RESEARCH PAPER DOI: 10.22080/frai.2026.32301.10751

Journal of Future Research in Artificial Intelligence and Internet of Things

Journal Home page: https://frai.journals.umz.ac.ir/

QT-MIL-FHR: A Context-Aware Multiple Instance Learning Framework with Quality- and Time-Conditioned Attention for Fetal Acidemia Prediction from Intrapartum CTG

Asghar Tareh [1], Ali Sheikhani [2*], Nader Jafarnia Dabanloo [3]

[1] *Department of Biomedical Engineering, SR.C., Islamic Azad University, Tehran, Iran.*

[2] *Department of Biomedical Engineering, SR.C., Islamic Azad University, Tehran, Iran, ORCID: 0000-0002-5231-5683.*

[3] *Department of Biomedical Engineering, SR.C., Islamic Azad University, Tehran, Iran, ORCID: 0000-0001-5823-8263.*

**Corresponding Author: a-sheikhani@srbiau.ac.ir*

Article Info

Received 3 July 2026

Accepted 24 Aug 2026

Available online 19 Sep 2026

Keywords:

Fetal Acidemia;

Cardiotocography;

Multiple Instance Learning;

Context-Aware Attention;

Signal Quality;

Time-to-Delivery.

Abstract:

Predicting fetal acidemia from intrapartum fetal heart rate (FHR) signals remains challenging due to subjective cardiotocography (CTG) interpretation and inter-observer variability. Obstetricians weight signal segments by contextual factors like signal quality and time-to-delivery, but most deep learning models ignore such context. We propose QT-MIL-FHR (Quality- and Time-Conditioned Multiple Instance Learning for Fetal Heart Rate), a context-aware MIL framework embedding clinical context into attention. The architecture has two branches: a CNN-BiGRU temporal branch for direct FHR analysis and a physiological branch using 11 heart rate variability (HRV) features, including nonlinear indices such as sample entropy and the SD1/SD2 ratio. Attention is conditioned on signal quality and time-to-delivery, with bidirectional cross-attention for branch interaction and a context-aware gating mechanism for adaptive integration. Evaluated on CTU-UHB (547 patients) via five-fold patient-level cross-validation, QT-MIL-FHR achieved a mean AUROC of 0.76 ± 0.04 with stable performance. Ablation studies showed that removing contextual conditioning reduced performance and increased variability; using only signal quality or time-to-delivery alone underperformed the full model. Attention analysis revealed emphasis on clinically relevant segments, and calibration showed acceptable agreement. Explicitly incorporating clinical context into attention improves predictive reliability and alignment with clinical interpretation. This work introduces clinical attention allocation as a framework for integrating clinical knowledge into deep learning for fetal monitoring.

© 2026 University of Mazandaran

1. Introduction

1.1. Clinical Problem Statement

Electronic fetal monitoring (EFM) has been fundamental to intrapartum care for over five decades, aiming to detect fetuses at risk of hypoxia, metabolic acidosis, and adverse neonatal outcomes, enabling timely intervention before irreversible injury [1-4]. Cardiotocography (CTG) remains the most accessible, cost-effective, and widely used method for continuous fetal assessment during labor, employed in over 85% of deliveries in the United States [4, 5]. By simultaneously recording fetal heart rate (FHR) and uterine contractions (UC), CTG provides indirect but clinically valuable information regarding fetal oxygenation, autonomic nervous system function, and placental–fetal interaction [6-8]. Despite its widespread clinical adoption, CTG interpretation continues to rely primarily on visual assessment by clinicians. This process is inherently subjective and is associated with considerable inter-observer and intra-observer variability [9]. Studies have shown that inter-observer agreement for CTG interpretation is only at a moderate level, with agreement ranges varying widely across different parameters. Such inconsistency has contributed to low specificity, high false-positive rates, and a significant increase in unnecessary obstetric interventions, particularly emergency cesarean deliveries [10, 11]. Misclassification of CTG traces may lead to both under-treatment, with the risk of fetal injury or death, and over-treatment, which includes unnecessary

Copyright © 2026 by the authors. Licensee FRAI, Babolsar, Mazandaran. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC-BY) license (https://creativecommons.org/licenses/by/4.0/deed.en)

ISSN 3060-6578

Please cite this paper as:A. Tareh, A. Sheikhani and N. Jafarnia Dabanloo, "QT-MIL-FHR: A Context-Aware Multiple Instance Learning Framework with Quality- and Time-Conditioned Attention for Fetal Acidemia Prediction from Intrapartum CTG", Future Research in AI and IOT, vol. 2, no. 2, pp. 30–49, 2026, doi:10.22080/frai.2026.32301.1075.

operative interventions that put both mother and baby at risk of complications. These interventions impose substantial economic, emotional, and iatrogenic burdens on mothers, neonates, and healthcare systems [12]. Fetal acidemia, objectively defined by umbilical artery pH measured immediately after birth, is one of the most clinically relevant markers of intrapartum hypoxia. A pH below 7.15 is strongly associated with increased risks of low Apgar scores, need for resuscitation, hypoxic-ischemic encephalopathy, and long-term neurodevelopmental sequelae [13-15]. Recent evidence suggests that umbilical cord arterial pH below 6.95 is significantly associated with cerebral palsy, epilepsy, and intellectual disability, whereas thresholds between 7.05 and 7.20 show weaker associations with these outcomes. In the present study, fetal acidemia is defined using the threshold pH < 7.15, a cutoff widely adopted in automated CTG analysis literature as striking a reasonable balance between sensitivity and specificity [15, 16].

1.2. Traditional and Classical Approaches to CTG Analysis

Early efforts in automated CTG analysis relied predominantly on handcrafted features and classical machine learning algorithms. From the late 1990s to the mid-2010s, researchers focused on extracting clinically meaningful features, such as baseline FHR, variability, accelerations, and decelerations, and feeding them into traditional classifiers [17-19]. Bernardes et al. were among the first to extract FIGO-based features and apply artificial neural networks, reporting better performance than visual interpretation [20]. Signorini et al. highlighted the diagnostic value of heart rate variability (HRV) indices such as short-term variability (STV), long-term variability (LTV), RMSSD, and NN50, demonstrating that linear and nonlinear parameters could capture distinct aspects of fetal autonomic function [21]. Subsequent studies by Georgoulas et al. using support vector machines achieved an AUROC of 0.84 for metabolic acidosis detection [22], while Warrick et al. reported an AUROC of 0.73 with dynamic modeling [23]. Fergus et al. conducted one of the most comprehensive classical studies, extracting over 40 features and achieving an AUROC of 0.91 with Random Forest, although performance declined significantly upon external validation [24]. Spilka et al. demonstrated the added value of nonlinear features such as sample entropy, reaching an AUROC of approximately 0.80 [25]. These classical approaches offered reasonable interpretability and low computational cost. However, they suffered from heavy reliance on manual feature engineering, limited capacity to model long-range temporal dependencies, high sensitivity to signal artifacts, and poor generalizability across different datasets and clinical settings [26, 27]. Studies have shown that many ML–CTG studies lack reproducibility and generalizability, with most using the same open-access database with a low number of subjects. These limitations paved the way for the transition toward deep learning, which promised to overcome these challenges through end-to-end learning.

1.3. Deep Learning and Modern Approaches

Since around 2017, deep learning has transformed the field by enabling end-to-end learning directly from raw FHR signals, eliminating the need for manual feature extraction. The rapid advancement of deep neural networks has demonstrated unparalleled proficiency in managing intricate tasks across multiple domains, with promising implications for fetal monitoring systems. Early models based on 1D-CNNs and recurrent networks showed promising results, achieving AUROCs in the range of 0.74–0.79 on retrospective cohorts [28, 29]. More recent architectures, such as Hybrid-FHR combining CNN and Transformer, Attention-based CNN-BiLSTM, and the DeepCTG series, have progressively improved performance, with reported AUROCs ranging from 0.74 to 0.79 [30]. Advanced fusion models like CTGFusionNet employing cross-attention have reached AUROC values around 0.81 [31]. However, despite these advances, state-of-the-art systems still face several challenges before they can be deployed in clinical practice, including the need for large, open, and anonymized multicentric databases, interpretable indicators that can be trusted by practitioners, and common standards for evaluation and validation. Despite these advances, most contemporary models still treat all temporal segments equally and largely ignore critical clinical contextual factors such as signal quality and temporal proximity to delivery. This represents a persistent gap between purely data-driven approaches and actual clinical decision-making, where clinicians implicitly prioritize segments with higher recording quality and those closer to the moment of delivery [32-34]. The visual interpretation of CTG signals by practitioners, even when following common guidelines, remains subject to high interobserver variability, and the efficiency of CTG monitoring is still debated.

1.4. Multiple Instance Learning and Attention Mechanisms

Multiple Instance Learning (MIL) is particularly suitable for CTG analysis because labels (umbilical pH) are available only at the patient (bag) level, while individual time windows (instances) remain unlabeled [35, 36]. In MIL, a label is assigned to a collection of instances (a bag), making it a natural choice for time series classification tasks where patterns of interest are sparse and localized within the signal. Attention-based MIL, introduced by Ilse et al. [34], has achieved remarkable success in digital pathology and other medical domains, enabling accurate classification with interpretable attention weights that highlight the most relevant instances within each bag [37, 38]. In FHR analysis, attention mechanisms Vaswani et al. [39] have been employed in various forms [39, 40]. However, most existing implementations are purely content-based, relying solely on features derived from the signal itself, and do not explicitly incorporate clinical contextual cues such as signal quality and time-to-delivery [32, 33]. This limitation is particularly relevant in the intrapartum setting, where the clinical significance of FHR pattern depends heavily on when it occurs during labor and how reliable the recording is.

1.5. Research Gaps and Proposed Innovation

The literature reveals three major methodological gaps that have hindered the translation of automated CTG analysis into clinical practice: (1) a structural mismatch between learning frameworks and the weakly-supervised nature of CTG data, where labels exist only at the patient level [28, 29]; (2) passive treatment of signal quality, which is typically used only for preprocessing rather than active weighting in the model's decision-making process [32, 41]; and (3) neglect of temporal context, despite its well-established clinical importance in guiding obstetric interventions [10, 33]. Studies have demonstrated

that ML approaches show potential in predicting fetal hypoxia from CTG, but more diverse datasets, standardization of hypoxia benchmarks, and enhancement of algorithms and features are needed for future clinical implementation. To address these limitations, the present study introduces the QT-MIL-FHR framework. The core innovation is the concept of Clinical Attention Allocation, which explicitly embeds clinicians' weighting logic, giving greater importance to higher-quality segments closer to delivery, into the model's attention mechanism. This approach builds upon the concept of attention-based multiple instance learning and extends it through contextual conditioning on quality and time. The key architectural contributions include: (i) reformulation as true patient-level MIL, (ii) Quality- and Time-Conditioned Attention (QT-Attention), (iii) dual-branch design with bidirectional cross-attention, and (iv) a Context-Aware Gate with entropy regularization. Main Hypothesis: Explicit and simultaneous integration of signal quality and time-to-delivery into the attention mechanism, in contrast to their isolated use, can lead to more stable, interpretable, and clinically aligned predictive performance. The remainder of this paper is organized as follows: Section 2 presents the methodology, Section 3 reports the experimental results, ablation studies, and supplementary analyses, and Section 4 discusses the implications, limitations, and future directions.

