A Lag-Attentive Transfer-Entropy Bottleneck for Quantifying UA-FHR Coupling from Cardiotocography

Mahdi Shamsi[1], Michael Kuzniewicz[2], Marie-Coralie Cornet[3], Yvonne Wu[3], John Parker[1], Lawrence Gerstley[2], Robert Kearney[5], Philip Warrick[1,5]

[1] PeriGen Inc., Montreal QC, Canada
[2] Kaiser Permanente Northern California, Oakland CA, USA
[3] University of California, San Francisco CA, USA
[5] McGill University, Montreal QC, Canada

Abstract

Intrapartum electronic fetal monitoring records fetal heart rate (FHR) together with uterine activity (UA). Conventional deep learning models in this context do not quantify how contraction history affects the fetal response. We present a causal forecasting model that produces two distributions over one latent state: one built from FHR history alone, and one that considers UA information through cross-attention over candidate delays in the last 6 minutes. A single shared decoder turns each distribution into a two-minute forecast of causal FHR time-frequency features. The model provides clinically relevant parameters: the gap between the log-likelihoods of the two forecasts measures what UA contributed; the divergence between the two distributions measures the magnitude of the UA influence on the latent state; and the attention weights estimate which delays the UA information came from. The model was trained on healthy deliveries with confirmed cord blood gas and evaluated on a held-out cohort that includes acidosis and hypoxic-ischaemic encephalopathy (HIE). The divergence rose steadily over the final 12 hours before delivery in all three classes.

1. Introduction

Cardiotocography (CTG) is widely used during labour to monitor FHR and UA, but visual interpretation is subjective and varies between observers [1,2]. Recently, deep-learning models have been developed to classify fetal status and estimate the risk of adverse neonatal conditions such as acidemia [3, 4]. Since labels for healthy CTG are abundant compared to the rare hypoxic-ischaemic encephalopathy (HIE) cases, we utilize healthy cases for a short-horizon forecasting of FHR as a self-supervised pre-training objective. It also consider this important, physiologically related question: does recent UA infromation, improve prediction beyond the information that already exists in FHR history? Transfer entropy (TE) formalizes that incremental predictability as the conditional mutual information $I(Y^+; X^- \mid Y^-)$, for source history X^- (here UA), target history Y^- (FHR) and target future Y^+ [5]. Estimating TE directly is difficult for noisy, high dimensional CTG; instead, the *transfer entropy bottleneck* approach learns a compressed source representation conditioned on target history [6]. Our earlier variational formulation used a single temporally encoded UA representation. The present model adds strictly backward-looking time-frequency inputs, causal Transformer history encoders, and cross-attention restricted to a set of candidate delays, such that a query built from FHR alone selects among delayed UA states. It reports three quantities: the *predictive gain*, the *divergence* between the two distributions, and that divergence *attributed across delays*. Only the first is evidence that UA is beneficial; the other two describe how far the distribution moved and where the model found the information.

2. Methods

Data and cohorts: We have used a large United States intrapartum CTG database sampled at 4 Hz, with one recording per delivery [7]. Recordings are divided into 20-minute segments drawn from the final 12 hours before delivery. Training uses *only* healthy deliveries that had cord blood-gas measurements; the other recordings, including adverse outcomes, and healthy deliveries with no cord blood-gas confirmation, were held out for evaluation and analysis alone, so the model never considered adverse outcomes during training. Acidosis is defined as umbilical artery pH < 7.0 or base deficit ≥ 10 mmol/L, and HIE as acidosis with neurological signs. We additionally stratified evaluation by availability of cord-blood gas measurements.

