of all CTGs had sufficient quality of the uterine activity signal). When applied to clinical practice, the quality of contraction signal would be checked before analysis, and if it were poor, only the FHR signal would be analyzed.

Batch normalization and dropout were also used through the network [32]. A *Softmax* transformation was included as the last layer of the network architecture, in order to get the class probability of each sample. The convolutional layer hyper-parameters (e.g., number of filters and filter length) were independently optimized for each layer, granting more flexibility during the network creation when compared to our prior model.

E. Stacked MCNN

The end of the CTG often coincides with the time of birth and thus, for classification, would be expected to yield the most relevant data for predicting outcome. But, from the clinical point of view, it is too late to alert the caregiver for the need of intervention.

To address this problem, we split the time series into two parts: (1) the last 60 minutes of the 1st stage of labor (900 FHR data points); and (2) the last 30 minutes of the 2nd stage of labor (450 FHR data points). The onset of 2nd stage of labor was documented by the attending clinician as part of standard clinical care, namely by full cervical dilatation. We only considered 30 minutes in the 2nd stage of labor because significant physiological changes are expected in a shorter time span and because often the second stage does not last longer than 30 minutes. Deliveries with less than 900 and 450 FHR data points for the 1st and 2nd stage respectively, were zero padded at the front.

In the Stacked MCNN, the class probability from the MCNN applied to the 1st stage of labor was used as additional input to the MCNN analyzing the 2nd stage of labor (Figure 4). The *Stacked* MCNN was then tested and, if the baby was delivered by intervention in the 1st stage of labor and thus had no monitoring in the 2nd stage, the probability output of the first MCNN was considered as the relevant MCNN's outcome prediction for this baby.

In particular, to investigate the effect of the stage of labor on the network performance, the MCNN was trained and tested only on data from the first or second stages separately. Secondly, we trained a simple *Stacked* MCNN as shown in Figure 4, using the MCNN model trained on the 1st stage data to generate the probability for compromise and then fed this as an additional feature into the 2nd stage MCNN (trained and tested on the 2nd stage with probability input from 1st stage when available). Second stage data was not available in 29% of the traces (i.e., there was a Cesarean section in the first stage), and the probability generated from the first stage analysis was used for the final classification.

F. Comparison Methods

We compared the models' performance to three other modalities of fetal monitoring: *Clinical Practice*, OxSys 1.5, and our prior work with the single channel CNN:

1) Clinical Practice:

The primary reason for operative delivery (*Cesarean, forceps* or *ventouse* delivery) was noted in the patient records by the attending clinician at the time of birth, when applicable. We used this to define true and false positive rates (TPR and FPR, respectively) as follows:

TPR – number of operative deliveries based on a clinical decision for 'presumed fetal compromise' as a proportion of the total number of babies with compromise;

FPR – number of operative deliveries based on a clinical decision for 'presumed fetal compromise' where there was no compromise as a proportion of the total number of normal cases.

