Detecting Intrapartum Fetal Hypoxia from Cardiotocography Using Machine Learning

Farah Francis[1], Honghan Wu[2], Saturnino Luz[1], Rosemary Townsend[1], Sarah Stock[1]

[1]University of Edinburgh, UK
[2]University College London, UK

Abstract

Intrapartum cardiotocography (CTG) can identify babies at risk of fetal hypoxia by detecting changes in fetal heart rate and uterine contractions during labour. However, variability in CTG interpretations affects intervention timings. Machine learning (ML) has been applied to this problem and shown success.

We proposed to use a 5-minute Apgar score as the benchmark for hypoxia in our ML algorithms as it has shown a high correlation with abnormal CTG and a better clinical support decision making than pH umbilical cord blood.

We used the CTU-UHB database containing 552 CTGs. We trained and compared five algorithms of decision tree (DT), random forest (RF), support vector machine (SVM), k-Nearest Neighbour (kNN) and artificial neural network (ANN). Performances were evaluated using precision, recall, F1 score and AUROC.

The ANN with 4 deep layers had the highest recall (100%), while the RF classifier had the highest F1 (97%), AUROC (99.73%) and precision (97%) (Table 1). The longest deceleration length is the most important features among a total of 21 features.

Apgar scores can be used as a surrogate hypoxia marker for classifying CTG, making the model results more informative for clinical decision making.

1. Introduction

Fetal hypoxia occurs when there is an interruption of constant oxygen supply to the baby during labour. Fetal hypoxic injury can cause intrapartum stillbirth, neonatal encephalopathy, neonatal death and disabilities [1-3]. While a level of hypoxic stress can be anticipated during labour when uterine contracts (UC), the main challenge is identifying that small number of babies where the natural physiological protective mechanisms fail to counteract for the hypoxic stress [4]. Fetal monitoring during labour is crucial to prevent the devastating effects of fetal hypoxia on babies and families. However, it must also be discriminatory enough to minimise unnecessary interventions in the form of surgical birth (caesarean section) that carry their own risks to both mother and baby [5].

Cardiotocography (CTG) have been widely used to as an electronic fetal monitoring device that can indicate fetal wellbeing in the uterus during labour. It is attached to the mother's uterus and it measures the fetal heart rate (FHR) changes in conjunction with UC. Clinicians will classify if the fetus's condition is reassuring, non-reassuring or pathological [6]. Based on the classification, clinicians can take steps to reduce the effect of hypoxia such as assisted birth to reduce the harmful effects on newborns [7].

Since the introduction of CTG in 1970, research have shown inconsistencies in the interpretation of visual CTG amongst clinicians can result in a delayed response due to the time taken to achieve an agreement [8, 9]. Furthermore, some decision making can be subjective and with some level of ambiguity which may contribute to discrepancies in CTG interpretation [10]. Due to the false positive cases – babies are deemed as hypoxic when they are not, there been a fivefold increase in caesarean sections rates while cerebral palsy prevalence remains unchanged [11]

To tackle the shortcomings of visual CTG, computerised CTG was introduced to aid in decision making for abnormal FHR by standardising interpretations allowing a quicker response to compromised fetuses. A randomised controlled trial and retrospective studies has shown that computerised CTG improved the quality of interpretations while minimising decision making time [12]. However, a meta-analysis of six studies showed no significant improvement in fetal well-being between visual and computerised CTG both in antenatal and intrapartum measurements [13].

Researchers who used machine learning (ML) on CTG data have demonstrated promising results in classifying fetal hypoxia. ML learning can improve fetal hypoxia detection while reducing interpretation variability between clinicians. Previous studies used varying pH umbilical cord blood level as a benchmark for hypoxia and showed promising outcomes. Umbilical cord blood cord pH are taken immediately when babies are born and

this does not reflect their ability to recover from birth stress [14]. Hence, we proposed using 5 minutes Apgar score as the surrogate marker of hypoxia in our ML algorithms. Low Apgar scores have shown a high correlation with hypoxic diagnosis and abnormal CTG, and it is a routine, standardised measurement of babies' physiology and condition after birth such as appearance, pulse, grimace, activity and respiration. Evidence showed that babies do recover from birth stress, where there are differences in the Apgar score taken at 1 minute and 5 minutes after birth. Studies had shown that low Apgar scores are associated with diagnosis of hypoxia and cerebral palsy [15, 16]. Therefore, 5 minutes Apgar after birth is a good indicator if babies have the ability to recover and a better clinical support decision making compared to pH umbilical cord blood [17].