Backward-looking time-frequency features: Each seg-

ment is described on a four-second grid ($T = 300$ steps) by scattering and phase-harmonic transforms computed separately from FHR and from UA [8, 9]. The filters for scattering transform are *one-sided*, looking only backwards in time, so the value stored at step t uses signal samples at or before t. Conventional two-sided wavelet features also use samples after t, so part of the future being predicted would already be present in the inputs. FHR contributes $c_y = 102$ channels (36 scattering, 66 phase-harmonic) and UA $c_x = 51$ (36 and 15); the letters y and x denote the FHR and UA streams, respectively. One-sided filtering has two drawbacks, and both can be seen in the simplest such filter, an average of the last N seconds. First, there is nothing to average at the start of a recording until N seconds have passed. Let W_c be the number of steps channel c must wait before its output is valid. Low-frequency channels average over long windows and wait longest, and because the transform is computed per segment, the wait is spent once every 20 minutes rather than once per recording. Keeping only channels with $W_c \leq 134$ steps (536 s) retains 98 of the 102 FHR channels. Second, a moving average describes the middle of its window, not its leading edge, even after the wait, the value stored at step t summarizes signal centred τ_c seconds earlier, where τ_c is the filter-bank group delay.

Anchors: An *anchor* t is the last observed step before the model forecasts the FHR coefficients at steps $t+1, \ldots, t+H$. Validity is handled differently for model inputs and forecast targets, rather than differently for FHR and UA. For both input streams, a channel-specific mask is zero before that channel's wait W_c has elapsed and one afterwards. The masked coefficient and the mask itself are supplied together to the corresponding encoder, so neither encoder treats the transform's start-of-segment transient as observed signal.

Two distributions over one latent state: Figure 1 summarizes the model at one forecast anchor. The central idea is to predict the same FHR future under two information sets: FHR history alone, and FHR history together with UA. Let Y_t^- be the FHR history available at anchor t. The FHR encoder converts it to a state $h_t^y = E_y(Y_t^-)$ that supports the FHR-only prediction. In parallel, the UA pathway produces a sequence of states $h_j^x = E_x(X_j^-)$, one for each earlier step j. Retaining a sequence instead of pooling UA into one summary lets the model choose which earlier UA interval is relevant. Both pathways are causal, and, as the separate upper paths in Fig. 1 emphasize, UA has no route into the FHR-only state.

The FHR state defines a 64-dimensional Gaussian distribution p_t over a latent state z_t. This is the model's FHR-only belief about the forthcoming FHR. A second distribution, q_t, describes how that belief changes when the model is allowed to use the UA history:

$$\begin{aligned} p(z_t \mid Y_t^-) &= \mathcal{N}(\mu_t^p, \mathrm{diag}[(\sigma_t^p)^2]), \\ q(z_t \mid Y_t^-, X_t^-) &= \mathcal{N}(\mu_t^q, \mathrm{diag}[(\sigma_t^q)^2]). \end{aligned} \tag{1}$$

To form q_t, a query derived only from the FHR-only mean μ_t^p searches the UA states $h_{t-\ell}^x$ over delays $\ell \in \{0, \ldots, 90\}$, or as far as 360 s before the anchor. Sparse *entmax* weights identify the delays used by each of four attention heads [10]. Each head modifies its own 16-dimensional part of the latent distribution, so the resulting change can later be attributed across delays. The modification is expressed as a bounded correction to p_t and is initialized at zero; consequently, $q_t = p_t$ before learning begins. In the middle of Fig. 1, lag cross-attention is therefore the only point at which UA enters the model.

Finally, the two latent representations pass through the same decoder. It produces an FHR-only forecast from p_t and a UA-informed forecast from q_t, each covering $H = 30$ future steps and 98 retained FHR channels, a two-minute 30×98 array. Sharing the decoder is essential: the two forecasts differ because UA changed the latent distribution, not because they were generated by different prediction networks. These two paths and their outputs, labelled D_0 and D_1, appear in the lower half of Fig. 1.

