Ensemble Models for Real-Time Fetal Monitoring Using Discrete Segmentation of Cardiotocography

Muhammad Usama Faheem[1], Rudy Lapeer[1], Beatriz De La Iglesia[1], Rahul Gore[2], Rowan Connell[2], and Wenjia Wang*[1]

[1] School of Computing Sciences, University of East Anglia, Norwich Research Park, Norwich, NR4 7TJ, United Kingdom
M.Faheem@uea.ac.uk, R.lapeer@uea.ac.uk, B.iglesia@uea.ac.uk, Wenjia.wang@uea.ac.uk*

[2] 5GoreConn, Fitzrovia House, 153-157 Cleveland Street, London, W1T 6QW, United Kingdom

Abstract. Accurate, real-time detection of fetal distress is critical during labour. While cardiotocography (CTG) is a standard for fetal monitoring, its manual interpretation is often subjective and error-prone. This study evaluates how CTG signal segment length influences the performance of machine learning models for timely fetal distress prediction. Using the CTU-UHB dataset of 552 intrapartum CTG recordings, signals were preprocessed and segmented into 90, 20, 10, and 5-minute windows. Features were extracted using eight open-source libraries and evaluated across seven classifiers. Notably, Random Forest combined with TSFEL time-domain features achieved peak accuracy (91.5%) on 10-minute segments. The last 20-minute segment also performed well (91.3%), indicating that shorter windows are sufficient for accurate prediction. Five-minute segments, however, showed reduced robustness due to data constraints. Findings suggest that 10–20 minute CTG segments are best for real-time fetal monitoring, with ensemble models offering high accuracy, reliability, and potential for clinical integration in labour management.

Keywords: Fetal Heart Rate, Fetal Distress, Cardiotocography, Ensemble Learning, Uterine Contraction

1 Introduction

Neonatal mortality remains a critical global health challenge. The WHO reported 2.3M of neonatal deaths in 2021 (25), i.e. 6,400 deaths per day on average. This highlights the need for new effective fetal surveillance techniques. For many decades to this date, fetal monitoring is still mostly relying on using Cardiotocography (CTG) - a non-invasive method that simultaneously records fetal heart rate (FHR) and uterine contractions (UCs) to assess fetal well-being. These signals provide crucial insights into fetal distress[3], particularly in detecting hypoxia,

[3] We decided to keep the traditional term of "fetal distress" although, it has been argued that "fetal compromise" may be more accurate.

acidemia, and abnormal labour patterns. However, the manual interpretation of CTG remains highly subjective, with significant interobserver variability, often leading to misclassification of fetal distress or unnecessary interventions (26).

This research aims to develop an AI-based algorithm to interpret CTG signals automatically and detect the distressful conditions of a baby during labour. Through comprehensively reviewing the relevant literature, we identified a key shortcoming in all the previous studies, that is, they all used the entire 90-minute length of the CTG signal in a benchmark dataset to train and test their models. The problem with this approach is in clinical practice; a predictive system should not wait for 90 minutes to produce its results.

In this paper we present the following four studies, **(1)** Investigate the influence of variable-length segments: Unlike prior studies that use fixed-duration CTG signals, We are the first to evaluate predictive performance across varying-length segments of CTG signals (5–90 minutes), reflecting real-world temporal constraints. The objective of this is to gain insights into optimal signal length for real time monitoring **(2)** Comparative analysis of ensemble models: This study demonstrates that Random Forest consistently outperforms all other models, including traditional classifiers such as Logistic Regression and SVM, confirming its reliability for real-time CTG analysis. **(3)** Comprehensive feature extraction comparison: Unlike previous studies that focus on manual feature selection or deep learning-based embeddings, this research systematically compares TSFEL, TSFresh, Catch22, and Sktime, identifying the most effective features for fetal distress classification. **(4)** Implications for clinical deployment in real-time: Our study provides evidence that shorter signals (10–20 mins) provide highest diagnostic information, reducing the computational burden and enabling faster decision-making in real time fetal monitoring systems leading to earlier planned interventions that could save lives and reduce complications.