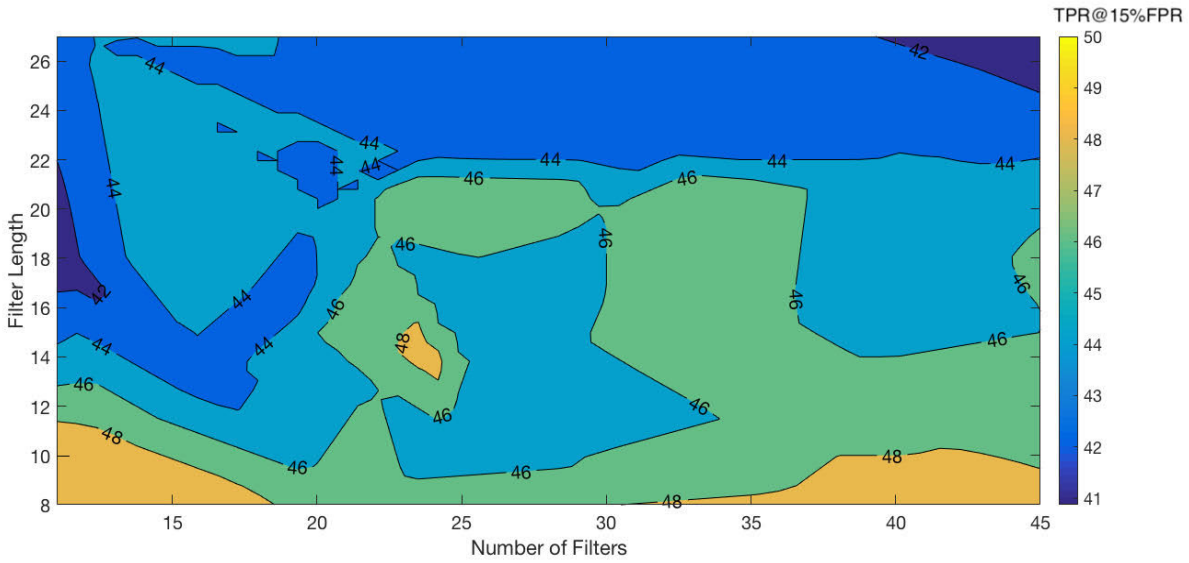


Figure 5 Optimization contour plot. To represent the ten dimensions into a 2-D plot, the x-axis and y-axis are the median number of filters and the median filter length across the five convolutional layers respectively. The color (and contour) represents the True Positive Rate (TPR) at 15%False Positive Rate (FPR) for every set of ten hyper-parameters chosen during the optimization.

2) OxSys 1.5:

This is a current prototype of the Oxford system for data-driven fetal monitoring in labor [18]. It uses only two FHR features and two clinical risk factors; and analyzes the entire FHR trace with a 15min sliding window (5min sliding step), and produces an alert if the risk for the fetus is high.

3) CNN:

For completeness, we also included a comparison with the single channel CNN from our preliminary work [23].

G. Performance metrics

Each of the proposed models was trained following a 3-fold cross validation schema to avoid overfitting and the reported median performance metrics were collected after running each algorithm five times.

Standard performance metrics for classification tasks were used to evaluate the networks: Area Under the ROC curve (AUC); TPR; and FPR. We present results for TPR with a fixed FPR of 5, 10, 15 and 20 percent, relating to the FPR of CTG analysis in clinical practice of 16%-21% [18] [15].

III. Results

A. Parameters Optimization

We allowed 40 iterations, with an initial random search of 10 samples. Ten hyper-parameters were optimized, representing the *number of filters* and *filter length* of each convolutional layer. The averaged results from the 3-fold cross validation are illustrated in Figure 5, which shows the hyper-parameters landscape after 40 iterations. To display the 10 hyper-parameters in a two-dimensional plot, we selected the median value across the five convolutional layers, for the *filter length* and the *number of filters* respectively. The color scheme and the contours in Figure 5 represent the TPR at a fixed 15% FPR for every set of chosen hyper-parameters in the [10, 50] interval for the *number of filters* and in the [5, 30] interval for the *filter length*. We observed that mainly the *filter length* (y-axis) contributed to the improvements of the fitness function.

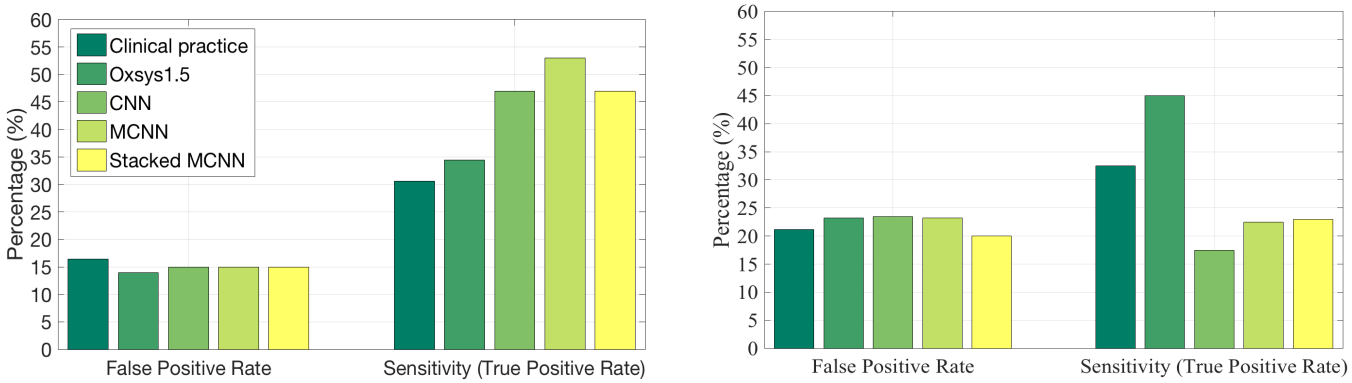
In particular, the network performed better using short filters (with a length smaller than 15 FHR sample points, i.e. 60 seconds). This led to the conclusion that the 'quicker' variations into fetal heart rate and contraction are more relevant than the long-term changes.