In summary, the key distinction between QT-MIL-FHR and prior attention-based MIL approaches lies in the source of the attention signal itself. Standard attention-based MIL (e.g., Ilse et al.[34]) and its content-only variants compute instance weights exclusively from the learned instance representation, implicitly treating all instances as equally trustworthy regardless of when or under what recording conditions they were captured. QT-MIL-FHR departs from this assumption by explicitly conditioning the attention weight of each instance on two independently measurable clinical variables, signal quality and time-to-delivery, in addition to signal content. This reformulation allows the network to learn not only what is informative in the signal, but also when and under what recording conditions that information should be trusted, a distinction that, to our knowledge, has not been explicitly modeled in prior CTG-based attention or MIL frameworks.

2. Methodology

2.1. Dataset and Outcome Definition

We used the CTU-UHB Intrapartum Cardiotocography Database (v1.0.0), widely regarded as the gold standard for automated FHR analysis due to open-access raw signals, objective outcome labeling (umbilical pH), and extensive literature use. It comprises 552 singleton records (≥37 weeks) with 4 Hz FHR and UC signals (up to 90 minutes before delivery), plus clinical and neonatal outcomes. Quality control: FHR outside 50–210 bpm was marked invalid; records with <50% valid samples were excluded, leaving 547 patients [42]. Primary outcome: fetal acidemia (umbilical pH < 7.15), the gold standard for acid-base status, strongly associated with hypoxia and adverse outcomes [42, 43]. This threshold follows common automated CTG practice [42, 43]. Of 547 patients, 105 (19.2%) were acidemic, and 442 (80.8%) were non-acidemic, a typical imbalance in perinatal studies, which challenges model training [44]. Table 3-1 presents the details of this metadata. Sensitivity analyses used stricter (pH < 7.10, 53 cases) and more lenient (pH < 7.20, 174 cases) thresholds to assess robustness to outcome definition and class imbalance [43, 44].

2.2. Signal Preprocessing

Signal quality is decisive for intelligent FHR analysis [45]. Real-world CTG often has artifacts, dropouts, and discontinuities degrading performance if not addressed [45]. Our preprocessing pipeline removes non-physiological data, manages gaps, and preserves quality information for the attention mechanism. FHR is sampled at 4 Hz; values outside 50–210 bpm are replaced with NaN (Not a Number) [42, 45]. Short gaps (<30 s) are imputed via linear interpolation; long gaps (≥30 s) are left unreconstructed to preserve quality metrics [45]. Notably, no smoothing filter (e.g., moving average) was applied, unlike many previous studies [42, 43]. Because key short-term variability indices (STV, RMSSD) depend on the natural, instantaneous fluctuations of the FHR signal, any smoothing operation would distort these physiologically meaningful variations and consequently impair the detection of subtle acid–base changes [42, 43].

2.3. Signal Segmentation and Instance Construction

Variable recording length and dynamic FHR changes pose challenges; acidemia patterns may appear only in limited segments [46, 47]. After preprocessing, each signal was segmented into overlapping 10-minute windows (2400 samples at 4 Hz), chosen for stable HRV estimation (STV, LTV, sample entropy), requiring 5–10 minutes. A 2-minute stride (~80% overlap) enhanced temporal coverage while balancing computational cost and feature stability [43, 47]. Figure 1 presents a schematic diagram of the sliding window strategy with a 10-minute window length and a 2-minute stride applied to a representative signal.

In the MIL formulation, each 10-minute window is treated as an instance, and the complete set of windows extracted from a single patient constitutes a bag. Labels are assigned exclusively at the bag level, while the QT-Conditioned Attention mechanism adaptively learns to weight individual instances according to their clinical relevance [42, 48]. Across the dataset, the bags exhibited a relatively homogeneous distribution, with a mean of 32.8 ± 3.1 instances per bag (range: 25–40). This uniformity is methodologically advantageous, as it mitigates the risk that bags of extreme size might disproportionately influence the learning process, a recognized challenge in MIL applications [48]. Figure 2 presents the histogram of the distribution of the number of windows per patient.

This bag-instance structure underpins QT-MIL-FHR: from each instance, raw temporal and physiological representations are extracted, interact via bidirectional cross-attention, and are aggregated by Multi-Head QT-Conditioned Attention (conditioned on quality and time). This enables salient window identification and patient-level prediction without manual segment labeling, balancing temporal preservation, computational cost, and MIL compatibility [42, 46, 48].

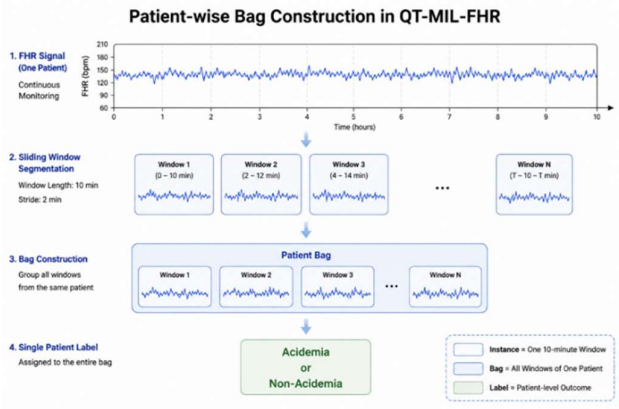

Figure 1. **Schematic of sliding window strategy (10-min window, 2-min stride)**

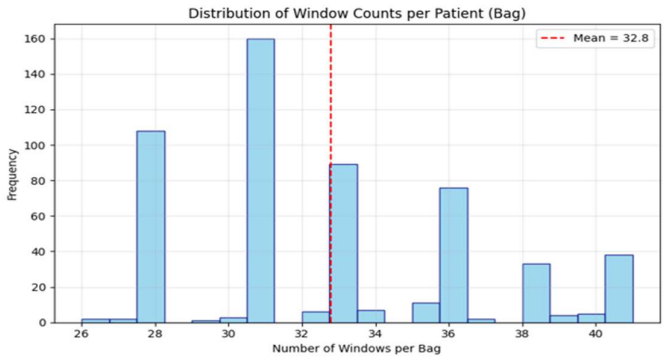

Figure 2. **Distribution of window counts per patient**

2.4. Contextual Information Extraction

Most existing automated FHR analysis weighs windows solely on signal content, ignoring quality and temporal location. In practice, clinicians implicitly consider both: signal reliability and pattern timing relative to delivery. QT-MIL-FHR, for the first time, extracts these contextual factors explicitly, quantitatively, and dynamically, integrating them into attention and fusion [45, 49].

Signal Quality Representation: Signal quality plays a critical role in both clinical interpretation and model performance. It has been reported that up to 30–40% of CTG recording time in real-world settings may be of poor quality, and that neglecting signal quality as an active variable represents a significant source of error in deep learning-based fetal monitoring [45].

For each window i, we define a 3D quality vector:

$$q_i = [VR_i, JR_i, LG_i] \tag{1}$$

Valid Ratio (VR):

$$VR_i = \frac{N_{valid}}{N_{total}} \tag{2}$$

Jump Ratio (JR):

$$JR_i = \frac{N_{jumps}}{N_{valid_diff}} \tag{3}$$

Longest Gap (LG):

$$LG_i = \frac{L_{max}}{f_s} \tag{4}$$

Relative Time-to-Delivery Encoding: The clinical significance of FHR patterns is strongly time-dependent. Late decelerations occurring in the final 30 minutes of labor have been associated with a 3.8-fold higher risk of acidemia [46].

For window i, relative time:

$$t_i = \frac{T_{end} - T_{center,i}}{60} \tag{5}$$

Typically 5–85 minutes. Standardized via Z-score for commensurability with other inputs.

2.5. Proposed Architecture

2.5.1. QT-MIL-FHR

The QT-MIL-FHR architecture integrates raw temporal and physiological branches via attention mechanisms and context-aware fusion, emulating clinical reasoning [38]. The following subsections detail the pipeline. Figure 3 presents the overall block diagram of this study. The remaining components of this pipeline are detailed in the following sections.

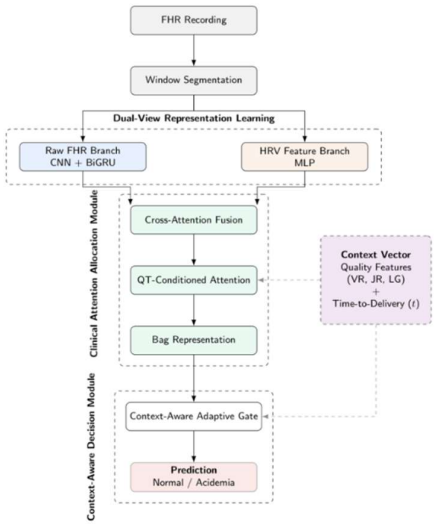

Figure 3. Overall block diagram of the QT-MIL-FHR framework

2.5.2. Raw Temporal Branch (RawEncoder)

The raw branch directly processes the FHR signal within each window. For a window containing L=2400samples (10 min at 4 Hz), the raw FHR vector x∈R^Lis first passed through a 1D convolutional backbone that extracts local morphological patterns:

$$x^{'}="Conv1D"(x)\in R^{(32\times 64)} \tag{6}$$

The convolutional block consists of two Conv1d layers with batch normalization and ReLU activations, interlaced with MaxPooling and Dropout (rate = 0.2). The first layer uses 16 filters (kernel size 5, stride 1, padding 2), while the second uses 32 filters. Adaptive average pooling compresses the temporal dimension to a fixed length of 64. The output x^' is then fed into a bidirectional GRU (hidden size = 48, thus 96 bidirectional) to model longer-range temporal dependencies:

$$h_"raw" ^{((i))}="BiGRU"(x^{'})\in R^{96} \tag{7}$$

The final instance-level representation is obtained by averaging the GRU outputs across the 64 time steps, yielding a fixed-length vector h_"raw" ^((i))∈R^96for the i-th window.

Rationale for the recurrent unit choice: A bidirectional GRU, rather than a bidirectional LSTM or a state-space sequence model (e.g., S4/Mamba), was selected for the temporal branch for three practical reasons. First, GRU cells have roughly one-third fewer trainable parameters than LSTM cells at an equivalent hidden size, which reduces the risk of overfitting given the modest size of the CTU-UHB cohort (547 patients, 105 positive cases). Second, several prior CTG studies employing GRU-based encoders have reported discrimination performance comparable to LSTM-based counterparts on similar FHR classification tasks while requiring shorter training time and fewer epochs to converge. Third, state-space sequence models, while promising for long-range dependency modeling, had not yet been extensively validated on short, noisy, irregularly sampled biomedical time series at the scale available here when this architecture was designed; we regard their application to FHR-based multiple instance learning as a promising direction for future work rather than an established choice for the present dataset size.

2.5.3. Physiological Branch (PhysEncoder)

The physiological branch encodes 11 hand-crafted HRV features computed from the same 10-minute window:

Time-domain: Median, STV (standard deviation of differences), LTV (standard deviation), RMSSD, NN50, pNN50, MAD, IQR, CV.

Nonlinear: Sample Entropy (SampEn, $m = 2$) and the Poincaré ratio SD1/SD2.

These 11 features are concatenated into a vector $p^{(i)} \in \mathbb{R}^{11}$and transformed by a two-layer MLP with ReLU activations and Dropout (0.2):

$$h_{phys}^{(i)} = MLP(p^{(i)}) \in \mathbb{R}^{96} \tag{8}$$

The MLP layers have dimensions $11 \rightarrow 48 \rightarrow 96$.

