Hybrid Scattering Transform - Long Short-Term Memory Networks for Intrapartum Fetal Heart Rate Classification

Derek Kweku Degbedzui[1], Michael Kuzniewicz[2], Cornet Marie-Coralie[3], Yvonne Wu[3]
Heather Forquer[2], Lawrence Gerstley[2], Emily Hamilton[4], Doina Precup[1]
Philip Warrick [1,2], Robert Kearney [1]

[1] McGill University, Montreal, Canada; [2] Kaiser Permanente, Northern California, USA
[3] University of California, San Francisco, USA; [4] PeriGen Inc., Montreal, Canada

Abstract

This study assessed the early detection of the risk of hypoxic ischemic encephalopathy using raw fetal heart rate and its transformation with scattering transform and a long short-term memory recurrent neural network. There was no significant difference between the two approaches. However, the use of scattering transform produced lower computational demands. Considering scalability to the large data in our database and computational efficiency, the experiments involving scattering transform coefficients will be selected to conduct subsequent experiments. Future works will address the limitations of this study, including the low model performance.

1. Introduction

The evaluation of both maternal and fetal health status is carried out during labor using cardiotocography (CTG) which measures both fetal heart (FHR) and uterine pressure (UP). Healthcare providers rely on the visual inspection and interpretation of these CTG signals to identify poor fetal oxygenation and take prompt actions to prevent neonatal mortality or adverse outcomes [1].

A reduction in oxygen supply (hypoxia) and blood flow (ischemia) to a newborn's brain caused by intrapartum events leads to hypoxic-ischemic encephalopathy (HIE) [2]. This condition is marked by seizures, diminished levels of consciousness, and respiratory challenges [3]. Neonates diagnosed with neonatal HIE experience significant long-term impairments such as hearing loss, cognitive impairments, and cerebral palsy [4]. The HIE incidence rate is approximately 1-3 cases per 1,000 live births in developed nations [5]. In contrast, low- and middle-income countries experience a higher incidence rate ranging from 10-20 cases per 1,000 live births [5].

The visual examination and interpretation of CTG signals poses some challenges. One notable issue is the significant inter- and intra-observer variability, which has produced an increased frequency of assisted deliveries and CS during childbirth [1]. Despite the availability of clinical guidelines, there is a lack of specific management recommendations for about 80% of FHR signals that are classified as “indeterminate risk” [6]. The above challenges, further exacerbated by the low specificity of current methods, contributes to missed diagnoses and unnecessary CS deliveries [7]. Unnecessary CS exposes the expectant mother and the unborn child to risk without commensurate benefits [8] while missed diagnoses can produce irreversible brain injury and worse outcomes for the child.

Traditional machine learning (ML) techniques have been examined to enhance the automated interpretation of CTG signals. Nevertheless, the application of feature engineering in ML has not yielded highly discriminative features for FHR classification. This challenge could be addressed using deep learning (Dl) methods. There have been some promising studies on the detection of fetal deterioration through DL. However, most of these studies did not investigate cases of HIE, instead, they focused only on hypoxia. Additionally, these studies employed relatively small datasets, which may not support the development of DL models, as these models typically require large datasets. Hence, there exists a need to investigate DL models on substantially large birth cohorts. Using our database, containing 250,000 CTG records, we explore the use of DL to improve the early detection of the risk of HIE. This study presents the preliminary results of assessing two approaches, using raw FHR and scattering transform (ST) coefficients computed for FHR signals, with a long short-term memory (LSTM) network to predict the intrapartum risk of HIE.

2. Method

2.1. Clinical data

The clinical data comprises up to 72 hours of FHR, UP and MHR signals obtained from singleton live births with $\geq$ 35 weeks of gestational age at 15 hospitals of Kaiser Permanente Northern California with pregnancy onset date from January 1, 2011, to December 31, 2018.

HIE was defined as the presence of both acidosis and encephalopathy. Neonatal encephalopathy was defined as a documented abnormal Sarnat score within the first 6 hours of age, while acidosis was defined as pH $<$ 7 or base deficit $\geq$ 10 mmol/L measured from the umbilical cord gas shortly after delivery. A further chart review by the clinicians was used to confirm the diagnosis of perinatal HIE. The healthy class was defined as the absence of both encephalopathy and acidosis, with no chest compression or intubation, discharged alive and an Apgar at 5 minutes $\geq$ 7. There were 39,213 healthy no acidosis cases, 3,223 acidosis cases and 417 HIE cases in our database. The low incidence rate of HIE accounts for the low number of HIE cases. This study focused on FHR signals from all 417 HIE cases and 418 randomly under-sampled healthy cases. A random undersampling approach was selected to address the highly imbalanced dataset to ensure the evaluation and selection of a good baseline performing model while preventing bias towards the majority healthy class.