B. Comparison with Clinical Practice, OxSys 1.5 and CNN (Oxford testing data)

From the 35429 CTGs studied here, 1786 (5%) did not have 60min of monitoring and required zero-padding at the front for the MCNN model (all of these had more than 20 minutes and about half had more than 40min valid signal just before birth). For the Stacked MCNN training and testing, a total of 33590 CTGs (94.8%) had some 1st stage and 25299 (71.4%) some 2nd stage. In these, zero-padding at the front was needed in 2441 (7.3%) and 4282 (16.9%) respectively. Those without any 1st stage (1839, 5.2%) were excluded from testing/training of the Stacked MCNN as per the methods section above (Section II.E).

The performance of MCNN trained on the last 60 minutes of CTG recording is shown in Figure 6. On the *Main Testing Set* (hold-out), MCNN outperformed *Clinical Practice*, *OxSys1.5* [18] and the *single-channel CNN* [23], increasing the TPR with the same or lower FPR.

For completeness, we present in Figure 6(b) the results on the *Additional Testing Set* where the deep learning models had inferior sensitivity, and the *OxSys1.5* was strikingly better than all other methods, including *Clinical Practice*. We believe that this is a result of the fact that babies with severe compromise without acidemia are a small and heterogeneous group, better detected with CTG interpretation that incorporates the clinical



(a) Testing Set (hold-out), n=4429 *(b) Additional Testing Set, n=885*

Figure 6 Performance on last 60min of CTG: Clinical Practice, Oxsys1.5, CNN, MCNN, Stacked MCNN (median of 5 runs). The FPR was fixed at 15% for the CNN and MCNN in order to be comparable to the FPR of Clinical Practice.

context, as in the *OxSys1.5*. Compromise may not always be 'visible' in this group as expected and discussed in Section IIA. Furthermore, Table 1 shows the results for networks trained on data from the two labor stages separately. The outcome of labor can only be assigned at birth. It is to be expected that a longer interval between an earlier CTG analysis and the time of birth will be associated with reduced diagnostic accuracy, as seen in Table 1 for the MCNN trained on the 1st stage data. If the fetus is exposed to stress in the first stage of labor, but compensates well, then its heart rate will be normal and correctly classified as such at that time.

On the other hand, when the MCNN model was trained and tested only on the last 30 minutes of the 2nd stage, the MCNN achieved AUC of 0.70 and *Sensitivity* of 42% for FPR of 15%. So, the *Stacked* MCNN improved on the individual MCNN performance in each labor stage, but remained slightly suboptimal when compared to the MCNN trained and tested on the last hour, regardless of stage (Table 1), AUC 0.74 vs 0.76 and *Sensitivity* for FPR at 15% of 47% vs 53%. Only the median values were reported here because all networks had very small performance variability over the five independent runs ($\pm$0.1 and $\pm$3.5 from the median for the AUC and TPR metrics respectively, when trained on the last 60 minutes of CTG trace; $\pm$0.2 and $\pm$3.5 when trained on last 60 minutes of the first stage).

We concluded that the best overall performance was achieved by the MCNN trained on the last 60 minutes of CTG (regardless of the stage of labor). Unsurprisingly, this indicated that the most relevant CTG information in connection to the labour outcome is contained in the last segments of monitoring – closest to the time of outcome evaluation.

TABLE 1 COMPARISON OF THE PROPOSED MODELS (MEDIAN OF 5 RUNS) ON THE *MAIN TESTING SET* (N=4429). COMPROMISE: ACIDEMIA (ARTERIAL CORD PH AT BIRTH <7.05); NORMAL: HEALTHY NEW-BORN WITH ARTERIAL CORD PH>7.15. FPR: FALSE POSITIVE RATE. 1ST LABOR STAGE: ESTABLISHED LABOR BEFORE THE ONSET OF PUSHING. 2ND LABOR STAGE: AFTER THE PUSHING BEGAN.

