Deep Learning for Continuous Electronic Fetal Monitoring in Labor

Alessio Petrozziello, Ivan Jordanov, Aris T. Papageorghiou, Christopher W.G. Redman, and Antoniya Georgieva

Abstract— **Continuous electronic fetal monitoring (EFM) is used worldwide to visually assess whether a fetus is exhibiting signs of distress during labor, and may benefit from an emergency operative delivery (e.g. Cesarean section). Previously, computerized EFM assessment that mimics clinical experts showed no benefit in randomized clinical trials. However, as an example of routinely collected 'big' data, EFM interpretation should benefit from data-driven computational approaches, such as deep learning, which allow automated evaluation based on large clinical datasets.**

Here we report our investigation of long short term memory (LSTM) and convolutional neural networks (CNN) in analyzing EFM traces from over 35,000 labors for the prediction of fetal compromise. Of these, 85% are used for training with cross-validation and the remainder are set aside for testing. The results are compared with *Clinical practice* (reason for operative delivery recorded as fetal distress) and an earlier prototype system for computerized analysis of EFM (OxSys 1.5), developed on the same data. We demonstrate that CNN outperforms LSTM, *Clinical practice*, and OxSys 1.5 in predicting fetal compromise, with a sensitivity of 42% (30%, 34%, and 36% for the others, respectively), at comparable or lower false positive rates. We also show that increasing the size of the training set improves the sensitivity and stability of CNN's performance on the testing set. When tested on a small open-access external database, CNN moderately improves on the performance of published feature extraction based methods.

We conclude that CNN could play an important role in the field of automated EFM analysis, but requires further work.

I. INTRODUCTION

Electronic fetal monitoring (EFM) is used in labor aiming to detect fetuses at risk of distress who might benefit from an emergency operative delivery (Caesarean or instrumental vaginal delivery). However, visual interpretation is unreliable and complex EFM graphs, that continuously display the fetal heart rate and uterine contractions, remain poorly understood (Fig. 1). This has a significant impact: in the UK alone, every year, about 220 healthy babies die [1] and about 1,000 sustain brain injury [2] during labor at term. Nearly 50% of the NHS litigation bill is due to obstetric claims (£3.1bn in 2000-10), most of which relate to shortcomings in labor management and electronic fetal monitoring (EFM) interpretation [3].

A few classic EFM patterns have been empirically identified and, for certain EFM patterns, the disagreement in visual interpretation between experts reaches 100% [4]. Computerized detection of such classic patterns, mimicking clinical visual assessment, is commercially available, but has not shown benefit in randomized clinical trials [5, 6].

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).
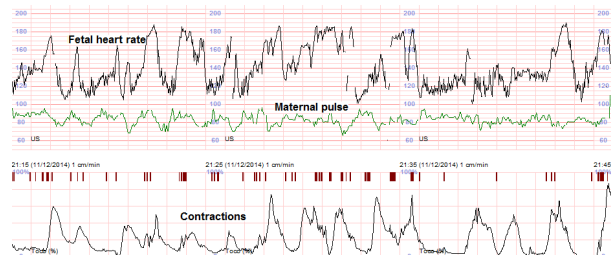

Figure 1.   Electronic fetal monitoring (EFM) in labor (a 30min snippet)

We have acquired a uniquely large and detailed cohort of routinely collected data during labor at Oxford (all monitored births between Apr'93 and Dec'11). We have already developed a basic computerized data-driven prototype for EFM evaluation: OxSys 1.5 [7]. It works comparable to clinical practice but is based only on a few clinical and EFM features and further improvements are needed.

In this study, we explore, for the first time, deep learning methods for the evaluation of EFM, without pre-defined feature extraction.

II. DATA AND METHODS

A. Data

We have previously discussed and published the main characteristics of the Oxford archive, including the definition of *Clinical practice* which uses the reason for operative delivery (fetal distress vs. other) to the define true and false positives [7]. Digitized EFM records were available in a linked form with relevant clinical details. For this study, we analyzed the set of 35,429 deliveries comprising singleton babies with gestation ≥36 weeks; EFM record in labor; cord gas analyses at birth; no congenital abnormalities or breech presentation. Signal loss was linearly interpolated and the data smoothed down to 0.25Hz. Validated cord gas analyses at birth gave details of fetal blood oxygenation at the time of birth. There were 1,470 compromised (either severe compromise or arterial cord pH<7.05) and 33,959 healthy newborns, yielding a 4.15% incidence rate (detailed definitions are given in [7]). We used 85% of the data (30,115 cases) for training and the remaining ones (5,314) were set aside for testing. The testing EFMs were identified by a random selection of 15% of cases within each outcome group, ensuring similar rates of compromise in training and testing.