2. Methodology

2.1. Dataset

We used raw CTG from the CTU-UHB database, which consists of 552 CTG recordings sampled at 4Hz, and the recording was taken no longer than 90 minutes during labour (second stage of labour). CTG records were taken between 2009 and 2012 at the University Hospital in Brno, Czech Republic [18]. The Apgar score ranges from 0 to 10, where 0 is very unhealthy and 10 is healthy. Our study used Apgar scores from 10 to 7 for healthy and 6 to 0 for hypoxic where our model is trained to classify between these two categories.

2.2. Feature Extraction

We used both FHR and UC for this study. 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 moving mean.

For morphological features, we extracted acceleration, deceleration, average baseline and long and short-term heart rate variability of FHR in conjunction with UC as recommended by the National Institute for Health and Care Excellence guidelines for CTG interpretations [19]. For time domain, frequency domain and non-linear features, we only used FHR signals. We extracted 21 features, which were used to build the classification model.

2.3. Classification

We used Scikit-learn for modelling, and five ML classifiers were used to compare the performances, which include decision tree (DT), random forest (RF), support vector machine (SVM), k nearest neighbours (kNN) and artificial neural network (ANN). Due to the small sample size, we used oversampled using the Synthetic Minority Oversampling Technique (SMOTE) to increase the number of samples. The data was split into two subsets: train (70%) and test (30%). We performed 5-fold cross validation on the training set. GridSearchCV was used for hyperparameter tuning on the training subset to boost the model performances where the best parameter was chosen for the final model [20]. The classification model was evaluated on a separate test subset.

2.4. Model Evaluation

We used the confusion matrix to measure the true positive (TP), true negative (TN), false positive (FP) and false negative (FN) values. TP represents the correct classification for positive samples, TN represents the correct classification of negative samples, FP represents the wrong classification for positive samples, and FN represents the wrong classification for negative samples [21]. Based on those values, we calculated precision, recall, F1 score and area under the receiving operator characteristics (AUROC).

$$\text{Precision (P)} = TP / (TP + FP) \quad (1)$$

$$\text{Recall (R)} = TP / (TP+FN) \quad (2)$$

$$\text{F1 score} = 2 \times (P \times R) / (P + R) \quad (3)$$

3. Results

By using SMOTE, the dataset was increased from 552 to 1066. The oversampled group is the hypoxic Apgar scores, where the sample size increased from 19 to 533 subjects (figure 1). Based on the performance metrics, ANN with four deep layers, rectified linear unit activation and ADAM optimiser has the highest recall (100%), while the RF classifier has the highest F1 score (97.00%), AUROC (99.73) and precision (97.00%) (Table 1). Other results from different classifiers are all recorded in table 1. In general, all five classifiers show promising results where most of the performance metrics score more than 75%, except for the F1 score, precision and AUROC for ANN. Using the RF algorithm, we identified the top three important features: longest FHR deceleration, Lempel-Ziv complexity and number of intrinsic mode functions.

Figure 1 shows the distribution between healthy and unhealthy babies in the training and test subset