2.5.4. QT-Conditioned Attention

The core innovation, QT-Conditioned Attention, computes weight α_i for each instance by conditioning on recording quality and time-to-delivery, rather than solely on signal content:

$$\alpha_i = \mathrm{softmax}\left(w^{\top} \tanh\left(Vh_i + U[q_i; t_i]\right)\right) \tag{9}$$

Here h_i is the instance representation, q_i the 3D quality vector, t_i the relative time-to-delivery, and U, V, W learned parameters. Building on attention-based MIL [35], this extends to contextual conditioning, mirroring clinicians' implicit weighting of high-quality, delivery-proximal segments [42]. Figure 4 illustrates the aforementioned mathematical relationships in the form of a block diagram.

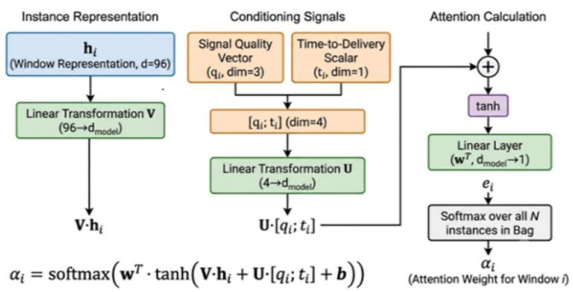

Figure 4. QT-Conditioned Attention mechanism diagram

2.5.5. Interaction via Cross-Attention

To facilitate information exchange between the raw temporal representation and the physiological feature representation, a bidirectional cross-attention mechanism was employed. Unlike self-attention, which models dependencies within the same representation space, cross-attention enables each branch to attend to complementary information contained in the other branch.

Specifically, the raw branch attends to the physiological branch according to:

$$Q_{raw} = H_{raw} W_Q^{raw} \tag{10}$$

$$K_{phys} = H_{phys} W_K^{phys} \tag{11}$$

$$V_{phys} = H_{phys} W_V^{phys} \tag{12}$$

$$\mathrm{Attention}_{raw \rightarrow phys} = \mathrm{softmax}\left(\frac{Q_{raw} K_{phys}^{T}}{\sqrt{d_k}}\right) V_{phys} \tag{13}$$

An analogous operation is performed in the opposite direction, allowing the physiological branch to attend to temporal information extracted from the raw FHR signal.

Residual connections are then applied to preserve branch-specific information:

$$H_{raw}^{*} = H_{raw} + \mathrm{Attention}_{raw \rightarrow phys} \tag{14}$$

$$H_{phys}^{*} = H_{phys} + \mathrm{Attention}_{phys \rightarrow raw} \tag{15}$$

Finally, the two enhanced representations are combined using adaptive fusion weights generated by the Context-Aware Gate:

$$H_{fused} = \mathrm{LayerNorm}\left(\beta_{raw} H_{raw}^{*} + \beta_{phys} H_{phys}^{*}\right) \tag{16}$$

where β_{raw} and β_{phys} are dynamically estimated for each patient and determine the relative contribution of the temporal and physiological representations.

2.5.6. Adaptive Context-Aware Fusion

To enable patient-specific fusion of the temporal and physiological representations, a Context-Aware Gate was employed. Unlike static fusion strategies, the proposed gate dynamically adjusts the relative contribution of each branch according to both feature representations and contextual information.

First, global representations of the raw branch, physiological branch, quality indicators, and temporal context are concatenated:

$$x_g = [\bar{h}_{raw}; \bar{h}_{phys}; \bar{q}; \bar{t}] \tag{17}$$

where $\bar{h}_{raw}$, $\bar{h}_{phys}$ denote pooled branch representations, $\bar{q}$ represents aggregated signal-quality information, and $\bar{t}$ denotes the normalized time-to-delivery context.

A gating vector is then computed as:

$$g = W_g x_g + b_g, g \in \mathbb{R}^2 \tag{18}$$

and converted into adaptive fusion weights through a softmax operation:

$$[\beta_{raw}, \beta_{phys}] = \text{softmax}(g) \tag{19}$$

subject to

$$\beta_{raw} + \beta_{phys} = 1 \tag{20}$$

Finally, the fused representation is obtained as:

$$z = \beta_{raw}\bar{h}_{raw} + \beta_{phys}\bar{h}_{phys} \tag{21}$$

where the fusion weights are estimated independently for each patient. This mechanism allows the model to dynamically determine whether the temporal representation or the physiological representation should receive greater emphasis, thereby incorporating both signal characteristics and contextual information into the fusion process [32].

2.5.7. Entropy Regularization

To prevent premature gate collapse, where the fusion mechanism excessively relies on a single representation branch, an entropy-based regularization term was incorporated into the training objective.

The entropy of the fusion weights is defined as:

$$H(\beta) = -\sum_{k=1}^{2} \beta_k \log(\beta_k) \tag{22}$$

where $\beta_k$denotes the adaptive fusion weight assigned to branch k.The final optimization objective is:

$$L_{total} = L_{BCE} - \lambda H(\beta) \tag{23}$$

where L_{BCE} denotes the binary cross-entropy loss and λ controls the strength of the entropy regularization term.

In this study, λ was selected within the range:

$$0.01 \leq \lambda \leq 0.05$$

To encourage a more balanced utilization of the temporal and physiological representations while maintaining stable optimization. By maximizing the entropy of the fusion weights, the model is discouraged from collapsing toward a single branch during training and is instead encouraged to exploit complementary information from both representations [37].

Figure 5 illustrates the interactions among the various modules of the proposed architecture and the mathematical reformulation governing the present proposed framework.

2.6. Training and Validation

Training employed patient-wise 5-fold cross-validation on 547 patients, partitioning into five folds with patient-level exclusivity to prevent leakage and ensure generalization. Acidemia prevalence (~19 %) was preserved per fold [39].

AdamW optimizer (weight decay λ_wd = 1e-4) mitigated overfitting. Initial LR η = 1e-3, reduced by 0.5 on plateau (3 epochs). Gradient clipping (max norm = 1.0) stabilized BiGRU training.

Loss: BCEWithLogitsLoss for numerical stability. Class imbalance (≈1:4.2) addressed via:

$$\text{"pos_weight"} = \text{N_"negative"} / \text{N_"positive"} \tag{24}$$

per fold, weighting minority errors more heavily.

Each fold ran ≤30 epochs with early stopping (patience=5, min delta=1e-3) based on training loss, restoring best parameters. Robustness was ensured via three random seeds: 42, 123, 2025. Optimal classification threshold determined per fold using Youden's index [42]:

$$\tau^* = \arg(\max)_\tau (\text{"Sens"}(\tau) + \text{"Spec"}(\tau) - 1) \tag{25}$$

applied to test data to avoid leakage and reflect clinical decision-making.

Evaluation Metrics. Evaluation Metrics: Model performance was assessed using a comprehensive set of metrics. Threshold-independent metrics included the area under the receiver operating characteristic curve (AUROC, primary) and the area under the precision–recall curve (AUPRC). Threshold-dependent metrics, sensitivity, specificity, precision, F1-score, Matthews Correlation Coefficient (MCC), and balanced accuracy were reported at both the default 0.5 threshold and the Youden-derived optimal threshold. Calibration was evaluated using the Expected Calibration Error (ECE) and the Brier Score. This protocol ensures a robust performance assessment and underpins the subsequent ablation and interpretability analyses [38].

Implementation details: The full QT-MIL-FHR model comprises approximately 138695 trainable parameters in total (raw temporal branch:73346; physiological branch: 52514; cross-attention modules: 55872; QT-conditioned attention and context-aware gate: 38596). All experiments were conducted using PyTorch 2.11.0 (CPU build) with Python 3.12.13, scikit-learn 1.6.1, NumPy 2.0.2, and Pandas 2.2.2, on Google Colab's free CPU runtime (no GPU acceleration was used). Mean training

time per fold was approximately 3.5 hours (≈18 epochs on average before early stopping out of a maximum of 30), and mean inference time per patient bag (≈33 windows) was approximately 50 milliseconds on the same CPU runtime. Patient-wise 5-fold cross-validation was implemented via scikit-learn's KFold (shuffle = True, random_state = 123).

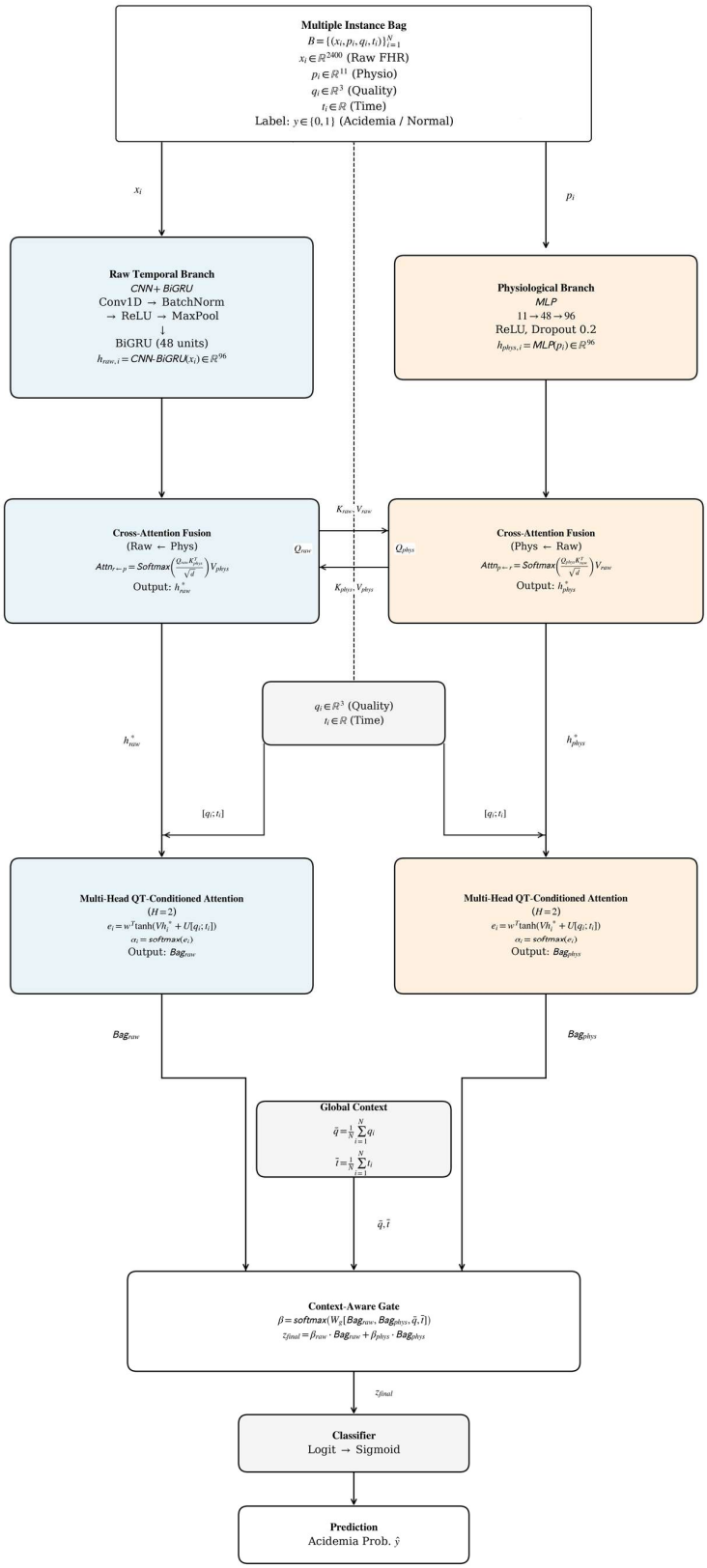

Figure 5. Detailed mathematical architecture diagram of QT-MIL-FHR

2.7. Ablation Studies and Supplementary Analyses

To assess each component's contribution and evaluate robustness, interpretability, and clinical coherence, supplementary experiments, ablation studies, sensitivity analysis, attention-weight analysis, gate-coefficient examination, and case studies were conducted.