Objective: Let D_0 and D_1 be the Gaussian negative log-likelihoods of the FHR-only and UA-informed forecasts, each summed over the $30 \times (32 + 66) = 2940$ coefficients in one forecast, 30 future steps for each of 32 retained scattering and 66 phase-harmonic FHR channels, and $K_t = D_{\mathrm{KL}}(q_t \,\|\, p_t)$ their Kullback-Leibler divergence. Averaged over valid anchors, with the subscript dropped for the averages, the optimized loss is

$$\mathcal{L} = D_0 + D_1 + \beta K + \beta_p R_p, \tag{2}$$

where $R_p = D_{\mathrm{KL}}[\mathcal{N}(\mu^p, \mathrm{diag}[(\sigma^p)^2]) \,\|\, \mathcal{N}(\mu^p, I)]$ keeps the variance of p_t from collapsing. In training only, coefficients carry per-channel weights in a $10:1$ scattering to phase-harmonic ratio (renormalized to sum to the channel count), since two thirds of the retained channels are phase-harmonic; the evaluation applies no weights, so its scores remain log densities. We increase the divergence weight β linearly from 0 to 20 over the first 50 epochs, set $\beta_p = 0.1$, and impose no lower bound on K. We optimize the model with AdamW at a learning rate of 3×10^{-4} after a 2000-step linear learning-rate warm-up. During training, we space anchors one forecast horizon apart ($S = H = 30$) so that their windows tile the segment rather than overlap, and we consider an anchor valid only when at least 90% of its forecast window is valid. During validation and evaluation, we use every valid anchor.

The three reported quantities: Whether UA improves the forecast is decided by the held-out predictive gain

$$G_t = D_0 - D_1, \tag{3}$$

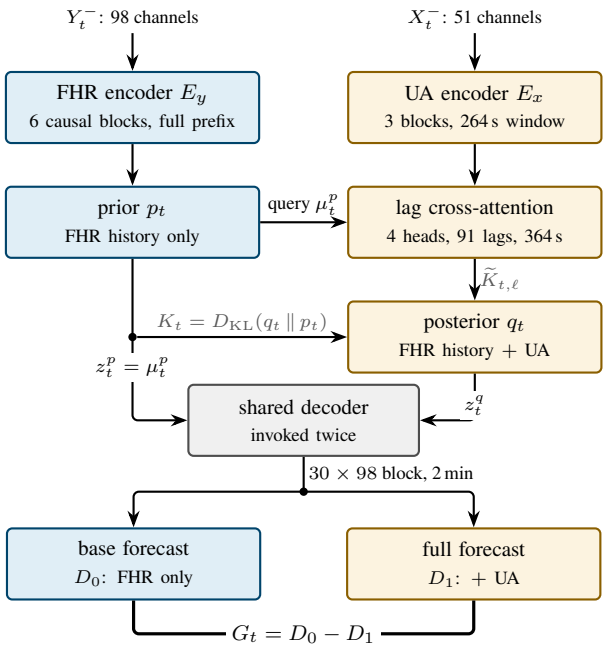

Figure 1. The model at one forecast anchor. Lag cross-attention is the single path by which UA reaches the UA-informed distribution, and one decoder is used for both. D_0 and D_1 are the two forecasts' negative log-likelihoods, G_t the predictive gain, K_t the divergence between the two and $\widetilde{K}_{t,\ell}$ its attribution across delays.

A positive G_t means that the UA-informed forecast assigns greater probability density to the observed future. Because both forecasts depend on a stochastic latent state, a score based on one draw would also reflect sampling noise. We therefore approximate each branch's marginal predictive density with $K = 8$ paired latent draws, reuse the same random draws for the two branches, and average their likelihoods before taking the logarithm. Multiple draws make the comparison less sensitive to any one latent sample, while pairing removes independent sampling noise from their difference. We fixed $K = 8$ as an a priori compromise between Monte Carlo variability and decoder cost, which increases linearly with K. Because latent group m is written by head m alone, the divergence splits additively as $K_t = \sum_m K_t^{(m)}$, a split the architecture itself enforces, and the same attention weights the model used distribute it over delays:

$$\widetilde{K}_{t,\ell} = \sum_m K_t^{(m)} \alpha_{t,\ell}^{(m)}, \qquad \sum_\ell \widetilde{K}_{t,\ell} = K_t. \tag{4}$$

No dropout is applied to the attention weights.

