Machine Learning to Classify Cardiotocography for Fetal Hypoxia Detection

Machine Learning to Classify Cardiotocography for Fetal Hypoxia Detection

Farah Francis, Saturnino Luz, Honghan Wu, Rosemary Townsend, Sarah S. Stock

***Abstract*— Fetal hypoxia can cause damaging consequences on babies' such as stillbirth and cerebral palsy. Cardiotocography (CTG) has been used to detect intrapartum fetal hypoxia during labor. It is a non-invasive machine that measures the fetal heart rate and uterine contractions. Visual CTG suffers inconsistencies in interpretations among clinicians that can delay interventions. Machine learning (ML) showed potential in classifying abnormal CTG, allowing automatic interpretation. In the absence of a gold standard, researchers used various surrogate biomarkers to classify CTG, where some were clinically irrelevant. We proposed using Apgar scores as the surrogate benchmark of babies' ability to recover from birth. Apgar scores measure newborns' ability to recover from active uterine contraction, which measures appearance, pulse, grimace, activity and respiration. The higher the Apgar score, the healthier the baby is.**

***We employ signal processing methods to pre-process and extract validated features of 552 raw CTG. We also included CTG-specific characteristics as outlined in the NICE guidelines. We employed ML techniques using 22 features and measured performances between ML classifiers. While we found that ML can distinguish CTG with low Apgar scores, results for the lowest Apgar scores, which are rare in the dataset we used, would benefit from more CTG data for better performance. We need an external dataset to validate our model for generalizability to ensure that it does not overfit a specific population.**

***Clinical Relevance*— This study demonstrated the potential of using a clinically relevant benchmark for classifying CTG to allow automatic early detection of hypoxia to reduce decision-making time in maternity units.**

I. Introduction

Fetal hypoxia occurs when the baby's continuous oxygen supply is disrupted during labor. Fetal hypoxia can cause stillbirth, neonatal encephalopathy and developmental disabilities [1-3]. During uterine contractions (UC), temporal hypoxia is expected due to babies' natural physiological responses. However, a small proportion of babies fail to recover from constant contractions of the uterus during labor [4]. Cardiotocography (CTG) is a non-invasive electronic fetal monitoring device that can indicate fetal well-being in the uterus during labor. It is attached to the mother's womb and measures fetal heart rate (FHR) changes and UC. From the CTG, obstetricians will intervene to remedy fetal hypoxia, such as emergency cesarean sections or assisted delivery [5]. However, CTG is not discriminatory enough and suffers from inconsistencies in the interpretation that can cause delayed response [6].

Furthermore, some decision-making can be subjective and ambiguous, which may contribute to discrepancies in CTG interpretation [7]. Since the introduction of CTG, there has been a five-fold increase in cesarean section rates, while cerebral palsy cases remain unchanged. This is a substantial number of false positive instances in which it can harm babies and women while increasing avoidable medical costs [8].

Computerized CTG was introduced to improve decision by enhancing interpretations to allow a quicker response to compromised fetuses. It works by alerting clinicians when there are any changes to the FHR, such as deceleration. However, computerized CTG relies on human input which a vulnerable towards bias and measurement errors. In addition, all changes are not pathological and current computerized CTG cannot differentiate between natural and harmful changes [9]. In addition, a meta-analysis of six studies showed no significant improvement in fetal outcomes between visual and computerized CTG during labor [10].

Machine learning (ML) demonstrated promising results in classifying abnormal CTG by reducing interpretation variability. Previous studies used varying pH umbilical cord blood levels and types of delivery as a benchmark for hypoxia [11]. However, pH levels do not reflect their ability to recover from birth stress, and some benchmarks used were clinically irrelevant [12]. Therefore, we proposed using 5 minutes Apgar score as the surrogate marker of hypoxia in our ML algorithms. Low Apgar scores have shown a high association with hypoxic diagnosis and abnormal CTG. It is a routine, standardized measurement of babies' physiology and condition after birth, such as appearance, grimace, activity, pulse and respiration. The scores range from 0 to 10; the higher the score, the healthier the baby [13]. Apgar score taken after birth is a good indicator if babies can recover and does not require resuscitation [14]. Therefore, we aim to use 5-minute Apgar scores as the benchmark of a

F. Francis is with Usher Institute, University of Edinburgh, NINE, 9 Little, France Road, EH16 4UX, Edinburgh, UK (e-mail: farah.francis@ed.ac.uk)