2 Related Work

The integration of machine learning (ML) and artificial intelligence (AI) techniques has gained significant attention in automating fetal health classification, aiming to address the limitations of manual CTG interpretation. Numerous studies have explored both traditional ML classifiers and deep learning architectures to improve diagnostic accuracy and reduce observer bias. Deep learning models, particularly Convolutional Neural Networks (CNNs), have demonstrated strong performance in fetal distress detection. Zhao et al. (26) introduced a CNN-based approach using recurrence plots to transform 1D FHR signals into 2D images, reporting an accuracy of 98.7% in detecting fetal hypoxia. However, their evaluation may suffer from data leakage, as the 21,000 images per class were generated from only 105 signals per class. This means that multiple representations of the same signal may appear in both training and test sets, potentially inflating performance by learning signal-specific rather than generalisable features. Similarly, Mendis et al. (21) proposed an input-length-invariant deep learning model, demonstrating robustness in fetal distress classification across varying time du-

rations. Daydulo et al. (14) employed Morse wavelet-based time–frequency representations, highlighting the effectiveness of wavelet transforms in capturing temporal patterns.

While deep learning models offer high accuracy, they present significant computational challenges, making them unsuitable for real-time clinical applications. These models require large labeled datasets, substantial computational resources, and extensive training times, making them difficult to deploy on CTG machines in hospital settings. Furthermore, deep learning-based methods lack interpretability, a major concern for clinicians who require transparent decision-making frameworks in medical diagnostics. Given these limitations, ensemble learning techniques, such as Random Forest (RF), XGBoost, and Gradient Boosting, have emerged as more computationally efficient and interpretable alternatives for fetal monitoring (24; 1).

Recent studies have demonstrated the effectiveness of ensemble models in improving fetal health classification. Aeberhard et al. (1) successfully implemented ensemble learning for classifying caesarean and vaginal deliveries, achieving high accuracy with interpretable decision boundaries. Similarly, Comert et al. (13) explored Random Forest and XGBoost for fetal distress prediction, outperforming traditional classifiers. In another study, Zou et al. (27) introduced a hybrid MGRU-XGB model, integrating multi-channel gated recurrent units with XGBoost, demonstrating superior performance in capturing temporal and nonlinear relationships in CTG data. However, a key limitation in many of these studies is their dependence on fixed-length CTG signals, which does not account for the variability in fetal heart rate dynamics across different labor stages.

Traditional ML models rely on feature extraction techniques to convert raw CTG signals into meaningful representations for classification. Several studies have explored handcrafted features based on clinical guidelines, but such methods may fail to capture the full complexity of FHR and UC patterns (22). To address this, automated feature extraction methods, such as recurrence plots, principal component analysis (PCA), and time-frequency analysis, have been explored (2; 26). Kuzu et al. (18) employed ensemble learning with extracted features from CTG signals, achieving 99.5% classification accuracy on the UCI CTG dataset (4) containing 2126 records with extracted features and no raw CTG signals. However, most studies have not systematically compared multiple feature extraction libraries to determine the most effective representations for fetal distress classification.

A major limitation in existing research is the assumption that longer CTG signals provide better classification performance. Many studies rely on 30 mins or 90 mins fixed-duration signals, without evaluating the efficacy of shorter segments. However, emerging research suggests that shorter signals (8–13 mins) may be sufficient for accurate predictions (26). Most AI-driven studies have not addressed the impact of variable-length segments, missing opportunities to optimize fetal monitoring for real-time decision-making.

This study aims to fill this gap by conducting a comparative analysis of classification performance across different CTG segment lengths (90 mins, 20 mins,

10 mins, and 5 mins segments). Instead of assuming that longer signals are superior, this research systematically evaluates whether shorter time segments can provide comparable accuracy, making them more suitable for real-time clinical applications. The results demonstrate that 10 mins and 20 mins signals yield accuracy levels comparable to the full 90 mins signal, challenging the assumption that longer durations necessarily enhance diagnostic performance.

By addressing the limitations of existing approaches such as reliance on fixed-length signals, lack of real-time feasibility, high computational demands, and the assumption that longer signals provide superior diagnostic value, this study sets a benchmark for AI-driven CTG classification by demonstrating that shorter CTG signals can yield reliable, interpretable, and computationally efficient predictions. The findings challenge conventional assumptions about signal length and feature selection, paving the way for future advancements in AI-driven fetal monitoring.

3 Methods and Materials

3.1 Data Background

The CTU-UHB dataset, available via PhysioNet, is the largest public repository of intrapartum cardiotocography (CTG) recordings (10). It includes 552 recordings collected at The University Hospital Brno, Czech Republic, each spanning 90 minutes before delivery. Signals capture fetal heart rate (FHR) and uterine contractions (UCs) at 4 Hz, covering both the first (60 mins) and second (30 mins) stages of labor for comprehensive fetal monitoring.