3. Results

Fig. 2 plots K_t on the held-out cohort over the final 12.5 hours before delivery. Two features are consistent. First, K_t rises as delivery approaches in every class: the healthy median doubles from 0.36 nats per anchor at twelve hours to 0.72 in the last full hour, and the HIE median moves from 0.43 to 0.80. Second, at every window down to the final hour the classes are ordered by severity, HIE above acidosis above healthy (0.54, 0.50 and 0.39 nats at six hours); in the final hour the three medians converge toward 0.8 nats. Cliff's delta, the probability that a recording from the more severe class exceeds one from the less severe minus the reverse, is positive for all three pairs in every window: 0.15-0.34 for HIE versus healthy, 0.11-0.25 for HIE versus acidosis and 0.04-0.18 for acidosis versus healthy. The contrasts weaken in the final hour, where the windows also hold the fewest recordings, so the late convergence may reflect a change in the recorded cohort as well as a change in coupling.

4. Discussion

The two distribution design provides three complementary measurements of UA's contribution. First, the predictive gain G_t tests whether including UA increases the probability assigned to the observed FHR future. Second, the divergence K_t measures how much the UA-informed latent distribution differs from the FHR-only distribution. Third, the lag attribution $\widetilde{K}_{t,\ell}$ assigns that latent change across the candidate UA delays. This separation is an advantage of the proposed design, a single joint FHR-UA forecaster produces only a combined forecast score and cannot isolate whether UA helped, how strongly it changed the forecast distribution, or when the relevant UA occurred. Fig. 2 supports two statements. UA moves the latent distribution further as labour advances, consistent with contractions growing stronger and more frequent. And the size of that movement separates the outcome classes twelve hours before delivery, even though the model never saw an adverse outcome during training. The ordering itself has two possible explanations: compromised fetuses may couple more strongly to uterine activity, or their recordings may simply lie further from the healthy-only training distribution. The trajectory alone cannot distinguish these; the predictive gain of Eq. 3, scored on the same anchors, can, and is the next quantity to report. Three limitations apply. First, K_t is a property of the fitted model, not of the underlying physiology: it is a surrogate for transfer entropy, not a measurement of it. Second, converting the attention's delay axis to physical lead time would require the difference between the group delays of the UA and FHR channels involved, which spans -389 to $+778$ s against a 360 s search; no such correc-