2.2. Data preprocessing

PeriCALM Patterns, proprietary software from PeriGen Inc., was used to remove noise, identity artifacts and repair the FHR signals, which were sampled at 4Hz. The repaired FHR signals were divided into 20-minute non-overlapping segments. For experiments involving the raw FHR (labeled *RAW4*), FHR were decimated by a factor of 4 to reduce computational time. In experiments involving the time-frequency transformation of FHR (labeled *ST64*), the sampling rate was reduced by a factor of 64.

2.3. Scattering transform

The scattering transform is a time-frequency representation that applies a series of wavelet transforms to a signal to produce translation-invariant, stable and informative signal representations [9]. A wavelet filter (ψ) is applied to a series of layers. At each layer, a signal in convoluted with wavelet function(s) and passed through a non-linear modulus function. The wavelet filter can be represented as

$$\psi_j(t) = 2^{-j/Q}\psi(2^{-j/Q}t) \tag{1}$$

where ψ is the mother wavelet, Q is a constant number of filters per octave, and the scale J represents an integer ranging 0 to J. We used the default mother wavelet, Morlet wavelet, with a quality factor Q = 1 and a maximum wavelet scale of $J = 11$. Using $J = 11$ produces 1 zeroth-order, 12 first-order and 63 second-order paths resulting in 76 order paths. The time scale (T), which controls the degree of time invariance was modified to T=64. We investigated other time scales and found no significant difference in the classification performance but selected T=64 to reduce computational load.

The first three orders of the scattering transform are:

$$\mathrm{S}_0 x = x \star \phi$$
$$\mathrm{S}_1(t, j_1) = |x \star \psi_{j1}| \star \phi$$
$$\mathrm{S}_2(t, j_1, j_2) = ||x \star \psi_{j1}| \star \psi_{j2}| \star \phi$$

where ϕ represents a low-pass filter, ψ represents the wavelet transform, $||$ represent complex modulus and $\star$ represents convolution product. We concatenated the coefficients of the first three order paths to produce a time series for subsequent classification experiments.

2.4. Classification

A 10-fold cross-validation technique was selected to provide a more reliable estimate of the model's performance. The indexes for training, validation, and test were randomly permuted for each fold and stratified by class, resulting in the generation of ten distinct train, validation and test sets. Each fold comprised 5 repetitions with different random initialization of the weights to aid the identification of unstable models or models that suffered from over-fitting.

Classification experiments were conducted using a three-layer LSTM model, striking a balance between model performance and complexity. The number of cells in each hidden layer was 128, 256 and 128, respectively. For both RAW4 and ST64, the train and validation records were used train and validate the models.

The best performing model from the 5 repetitions of each fold was selected using the sensitivity and specificity performance for the validation set. Initially, models with average specificity exceeding 0.7 were selected, and subsequently, the model with the highest average sensitivity was selected for each fold. In the absence of average specificity exceeding 0.7, we selected the model with the highest average specificity and the highest average sensitivity. The approach of selecting the highest average sensitivity based on a minimum specificity was to ensure that the performance of the selected model did not exceed the current CS rate of 32% while detecting a substantial portion of pathological fetuses and keeping the false positives to a minimum.

To prevent overfitting, an early stopping criterion of 30 epochs was implemented. Training was configured with a maximum of 1000 epochs, a batch size of 32, the binary cross-entropy loss function and the Adam optimizer with learning rate of 0.0001.

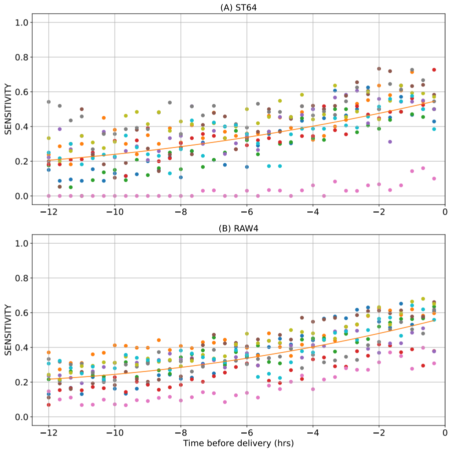

Figure 1. Sensitivity as a function of time before birth for classification experiments using (A) scattering coefficients (ST64), and (B) raw FHR (RAW). The solid line is a polynomial line is fitted to the 10-fold results for each dataset.