We analyzed the last hour of the EFM recording (900 data points for fetal heart rate and 900 data points for uterine contractions at 0.25Hz). There were 1,796 traces shorter than one hour and zeros were added to these at the beginning of the trace to obtain 900 data points. Contraction signals of the EFM commonly have poor quality due to technical limitations of the monitoring devices. To avoid overloading the network with noise, we assessed the contraction signals with an established method [8] and imposed a restriction: >40min of acceptable quality, of which >20min is of excellent quality.

Only 24% of EFM records met this condition and were used in the networks, whereas the remainder were inputted as zero.

To avoid overfitting, we adopted a five-fold cross validation during training. The average prediction of the five models was used as outcome for the test set patterns. To tackle the problem of unbalanced training dataset (4% compromised babies vs. 96% healthy ones), a weighted error function was adopted, such that an error in classifying one compromised fetus counts as heavily as 24 misclassified healthy fetuses.

The methods were evaluated using widely accepted performance metrics: Area under the ROC curve (AUC); True Positive Rate (TPR); and False Positive Rate (FPR). Finally, we tested the models on the 552 traces in the *Prague/Brno* open-access database (CTU-UHB) [9], created to provide researchers with independent EFM data.

B. Long Short Term Memory Networks

We adopted the Long Short Term Memory (LSTM) model presented in [10]. The main advantage of this architecture is its ability to capture both long and short time dependencies in time series, which have proven to be effective in many domains (including medical applications) [11]. We used a single layer LSTM with two inputs: fetal heart rate and contraction signals; and two outputs: healthy and compromised newborns. The LSTM architecture included hyperbolic tangent as a hidden activation function and a hard sigmoid as a recurrent activation (default activation functions for LSTM, as advised in [12]). To get a binary class probability, a *softmax* function was used in the output layer. The data was normalized before being inputted in the LSTM.

C. Convolutional Neural Networks

The use of Convolutional Neural Networks (CNN) is another well-studied branch of deep learning, particularly suitable when learning from large amounts of *raw* data [13, 14]. To exploit the time structure of the input data, we implemented CNN that perform convolutions on overlapping sliding windows. The input layer size of 2x900 points corresponded to the last hour of fetal monitoring at 0.25Hz. A 12-layer CNN was chosen, given that the input size is halved by a *max-pooling* layer for every two layers. The model comprised five convolutional layers (with a ReLU default activation function, as advised in [15]), interleaved by five max pooling layers. This also determined maximum length of 29 kernels (input: 900; 1st conv: 450; 2nd conv: 225; 3rd conv: 113; 4th conv: 57, and the 5th conv: 29). The last max-pooling layer was then flattened and fed as input to a fully connected layer. As in the LSTM, to get a class probability, a *softmax* function was used in the output layer. Furthermore, *dropout* (to avoid overfitting) and *batch normalization* (normalizing the outputs of each activation layer) were used as in a standard CNN [15].

D. Bayesian Hyper-Parameters Optimization

Bayesian optimization uses previous observations of a loss function f, to determine the next (optimal) point to sample [16]. Firstly, using previously evaluated points $x_{1:n}$, a posterior expectation of the landscape of f is computed. Secondly, the loss f at a new point x_{new} that maximizes some utility of the expectation of f is sampled. The utility specifies which regions of the domain of f are optimal to sample from. These steps are repeated until a convergence criterion is met. To compute a posterior expectation, we needed a likelihood model for the samples from f and a prior probability model on f. In the Bayesian search, we assumed a normal likelihood with noise. For the prior distribution, we assumed that the loss function f can be described by a *Gaussian process (GP)*. GP is a popular probability model, because it induces analytically tractable posterior distribution over the loss function. This allows updating our beliefs of f's landscape, after the loss for a new set of hyper-parameters is computed.

For both models we used Bayesian optimization with GP to maximize the models' TPR at 15% FPR. This specific value for the FPR was chosen as the maximum allowed, in order not to exceed the FPR seen in clinical practice [7]. The optimization ran for 40 iterations, with an initial random search of 10 samples. We report the results from the best performing model on a 5-fold cross validation average.