	AUC	True Positive Rate (TPR %)			
		At 5% FPR	At 10% FPR	At 15% FPR	At 20% FPR
Test on last 60min of CTG, regardless of labor stage (4429 CTGs)					
MCNN (trained on last 60min of CTG)	0.77	32	44	53	58
Test on last 60min of 1st stage (subset of 4177 CTGs)					
MCNN (trained on last 60min of 1st stage)	0.65	17	27	33	40
Test on last 30min of 2nd stage (subset of 3138 CTGs)					
MCNN (trained on last 30min of 2nd stage)	0.71	22	36	43	47
Test on last 60min of 1st stage and/or the last 30min of 2nd stage as available (4348 CTGs)					
Stacked MCNN (trained on last 60min of 1st stage and last 30min of 2nd stage)	0.67	23	36	43	47
Stacked MCNN (trained on last 60min of CTG and last 30min of 2nd stage)	0.73	28	41	47	53

C. *Effect of the fetal heart rate signal quality on the classification threshold and the MCNN performance*

We aimed to examine the influence of signal loss (after de-noising) on the performance of our best model (MCNN trained on the last 60 minutes of CTG). To achieve this, we defined four groups of heart rate signal quality (described in Table 2), based on the quality score vector (which consists of 10 values for the 60min monitoring corresponding to each 15min window moving with a 5min step). Each value is the ratio of valid signal and missing signal in the 15min window. We found that MCNN had consistently higher number of 'alerts' (i.e. high-risk classifications) when there was more signal loss/noise (i.e. poorer signal quality), regardless of the labor outcome. Importantly, for every quality group, there was a different cut-off point in order to obtain FPR at 15%. There was an association between signal quality and performance as the AUC was particularly low for the group with poorest signal quality (Figure 7). Table 2 shows that, when using the same classification threshold for MCNN, the number of traces classed as high risk increases as the signal quality deteriorates from excellent to mediocre, resulting in higher TPR and higher FPR. Nevertheless, the ROC curves are similar for these signal quality groups (Figure 7).

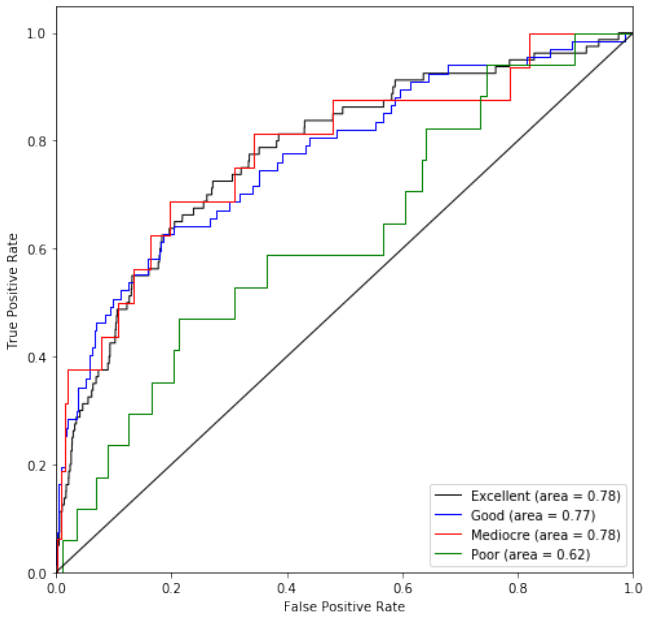


Figure 7 ROC curves for the four FHR signal quality groups as defined in Table 2 (Hold-out Testing Set, n = 4429).

TABLE 2 QUALITY GROUPS ON THE HOLD-OUT TESTING SET. REPORTED IS THE PERCENTAGE OF DATA BELONGING TO EACH QUALITY GROUP AND THE RESPECTIVE TPR/FPR MEDIAN (MIN – MAX) FOR THE FIVE RUNS OF THE MCNN MODEL (TRAINED ON THE LAST 30MIN OF THE 2ND STAGE).