Classification was based on umbilical artery blood pH measured post-delivery: recordings with pH $\leq$ 7.15 were labeled pathological, and those with pH $>$ 7.15 as normal (11), avoiding subjectivity from visual annotations or APGAR scores. This yielded 445 normal (80.9%) and 105 compromised (19.1%) cases, reflecting real-world class imbalance. The inherent skew highlights the need for robust techniques to address imbalance in predictive modeling. The dataset is openly accessible via PhysioNet (17). The CTU-UHB dataset is based on a single geographic cohort; therefore, generalization to broader populations remains a limitation.

3.2 Metadata and Signal Preprocessing

Signal preprocessing is crucial in biomedical signal analysis, directly impacting the accuracy and reliability of downstream models (12). In CTG, signals are acquired using external ultrasound probes or internal electrodes, which are prone to noise and artifacts due to fetal/maternal movement, poor sensor placement, or clinical interventions. Effective preprocessing is therefore essential to ensure signal integrity. CTG signals commonly exhibit missing values (zeros) and outliers (6). Preprocessing begins by cleaning both signal and metadata. Postnatal variables such as pH, BE, BDecf, pCO2, and Apgar scores are excluded to preserve real-time applicability. Columns with only null or zero values are dropped, and missing metadata entries are imputed using the mode.

Signal-specific preprocessing involves removing segments with >15 seconds of zeros, linearly interpolating shorter ones, and eliminating unstable segments (BPM differences >25) through interpolation. Extreme values (<50 or >200 BPM) are replaced using Hermite spline interpolation (23). To address class imbalance (105 pathological vs. 447 normal cases), the Synthetic Minority Oversampling Technique (SMOTE) was applied. SMOTE generates synthetic samples for the minority class, balancing the dataset and improving model performance by reducing bias toward the majority class (7).

3.3 Feature Extraction techniques

Feature extraction plays a critical role in time series analysis of biomedical signals by transforming raw data into meaningful inputs for machine learning models. Traditional fetal heart rate (FHR) analysis involves hand-crafted features like accelerations, decelerations, and baseline estimates, requiring expert knowledge (16; 5). Modern computational approaches allow systematic, scalable, and automated feature extraction.

In this study, we evaluate eight state-of-the-art feature extraction libraries on the CTU-UHB dataset, offering a novel perspective on fetal monitoring. Unlike previous studies that relied on limited manual features, we adopt automated methods and assess their comparative performance.

Overview of Feature Extraction Libraries

- **TSFEL (Time Series Feature Extraction Library):** Extracts 390+ features from time, statistical, and spectral domains, including mean, variance, and power spectral density (PSD) (3). Variants include TSFEL All, Time, Statistical, and Spectral.
- **TSFresh:** Offers 750+ statistical and time-series features, with automated relevance testing (e.g., autocorrelation, peaks, spectral entropy) (8).
- **Sktime.TSFresh:** Integrates TSFresh into the Sktime framework for classification and forecasting (19).
- **Catch22:** Extracts 22 interpretable and efficient features (e.g., entropy, fluctuation analysis) (20).
- **Sktime Rocket:** Uses random convolutional kernels to summarize local patterns with minimal tuning (15).

3.4 Variable window length technique

Signal Characteristics and Segmentation Rationale This study uses 552 CTG recordings in which signal length is ranging from 60 to 90 minutes, with most signals clustered around 65–75 minutes in length. The variability in signal lengths challenges consistent feature extraction and modeling. To standardize inputs, signals were segmented into fixed 20-, 10-, and 5-minute windows, enabling uniform comparisons across models. Segmenting into shorter windows

reflects real-world clinical needs, where timely decisions are essential. The study uniquely explores the minimal time window necessary to maintain high model accuracy and robustness, setting a new benchmark for efficient, real-time CTG analysis and fetal monitoring.

3.5 Experimental Design

We conducted a two-phase comparative analysis to evaluate the impact of signal length on classification performance: (1) using the full 90-minute CTG signals, and (2) segmenting signals into shorter windows of 20, 10, and 5 minutes to simulate real-time decision-making.

Full 90-Minute Signal The complete 90-minute CTG signals were used as a baseline, capturing the full progression of labor. While comprehensive, such long windows are impractical for real-time clinical decision-making, where prompt assessments are essential.