S. Luz is with Usher Institute, University of Edinburgh, NINE, 9 Little, France Road, EH16 4UX, Edinburgh, UK (e-mail: s.luz@ed.ac.uk)

H. Wu is with Institute of Health Informatics, University College London, 222, Euston Road, NW1 2DA, London, UK (e-mail: honghan.wu@ucl.ac.uk)

Rosemary Townsend is with Usher Institute, University of Edinburgh, NINE, 9 Little, France Road, EH16 4UX, Edinburgh, UK (e-mail: rtownse2@exseed.ed.ac.uk)

S. S. Stock is with Usher Institute, University of Edinburgh, NINE, 9 Little, France Road, EH16 4UX, Edinburgh, UK (e-mail: Sarah.Stock@ed.ac.uk)

newborn's ability to recover from hypoxia during active contraction of the uterus.

II. Methods

A. Dataset

We used raw CTG from the open-access CTU-UHB database with 552 CTG recordings sampled at 4Hz. The recording was taken no longer than 90 minutes during labor (second stage of labor). CTG records were taken between 2009 and 2012 at the University Hospital in Brno, Czech Republic. This database was approved by the Institutional Review Board of University Hospital Brno and all women signed the informed consent [15].

B. Feature Extraction

Before feature extraction, CTG signals were denoised to remove unwanted artefacts and missing recordings due to fetal and maternal movements. Missing beats were interpolated, and the signal was smoothed with a moving mean of 30 windows. Pre-processed CTGs were shown below in figure 1. For morphological features, we extracted FHR in conjunction with UC as recommended by the National Institute for Health and Care Excellence guidelines for CTG interpretations (NICE, 2014). For the time domain, frequency domain and non-linear features, we only used FHR signals. We extracted 22 features, which were included in the ML models.

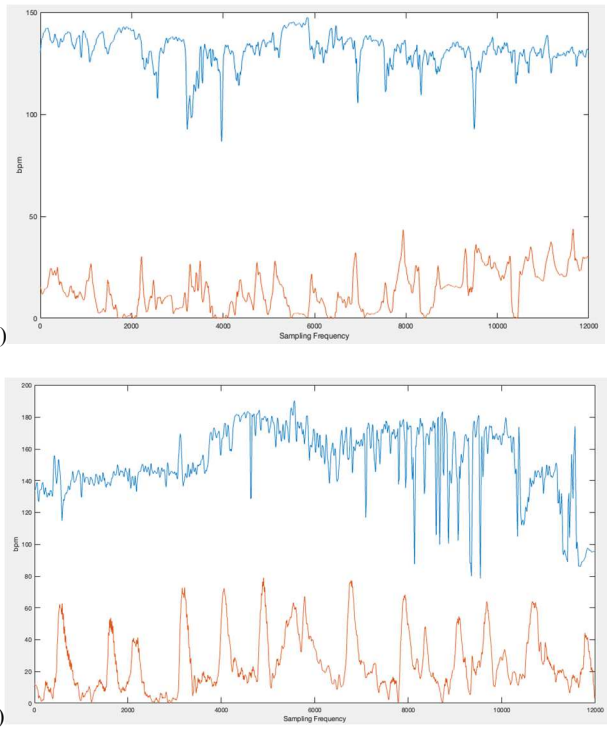

Figure 1 shows the comparison between pre-processed CTG where recordings (a) have high Apgar score and (b) have low Apgar score. bpm stands for beats per minute. The top line in blue represents the FHR and the bottom line in orange represents the UC.

C. Label categories

Using the same original dataset, we created five different subsets with various low Apgar score boundaries to investigate how changes in the boundaries affect the performance metrics in classifying abnormal CTG. Categories of low Apgar scores were: 1) less than ten*, 2)* less than nine, 3) less than eight and 4) less than seven. We also classified CTG using pH as a benchmark to compare the results between Apgar scores and pH for classifying hypoxia using CTG. We set a pH of less than 7.05 as low pH.

D. Classification

We used Scikit-learn for modelling the random forest (RF) and multi-level perceptron (MLP). The data was split into train and test subsets using stratified k fold (k = 3). We used the synthetic minority oversampling technique to increase the number of samples. We only oversampled the training set and the test set remained imbalanced (table 1). Grid search with cross-validation (k = 3) (GridsearchCV) was used for hyperparameter tuning on the training subset to boost the model performances; the best parameter was chosen for the final model [16]. The classification model was evaluated on a separate test subset.