Groups of cases according to their Fetal Heart Rate quality (% of CTGs in the Testing Set, *n* = 4429)	TPR (%)	FPR (%)
Excellent (52%) (at least 5 windows >= 0.9 and at least 3 windows >= 0.5)	43 (41 – 44)	9 (8 – 10)
Good (31%) (at least 3 windows >= 0.7 and at least 3 windows >= 0.5)	57 (57 – 61)	15 (14 – 18)
Mediocre (9%) (at least 5 windows >= 0.5)	69 (56 – 69)	25 (25 – 29)
Poor (8%) (at least 5 windows < 0.5)	59 (59 – 76)	37 (29 – 44)

D. *Testing on external data*

The MCNN and the *Stacked* MCNN were also tested on two external datasets, for which simulations the results are shown in Table 3 and Table 4. For the particular FPR values previously reported on the CTU-UHB dataset (see Section II), the TPR was substantially better for our deep learning approaches: it was 58% (53%-60%) at 14%FPR with MCNN and 80% (75%-85%) at 35%FPR; with the *Stacked* MCNN it was 55% (53%-60%) at 14%FPR and 83% (75%-88%) at 35%FPR.

TABLE 3 TESTING ON THE SPAM'17 DATASET (HTTP://USERS.OX.AC.UK/~NDOG0178/SPAM2017.HTM). REPORTED IS THE MEDIAN PERFORMANCE FOR FIVE MODELS.

	AUC	True Positive Rate (TPR %)			
		At 5% FPR	At 10% FPR	At 15% FPR	At 20% FPR
MCNN (Lyon)	0.92	63	70	78	83
MCNN (Brno)	0.82	35	50	55	65
Stacked MCNN (Lyon)	0.91	60	70	75	80
Stacked MCNN (Brno)	0.77	30	40	50	60

TABLE 4 TESTING ON THE CTU-UHB DATASET *[27]*. NOTE THIS DATASET ALSO COMES FROM THE SAME BRNO HOSPITAL BUT THERE IS NO OVERLAP WITH THE SPAM'17 DATA.

	AUC	True Positive Rate (TPR %)			
		At 5% FPR	At 10% FPR	At 15% FPR	At 20% FPR
MCNN	0.81	33	48	58	65
Stacked MCNN	0.82	33	45	58	65

IV. DISCUSSION

Cardiotocography (CTG) analysis during labor still relies on visual examination of long and complex heart rate patterns [2]. Here, we present our work on deep learning methods employing more than 35000 CTGs. We investigated a *Multimodal Convolutional Neural Network* (MCNN) and a *Stacked* MCNN model for the prediction of fetal compromise, using CTG traces from over 35000 labors (85% for training and 15% for hold-out testing). The *Stacked* MCNN can be considered as a more clinically relevant model, allowing analysis of the CTGs from the first and second labor stages separately. This is achieved by feeding the estimated probability of compromise from the first stage of labor into the analysis of the second stage. In addition to the fetal heart rate (FHR) and contraction signals, we incorporated into the network architecture a signal quality vector with the proportion of signal loss in the fetal heart rate trace.

The MCNNs' convolutional layer hyper-parameters (i.e., number of filters and filter length) were independently optimized for each layer, allowing full flexibility during the network optimization. We found that MCNN worked better

when using many short filters (Figure 5), whereas the CNN reported in [23] worked better with few large filters. This finding could be explained by the different architecture proposed here, where each input is processed separately before reaching the 'fully connected layer'.

On the Oxford *Main Testing Set* (hold-out) of 4429 CTGs (Figure 2), we compared the results of our models in predicting acidemia (cord pH<7.05, with or without severe compromise) with the clinical assessment in practice (*Clinical Practice*); the current Oxford prototype system *OxSys 1.5* [18]; and our pilot work with Convolutional Neural Network (CNN, [23]). All neural networks performed substantially better than the *Clinical Practice* and *OxSys 1.5*, with True Positive Rates (TPR) significantly higher than that of the *Clinical Practice,* for the same or lower False Positive Rate (FPR). The TPR was 53% and 31% for the MCNN and *Clinical Practice*, respectively (Figure 6a). The best performing model was our newly proposed MCNN, trained on the last 60 minutes of CTG, regardless of the stage of labor (Figure 6a and Table 1). This MCNN also outperformed the *Stacked* MCNN, achieving higher sensitivity (Figure 6a). There are several possible explanations for this: the main challenge of analyzing the second stage of labor separately is the different durations of each labor (in our data, 71% of the women had a second stage of more than 30 minutes); the proposed *Stacked* MCNN analyzed strictly only the last 30 minutes of the second stage (if available), which introduced a gap in the second stage of labor's CTG data that was not analyzed by the *Stacked* MCNN model (and potentially losing information). In addition, as presented in Section IIIB, there was more 'zero-padding' of missing signal data points in the *Stacked* MCNN model and the effect of this may be a contributor to poorer performance in this model. We plan to investigate in the future more flexible models of the stacked approach, allowing iterative analysis of the entire CTG available, as well as more flexibility in processing segments with missing data points.