III. RESULTS

A. Parameter optimization

Table I shows the parameters used during the *Bayesian Hyper-Parameters Optimization* for the LSTM and CNN models. The optimal *Wavelet De-noising* was found to be small for the LSTM, indicating that the initial signal was smoothed additionally. On the other hand, for the CNN, this was zero, indicating that no smoothing was required.

TABLE I. PARAMETERS FOR THE NEURAL NETWORK MODELS.

Parameter	Range optimized on	Optimal value (model)
Number of Convolutional Kernels	[5, 50]	17 (CNN)
Kernel Length in data points	[5, 29]	29 (CNN)
Wavelet De-noising	[0, 2]	0.00 (CNN) 0.52 (LSTM)
Elastic net regularization	[0.001, 0.01]	0.006 (LSTM)
Number of neurons in the recurrent layer	[10, 125]	124 (LSTM)
Lookback - past minutes used during the training phase	[5, 7.5, 10]	7.5 (LSTM)

B. Training and testing (Oxford data)

Table II shows the AUC and TPR (at a fixed 15% and 20% FPR) for training and testing sets. The CNN outperformed the LSTM in all proposed metrics and both models performed on the testing set similarly to their performance on the training set, showing they generalize well on unseen data. The ROC curves in Fig. 2 present a small gap in performance in the first 0.1 FPR (easy to predict cases), while subsequently increasing difference for FPR of 0.1 to 0.2. The proposed deep learning architectures are also compared with the current clinical practice performance and a state-of-the-art algorithm in fetal monitoring (OxSys 1.5 [7]). Fig. 3 illustrates the FPR and TPR for the four compared techniques. The CNN showed better sensitivity, at a lower false positive rate, compared to clinical practice and OxSys 1.5. Lastly, an empirical experiment was carried out to test robustness and validate the importance of the amount of CNN training data. Fig. 4 shows the test sensitivity at fixed FPR achieved when using 10%, 50%, and 100% of the data during training.
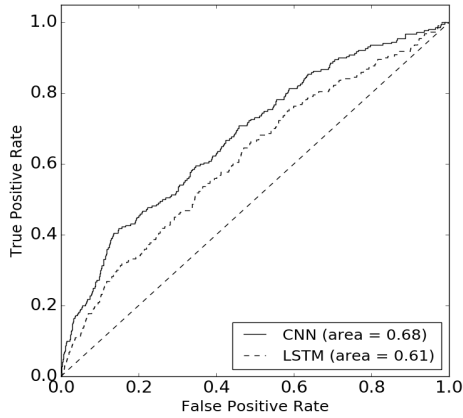
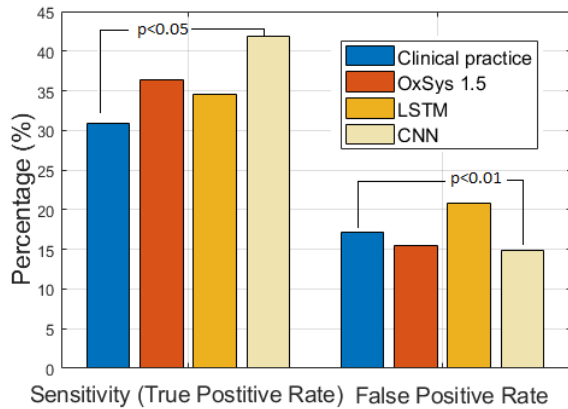Figure 2.   ROC curves (Oxford test set: 5,314 cases).

Figure 3.   Comparison of the two deep learning models, OxSys 1.5 and *Clinical practice* o the test set (χ^2 test for comparison of proportions).
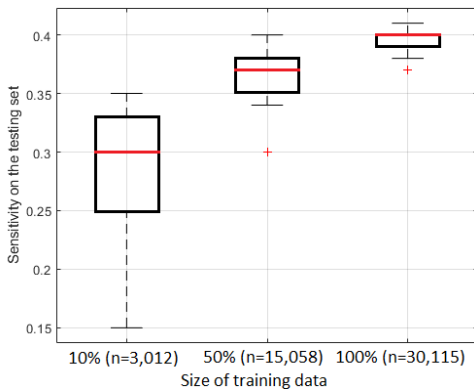

Figure 4.   Robustness of CNN with respect to the size of training dataset over 30 runs (FPR is fixed at 15%).