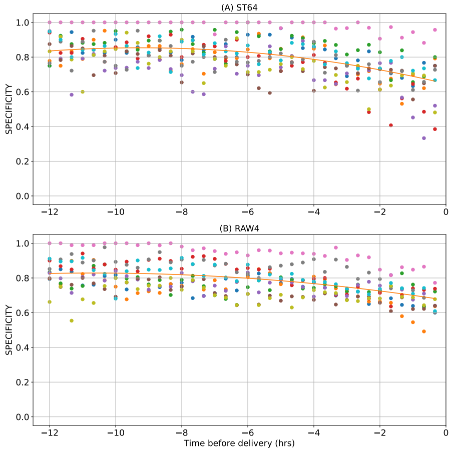

Figure 2. Specificity as a function of time before birth for classification experiments involving (A) scattering coefficients (ST64), and (B) raw FHR (RAW).

	Coefficient #1	Coefficient #2
	Mean [95% CI]	Mean [95% CI]
RAW4	-0.029 [-0.040, -0.019]	-0.001 [-0.002, -0.000]
ST64	-0.045 [-0.061, -0.030]	-0.002 [-0.003, -0.001]

Table 1. The coefficients of quadratic polynomial computed using bootstrapping technique on the 10-fold specificity performance.

All segments within the last 6 hrs of FHR before delivery were used for training and validation, while the trained models were evaluated using independent segments within the last 12 hrs. We saw no significant difference in performance between experiments conducted with either 12 hrs or 6 hrs, hence we restricted our training experiments to the last 6 hrs of FHR to reduce computational load.

To compare the performance of using RAW4 and ST64, the selected performance metrics (sensitivity, specificity, AUROC) were computed on the independent test datasets for the 10-fold cross validation. For each performance metric, a polynomial regression with an appropriate order was fitted to the metric performance across time. The optimal degree for each polynomial regression fitted to a performance metric was selected using minimum description length (MDL) principle. The rationale for using MDL is based on finding a trade-off between an appropriate polynomial order that fits the data well while avoiding overfitting. Increasing the degree of a polynomial regression can produce a complex model that fits to the noise within the data rather than the true underlying pattern.

A bootstrapping approach was used to estimate the mean and 95% confidence interval (CI) of the polynomial coefficients. Non-overlapping CI indicates statistical significance, suggesting a significant difference between the compared models. Bootstrapping was chosen due to the small sample size and non-normal distribution of performance metrics.

3. Results

Figures 1 and 2 show the various fold performance of sensitivity and specificity for experiments using ST64 and RAW4 on the test data as functions of time before birth, respectively. In the scatter plots, each color represents one of the 10 folds across the time of delivery. For both ST64 and RAW4, the quadratic polynomial curve for sensitivity initiated at approximately 0.2, displaying an upward-opening parabolic trend. Conversely, the quadratic polynomial curve for specificity commenced slightly above 0.8 and exhibited a downward-opening parabolic pattern towards delivery.

MDL consistently selected degree 1 and degree 2 for AUROC and specificity for both RAW4 and ST64, respectively. However, for sensitivity, degree 2 was selected for RAW4 while degree 1 was selected for ST64. Degree 2 was selected to analyse the sensitivity due to the visible non-linear trend of Fig. 1. Next, bootstrapping was used to compute the mean and 95% CI of the coefficients of the polynomial regression. The summary of the results is reported in Tables 1 and 2 for specificity and sensitivity performances of RAW4 and ST16. The overlap of the 95% CI for RAW4 and ST64 leads to the conclusion that there are no significant differences between the two approaches. We reached a similar conclusion for AUROC (results not shown).

The training times with the ST64 were very low compared to the training time for RAW4. The mean and stan-

	Coefficient #1	Coefficient #2
	Mean [95% CI]	Mean [95% CI]
RAW4	0.048 [0.037, 0.059]	0.002 [0.001, 0.003]
ST64	0.044 [0.034, 0.052]	0.001 [0.001, 0.002]

Table 2. The coefficients of quadratic polynomial computed using bootstrapping technique on the 10-fold sensitivity performance.

dard deviation of the training epoch duration for all repetitions of each fold using RAW4 and ST64 was 742.6 sec ± 202.1 sec and 10.0 sec ± 2.2 sec, respectively. This is attributed to the dimensionality reduction and extraction of relevant time-frequency features provided by scattering transform. Experiments involving RAW4 required greater number of epochs to converge due the reliance on the LSTM alone to generate discriminative representations.