Even though it could not outperform the MCNN trained on the last 60min labor, our *Stacked* MCNN model performed comparably. Moreover, to the best of our knowledge, this is a first attempt to analyze the CTG by estimating the probability of compromise at a time point, by using probability estimates from CTG data at an earlier time. We believe that such stacked models, after further developments, could be clinically relevant and suitable approaches for use at the bedside, building on the time-series nature of the CTG. These methods require significant computational resources and time 'offline' for development, optimization and training. But once trained, the *Stacked* MCNN could provide analysis of a new CTG trace in the matter of milliseconds and is thus entirely suitable for use at the bedside.

Furthermore, we showed that signals of poor quality adversely affect the performance of all models (Figure 7, Table 2). Thus, future models could benefit from adjusting the classification thresholds to the level of signal loss and further developments should ensure the MCNN models account for signal quality more flexibly. In particular, further work is needed to investigate the best way to handle any segments that are shorter than a predefined duration.

For the *Additional Testing Set* (n = 885, Figure 6b) and the detection of severe compromise without acidemia, all neural networks had low TPR. *OxSys 1.5* was the best with 45% TPR, followed by *Clinical Practice* with 33% and the deep learning models around 20% TPR, for the same FPR. Newborns with severe compromise without acidemia are a heterogeneous and challenging group to detect and are typically excluded from analysis and CTG datasets [12] [15] [16] [17] [27] [28] [29]. Such cases seem better suited to detection by tailored diagnostic rules, such as the ones of *OxSys 1.5*, which incorporates clinical risk factors and analyzes the entire CTG trace from the very beginning. Thus, for example, certain pre-existing fetal injuries are detected by *OxSys 1.5* early on in the CTG but are irrelevant to the proposed here models. In particular, we are working towards a new generation *OxSys* system, incorporating the best of both – the deep learning models and the heuristic, domain-based knowledge. Finally, our MCNN models convincingly outperformed the other automated methods when tested on the two external datasets (SPaM and CTU-UHB).

V. Conclusion

We demonstrated that deep learning methods applied to CTG analysis can strengthen our ability to detect fetal compromise during labor. The reported results showed the proposed Multimodal Convolutional Neural Network (MCNN, trained on the last 60 minutes of more than 30000 CTGs) as the best performing automated model for the detection of cord pH<7.05 achieved to date. It outperformed existing computerized approaches and clinical assessment, when tested on internal and external data.

Nevertheless, the model is still at an early stage of development and we anticipate that substantial future research (including the addition of more data) should improve its performance in the following ways:

- The proposed multimodal architecture will permit the introduction of new inputs, for example, more suitably structured information about signal quality and clinical risk factors/characteristics;
- Further experiments and simulations with the network's architectures;
- The MCNN and especially the *Stacked* MCNN could underpin a *Recurrent* MCNN, where the network is updated in real time (for example, every minute) with new available data and the latest available prediction;
- Developing hierarchical/stacked LSTM models, for example, using the MCNN risk estimates at different times as inputs;
- Combining deep learning methods with domain-specific knowledge and/or existing algorithms that complement each other to yield risk assessment for different types of fetal compromise.

Importantly, our deep learning models are currently not suitable for the detection of severe fetal injury in the absence of cord acidemia – a heterogeneous, small, and poorly understood

group. We suggest that hybrid approaches to CTG interpretation in labor, in which different diagnostic models can estimate the risk for different types of fetal compromise, incorporating clinical knowledge with data-driven analyses, are the most promising way forward.

REFERENCES

[1] W. H. Organization, *WHO recommendation on intermittent fetal heart rate auscultation during labour,* 2018.