Seven configurations were defined and compared to the full model to isolate contextual conditioning, dual-branch interaction, and adaptive fusion:

Removed quality and time (content-based attention).

Quality conditioning only.

Time conditioning only.

Removed physiological branch (raw-only).

Removed raw branch (physiological-only).

Disabled cross-attention (concatenated).

Replaced gate with simple averaging.

Each was run under the identical protocol, enabling extraction of each component's unique contribution.

Model performance was tested against stricter (pH < 7.10) and broader (pH < 7.20) thresholds, repeating 5-fold CV independently. This demonstrates robustness to outcome definition and class-imbalance variations.

To understand decision-making and clinical alignment, interpretability analyses targeted temporal attention allocation, gate behavior, and focus patterns. Spearman correlation between average attention weights (raw and physiological) and temporal position assessed monotonic relationship; distributions plotted across the test population. Gate coefficients β_raw and β_phys distributions visualized via violin plots; stratified analysis compared academic vs. non-academic subgroups to detect fusion-strategy shifts under critical conditions. Top 10 % attention weights per patient located relative to delivery; cumulative histogram assessed concentration in final labor stages versus dispersed patterns.

3. Results

This section reports the experimental evaluation of QT-MIL-FHR. In addition to predictive performance obtained through five-fold patient-wise cross-validation, the contribution of individual architectural components is investigated through ablation studies. Supplementary analyses, including sensitivity analyses, calibration assessment, adaptive gate evaluation, attention behavior analysis, and representative case studies, are further presented to examine the robustness, interpretability, and clinical consistency of the proposed framework.

3.1. Descriptive Characteristics of the Dataset

After exclusions (records with >50% invalid data), the final dataset comprised 547 singleton deliveries from CTU-UHB. Table 1 summarizes demographic and clinical characteristics.

Table 1. Characteristics of the CTU-UHB Database

Variable	Value (Mean ± SD or Percentage)
Maternal age (years)	29.8 ± 5.3
Gestational age (weeks)	39.9 ± 1.3
Birth weight (g)	3408 ± 467
Umbilical artery pH	7.28 ± 0.09
Base excess (mmol/L)	-4.7 ± 3.1
Acidemia (pH < 7.15)	19.2%

Acidemia prevalence (19.2%) aligns with prior CTU-UHB studies, reflecting the inherent class imbalance characteristic of perinatal outcome research. This imbalance, while clinically realistic, poses challenges for model training and evaluation, necessitating appropriate strategies such as weighted loss functions and careful interpretation of AUPRC alongside AUROC. Mean FHR recording length: 74.1 ± 7.5 min. Sliding window (10-min window, 2-min stride) yielded 32.8 ± 3.1 windows/patient. Figure 6 illustrates pH distribution skewness toward normal values and low acidemia prevalence, consistent with the epidemiology of fetal compromise.

3.2. Overall Performance of QT-MIL-FHR

Table 2 summarizes the overall performance of QT-MIL-FHR across five patient-wise cross-validation folds. The model achieved a mean AUROC of 0.766 ± 0.047 and a mean AUPRC of 0.421 ± 0.051. At the default 0.5 threshold, the mean sensitivity and specificity were 0.433 ± 0.289 and 0.841 ± 0.132, respectively; at the fold-specific Youden-optimal threshold ($\tau^* = 0.212 \pm 0.200$), sensitivity improved to 0.796 ± 0.102 (see Table 3).

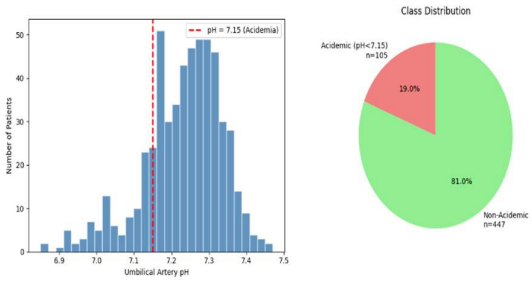

Figure 6. **pH distribution histogram and acidemia class pie chart**

The corresponding Precision, F1-score, MCC, and Balanced Accuracy were 0.441 ± 0.117, 0.352 ± 0.131, 0.271 ± 0.081, and 0.637 ± 0.081, respectively. Performance variability across folds remained relatively limited, with standard deviations ranging from 0.047 for AUROC to 0.100 for specificity.

Overall, the results indicate consistent predictive performance across the five validation folds, with AUROC values ranging from 0.7199 to 0.8100 and Balanced Accuracy values ranging from 0.662 to 0.747.

The per-fold and aggregated performance metrics, detailed in Table 2, show that the model achieved a mean AUROC of 0.766 and a mean AUPRC of 0.421, with moderate variability across folds (AUROC range: 0.720–0.810). To provide a transparent picture of this variability and the associated trade-offs, the receiver operating characteristic curves for all five folds are displayed in Figure 7 together with the mean curve and ±1 standard deviation band, while Figure 8 presents the corresponding precision–recall curves. The ROC curves illustrate how sensitivity and specificity co-vary across folds and allow the consistency of the decision threshold behavior to be judged, whereas the precision–recall curves highlight the model's ability to identify the minority academic class in the presence of substantial class imbalance. Together, these visualizations confirm that the model's discriminative performance remains stable across data splits, with no fold exhibiting a marked performance collapse.

Table 2. **Summary Performance Statistics of QT-MIL-FHR**

Metric	Value (Mean ± SD)
AUROC	0.766 ± 0.047
AUPRC	0.421 ± 0.051
Sensitivity	0.433 ± 0.289
Specificity	0.841 ± 0.132
Precision	0.441 ± 0.117
F1-Score	0.352 ± 0.131
MCC	0.271 ± 0.081
Balanced Accuracy	0.637 ± 0.081

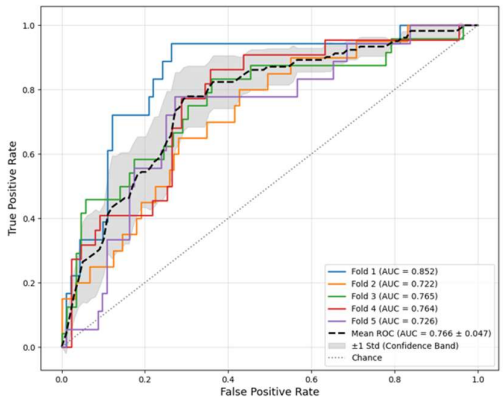

Figure 7. **ROC curves across all five folds with mean curve and ±1 SD band**

3.3. Clinical Evaluation and Threshold Selection

Model performance was evaluated at both a fixed decision threshold of 0.5 and the Youden-derived optimal threshold, with the full set of results reported in Table 3. At the default 0.5 threshold, the model exhibited high specificity (0.841 ± 0.132) but limited sensitivity (0.433 ± 0.289). Shifting to the Youden optimal threshold ($\tau^* = 0.212 \pm 0.200$) markedly improved sensitivity to 0.796 ± 0.102, while balanced accuracy reached 0.721 ± 0.059. This substantial gain in sensitivity, achieved at the expense of specificity, underscores the critical role of threshold calibration in striking an appropriate balance between the

two, an issue of particular relevance in fetal monitoring, where the clinical imperative to minimize false negatives often justifies a degree of over-calling. The relatively large standard deviation of τ indicates that the optimal operating point varies across data folds, reinforcing the importance of fold-specific threshold selection within cross-validation to avoid optimistically biased performance estimates.

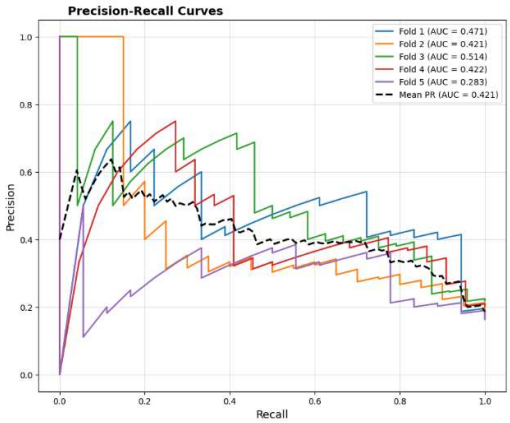

Figure 8. **Precision-Recall curves across all five folds**

Table 3. **Performance at fixed (0.5) and Youden-optimal thresholds**

Metric	Threshold 0.5	Youden (τ* = 0.212 ± 0.200)
Sensitivity	0.433 ± 0.289	0.796 ± 0.102
Specificity	0.841 ± 0.132	0.647 ± 0.056
Precision	0.441 ± 0.117	0.342 ± 0.052
F1	0.352 ± 0.131	0.476 ± 0.063
MCC	0.271 ± 0.081	0.348 ± 0.089
Balanced Acc.	0.637 ± 0.081	0.721 ± 0.059

3.4. Ablation Studies

A comprehensive ablation study was conducted to isolate the contribution of each architectural component. The results are summarized in Table 4, which reports the primary performance and stability metrics for the full QT-MIL-FHR model and the seven ablated variants.

The full model achieved the highest mean AUROC of 0.766 ± 0.047 and AUPRC of 0.421 ± 0.051 among all configurations, while simultaneously exhibiting the greatest stability, with the lowest standard deviation of 0.042, the narrowest AUROC range of 0.718–0.811, and a coefficient of variation of only 5.5%. This confirms that the model's superior discriminative performance was not attained at the expense of robustness.

Impact of contextual conditioning: Removing both quality and time information from the attention mechanism led to a pronounced decline in performance from 0.766 to 0.683 and a marked increase in variability, from 5.5% to 15.4%, demonstrating that context awareness benefits both accuracy and stability. When the two contextual signals were disentangled, time-only conditioning achieved a substantially higher AUROC of 0.710 than quality-only conditioning 0.641, and the latter also recorded the lowest balanced accuracy among all variants 0.574. This differential impact indicates that proximity to delivery carries a stronger predictive signal than signal quality alone, aligning with clinical intuition.

Impact of cross-attention: Removing the cross-attention module reduced the AUROC to 0.705 and caused the MCC to drop from 0.451 to 0.312. The substantial decline in MCC underscores that bidirectional interaction between the two branches contributes beyond simple additive effects and enhances the agreement between predictions and true labels, an important property for clinical trustworthiness.

Impact of the adaptive gate: Replacing the context-aware gate with simple averaging (AUROC 0.712 or concatenation (AUROC 0.689) degraded performance relative to the full model. The averaging strategy preserved greater discriminative power, yet the gap to the full model confirms that the gate functions beyond a static combiner, adaptively weighting branches based on patient-specific context.