4. Discussion

Sensitivity, which measures the model's detection of HIE cases, improved for RAW4 and ST64 as delivery approached. The highest sensitivity of approximately 0.60 achieved close to delivery, suggests that 4 out of every 10 HIE cases may still be misdiagnosed at this late stage of labour. Such a rate of misdiagnosis is undesirable within a clinical context. Furthermore, the inherent issue of class imbalance and the resulting low sensitivity may produce false negatives, highlighting the need for improvement. To maintain an acceptable false positive rate (FPR), the minimum acceptable specificity was set at the current CS rate. In future work, we will further explore the criteria to select the repetition with the best model for each fold. This will involve the investigation of the threshold applied to the softmax output probabilities in the final dense layer of our deep learning models, ensuring that it results in FPR lower than the current CS rate.

There were no statistically significant differences in the performance of experiments conducted using RAW4 and ST16. Our current experiments were constrained to a dataset comprising 417 HIE cases and 418 randomly undersampled healthy cases. This does not fully leverage the potential of our large dataset. In subsequent experiments, we will utilize ST64 as it imposes lower computational demands and offers a significant advantage in terms of scalability, especially when our future experiment extends to our larger dataset and the requirement for experiments with a shorter time frame becomes more evident.

5. Conclusion

This study aimed to assess the performance of two approaches with an LSTM network: one using raw FHR and the other using scattering transform coefficients of FHR signals . There was no statistically significant differences observed between these two approaches. Nevertheless, in consideration of computational efficiency and scalability, we favor conducting subsequent experiments using the scattering transform coefficients due to their lower computational demands. Thus, this study has limitations, particularly related to the observed low model performance. We plan to focus on improving model performance and addressing these limitations in our future work.

Acknowledgments

The research was supported by the Bill & Melinda Gates Foundation and the National Institutes of Health.

References

[1] Chauhan SP, Klauser CK, Woodring TC, Sanderson M, Magann EF, Morrison JC. Intrapartum nonreassuring fetal heart rate tracing and prediction of adverse outcomes: interobserver variability. American journal of obstetrics and gynecology 2008;199(6):623–e1.

[2] Verklan MT. The chilling details: hypoxic-ischemic encephalopathy. The Journal of Perinatal Neonatal Nursing 2009;23(1):59–68.

[3] American Academy of Pediatrics. Neonatal encephalopathy and neurologic outcome, 2^{nd} ed. report of the american college of obstetricians and gynecologists' task force on neonatal encephalopathy. Pediatrics 2014;133(5):e1482–e1488.

[4] Adstamongkonkul D, Hess DC. Ischemic conditioning and neonatal hypoxic ischemic encephalopathy: a literature review. Conditioning medicine 2017;1(1):9.

[5] Greco P, Nencini G, Piva I, Scioscia M, Volta C, Spadaro S, Neri M, Bonaccorsi G, Greco F, Cocco I, et al. Pathophysiology of hypoxic–ischemic encephalopathy. Acta Neurologica Belgica 2020;120(2):277–288.

[6] Macones GA, Hankins GD, Spong CY, Hauth J, Moore T. The 2008 national institute of child health and human development workshop report on efm: update on definitions, interpretation, and research guidelines. Journal of Obstetric Gynecologic Neonatal Nursing 2008;37(5):510–515.

[7] Hayashi M, Nakai A, Sekiguchi A, Takeshita T. Fetal heart rate classification proposed by the perinatology committee of the japan society of obstetrics and gynecology. Journal of Nippon Medical School 2012;79(1):60–68.

[8] Sung S, Mahdy H. Cesarean section. StatPearls 2019;.

[9] Chudáček V, Andén J, Mallat S, Abry P, Doret M. Scattering transform for intrapartum fetal heart rate variability fractal analysis: a case-control study. IEEE Transactions on Biomedical Engineering 2013;61(4):1100–1108.

Address for correspondence:

Derek Kweku DEGBEDZUI
Department of Biomedical Engineering, McGill University
derek.degbedzui@mail.mcgill.ca