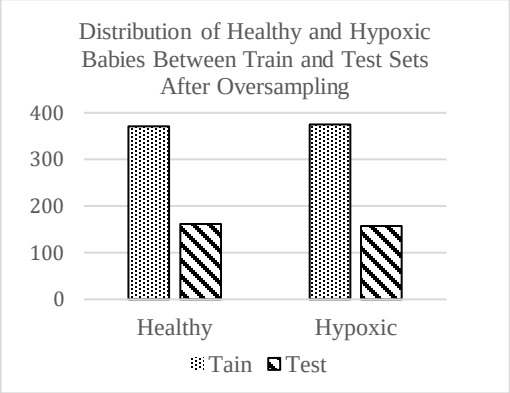
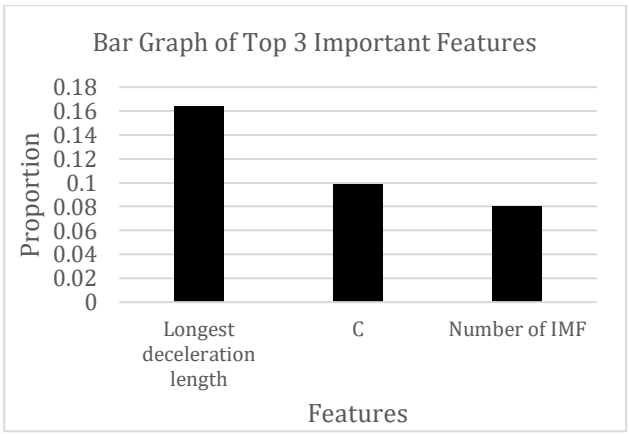


Table 1 shows the comparison of performances metrics between classifiers

Classifiers	P (%)	R (%)	F1 (%)	AUROC (%)
DT	93	88	90	94.8
RF	**98**	98	**98**	**99.8**
SVM	72	71	71	76.9
kNN	70	85	77	81.1
ANN	51	**100**	67	0.5

Figure 2 shows the top three important feature calculated using the RF algorithm



4. Discussion

Compared with previous studies, our performances are as high as those that used pH levels as a surrogate marker for hypoxia [22, 23]. This indicates that Apgar scores are as good as pH levels in classifying hypoxia for this dataset.

Interestingly, two of the top features are from the time domain (longest FHR deceleration and average baseline), and one is from the non-linear domain (number of intrinsic mode functions) (figure 2). This shows that other domains of CTG are useful in distinguishing hypoxic fetuses compared to the traditional morphological changes suggested by clinical guidelines and discrete signal processing techniques is crucial in interpreting CTGs.

One of the limitations of this study is the number of samples. While we employed oversampling techniques, the small sample size is still small for an ML study. Future studies would benefit from a larger sample size and a mixture of geographical regions to increase model generalisability. Next, we oversampled the train and test set, where there is an equal number of healthy and hypoxic fetuses. However, in real life, the number of hypoxic fetuses is very small, demonstrating a severe imbalance between healthy and cases of hypoxia. Therefore, we need to create a detection model that can be implemented in real-life situations and is relevant in clinical settings. In addition, we tried to compare previous studies that used pH umbilical cord blood and we found it difficult as previous studies used various pH benchmarks and selective performance measures when reporting their outcomes.

5. Conclusion

Our study shows that 5 minutes Apgar score can be used to distinguish between hypoxic and healthy CTGs for this dataset and achieved performances as high as studies using pH levels. Since Apgar scores reflect babies' ability to recover from intrapartum hypoxia, it is a more relevant surrogate marker to distinguish unhealthy babies. We can benefit from an external validation dataset to make our model clinically relevant and more generalisable for the overall population. We also plan to integrate other obstetrics factors to improve classifications make our model more clinically relevant.

References

1. Petterson B, Bourke J, Leonard H, Jacoby P, Bower C. Co-occurrence of birth defects and intellectual disability. Paediatr Perinat Epidemiol. 2007;21(1):65-75.