Segmented Signal Analysis CTG signals were segmented into non-overlapping windows:

- **20 mins:** Early (0–20 mins), Mid (40–60 mins), and Last 20 mins (scaled to end).
- **10 mins and 5 mins:** Each 20-minute segment was further divided into two 10-minute or four 5-minute windows to assess finer temporal performance.

This segmentation approach ensured standardized input lengths, handled variability in signal durations (60–90 mins), and enabled consistent comparisons across labor stages.

Feature Extraction and Modeling Eight feature extraction libraries were applied to each time window: TSFEL (All, Time, Spectral, Statistical), TSFresh, sktime.TSFresh, Catch22, and ROCKET. These provided diverse features across statistical, spectral, and time domains.

Models included both baseline classifiers (Logistic Regression, SVM, k-NN, Naive Bayes) and ensemble models (Random Forest, Gradient Boosting, AdaBoost). Ensemble approaches were prioritized due to their robustness on noisy, imbalanced biomedical data and real-time feasibility, unlike deep learning methods which require significant computational resources and much larger datasets than the CTU-UHB dataset.

Training and Evaluation Data was split into 80% training and 20% testing sets. The training experiment was repeated five times with a different training and test set each time and the mean and std were calculated and reported.

4 Results

This section presents a comprehensive evaluation of fetal distress classification performance across varying CTG signal lengths 90, 20, 10, and 5 minutes using multiple machine learning models and feature extraction libraries. The goal was to identify the shortest possible window that delivers high predictive accuracy while supporting real-time clinical decision-making. Performance was assessed based on accuracy, sensitivity, and specificity.

4.1 90 min signal

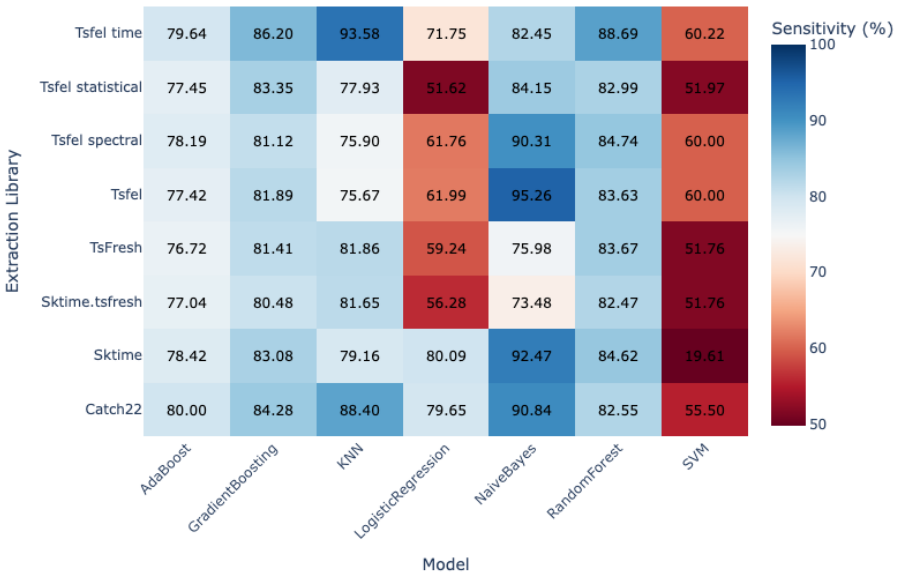

Fig. 1. Sensitivity heat map for 90 min signal

Best-Performing Model The full-length 90-minute signal served as a baseline. Among all models and feature libraries, Random Forest paired with TSFEL Time-domain features achieved the highest performance, with 88.94% accuracy, 88.69% sensitivity, and 89.04% specificity. Other ensemble models, such as Gradient Boosting and AdaBoost, performed well as well. Overall ensembles were not only more accurate but also more consistent with smaller variance. Baseline classifiers like Logistic Regression and SVM showed poor generalization (see

Figure 2, particularly with high-dimensional features (e.g., TSFresh), affirming the need for models that can handle complex, nonlinear time-series patterns. See Figure 1 for 90 min sensitivity heatmap.

While 90-minute recordings offer a fairly complete (as labour can be much longer than 90 mins) view of labor, they are not suitable for real-time clinical use due to signal degradation in later stages, increased imputation, and decision-making latency—points echoed by researchers. These findings establish the upper bound for classification performance but highlight the limitations of relying on extended monitoring in real-world settings.