[2] A. E. Timmins and S. L. Clark, "How to approach intrapartum category II tracings," *Obstetrics and Gynecology Clinics,* vol. 42, no. 2, pp. 363-375, 2015.

[3] E. F. Hamilton and P. A. Warrick, "New perspectives in electronic fetal surveillance," *Journal of perinatal medicine,* vol. 41, no. 1, pp. 83-92, 2013.

[4] J. J. Kurinczuk, M. White-Koning and N. Badawi, "Epidemiology of neonatal encephalopathy and hypoxic-ischaemic encephalopathy," *Early human development,* vol. 86, no. 6, pp. 329-338, 2010.

[5] C. A. Walsh, M. B. McMenamin, M. E. Foley, S. F. Daly, M. S. Robson and M. P. Geary, "Trends in intrapartum fetal death, 1979-2003," *American Journal of Obstetrics & Gynecology,* vol. 198, no. 7, pp. 47.e1-47.e7, 2008.

[6] "Perinatal Mortality 2008," Centre for Maternal and Child Enquiries, United Kingdom, London, 2010.

[7] J. E. Lawn, H. Blencowe, P. Waiswa, A. Amouzou, C. Mathers, D. Hogan, V. Flenady, J. F. Froen, Z. U. Qureshi and C. Calderwood, "Stillbirths: rates, risk factors, and acceleration towards 2030," *The Lancet,* vol. 387, no. 10018, pp. 587-603, 2016.

[8] E. Hon and E. Quilligan, "The classification of fetal heart rate," *Childbirth: The medicalization of obstetrics,* vol. 2, no. 1, p. 339, 1996.

[9] D. Ayres-de-Campos, P. Sousa, A. Costa and J. Bernardes, "Omniview-SisPorto® 3.5–a central fetal monitoring station with online alerts based on computerized cardiotocogram+ ST event analysis," *Journal of Perinatal Medicine,* vol. 36, no. 3, pp. 260-264, 2008.

[10] P. Brocklehurst, D. J. Field, E. Juszczak, S. Kenyon, L. Linsell, M. Newburn, R. Plachcinski, M. Quigley, L. Schroeder and P. Steer, "The INFANT trial," *The Lancet,* vol. 390, no. 10089, p. 28, 2017.

[11] A. Georgieva, P. Abry and V. Chudacek, "Computer-based intrapartum fetal monitoring and beyond: A review of the 2nd Workshop on Signal Processing and Monitoring in Labor (October 2017, Oxford, Uk)," *Aca Obstet Gynecol Scand,* vol. 1, no. 1, pp. 1-11, 2019.

[12] S. Dash, J. G. Quirk and P. M. Djuric, "Fetal heart rate classification using generative models," *IEEE Transactions on Biomedical Engineering,* vol. 61, no. 11, pp. 2796-2805, 2014.

[13] L. Xu, C. W. Redman, S. J. Payne and A. Georgieva, "Feature selection using genetic algorithms for fetal heart rate analysis," *Physiological measurement,* vol. 35, no. 7, p. 1357, 2014.

[14] A. Georgieva, S. J. Payne, M. Moulden and C. W. Redman, "Artificial neural networks applied to fetal monitoring in labour," *Neural Computing and Applications,* vol. 22, no. 1, pp. 85-93, 2013.

[15] P. Abry, J. Spilka, R. Leonarduzzi, V. Chudavcek, N. Pustelnik and M. Doret, "Sparse learning for Intrapartum fetal heart rate analysis," *Biomedical Physics & Engineering Express,* vol. 4, no. 3, 2018.

[16] J. Spilka, J. Frecon, R. Leonarduzzi, N. Pustelnik, P. Abry and M. Doret, "Sparse support vector machine for intrapartum fetal heart rate classification," *IEEE journal of biomedical and health informatics,* vol. 21, no. 3, pp. 664-671, 2017.

[17] C. Elliott, P. A. Warrick, E. Graham and E. F. Hamilton, "Graded classification of fetal heart rate tracings: association with neonatal metabolic acidosis and neurologic morbidity," *American journal of obstetrics and gynecology,* vol. 202, no. 3, p. 258, 2010.