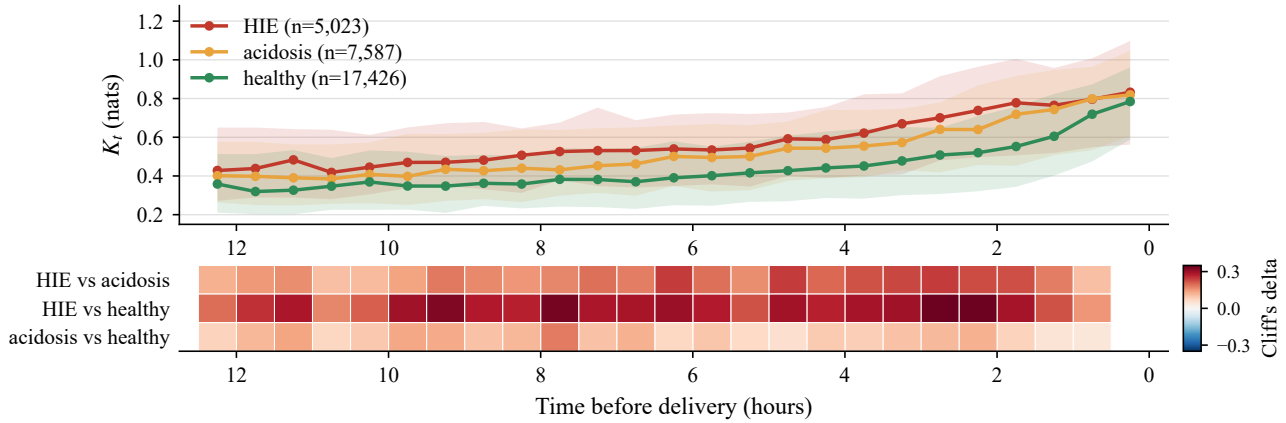

Figure 2. Median (line) and interquartile range (band) of the divergence K_t across held-out recordings, in half-hour windows of time before delivery; a recording contributes the mean over its anchors in the window. The legend sums the per-window recording counts, which range from 132 to 981; the final half-hour retains only 14-80 per class. Bottom: Cliff's delta per window, positive when the more severe class has higher values.

tion is applied, so an attention peak is not a contraction-to-deceleration latency. Third, the forecasts live in the feature domain, so per-channel accuracy measures prediction of a representation, not an error in beats per minute. If the gain confirms a UA-specific contribution, the next steps are a physical-time delay axis, validation on synthetic systems with known directed information, and testing whether these latents improve HIE-risk classification. External validation is required before we can claim specific clinical interpretations.

5. Conclusion

We introduced a causal forecasting model that predicts two minutes of FHR features from two distributions over one latent state, one built from FHR history alone and one that may also read UA. Sharing a single decoder makes the gap between their forecasts a direct test of whether UA is beneficial. On held-out recordings, the divergence between the two distributions rose toward delivery and separated the outcome classes twelve hours before delivery, despite the model being trained on healthy recordings alone.

References

[1] Ayres-de Campos D, Bernardes J, Costa-Pereira A, Pereira-Leite L. Inconsistencies in classification by experts of cardiotocograms and subsequent clinical decision. British Journal of Obstetrics and Gynaecology 1999; 106(12):1307–1310.

[2] Mohan M, Ramawat J, La Monica G, Jayaram P, Abdel Fattah S, Learmont J, Bryan C, Zaoui S, Pullattayil AK, Konje J, Lindow S. Electronic intrapartum fetal monitoring: A systematic review of international clinical practice guidelines. AJOG Global Reports 2021;1(2):100008.

[3] Zhou Z, Zhao Z, Zhang X, Zhang X, Jiao P, Ye X. Identifying fetal status with fetal heart rate: Deep learning approach based on long convolution. Computers in Biology and Medicine 2023;159:106970.

[4] McCoy JA, Levine LD, Wan G, Chivers C, Teel J, La Cava WG. Intrapartum electronic fetal heart rate monitoring to predict acidemia at birth with the use of deep learning. American Journal of Obstetrics and Gynecology 2025; 232(1):116.e1–116.e9.

[5] Schreiber T. Measuring information transfer. Physical Review Letters 2000;85(2):461–464.

[6] Kalajdzievski D, Mao X, Fortier-Poisson P, Lajoie G, Richards BA. Transfer entropy bottleneck: Learning sequence to sequence information transfer. Transactions on Machine Learning Research, 2023.

[7] Kearney RE, Wu YW, Vargas-Calixto J, Kuzniewicz MW, Cornet MC, Forquer H, Gerstley L, Hamilton E, Warrick PA. Construction of a comprehensive fetal monitoring database for the study of perinatal hypoxic ischemic encephalopathy. MethodsX 2024;12:102664.

[8] Mallat S. Group invariant scattering. Communications on Pure and Applied Mathematics 2012;65(10):1331–1398.

[9] Mallat S, Zhang S, Rochette G. Phase harmonic correlations and convolutional neural networks. Information and Inference A Journal of the IMA 2020;9(3):721–747.

[10] Peters B, Niculae V, Martins AFT. Sparse sequence-to-sequence models. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019; 1504–1519.

Address for correspondence:

Mahdi Shamsi
PeriGen Inc., Toronto ON, Canada
mahdi.sh.sci@gmail.com