Independent branch contribution: The single-branch variants Raw Only + QT with AUROC 0.678 and Phys Only + QT with AUROC 0.665 both underperformed the full model and exhibited higher variability. The raw branch's slightly superior performance suggests that end-to-end learning from the raw signal captures information not fully encoded in handcrafted HRV features, consistent with the motivation for deep learning approaches in this domain.

To facilitate direct comparison and visual interpretation of the ablation outcomes, Figure 9 presents a consolidated bar chart summarizing the key performance metrics across all evaluated configurations.

Table 4. Ablation study results

Model Variant	AUROC	AUPRC	MCC	Std	CV (%)
QT-MIL-FHR (full)	0.766 ± 0.047	0.421 ± 0.051	0.271 ± 0.081	0.042	5.5
No Context Conditioning	0.683 ± 0.105	0.342 ± 0.083	0.184 ± 0.143	0.105	15.4
Time Only Conditioning	0.710 ± 0.082	0.396 ± 0.082	0.254 ± 0.135	0.082	11.5
Quality Only Conditioning	0.641 ± 0.097	0.325 ± 0.076	0.132 ± 0.128	0.097	15.1
No Cross Attention	0.705 ± 0.053	0.405 ± 0.066	0.312 ± 0.101	0.053	7.5
No Gate – Concatenation	0.689 ± 0.081	0.371 ± 0.092	0.241 ± 0.136	0.081	11.8
Raw Only + QT	0.678 ± 0.096	0.346 ± 0.088	0.212 ± 0.151	0.096	14.2
Phys Only + QT	0.665 ± 0.101	0.332 ± 0.095	0.163 ± 0.169	0.101	15.2

Stability Analysis: The stability of model performance across folds was evaluated as an important indicator of robustness. The complete QT-MIL-FHR model showed the lowest variability among all configurations, with a mean AUROC of 0.76۶ ± 0.047 (coefficient of variation =5.5%, range: 0.718–0.811). In comparison, the model without contextual conditioning exhibited notably higher variability (SD = 0.105, CV = 15.4%). The single-context variants also demonstrated greater fluctuation, with CV values of approximately 11.5% for Time-Only and 15.1% for Quality-Only conditioning. While certain ablated models achieved competitive AUROC in some individual folds, they consistently displayed larger performance dispersion across the five folds. These results indicate that the full model provided more consistent predictions across different patient partitions.

Statistical assessment of ablation comparisons. To complement the descriptive comparisons above, we computed 95% confidence intervals for the AUROC of each configuration using a t-distribution (df = 4) applied to the reported per-fold mean and SD, and we performed approximate two-sided significance tests (Welch's t-test on the summary statistics) comparing each ablated variant against the full model. The full model's AUROC 95% CI was [0.708, 0.824]. Relative to the full model, the Quality-Only-conditioning variant showed the largest and only conventionally significant difference (AUROC = 0.641, 95% CI [0.521, 0.761]; p = 0.042), while the remaining comparisons — No-Context-Conditioning (AUROC = 0.683, 95% CI [0.553, 0.813]; p = 0.162), Time-Only-Conditioning (0.710, [0.608, 0.812]; p = 0.231), No-Cross-Attention (0.705, [0.639, 0.771]; p = 0.091), No-Gate/Concatenation (0.689, [0.588, 0.790]; p = 0.112), Raw-Only+QT (0.678, [0.559, 0.797]; p = 0.117), and Phys-Only+QT (0.665, [0.540, 0.790]; p = 0.092) — showed directionally consistent but not individually significant reductions at this sample size. Because this test treats the five folds as unpaired samples rather than exploiting their shared partition across configurations, it is a conservative approximation of the paired comparison that raw per-fold predictions would permit; we therefore interpret these results as supportive rather than confirmatory evidence for the contribution of each architectural component, and note that a formal paired analysis is a natural direction for future work.

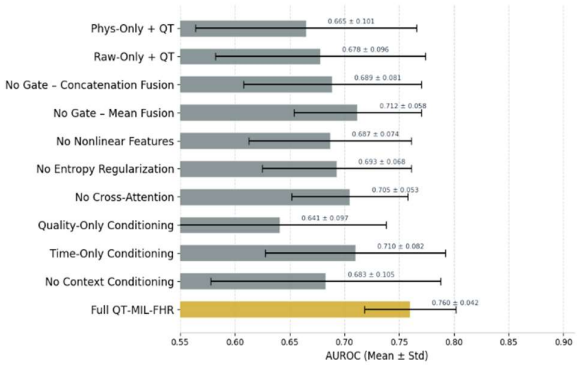

Figure 9. Comparative bar chart of ablation results (AUROC ± Std)

3.5. Supplementary Analyses

3.5.1. Sensitivity Analysis to pH Threshold

As a supplementary analysis designed to assess the model's robustness to variations in the clinical definition of acidemia, performance was examined across three commonly used pH thresholds, with the results summarized in Table 5. At pH < 7.15, the model achieved the highest AUROC (0.76 ± 0.04) together with an AUPRC of 0.48. Applying the stricter threshold of pH < 7.10 (n = 53) led to a marked drop in both AUROC (0.63 ± 0.12) and AUPRC (0.25), a decline attributable to the limited number of positive cases and the resulting estimation instability. When the broader threshold pH < 7.20 (n = 174) was adopted, the AUROC decreased to 0.71 ± 0.07 while the AUPRC remained at 0.48, reflecting the higher prevalence of the positive class. The lower AUROC at this threshold likely stems from increased class overlap, as borderline cases (pH 7.15–7.20) present milder physiological alterations that are inherently more difficult to distinguish from normal tracings. Taken together, these findings indicate that pH < 7.15 offers the most favorable trade-off between pathological severity and sample adequacy for model training and evaluation.

Table 5. Performance across pH thresholds

pH Threshold	Acidemic Cases	AUROC	AUPRC
pH < 7.10	53	0.63 ± 0.12	0.25
pH < 7.15	105	0.76 ± 0.04	0.42
pH < 7.20	174	0.71 ± 0.07	0.48

3.5.2. Sensitivity Analysis to Window Stride

To further probe the framework's resilience to variations in data preparation, the sensitivity of model performance to temporal sampling density was examined by repeating experiments with strides of 1, 2, and 5 minutes. The results are summarized in Table 6. Varying the stride had only a limited impact on discriminative performance. A stride of 1 minute (mean 65 windows per patient) yielded an AUROC of 0.75 ± 0.05, while a stride of 2 minutes (33 windows) achieved the best performance (0.76 ± 0.04). Notably, even with a 5-minute stride, reducing the average number of windows to 13- an approximately 80% reduction- the AUROC remained at 0.73 ± 0.06, a drop of merely 0.03. The similar standard deviations across settings further indicate that stability was not meaningfully affected by the choice of stride. These findings suggest that the MIL framework is robust to substantial changes in temporal resolution, and that a stride of up to 5 minutes may offer a practical trade-off between predictive performance and computational cost.

Table 6. Performance across window strides

Stride (min)	Mean Windows	AUROC
1	65	0.75 ± 0.05
2	33	0.76 ± 0.04
5	13	0.73 ± 0.06

3.5.3. Model Calibration Analysis

According to the results obtained in Table 7, probability calibration improved substantially after Platt scaling. The Expected Calibration Error (ECE) decreased from 0.1494 to 0.0357 (76.2% reduction), while the Brier Score improved from 0.1721 to 0.1406 (18.3% reduction). These findings indicate that the calibrated probabilities were more consistent with the observed event frequencies. Although calibration improved markedly, the reliability diagrams still revealed residual deviations from perfect calibration, suggesting that additional calibration or uncertainty-aware methods may further enhance probabilistic reliability. This analysis represents an important strength of the present study, as many previous CTG-based deep learning models have focused primarily on discrimination metrics such as AUROC without explicitly evaluating calibration. Well-calibrated probabilities are particularly important in clinical decision-making, where predicted risk estimates may influence intervention thresholds and support more informed obstetric management.

Table 7. Calibration metrics before and after Platt scaling

Metric	Before Calibration	After Calibration	Improvement
ECE	0.1494	0.0357	76.2%
Brier Score	0.1721	0.1406	18.3%

3.5.4. Attention Weights vs. Time-to-Delivery

Figure 10 illustrates the relationship between attention weights and normalized time-to-delivery across all windows in the test set. A strong negative association was observed between the two variables (Spearman r = -0.976, Pearson r = -0.833; both $p < 0.001$). Attention weights generally increased as time-to-delivery decreased, with earlier windows receiving lower weights than later windows. Both the LOESS smoothing curve and the moving-average curve exhibited a consistent increasing trend toward delivery, indicating that this pattern was maintained across the entire dataset despite local variability among individual windows.

3.5.5. Attention Weights vs. Signal Quality Indices

Figure 11 presents the relationships between attention weights and the three signal quality indicators. Statistically significant correlations were observed for all quality metrics (all $p < 0.001$), including Valid Ratio (r = -0.110), Jump Ratio (r = 0.178), and Longest Gap (r = 0.138). Compared with the correlation observed for normalized time-to-delivery (Figure 10), these associations were weaker in magnitude. Nevertheless, all three quality indicators exhibited consistent, non-zero relationships with the attention weights. The scatter plots showed substantial dispersion around the fitted regression lines, indicating considerable variability in attention allocation across different quality conditions while maintaining statistically significant overall trends.

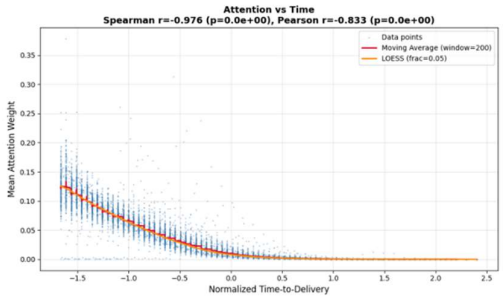

Figure 10. **Attention weights vs. time to delivery**

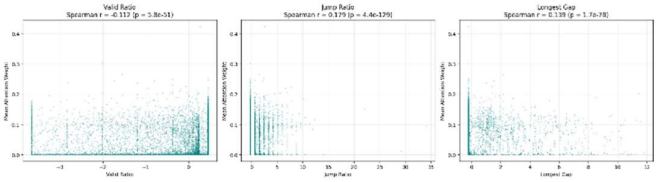

Figure 11. **Attention weights vs. signal quality**

3.5.6. Adaptive Gate Mechanism Analysis

Figure 12 presents the distribution of branch weights across the test set. Overall, the model assigned a higher average weight to the physiological branch (β_phys = 0.7253) versus the raw temporal branch (β_raw = 0.2747). However, the pattern varied across clinical groups: non-acidemic patients relied predominantly on HRV-based physiological representation; in acidemic patients, raw branch contribution increased markedly. Physiologically, HRV indices reflect stable averaged patterns, sufficient in normal cases, but acidemic conditions may manifest as transient decelerations or abrupt fluctuations not fully captured by extracted features. The model adaptively assigns greater weight to the raw branch for higher-resolution temporal information. Thus, the gate functions as an adaptive representation selection mechanism, making patient-specific trust decisions between physiological features and raw signal. This dynamic behavior mirrors clinical practice, where obstetricians adjust their interpretive strategy based on the clinical context and signal characteristics.