[18] A. Georgieva, C. W. Redman and A. T. Papageorghiou, "Computerized data-driven interpretation of the intrapartum cardiotocogram: a cohort study," *Acta obstetricia et gynecologica Scandinavica,* vol. 96, no. 7, pp. 883-891, 2017.

[19] A. Georgieva, "Advances in computing are driving progress in fetal monitoring," *BJOG: An International Journal of Obstetrics & Gynaecology,* vol. 123, no. 12, pp. 1955-1955, 2016.

[20] Y. LeCun, Y. Bengio and G. Hinton, "Deep Learning," *Nature,* vol. 521, no. 7553, p. 436, 2015.

[21] D. Nie, R. Trullo, J. Lian, L. Wang, C. Petitjean, S. Ruan, Q. Wang and D. Shen, "Medical Image Synthesis with Deep Convolutional Adversarial Networks," *IEEE Transactions on Biomedical Engineering,* vol. 65, no. 12, pp. 2720-2730, 2018.

[22] T.-E. Chen, S.-I. Yang, L.-T. Ho, K.-H. Tsai, Y.-H. Chen, Y.-F. Chang, Y.-H. Lai, S.-S. Wang, Y. Tsao and C.-C. Wu, "S1 and S2 heart sound recognition using deep neural networks," *IEEE Transactions on Biomedical Engineering,* vol. 64, no. 2, pp. 372-380, 2018.

[23] A. Petrozziello, I. Jordanov, A. Papageorghiou, R. C.W.G and A. Georgieva, "Deep Learning for Continuous Electronic Fetal Monitoring in Labor," in *40th International Engineering in Medicine and Biology Conference*, Honolulu, 2018.

[24] W. Rawat and Z. Wang, "Deep convolutional neural networks for image classification: A comprehensive review," *Neural computation,* vol. 29, no. 9, pp. 2352-2449, 2017.

[25] A. O'Shea, G. Lightbody, G. Boylan and A. Temko, "Investigating the Impact of CNN Depth on Neonatal

Seizure Detection Performance," in *Engineering in Medicine and Biology Society (EMBC), 2018 40th Annual International Conference of the IEEE*, Honolulu, 2018.

[26] A. Georgieva, M. Moulden and C. W. Redman, "Umbilical cord gases in relation to the neonatal condition: the EveREst plot," *European Journal of Obstetrics & Gynecology and Reproductive Biology,* vol. 168, no. 2, pp. 155-160, 2013.

[27] B. Chudacek, J. Spilka, M. Bursa, P. Janku, L. Hruban, M. Huptych and L. Lhotska, "Open access intrapartum CTG database," *BMC pregnancy and childbirth,* vol. 14, no. 1, p. 16, 2014.

[28] J. Spilka, V. Chudáček, M. Huptych, R. Leonarduzzi, P. Abry and D. M, "Intrapartum fetal heart rate classification: Cross-database evaluation," in *XIV Mediterranean Conference on Medical and Biological Engineering and Computing*, 2016.

[29] G. Georgoulas, P. Karvelis, J. Spilka, V. Chudáček, C. Stylios and L. L, "Investigating pH based evaluation of fetal heart rate (FHR) recordings," *Health and technology,* vol. 7, no. 1, pp. 2-3, 2017.

[30] J. S. Bergstra, R. Bardenet, Y. Bengio and B. Kegl, "Algorithms for hyper-parameter optimization," *Advances in neural information processing systems,* pp. 2546-2554, 2011.

[31] S. Cazares, M. Moulden, C. W. Redman and L. Tarassenko, "Tracking poles with an autoregressive model: a confidence index for the analysis of the intrapartum cardiotocogram," *Medical Engineering and Physics,* vol. 23, no. 9, pp. 603-614, 2001.

[32] L. Wan, M. Zeiler, S. Zhang, Y. Le Cun and R. Fergus, "Regularization of neural networks using dropconnect," in *International Conference on Machine Learning*, 2013.

[33] I. Nunes, D. Ayres-de-Campos, A. Ugwumadu, P. Amin, P. Banfield, A. Nicoll, S. Cunningham, P. Sousa, C. Costa-Santos and J. Bernardes, "Central fetal monitoring with and without computer analysis: a randomized controlled trial," *Obstetrics & Gynecology,* vol. 129, no. 1, pp. 83-90, 2017.