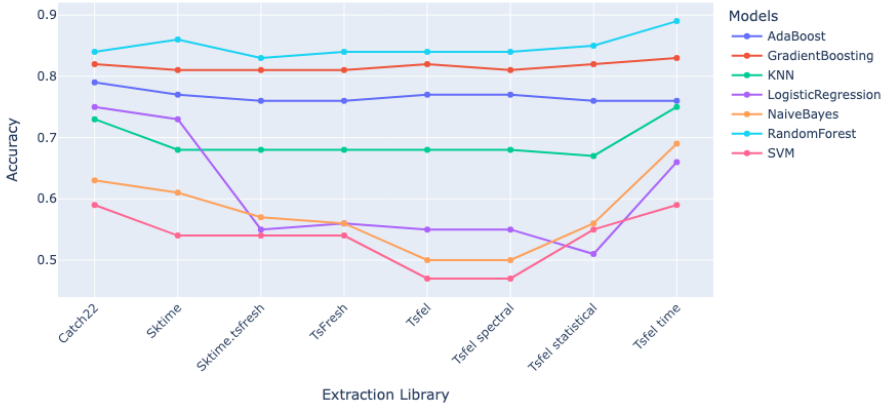

Fig. 2. Machine learning model performance across libraries on the 90 mins signal

4.2 20 min signal

To simulate progressive labor stages, we segmented signals into non-overlapping 20-minute windows. The last 20-minute segment (relative to delivery) outperformed all others, with 91.3% accuracy, 91.3% sensitivity, and 91.4% specificity using Random Forest and TSFEL Time features. This suggests that fetal physiological stress near delivery generates stronger predictive signals. However, this segment is also more prone to missing data, as there is more movement and hence signal loss in the last minutes before birth in the dataset, leading to greater reliance on imputation and potentially reduced robustness.

The first (0–20 mins) and mid (40–60 mins) segments achieved 89.5% and 86.8% accuracy, respectively, indicating that early-stage signals are already informative, while mid-labor may exhibit more stable but less discriminative patterns (9).

Table 1 highlights the performance of the top models across different 20 mins signal segments.

Table 1. Performance of top models on various 20 mins signal segments

Segment	Feature Set	Model	Acc	Sens	Spec	Std(Acc)
First 20 min (0–20)	TSFEL Time	Random Forest	89.5	90.7	88.2	0.0147
Mid-Segment (40–60)	TSFEL Time	Random Forest	86.8	87.0	86.6	0.0084
Last 20 min (relative to end)	TSFEL Time	Random Forest	**91.3**	**91.3**	**91.4**	0.0199

4.3 10 min signal

Reducing the window to 10 minutes maintained excellent classification results. The first 10-minute segment produced the best performance overall: 91.5% accuracy, 91.6% sensitivity, and 91.5% specificity. Other segments performed consistently well, except those beyond the 70-minute mark, where signal sparsity and data imbalance resulted in declining performance. Figure 3 demonstrates that the best-performing model remains Random Forest, mostly utilising TSFEL Time characteristics, except for the 40-50 mins section, when TSFresh performs marginally better. Table 2 shows classification Report of 0-10 min signal.

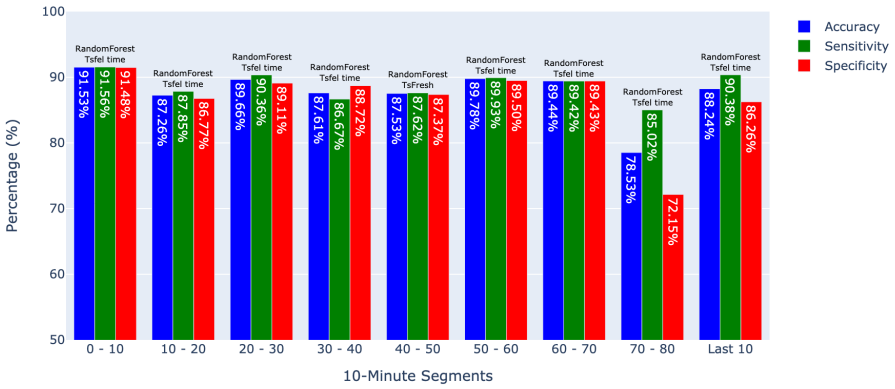

Fig. 3. Accuracy, Sensitivity and Specificity performance on all 10 mins segment experiments

These findings demonstrate that prolonged monitoring is not required to achieve accurate classification. A 10-minute window captures sufficient diagnos-

tic information, making it well-suited for real-time implementation with minimal computational overhead and latency. Notably, the last 10-minute segment, dynamically scaled to each signal's end, maintained 88.2% accuracy, reaffirming the diagnostic relevance of end-of-labor monitoring despite inherent data limitations.