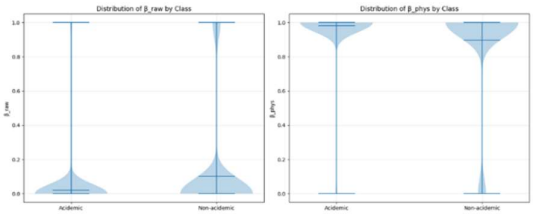

Figure 12. **Distribution of gate coefficients β_raw and β_phys by clinical group (violin plots)**

3.5.7. Patient Case Studies

To qualitatively validate the Clinical Attention Allocation hypothesis, two representative patients were selected from the test set: a True Positive acidemic patient (ID=1071, pH=6.98, predicted probability=0.991) and a True Negative non-acidemic patient (ID=1409, pH=7.21, predicted probability=0.390). Figure 13 presents, for each patient, the raw FHR signal, the preprocessed signal, the per-window attention weights (α), the normalized signal quality (Valid Ratio, Z-scored), and the Context-Aware Gate coefficients (β_raw, β_phys).

True Positive Acidemic Patient (pH=6.98): The raw FHR signal exhibits marked variability loss and multiple abrupt signal drops, particularly after minute 45, consistent with severe fetal compromise. After preprocessing, long gaps are flagged in the quality vector while short gaps are reconstructed via linear interpolation. The attention weight profile reveals a bimodal distribution: elevated weights in the early segment (minutes 5–20) followed by a mid-recording trough, then a systematic and progressive increase toward delivery, peaking in the final windows (minutes 60–65). The early attention peak coincides with a period of low signal quality (negative Z-scored Valid Ratio), suggesting the model identified pathologically informative morphological patterns in the raw signal despite degraded recording conditions — a behavior consistent with the cross-attention mechanism extracting complementary information between the raw and physiological branches. Critically, the Context-Aware Gate assigned β_raw = 1.000 and β_phys = 0.000, indicating complete reliance on the raw temporal branch. This is physiologically interpretable: in severe acidemia (pH = 6.98), pathological FHR patterns manifest as transient decelerations and loss of beat-to-beat variability that are not fully captured by aggregate HRV statistics, making the raw temporal representation informationally superior. The model's predicted probability of 0.991 reflects high discriminative confidence consistent with the severity of the clinical outcome.

True Negative Non-Acidemic Patient (pH = 7.21): The raw FHR signal shows physiologically normal variability with multiple artifact episodes visible as near-zero drops, which are effectively managed by preprocessing. The attention weight distribution is markedly different from the TP case: weights follow a near-monotonic increasing pattern from the recording onset toward delivery, with the highest weights concentrated in the final 20 minutes (minutes 60–80). This temporal gradient is the clearest visual confirmation of the QT-Conditioned Attention mechanism operating as designed, systematically prioritizing delivery-proximal windows in accordance with established clinical evidence that FHR patterns near delivery carry the greatest diagnostic significance [8]. The Context-Aware Gate produced balanced coefficients (β_raw = 0.455, β_phys = 0.545), with a slight dominance of the physiological branch. This is consistent with the population-level finding (β_phys ≈ 0.73 across non-acidemic patients) and reflects the model's reliance on stable HRV-derived features — which adequately characterize normal autonomic regulation — rather than raw morphological detail. The predicted probability of 0.390 correctly identifies this patient as low-risk, avoiding a false alarm.

Comparative Interpretation: The juxtaposition of these two cases illuminates the adaptive behavior of the Context-Aware Gate as a representation selection mechanism rather than a fixed fusion operator. In pathological conditions (TP, pH = 6.98), the gate shifts decisively toward raw temporal information to capture transient signal morphology; in physiologically normal conditions (TN, pH = 7.21), it maintains a balanced or phys-dominant allocation, leveraging the interpretive power of HRV indices. Notably, both patients exhibit the same qualitative temporal attention trend, increasing weight toward delivery, confirming that temporal prioritization is a universal property of the learned attention mechanism, independent of the clinical outcome. The key discriminating factor between the two cases lies not in the temporal pattern of attention but in the gate behavior and the absolute probability assigned, reflecting the model's ability to disentangle temporal context (shared) from pathological content (outcome-specific).

4. Discussion

The primary objective of this study was not merely to develop another deep learning classifier for fetal acidemia prediction, but rather to investigate whether a clinically inspired attention allocation strategy could be explicitly incorporated into a weakly supervised multiple-instance learning framework [15, 35]. Specifically, we hypothesized that fetal heart rate (FHR) segments should not contribute equally to the final decision. Instead, their diagnostic importance should depend on clinically meaningful contextual information, particularly signal quality and temporal proximity to delivery. The proposed QT-MIL-FHR framework was therefore designed as a proof-of-concept to test whether these contextual cues could guide attention allocation in a manner that more closely resembles clinical interpretation.

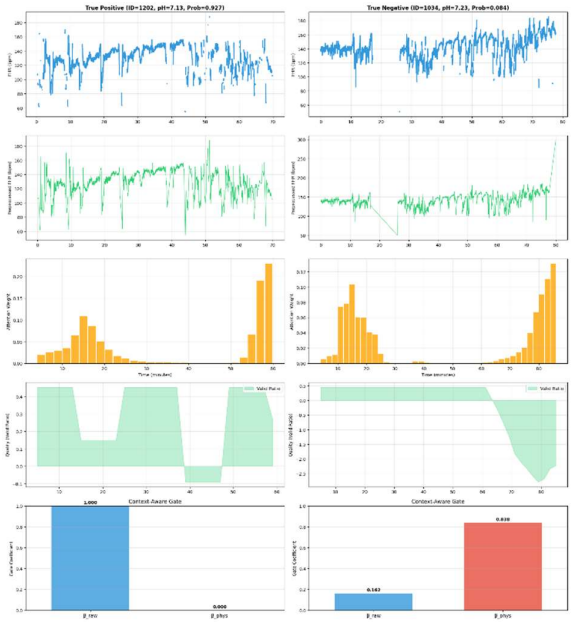

Figure 13. Case study illustrating the attention allocation behavior of QT-MIL-FHR for a representative patient

The experimental results support this hypothesis. Using patient-wise five-fold cross-validation on the CTU-UHB database [5], QT-MIL-FHR achieved a mean AUROC of 0.76 while maintaining relatively stable performance across folds (coefficient of variation approximately 5.5%). Although this level of discrimination is comparable to several recent deep learning approaches rather than substantially exceeding them, the principal contribution of the proposed framework lies in its design philosophy rather than incremental gains in predictive performance. Unlike conventional attention mechanisms that rely solely on signal content, QT-MIL-FHR explicitly incorporates contextual clinical information into the attention generation process, enabling the network to allocate diagnostic importance according to clinically relevant conditions.

The ablation studies further reinforce this interpretation. Removing Quality-Time conditioning consistently reduced discrimination performance, while eliminating Cross-Attention, nonlinear physiological features, entropy regularization [37],

or adaptive fusion also degraded model performance. These observations indicate that the proposed architecture does not derive its performance from architectural complexity alone. Instead, each component contributes complementary information, and their combined interaction produces the most robust results. Particularly noteworthy is the degradation observed after removing contextual conditioning, suggesting that explicit incorporation of clinical context provides additional information beyond that contained in the FHR waveform itself.

Attention analysis provides further insight into how the proposed framework reaches its decisions. Across the entire test cohort, attention weights exhibited a strong monotonic increase as delivery approached. Spearman correlation analysis demonstrated an almost monotonic relationship between normalized time-to-delivery and attention allocation, whereas Pearson correlation suggested a similarly strong overall trend. Importantly, these observations should not be interpreted as indicating that temporal proximity alone determines model decisions. Rather, they demonstrate that the model systematically assigns greater importance to observations recorded closer to birth, consistent with the well-established clinical understanding that late intrapartum FHR patterns often provide the most informative representation of fetal condition. Similar emphasis on late-stage recordings has long been reflected in intrapartum fetal surveillance guidelines, where clinical interpretation increasingly focuses on patterns immediately preceding delivery rather than treating all portions of the tracing equally [4, 21, 22, 25].

Signal quality also contributed to attention allocation, although its influence appeared substantially weaker than that of temporal context. Correlation analysis demonstrated statistically significant but relatively modest associations between attention weights and the three quality indicators (Valid Ratio, Jump Ratio, and Longest Gap) [18, 30, 32]. The corresponding scatter plots showed considerable dispersion, indicating that quality does not independently determine attention allocation. Instead, the model appears to treat quality as a complementary contextual variable whose influence depends on the underlying physiological content of each segment. This finding agrees with clinical practice, where poor-quality recordings do not automatically lose diagnostic value but instead require more cautious interpretation. Consequently, the proposed framework does not prioritize time at the expense of signal quality; rather, it combines both contextual variables to determine how much confidence should be placed in each temporal segment.

Another notable finding concerns the adaptive fusion mechanism. Analysis of the Context-Aware Gate demonstrated that the model dynamically adjusts the relative contribution of the raw temporal representation and the physiological HRV representation on a patient-specific basis [40]. Rather than maintaining fixed fusion coefficients throughout inference, the network learns to emphasize whichever representation appears more informative for a particular recording. This behavior suggests that the gate functions as a representation selection mechanism rather than a simple feature aggregation module. Such adaptive behavior is conceptually consistent with clinical reasoning, where physicians rarely rely on a single source of information but instead integrate multiple complementary observations depending on the clinical situation.

From a broader perspective, the proposed framework differs conceptually from most previous CTG-based deep learning models [23, 38]. Numerous recent studies have reported competitive discrimination performance using convolutional neural networks, recurrent neural networks, Transformer-based architectures, or hybrid feature fusion strategies. For example, DeepCTG v2 [48] and CTGFusionNet [42] reported AUROC values broadly comparable to those obtained in the present study. More recently, foundation models trained using large-scale self-supervised learning have demonstrated even higher discrimination by leveraging thousands of hours of unlabeled CTG recordings. However, despite differences in architecture, most existing approaches remain fundamentally signal-driven: attention mechanisms, when present, are generally derived from latent signal representations without explicitly incorporating clinically meaningful contextual information.

In contrast, QT-MIL-FHR attempts to embed clinical context directly into the attention generation process. The proposed Quality-Time conditioning mechanism represents an explicit modeling choice motivated by clinical interpretation rather than purely data-driven feature extraction. Accordingly, the principal contribution of this work should not be interpreted as achieving the highest AUROC reported in the literature, but rather as demonstrating the feasibility of incorporating clinically inspired contextual reasoning into a modern MIL architecture. From this perspective, the present work provides evidence supporting a new design direction for interpretable deep learning models in fetal monitoring [45].

An additional strength of the present study is the explicit evaluation of probability calibration. Many published CTG classification studies primarily emphasize discrimination metrics such as AUROC while providing little information regarding the reliability of predicted probabilities. In clinical decision support systems, however, calibrated probabilities are often as important as discrimination because intervention thresholds depend on estimated risk rather than ranking alone. Accordingly, calibration performance was evaluated using Expected Calibration Error (ECE), Brier Score, and reliability diagrams. Platt scaling substantially improved calibration, reducing ECE by approximately 76% and improving the Brier Score by approximately 18%. Nevertheless, the post-calibration reliability diagrams still demonstrated residual deviations from perfect calibration, indicating that further improvement remains possible. Reporting these calibration results transparently represents a methodological strength of the present work, while simultaneously identifying an important direction for future research involving more advanced calibration approaches, such as temperature scaling, Bayesian uncertainty estimation, or uncertainty-aware deep learning models .