E. Performance metrics

A confusion matrix was used to measure the true positive (TP), true negative (TN), false positive (FP) and false negative (FN) values. TP represents the correct classification of positive samples, TN represents the correct classification of negative samples, FP represents the wrong classification of positive samples, and FN represents the incorrect classification of negative samples (Hicks et al., 2022). Based on those values, we calculated the area under the receiving operator characteristic (AUROC), precision (P), recall (R) and f1 for each classifier.

III. Results

A. Extracted features

From the signal processing step, we extracted a total of 22 features from each CTG. The NICE guidelines features were the number of accelerations, number of decelerations, longest acceleration length, longest deceleration length, average baseline, short-term and long-term variability.

The FHR features extracted were the mean, standard deviation, sample entropy, approximate entropy, cross-entropy, Lempel-Ziv, delta, total delta, number of intrinsic mode functions, power spectral density (PSD) of very low frequency (0 – 0.03 Hz), PSD of low frequency (0.03-0.15 Hz), PSD of the medium frequency (0.15-0.5 Hz), PSD of high frequency (0.5-1 Hz), the ratio of PSD of low and PSD high frequency, the ratio of PSD of low frequency and PSD of medium and high frequency.

B. Parameter optimization

By using GridsearchCV, four RF parameters were optimized: 1) the number of trees (50), 2) criteria to measure the quality of split ('gini'), 3) the depth of the tree (5) and 4) the method to select the number of features considered when looking for the best split (square root). Other parameters were set to default. For MLP, four parameters were optimized: 1) the number of hidden layers (350), 2) the method of activation of hidden layers (rectified linear unit function), 3) solver weight for optimization ('lbfgs' - optimizer in the family of

quasi-Newton methods) and 4) learning rate schedule for weight updates (constant). Other parameters were set to default. Apgar scores and pH levels resulted in the same optimized parameter.

C. Performance comparisons

The number of CTG with low Apgar scores changes when we change the boundaries as we did not oversample the test set (table 1). The best AUROC score for MLP (60.47%) and RF (68.12%) was achieved when we categorized low Apgar scores as less than 9. When the low Apgar score was set to less than 10, our model had the best overall performance: RF (P – 68.31%, R – 59.33% and f1 – 63.28%) and MLP (P – 61.68%, R – 57.78% and f1 – 59.61%). The performances were the lowest when a low Apgar was less than 7. When compared to pH as the benchmark, the RF classifier had the highest AUROC, which can discriminate well between CTGs that have low pH (table 2).

Label	Low Apgar Boundary	Number of good	Number of low
Apgar	<10	107	77
	<9	145	39
	<8	168	16
	<7	177	7
pH	<7.05	170	14

Table 1: Distribution of low Apgar and good Apgar scores in the test set using different boundaries to define low Apgar scores and distribution of low and good pH level.

Boundary for low	Model	f1 (%)	P (%)	R (%)	AUROC (%)
Apgar <10	RF	63.28	68.31	59.33	63.21
	MLP	59.61	61.68	57.78	56.31
Apgar <9	RF	41.84	36.7	49.17	68.12
	MLP	34.37	35.93	33.93	60.47
Apgar <8	RF	10.62	5.29	11.11	50.23
	MLP	15.41	12.22	8.89	52.91
Apgar <7	RF	15.24	6.23	27.78	56.99
	MLP	10.52	2.56	5.56	52.39
pH <7.05	RF	62.31	56.75	58.33	**76.18**
	MLP	34.43	35.50	34.65	**75.09**

Table 2: Performances of RF and MLP with different boundaries of Apgar score and comparison to cord blood pH

IV. Discussion

We are the first to use the Apgar score as a benchmark for classifying hypoxic CTGs. Our previous study demonstrated how oversampling of both training and train subsets massively improved CTG classification [17]. However, in this study, we only oversampled the training set to mimic real-life events, since there is a significantly small number of hypoxia cases compared to healthy fetuses.

We showed that the Apgar score defined as less than 10 showed higher performances in f1, precision and recall, among other boundaries of Apgar scores, while the highest AUROC was obtained when the low Apgar score is defined as less than 9. This is because there were more samples of low Apgar scores when "low" was defined as less than 10 or less than 9 compared to scores of less than 8 and less than 7 (table 1). Modelling was challenging because there were so few cases of CTG with Apgar scores < 8. P, R and f1 were particularly low for Apgar scores of less than 7 as the test set was severely imbalanced. There were only 7 CTG in the low Apgar group in the test set which provided an insufficient number of samples for learning this class.