2. Bogdanovic G, Babovic A, Rizvanovic M, Ljuca D, Grgic G, Djuranovic-Milicic J. Cardiotocography in the prognosis of perinatal outcome. Medical archives (Sarajevo, Bosnia and Herzegovina). 2014;68(2):102-5.
3. Wood CE, Keller-Wood M. 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. 2019;317(1):R1-R13.
4. Thompson LP, Crimmins S, Telugu BP, Turan S. Intrauterine hypoxia: clinical consequences and therapeutic perspectives. Research and reports in neonatology. 2015;5:79-89.
5. Sandall J, Tribe RM, Avery L, Mola G, Visser GHA, Homer CSE, et al. Short-term and long-term effects of caesarean section on the health of women and children. The Lancet. 2018;392(10155):1349-57.
6. Ayres-de-Campos D, Spong CY, Chandraharan E, Panel FIFMEC. FIGO consensus guidelines on intrapartum fetal monitoring: Cardiotocography. International Journal of Gynecology & Obstetrics. 2015;131(1):13-24.
7. Alfirevic Z, Devane D, Gyte GML, Cuthbert A. Continuous cardiotocography (CTG) as a form of electronic fetal monitoring (EFM) for fetal assessment during labour. Cochrane Database of Systematic Reviews2017.
8. Blackwell SC, Grobman WA, Antoniewicz L, Hutchinson M, Gyamfi Bannerman C. Interobserver and intraobserver reliability of the NICHD 3-Tier Fetal Heart Rate Interpretation System. American Journal of Obstetrics and Gynecology. 2011;205(4):378.e1-.e5.
9. Gyllencreutz E, Hulthén Varli I, Lindqvist PG, Holzmann M. Reliability in cardiotocography interpretation - impact of extended on-site education in addition to web-based learning: an observational study. Acta Obstet Gynecol Scand. 2017;96(4):496-502.
10. Das S, Mukherjee H, Roy K, Saha CK. Shortcoming of Visual Interpretation of Cardiotocography: A Comparative Study with Automated Method and Established Guideline Using Statistical Analysis. SN Computer Science. 2020;1(3):179.
11. MacLennan AH, Thompson SC, Gecz J. Cerebral palsy: causes, pathways, and the role of genetic variants. Am J Obstet Gynecol. 2015;213(6):779-88.
12. Dawes GS, Lobb M, Moulden M, Redman CW, Wheeler T. Antenatal cardiotocogram quality and interpretation using computers. Bjog. 2014;121 Suppl 7:2-8.
13. Grivell RM, Alfirevic Z, Gyte GML, Devane D. Antenatal cardiotocography for fetal assessment. Cochrane Database Syst Rev. 2015;2015(9):CD007863-CD.
14. Yeh P, Emary K, Impey L. 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. 2012;119(7):824-31.
15. Hogan L, Ingemarsson I, Thorngren-Jerneck K, Herbst A. How often is a low 5-min Apgar score in term newborns due to asphyxia? Eur J Obstet Gynecol Reprod Biol. 2007;130(2):169-75.
16. Persson M, Razaz N, Tedroff K, Joseph KS, Cnattingius S. Five and 10 minute Apgar scores and risks of cerebral palsy and epilepsy: population based cohort study in Sweden. BMJ. 2018;360:k207.
17. Simon LV, Hashmi MF, Bragg BN. APGAR score. 2017.
18. Chudáček V, Spilka J, Burša M, Janků P, Hruban L, Huptych M, et al. Open access intrapartum CTG database. BMC Pregnancy and Childbirth. 2014;14(1):16.
19. NICE. National Institute for Health and Care Excellence: Clinical Guidelines. Intrapartum Care: Care of Healthy Women and Their Babies During Childbirth. London: National Institute for Health and Care Excellence (UK)
Copyright © 2014 National Collaborating Centre for Women's and Children's Health.; 2014.
20. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. the Journal of machine Learning research. 2011;12:2825-30.
21. Hicks SA, Strümke I, Thambawita V, Hammou M, Riegler MA, Halvorsen P, et al. On evaluation metrics for medical applications of artificial intelligence. Scientific Reports. 2022;12(1):5979.
22. Spilka J, Frecon J, Leonarduzzi R, Pustelnik N, Abry P, Doret M. Sparse Support Vector Machine for Intrapartum Fetal Heart Rate Classification. IEEE Journal of Biomedical and Health Informatics. 2016;21:1-.
23. Cömert Z, Kocamaz A. A Comparison of Machine Learning Techniques for Fetal Heart Rate Classification2016.

Address for correspondence:

My Name: Farah Francis
My Full postal address: Usher Institute, 9 Little France Road, Edinburgh EH16 4UX, UK
My E-mail address: farah.francis@ed.ac.uk