Several limitations should also be acknowledged. First, model development and evaluation were performed exclusively on the CTU-UHB database [5]. Although this dataset remains one of the most widely used public benchmarks for intrapartum CTG analysis, its relatively limited size and single-center origin restrict conclusions regarding external generalizability [56]. Second, the relatively small number of academic cases limits statistical power and increases sensitivity to sampling variability across folds. Third, despite considerable improvement following Platt scaling, calibration remains imperfect, suggesting that probability estimation could benefit from additional refinement. Fourth, the current framework does not explicitly quantify

predictive uncertainty, an increasingly important consideration for high-risk clinical decision support. Finally, unlike recent foundation models based on large-scale self-supervised pretraining, QT-MIL-FHR was developed under a data-limited supervised learning scenario. Whether contextual attention allocation can further enhance large-scale pretrained representations remains an important open question.

A further consideration concerns generalization beyond the single-center CTU-UHB cohort. Differences across hospitals and monitoring devices in FHR acquisition electronics, sampling rate, artifact characteristics, and clinical protocols for labor management could all affect the distribution of both the raw signal and the quality indicators (Valid Ratio, Jump Ratio, Longest Gap) that QT-MIL-FHR relies on. In particular, the quality vector was calibrated on CTU-UHB's specific artifact patterns and may require re-normalization, or in some cases recalibration, when applied to signals from different CTG monitors. Similarly, the time-to-delivery encoding assumes a labor-duration distribution broadly consistent with the source cohort; populations with systematically different labor durations, for example due to differing induction or intervention practices, may shift the effective range over which this variable operates. We therefore regard the performance reported here as representative of within-database generalization only, and consider multi-center external validation, ideally spanning multiple monitor manufacturers and clinical protocols, a necessary next step before any claim of cross-institutional generalizability can be made.

Future work should therefore focus on several complementary directions. External validation using independent multicenter cohorts will be essential to assess generalizability across different acquisition protocols and patient populations [56]. Incorporating self-supervised representation learning may strengthen the raw temporal branch while preserving the clinically inspired attention mechanism proposed here. Integration of uncertainty estimation and improved calibration techniques could further increase clinical reliability. Finally, prospective evaluation within real-time intrapartum monitoring systems would provide valuable evidence regarding the practical utility of clinically conditioned attention mechanisms in routine obstetric practice [46].

Overall, the findings of this study suggest that explicitly incorporating clinical context into attention allocation represents a promising direction for interpretable fetal monitoring systems. Rather than attempting to replicate the entire clinical reasoning process, QT-MIL-FHR models two measurable components of clinical attention—signal quality and temporal proximity to delivery—and demonstrates that these factors can be successfully integrated into a weakly supervised deep learning framework [15]. Consequently, the principal contribution of this work is not simply the development of another CTG classifier, but the validation of a clinically inspired design philosophy that may guide future development of context-aware medical artificial intelligence systems.

5. Conclusion

This study presented QT-MIL-FHR, a context-aware multiple instance learning framework for fetal acidemia prediction from intrapartum fetal heart rate recordings. Unlike conventional signal-driven approaches, the proposed framework explicitly incorporates two clinically meaningful contextual variables—signal quality and time-to-delivery—into the attention allocation process. Rather than treating all temporal segments equally, the model learns to assign patient-specific attention weights according to both signal content and clinical context.

Experimental evaluation on the CTU-UHB database demonstrated that QT-MIL-FHR achieved competitive discrimination performance while maintaining stable patient-wise cross-validation results. Ablation studies consistently showed that the proposed components, including Quality-Time conditioning, bidirectional cross-attention, adaptive fusion, and entropy-regularized gating, each contributed to the overall performance. Furthermore, interpretability analyses indicated that the learned attention distribution follows clinically plausible temporal patterns, while signal quality acts as a complementary contextual factor. The adaptive fusion mechanism also revealed that the relative contributions of raw temporal representations and physiological HRV features vary across patients, suggesting that the model dynamically selects complementary representations rather than relying on a fixed feature hierarchy.

An additional strength of this work is the explicit evaluation of probability calibration using Expected Calibration Error (ECE), Brier Score, and Reliability Diagrams. While calibration improved substantially after Platt scaling, residual deviations indicate that further calibration refinement remains an important direction for future research.

Overall, the principal contribution of this work is not merely the development of another deep learning classifier, but the demonstration that clinically inspired attention allocation can be explicitly embedded into a weakly supervised MIL framework. This proof-of-concept provides a promising foundation for future context-aware and clinically interpretable artificial intelligence systems for intrapartum fetal monitoring.

6. Statements & Declarations

6.1. Author Contributions

Conceptualization, A.Sheykhani. and A.Tareh.; methodology, A.Tareh.; software, A.Tareh.; validation, A.Tareh., A. Sheykhani. and N.Jafarnia.D.; formal analysis, A.Tareh.; investigation, A.Tareh.; resources, A. Sheykhani. and N.Jafarnia.D.; data curation, A.T.; writing—original draft preparation, A.T.; writing—review and editing, A.S. and N.J.D.; visualization, A.Tareh.; supervision, A. Sheykhani. and N.Jafarnia.D.; project administration, A. Sheykhani. All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

Publicly available datasets were analyzed in this study. These data can be found here: https://physionet.org/content/ctu-uhb-ctgdb/1.0.0/

6.3. Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

6.4. Acknowledgments

The authors acknowledge the open-access provision of the CTU-UHB database through PhysioNet.

6.5. Declaration of Generative AI and AI-assisted Technologies in the Writing Process

During the preparation of this manuscript, the authors used Claude (Anthropic) for language editing, grammar checking, and improving readability. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

6.6. Ethics Approval

Not applicable. This study used a publicly available, de-identified dataset (CTU-UHB), and no new human or animal data were collected.

6.7. Informed Consent

Not applicable.

6.8. Consent for Publication

Not applicable.

6.9. Conflicts of Interest

The authors declare no conflict of interest.

7. References

[1] E. H. Hon and S. T. Lee, "Electronic evaluation of the fetal heart rate. VIII. Patterns preceding fetal death, further observations", American Journal of Obstetrics and Gynecology, vol. 87, pp. 814–826, Nov. 1963.

[2] M. J. Stout and A. G. Cahill, "Electronic fetal monitoring: Past, present, and future", Clinics in Perinatology, vol. 38, no. 1, pp. 127–142, Mar. 2011, doi:10.1016/j.clp.2010.12.002.

[3] D. Ayres-de-Campos, "Introduction: Why is intrapartum foetal monitoring necessary—Impact on outcomes and interventions", Best Practice & Research Clinical Obstetrics & Gynaecology, vol. 30, pp. 3–8, Jan. 2016, doi:10.1016/j.bpobgyn.2015.06.004.

[4] D. Ayres-de-Campos, C. Y. Spong, and E. Chandraharan, "FIGO consensus guidelines on intrapartum fetal monitoring: Cardiotocography", International Journal of Gynecology & Obstetrics, vol. 131, no. 1, pp. 13–24, Oct. 2015, doi:10.1016/j.ijgo.2015.06.020.

[5] G. Macones, G. Hankins, C. Spong, J. Hauth, and T. Moore, "The 2008 National Institute of Child Health and Human Development report on fetal heart rate monitoring: Update on definitions, interpretation, and research guidelines", Obstet. Gynecol., vol. 112, no. 3, pp. 661–666, Sep. 2008, doi:10.1097/AOG.0b013e3181841395.

[6] S. B. Mulkey, A. Plessis, "The critical role of the central autonomic nervous system in fetal-neonatal transition", Seminars in Pediatric Neurology, vol. 28, pp. 29–37, Dec. 2018, doi:10.1016/j.spen.2018.05.004.

[7] C. Garabedian, J. De Jonckheere, L. Butruille, P. Deruelle, L. Storme, and V. Houfflin-Debarge, "Understanding fetal physiology and second line monitoring during labor", Journal of Gynecology Obstetrics and Human Reproduction, vol. 46, no. 2, pp. 113–117, Feb. 2017, doi:10.1016/j.jogoh.2016.11.005.

[8] E. Chandraharan, S. A. Evans, D. Krueger, S. Pereira, S. Skivens, and A. Zaima, "Physiological interpretation: Intrapartum fetal monitoring guideline", Physiological CTG Interpretation. Intrapartum fetal monitoring guideline, 2018.

[9] SP Chauhan, CK Klauser, TC Woodring, M Sanderson, EF Magann, JC Morrison, "Intrapartum nonreassuring fetal heart rate tracing and prediction of adverse outcomes: Interobserver variability" , American journal of obstetrics and gynecology, vol. 199, no. 6, pp. 623.e1–623.e5, Dec. 2008, doi:10.1016/j.ajog.2008.06.027.

[10] SL Clark, EF Hamilton, TJ Garite, A Timmins, PA Warrick, S Smith, "The limits of electronic fetal heart rate monitoring in the prevention of neonatal metabolic acidemia", American Journal of Obstetrics and Gynecology, vol. 216, no. 2, pp. 163.e1–163.e6, Feb. 2017, doi:10.1016/j.ajog.2016.10.009.

[11] B. M. Yli, J. Kessler, T. Eikeland, B. L. Hustad, W. Dragnes, and T. Henriksen, "What is the gold standard for intrapartum fetal monitoring?", Acta Obstetricia et Gynecologica Scandinavica, vol. 91, no. 9, pp. 1011–1014, Sep. 2012, doi:10.1111/j.1600-0412.2012.01475.x.

[12] L. Heelan, "Fetal monitoring: Creating a culture of safety with informed choice", Journal of Perinatal Education, vol. 22, no. 3, pp. 156–165, 2013, doi:10.1891/1058-1243.22.3.156.

[13] L. C. Gilstrap, III, K. J. Leveno, J. Burris, M. Lynne Williams and B. B. Little, "Diagnosis of birth asphyxia on the basis of fetal pH, Apgar score, and newborn cerebral dysfunction", American Journal of Obstetrics & Gynecology, vol. 161, no. 3, pp. 825–830, Sep. 1989, doi:10.1016/0002-9378(89)90410-9.

[14] M. L. Kitlinski, K. Källén, K. Marsál, and P. Olofsson, "Gestational age-dependent reference values for pH in umbilical cord arterial blood at term", Obstetrics & Gynecology, vol. 102, no. 2, pp. 338–345, Aug. 2003, doi:10.1016/S0029-7844(03)00512-X.

[15] P. Olofsson, "Umbilical cord pH, blood gases, and lactate at birth: Normal values, interpretation, and clinical utility", American Journal of Obstetrics & Gynecology, vol. 228, no. 5S, pp. S1222–S1240, May 2023, doi:10.1016/j.ajog.2022.07.001.

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

[17] F. Francis, S. Luz, H. Wu, S. J. Stock, and R. Townsend, "Machine learning on cardiotocography data to classify fetal outcomes: A scoping review", Computers in Biology and Medicine, vol. 172, p. 108220, Apr. 2024, doi:10.1016/j.compbiomed.2024.108220.

[18] A. M. Ponsiglione, C. Cosentino, G. Cesarelli, F. Amato, and M. Romano, "A comprehensive review of techniques for processing and analyzing fetal heart rate signals", Sensors, vol. 21, no. 18, p. 6136, Sep. 2021, doi:10.3390/s21186136.