TABLE II.   SELECTED TRAINING AND TESTING RESULTS FOR THE TWO NEURAL NETWORKS MODELS

Model	Data	AUC	TPR@15% FPR	TPR@20% FPR
CNN	Train	0.73	0.44	0.52
	Test	0.68	0.36	0.45
LSTM	Train	0.68	0.41	0.46
	Test	0.61	0.30	0.34

AUC: Area Under the ROC Curve; TPR: True Positive Rate; FPR: False Positive Rate

C. *Testing (external open-access dataset, CTU-UHB)*

In order to test and validate our approach on publicly available data (Oxford data has restricted access) and compare the results with other authors, we report our findings on the CTU-UHB dataset. Table III presents the LSTM and CNN results from two published methods [17, 18]. The deep learning methods outperformed them but it is important to note that no direct comparison is possible due to: a smaller subset of 420 EFMs used in [17]; training done with 'leave-2%-out' and no separate data used for testing in [18]. Fig. 5 shows the ROC curves for our two models.

TABLE III.   COMPARISON ON THE CTU-UHB DATASET (552 LABORS).

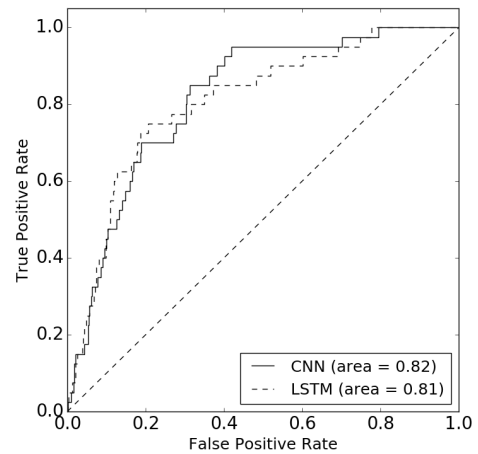| Model	True Positive Rate (TPR)	False Positive Rate (FPR)
CNN	0.55	0.14
LSTM	0.60	0.14
Spilka et al. [17] ($MAD_{dtrd, b0, H}$)	0.40	0.14
CNN	0.63	0.17
LSTM	0.65	0.17
Spilka et al. [17] ($MAD_{dtrd, b0, c1}$)	0.60	0.17
CNN	0.70	0.22
LSTM	0.72	0.22
Georgoulas et al. [18] (MMC)	0.68	0.22
CNN	0.85	0.35
LSTM	0.80	0.35
Georgoulas et al. [18] (F-measure)	0.72	0.35

Figure 5.   ROC curves (CTU-UHB testing set: 552 cases).

IV.   DISCUSSION

This work presents a first application of deep learning methods on data from continuous EFM during labor. The motivation was to move away from classic feature extraction approaches and examine whether deep learning has the potential to detect information in the EFM that is currently 'hidden'. EFMs of over 30,000 births were used for training and over 5,000 were set aside for testing. A small external open-access database was also used for external testing and comparison with published methods from other groups.

Two deep learning architectures were investigated: LSTM and CNN. Both methods showed good generalizability – retaining similar performance on testing and training (Table II). On the Oxford data, CNN performed better than the

LSTM (Table II, Fig. 2, Fig. 3), but LSTM was slightly better on the CTU-UHB data (Table III). Also, on the Oxford testing set, CNN outperformed the results of clinical practice (Fig.3) and showed robust performance (Fig. 4). As can be seen from Fig. 4, the sensitivity of CNN increased and its variance decreased with larger size of the training set. Direct comparison with OxSys 1.5 suggested that CNN achieved better results, but this must be interpreted with caution because OxSys 1.5 is based on the entire dataset and analyzes the entire EFM trace, incorporating clinical risk factors. The CNN's TPR of 44% @15%FPR (Fig. 3) is significantly higher than the TPR in clinical practice (31%) but needs to exceed 60% if we want tangible clinical benefits.

Both LSTM and CNN performed better on the external testing set CTU-UHB (552 births) (Table III) than on the Oxford testing set (5,314 births) (Table II). The CTU-UHB dataset is 'easier' than the Oxford one for the automated methods to analyze: firstly, because it is smaller and less heterogeneous (the Oxford data span many years of varying clinical practice); and secondly, because it defined compromise only as acidemia (while the Oxford data defined compromise as a more clinically relevant outcome of acidemia and/or severe compromise [7]).

The maximal CNN kernel length was set to 29 data points (Table II) and future work will focus on changing the architecture to allow larger kernels. And for LSTM, the optimal number of neurons here was 124 (where 125 was the maximum allowed) thus higher values may be beneficial.