From the results, it seems ideal to set the low Apgar scores of less than 10 as the benchmark of hypoxia, in terms of prediction performance. In clinical practice, however, scores between 7 to 10 are considered healthy [18, 19]. This means that scores of 7, 8 and 9 will be misclassified as hypoxic, increasing the number of cases of false positives and exposing mothers and babies to the unnecessary risk of cesarean section and other interventions, making the model clinically irrelevant.

Compared to pH as a benchmark, results show higher discrimination against hypoxic CTG. Compared to previous studies that used pH as the benchmark and the same data source, their performances were much higher than ours. However, due to the lack of a gold standard, these studies used various pH range boundaries between pH less than 7.01 and pH less than 7.20. Other studies also used different classifiers and various methods to pre-process and extract CTG features [20-22]. Therefore, it is difficult to compare those studies with our results.

Our study is limited by its small sample size as we used an open-access database with a relatively small number of patients. We acknowledge that our dataset is very imbalanced, especially when low Apgar scores were less than 7. While we used oversampling techniques, the sample size is still small to utilize ML algorithms fully. Future studies would benefit from a larger sample size and a mixture of geographical regions. In addition, this study is not externally validated. We plan to collect CTG data from hospitals to enable external validation to improve the quality of the study and to increase model generalizability. We also plan to investigate cost-sensitive methods to assess imbalanced classification in terms of clinical requirements and costs. We will explore the use of other deep learning methods such as the recurrent neural network and the long-short term memory to investigate if those methods can improve classification performance. FHR and UC measurements are time-dependent, and their classification would likely benefit from those deep learning methods. Lastly, future studies should use relevant techniques to determine the important feature of CTG classification.

V. Conclusion

We demonstrated that 5 minutes Apgar score could be used to distinguish between hypoxic and healthy CTGs for this dataset. We also showed how the degree of data imbalance in the test set affects the performance metrics. We highlighted the lack of gold standards in benchmarking CTGs. Since Apgar scores reflect babies' ability to recover from intrapartum hypoxia, it is a more relevant surrogate marker to distinguish unhealthy babies compared to pH cord blood, a one-off measure. The challenge of modelling healthcare data is that the number of diseases or cases is significantly imbalanced.

References

[1] B. Petterson, J. Bourke, H. Leonard, P. Jacoby, and C. Bower, "Co-occurrence of birth defects and intellectual disability," (in eng), *Paediatr Perinat Epidemiol,* vol. 21, no. 1, pp. 65-75, 2007/01// 2007, doi: 10.1111/j.1365-3016.2007.00774.x.

[2] G. Bogdanovic, A. Babovic, M. Rizvanovic, D. Ljuca, G. Grgic, and J. Djuranovic-Milicic, "Cardiotocography in the prognosis of perinatal outcome," (in eng), *Medical archives (Sarajevo, Bosnia and Herzegovina),* vol. 68, no. 2, pp. 102-105, 2014, doi: 10.5455/medarh.2014.68.102-105.

[3] C. E. Wood and M. Keller-Wood, "Current paradigms and new perspectives on fetal hypoxia: implications for fetal brain development in late gestation," *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology,* vol. 317, no. 1, pp. R1-R13, 2019, doi: 10.1152/ajpregu.00008.2019.

[4] L. P. Thompson, S. Crimmins, B. P. Telugu, and S. Turan, "Intrauterine hypoxia: clinical consequences and therapeutic perspectives," *Research and reports in neonatology,* vol. 5, pp. 79-89, 2015.

[5] Z. Alfirevic, D. Devane, G. M. L. Gyte, and A. Cuthbert, "Continuous cardiotocography (CTG) as a form of electronic fetal monitoring (EFM) for fetal assessment during labour," in *Cochrane Database of Systematic Reviews* vol. 2017, ed, 2017.

[6] E. Gyllencreutz, I. Hulthén Varli, P. G. Lindqvist, and M. Holzmann, "Reliability in cardiotocography interpretation - impact of extended on-site education in addition to web-based learning: an observational study," (in eng), *Acta Obstet Gynecol Scand,* vol. 96, no. 4, pp. 496-502, Apr 2017, doi: 10.1111/aogs.13090.

[7] S. Das, H. Mukherjee, K. Roy, and C. K. Saha, "Shortcoming of Visual Interpretation of Cardiotocography: A Comparative Study with Automated Method and Established Guideline Using Statistical Analysis," *SN Computer Science,* vol. 1, no. 3, p. 179, 2020/05/22 2020, doi: 10.1007/s42979-020-00188-x.