[19] M. Ribeiro, J. Monteiro-Santos, L. Castro, L. Antunes, C. Costa-Santos, A. Teixeira, and T. S. Henriques, "Non-linear methods predominant in fetal heart rate analysis: A systematic review", Frontiers in Medicine, vol. 8, p. 661226, Nov. 2021, doi:10.3389/fmed.2021.661226.

[20] G. S. Dawes, M. Moulden, and C. W. G. Redman, "System 8000: Computerized antenatal FHR analysis", Journal of Perinatal Medicine, vol. 19, no. 1–2, pp. 47–51, 1991, doi:10.1515/jpme.1991.19.1-2.47.

[21] M. G. Signorini, G. Magenes, S. Cerutti, and D. Arduini, "Linear and nonlinear parameters for the analysis of fetal heart rate signal from cardiotocographic recordings", IEEE Transactions on Biomedical Engineering, vol. 50, no. 3, pp. 365–374, Mar. 2003, doi:10.1109/TBME.2003.808824.

[22] E. Salamalekis Assoc Professor, E. Hintipas, I. Salloum, G. Vasios, C. Loghis, N. Vitoratos, et al., "Computerized analysis of fetal heart rate variability using the matching pursuit technique as an indicator of fetal hypoxia during labor", Journal of Maternal-Fetal & Neonatal Medicine, vol. 19, no. 3, pp. 165–169, 2006, doi:10.1080/14767050500233290.

[23] J. van Laar, C. H. L. Peters, R. Vullings, S. Houterman, J. W. M. Bergmans and S. G. Oei, "Fetal autonomic response to severe acidaemia during labour", BJOG: An International Journal of Obstetrics & Gynaecology, vol. 117, no. 4, pp. 429–437, Mar. 2010, doi:10.1111/j.1471-0528.2009.02456.x.

[24] L. Mendis, M. Palaniswami, F. Brownfoot, and E. Keenan, "Computerised cardiotocography analysis for the automated detection of fetal compromise during labour: A review", Bioengineering, vol. 10, no. 9, p. 1007, Aug. 2023, doi:10.3390/bioengineering10091007.

[25] V. Chudáček, J. Spilka, M. Burša, P. Janků, L. Hruban, M. Huptych and L. Lhotská, "Open access intrapartum CTG database", BMC Pregnancy Childbirth, vol. 14, p. 16, Jan. 2014, doi:10.1186/1471-2393-14-16.

[26] Costa, D. Ayres-de-Campos, F. Costa, C. Santos, and J. Bernardes, "Prediction of neonatal acidemia by computer analysis of fetal heart rate and ST event signals," American Journal of Obstetrics & Gynecology, vol. 201, no. 5, pp. 464.e1–464.e6, Nov. 2009, doi:10.1016/j.ajog.2009.04.033.

[27] A. Georgieva, C. W. G. Redman, and A. T. Papageorghiou, "Computerized data-driven interpretation of the intrapartum cardiotocogram: A cohort study", Acta Obstetricia et Gynecologica Scandinavica, vol. 96, no. 7, pp. 883–891, Jul. 2017, doi:10.1111/aogs.13136.

[28] N. Baghel, R. Burget, and M. K. Dutta, "1D-FHRNet: Automatic diagnosis of fetal acidosis from fetal heart rate signals", Biomedical Signal Processing and Control, vol. 71, p. 102794, Jan. 2022, doi:10.1016/j.bspc.2021.102794.

[29] 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, p. 286, Dec. 2019, doi:10.1186/s12911-019-1007-5.

[30] Z. Zhao, Y. Zhang, Z. Comert, and Y. Deng, "Computer-aided diagnosis system of fetal hypoxia incorporating recurrence plot with convolutional neural network", Frontiers in Physiology, vol. 10, p. 255, Mar. 2019, doi:10.3389/fphys.2019.00255.

[31] P. P. Aswathi Mohan, V. Uma, R. Sasirekha, and V. Hamsika, "CTGFusionNet: Fusion of deep learning models for predicting fetal distress—A multimodal approach", Health Information Science and Systems, vol. 13, p. 66, 2025, doi:10.1007/s13755-025-00383-1.

[32] P. Hamelmann, R. Vullings, A. F. Kolen, J. W. M. Bergmans, J. O. E. H. v. Laar, P. Tortoli and M. Mischi, "Doppler ultrasound technology for fetal heart rate monitoring: A review", IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control, vol. 67, no. 2, pp. 226–238, Feb. 2020, doi: 10.1109/TUFFC.2019.2943626.

[33] M. Tarvonen, J. Markkanen, V. Tuppurainen, R. Jernman, V. Stefanovic and S. Andersson, "Intrapartum cardiotocography with simultaneous maternal heart rate registration improves neonatal outcome", American Journal of Obstetrics and Gynecology, vol. 230, no. 4, pp. 379.e1–379.e12, Apr. 2024, doi:10.1016/j.ajog.2024.01.011.

[34] M. Ilse, J. M. Tomczak, and M. Welling, "Attention-Based Deep Multiple Instance Learning," in Proceedings of the 35th International Conference on Machine Learning, vol. 80, Stockholm, Sweden, Jul. 2018, pp. 2127–2136.

[35] Y. Xiao, Y. Lu, M. Liu, R. Zeng, and J. Bai, "A deep feature fusion network for fetal state assessment", Frontiers in Physiology, vol. 13, p. 969052, Nov. 2022, doi:10.3389/fphys.2022.969052.

[36] P. Brocklehurst, D. Field, K. Greene, E. Juszczak, R. Keith, S. Kenyon, et al., "Computerised interpretation of fetal heart rate during labour (INFANT): A randomised controlled trial", The Lancet, vol. 389, no. 10080, pp. 1719–1729, Apr. 2017, doi:10.1016/S0140-6736(17)30568-8.

[37] A. D. J. Hulsenboom, M. B. Van der Hout-van der Jagt, E. S. A. van den Akker, P. Bakker, E. van Beek, A. P. Drogtrop, et al., "New possibilities for ST analysis—A post-hoc analysis on the Dutch STAN RCT", Early Human Development, vol. 166, p. 105537, Mar. 2022, doi:10.1016/j.earlhumdev.2021.105537.

[38] I. Ben M'Barek, G. Jauvion, J. Merrer, M. Koskas, O. Sibony, P.-F. Ceccaldi, E. Le Pennec, and J. Stirnemann, "DeepCTG® 2.0: Development and validation of a deep learning model to detect neonatal acidemia from cardiotocography during labor," Computers in Biology and Medicine, vol. 184, p. 109448, Jan. 2025, doi:10.1016/j.compbiomed.2024.109448.

[39] G. Li, C. Li, G. Wu, D. Ji, and H. Zhang, "Multi-view attention-guided multiple instance detection network for interpretable breast cancer histopathological image diagnosis", IEEE Access, vol. 9, pp. 79671–79684, 2021, doi: 10.1109/ACCESS.2021.3084360.

[40] D. Ayres-de-Campos, P. Sousa, A. Costa, and J. Bernardes, "Omniview-SisPorto® 3.5—A central fetal monitoring station with online alerts based on computerized cardiotocogram+ST event analysis", Journal of Perinatal Medicine, vol. 36, no. 3, pp. 260–264, May 2008, doi:10.1515/JPM.2008.030.

[41] J. Cai, Y. Li, B. Liu, Z. Wu, S. Zhu, Q. Chen, Q. Lei, H. Hou, Z. Guo, H. Jiang, S. Guo, F. Wang, S. Huang, S. Zhu, X. Fan, and S. Tao, "Developing Deep LSTMs With Later Temporal Attention for Predicting COVID-19 Severity, Clinical Outcome, and Antibody Level by Screening Serological Indicators Over Time", IEEE Journal of Biomedical and Health Informatics, vol. 28, no. 7, pp. 4204–4215, Jul. 2024, doi:10.1109/JBHI.2024.3384333.

[42] L. Zhou, Z. Liu, Z. Shangguan, X. Yuan, Y. Li, and B. Hu, "TAMFN: Time-Aware Attention Multimodal Fusion Network for Depression Detection," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 31, pp. 669–679, 2023, doi:10.1109/TNSRE.2022.3224135.

[43] S. Heo, E.-A. Kang, J. Y. Yu, H. R. Kim, S. Lee, K. Kim, Y. Hwangbo, R. W. Park, H. Shin, K. Ryu, C. Kim, H. Jung, Y. Chegal, J.-H. Lee, and Y. R. Park, "Time Series AI Model for Acute Kidney Injury Detection Based on a Multicenter Distributed Research Network: Development and Verification Study," JMIR Medical Informatics, vol. 12, p. e47693, Jul. 2024, doi:10.2196/47693.

[44] Francis, F., Wu, H., Luz, S., Townsend, R., & Stock, S. (2022). Detecting intrapartum fetal hypoxia from cardiotocography using machine learning. In 2022 Computing in Cardiology (CinC) (Vol. 49, pp. 1–4). Z. Li, J. Li, K. Kuang, B. Jing, B. Du, F. Liao, and J. Yang, "SimTA++: Simple Attention Neural Network for Clinical Asynchronous Time Series," Neural Networks, vol. 191, p. 107735, Nov. 2025, doi:10.1016/j.neunet.2025.107735.

[45] A. Petrozziello, C. W. G. Redman, A. T. Papageorghiou, I. Jordanov, and A. Georgieva, "Multimodal convolutional neural networks to detect fetal compromise during labor and delivery", IEEE Access, vol. 7, pp. 112026–112036, Aug. 2019, doi:10.1109/ACCESS.2019.2933368.

[46] Esteban, L. M., Castán, B., Esteban-Escaño, J., Sanz-Enguita, G., Laliena, A. R., Lou-Mercadé, A. C., Chóliz-Ezquerro, M., Castán, S., & Savirón-Cornudella, R. (2023). Machine learning algorithms combining slope deceleration and fetal heart rate features to predict acidemia. Applied Sciences, 13(13), 7478. https://doi.org/10.3390/app13137478G. Mertes, Y. Long, Z. Liu, Y. Li, Y. Yang, and D. A. Clifton, "A Deep Learning Approach for the Assessment of Signal Quality of Non-Invasive Foetal Electrocardiography", Sensors, vol. 22, no. 9, p. 3303, Apr. 2022, doi:10.3390/s22093303.

[47] K. Barnova et al., "Artificial intelligence and machine learning in electronic fetal monitoring", Archives of Computational Methods in Engineering, vol. 31, pp. 2557–2588, 2024, doi:10.1007/s11831-023-10055-6.

[48] S. Das, H. Mukherjee, S. M. Obaidullah, K. C. Santosh, K. Roy, and C. K. Saha, "Recurrent Neural Network Based Classification of Fetal Heart Rate Using Cardiotocograph," in Recent Trends in Image Processing and Pattern Recognition: Second International Conference, RTIP2R 2018, Revised Selected Papers, Part II, Communications in Computer and Information Science, vol. 1037, pp. 226–234, Springer, 2019, doi:10.1007/978-981-13-9184-2_20.

[49] Baghel, N., Burget, R., & Dutta, M. K. (2022). 1D-FHRNet: Automatic diagnosis of fetal acidosis from fetal heart rate signals. Biomedical Signal Processing and Control, 71, 103179. https://doi.org/10.1016/j.bspc.2021.102794.