LSTM are widely used for time series forecasting, but have two important limitations in our setting. Firstly, we require one risk assessment at the end of the series (i.e. classification task) and the error only depends on the last output of the LSTM whereas, in forecasting, a prediction is required for every time step in the series and the learning is based on the error calculated through the whole time span. Secondly, the EMF records are large and cause problems such as vanishing gradients during backpropagation. On the other hand, CNN have proven suitable for both spatial and temporal data. The CNN is able to handle long time series using moving filters and max-pooling (i.e., the size of the input is halved at each convolution allowing more compact representation of the feature space to be learned each time).

Limitations of this work are the use of EFM signals at 0.25Hz; analysis only of the last hour of EFM; no account of the labor stage (a known confounder). Even with these limitations, the deep learning models demonstrate potential to improve the results of OxSys 1.5 and other feature-based data-driven techniques. In particular, we expect further significant improvements with larger training datasets and network sizes; and with incorporation of clinical risk factors.

V. ACKNOWLEDGMENT

VI. REFERENCES

[1] MBRRACE-UK, "Perinatal Mortality Surveillance Enquiry - Term, Singleton, Intrapartum, Stillbirth and Intrapartum Neonatal Death," 2017.

[2] 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.

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

[4] S. L. Sholapurkar, "Interpretation of British experts' illustrations of fetal heart rate (FHR) decelerations by Consultant Obstetricians, Registrars and midwives: A prospective observational study-Reasons for major disagreement with experts and implications for clinical prac," *Open Journal of Obstetrics and Gynecology,* vol. 3, no. 6, pp. 454-465, 2013.

[5] I. Nunes, D. Ayres-de-Campos, A. Ugwumadu, P. Amin, P. Banfield and A. Nicoll, "Central Fetal Monitoring With and Without Computer Analysis: A randomized Controlled Trial," *Obstet. Gynecol.,* vol. 129, no. 1, pp. 83-90, 2017.

[6] "Computerised Interpretation of fetal heart rate during labour (INFANT): a randomis controlled trial," Lancet, 2017.

[7] A. Georgieva, C.W.G. Redman and A. Papageorghiou, "Computerized data-driven interpretation of the intrapartum cadiotocogram: a cohort study". *Acta Obstet Gynecol Scand*, vol. 96, no. 7, pp. 883-338, 2017.

[8] S. Cazares, M. Moulden, C.W. Redman, L. Tarassenko, "Tracking poles with an autoregressive model: a confidence index for the analysis of the intrapartum cardiotocogram," *Med Eng Phys* vol. 23, pp. 603-14, 2001.

[9] V. Chudáček, J. Spilka, M. Burša, P. Janků, L. Hruban, M. Huptych, L. Lhotská, "Open access intrapartum CTG database". *BMC pregnancy and childbirth*, vol. 14, no. 1, pp. 16-27, 2014.

[10] A. Graves, "Generating sequences with recurrent neural networks," *arXiv preprint arXiv:1308.0850,* 2013.

[11] E. Choi et al., "Using recurrent neural network models for early detection of heart failure onset," *Journal of the American Medical Informatics Association*, vol. 24, no. 2, pp. 361-370, 2016

[12] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation,* vol. 9, no. 8, pp. 1735--1780, 1997.

[13] S. M. Plis et al., "Deep learning for neuroimaging: a validation study," *Frontiers in neuroscience,* vol. 8, 2014.

[14] L. Deng and et al., "Recent advances in deep learning for speech research at Microsoft," *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE Intern. Conference,* pp. 8604--8608, 2013.

[15] L. Wan et al., "Regularization of neural networks using dropconnect," *Machine Learning International Conference on*, pp. 1058-1066, 2013.

[16] 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.

[17] J. Spilka, V. Chudáček, M. Huptych, R. Leonarduzzi, P. Abry, M. Doret, "Intrapartum fetal heart rate classification: Cross-database evaluation." in *2016 Proc. XIV Mediterranean Conference on Medical and Biological Engineering and Computing*, pp. 1199-204.

[18] G. Georgoulas, P. Karvelis, J. Spilka, V. Chudáček, C.D. Stylios, L. Lhotská, "Investigating pH based evaluation of fetal heart rate (FHR) recordings." *Health and technology*, vol. 7, no. 2-3, pp. 241-54, 2017.