[8] A. H. MacLennan, S. C. Thompson, and J. Gecz, "Cerebral palsy: causes, pathways, and the role of genetic variants," (in eng), *Am J Obstet Gynecol,* vol. 213, no. 6, pp. 779-88, Dec 2015, doi: 10.1016/j.ajog.2015.05.034.

[9] E. Mullins, C. Lees, and P. Brocklehurst, "Is continuous electronic fetal monitoring useful for all women in labour?," *BMJ,* vol. 359, p. j5423, 2017, doi: 10.1136/bmj.j5423.

[10] R. M. Grivell, Z. Alfirevic, G. M. L. Gyte, and D. Devane, "Antenatal cardiotocography for fetal assessment," (in eng), *Cochrane Database Syst Rev,* vol. 2015, no. 9, pp. CD007863-CD007863, 2015, doi: 10.1002/14651858.CD007863.pub4.

[11] P. Fergus, M. Selvaraj, and C. Chalmers, "Machine learning ensemble modelling to classify caesarean section and vaginal delivery types using Cardiotocography traces," *Computers in Biology and Medicine,* vol. 93, pp. 7-16, 2018/02/01/ 2018, doi: https://doi.org/10.1016/j.compbiomed.2017.12.002.

[12] P. Yeh, K. Emary, and L. Impey, "The relationship between umbilical cord arterial pH and serious adverse neonatal outcome: analysis of 51 519 consecutive validated samples," *BJOG: An International Journal of Obstetrics & Gynaecology,* vol. 119, no. 7, pp. 824-831, 2012, doi: https://doi.org/10.1111/j.1471-0528.2012.03335.x.

[13] V. Apgar, "A Proposal for a New Method of Evaluation of the Newborn Infant," *Anesthesia & Analgesia,* vol. 32, no. 4, pp. 260-267, 1953. [Online]. Available: https://journals.lww.com/anesthesia-analgesia/Fulltext/1953/07000/A_Proposal_for_a_New_Method_of_Evaluation_of_the.6.aspx.

[14] L. V. Simon, M. F. Hashmi, and B. N. Bragg, "APGAR score," 2017.

[15] V. Chudáček *et al.*, "Open access intrapartum CTG database," *BMC Pregnancy and Childbirth,* vol. 14, no. 1, p. 16, 2014/01/13 2014, doi: 10.1186/1471-2393-14-16.

[16] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *the Journal of machine Learning research,* vol. 12, pp. 2825-2830, 2011.

[17] F. Francis, H. Wu, S. Luz, R. Townsend, and S. Stock, "Detecting Intrapartum Fetal Hypoxia from Cardiotocography Using Machine Learning," in *2022 Computing in Cardiology (CinC)*, 4-7 Sept. 2022 2022, vol. 498, pp. 1-4, doi: 10.22489/CinC.2022.339.

[18] A. C. o. O. G. A. A. o. Pediatrics, "Committee Opinion No. 644: The Apgar Score," *Obstet Gynecol*, vol. 126, no. 4, pp. e52-5.

[19] I. Aliyu, T. Lawal, and B. Onankpa, "Hypoxic-ischemic encephalopathy and the Apgar scoring system: The experience in a resource-limited setting," *Journal of Clinical Sciences,* Original Research Report vol. 15, no. 1, pp. 18-21, January 1, 2018 2018, doi: 10.4103/jcls.jcls_102_17.

[20] G. Georgoulas, P. Karvelis, J. Spilka, V. Chudáček, C. D. Stylios, and L. Lhotská, "Investigating pH based evaluation of fetal heart rate (FHR) recordings," (in eng), *Health Technol (Berl),* vol. 7, no. 2, pp. 241-254, 2017, doi: 10.1007/s12553-017-0201-7.

[21] Z. Zhao, Y. Deng, Y. Zhang, Y. Zhang, X. Zhang, and L. Shao, "DeepFHR: intelligent prediction of fetal Acidemia using fetal heart rate signals based on convolutional neural network," *BMC Medical Informatics and Decision Making,* vol. 19, no. 1, p. 286, 2019/12/30 2019, doi: 10.1186/s12911-019-1007-5.

[22] Z. Cömert, A. Şengür, Ü. Budak, and A. F. Kocamaz, "Prediction of intrapartum fetal hypoxia considering feature selection algorithms and machine learning models," (in eng), *Health Inf Sci Syst,* vol. 7, no. 1, p. 17, Dec 2019, doi: 10.1007/s13755-019-0079-z.