Table 2. Classification Report of 0-10 min segment on best model based on accuracy

Metric	Best Model	Best Library	Value (%)	Std
Accuracy	RandomForest	Tsfel time	91.53	0.0286
Sensitivity	RandomForest	Tsfel time	91.56	0.0313
Specificity	RandomForest	Tsfel time	91.48	0.0278

4.4 5 min signal

Five-minute segments were evaluated to test the limits of minimal-data predictions. Performance varied significantly across segments, with accuracies ranging from 83.0% to 90.3%. The 55–60 minute and last 5-minute segments achieved the highest results, with the latter recording 88.9% accuracy, 91.7% sensitivity, and 86.5% specificity.

However, several segments exhibited notable performance drops due to signal loss, limited data points, and imputation artifacts. Despite Random Forest's strong stability, 5-minute segments are less reliable for clinical deployment, particularly in settings where continuous data integrity cannot be ensured. Figure 4 displays the best performing segments.

4.5 Deep Learning Baseline Comparison

To compare with deep learning, we trained a simple RNN on TSFEL time-domain features from the first 10-minute segment. The RNN achieved **67.5% accuracy**, with **65.6% precision**, **72.8% recall** for fetal distress—substantially lower than the **91.5% accuracy** of the **Random Forest** model. This suggests that, in feature-based settings with limited data, ensemble models offer better performance, interpretability, and efficiency than deep learning.

Key Insights and Implications

- A 10-minute CTG segment is sufficient to classify fetal distress with high accuracy, matching or exceeding performance from longer recordings.
- Random Forest, especially with TSFEL Time-domain features, consistently delivered top-tier performance across all segment lengths.
- While 20-minute segments performed slightly better in some cases, they come with greater computational cost and signal variability.

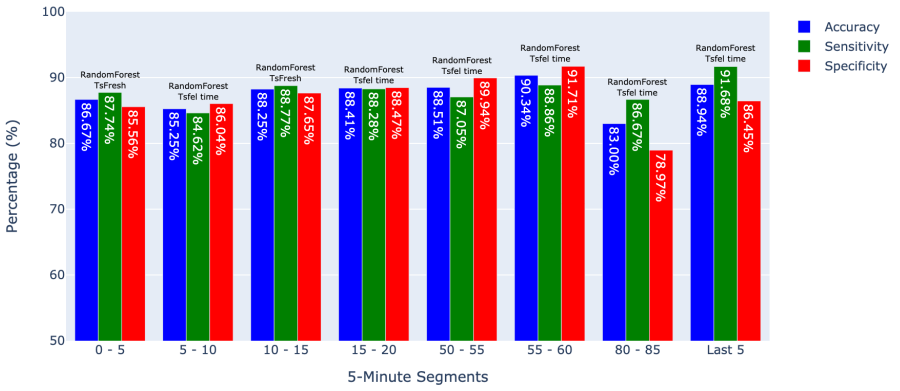

Fig. 4. Accuracy, Sensitivity and Specificity performance on all 5 mins segment performance

- 5-minute windows, though promising in select scenarios, lack robustness due to higher data sparsity and signal inconsistency.
- Random Forest models trained on 10-minute CTG segments achieved higher mean sensitivity (0.9044) than those trained on 5-minute segments (0.8667). **This difference was statistically significant (paired t-test $p = 0.0267$; Wilcoxon $p = 0.0488$)**, suggesting 10-minute segments better capture fetal distress.
- Importantly, although labor can last over 24 hours, our findings show that only 10 minutes of clean signal data is necessary for accurate prediction, supporting the feasibility of real-time decision-making in clinical settings.

5 Discussion and Conclusion

This study systematically explored the impact of CTG signal segment length on the classification performance of fetal heart rate (FHR) and uterine contraction (UC) signals. Contrary to expectations that longer recordings or later-stage segments where fetal activity typically increases would yield better performance, our findings show that classification accuracy remains relatively stable across different segment lengths, with only 3% variation. This challenges prior assumptions (12) and demonstrates that early-stage monitoring can be highly informative.

Among all classifiers tested, Random Forest consistently outperformed others, achieving the highest accuracy, sensitivity, and specificity, particularly on 10-minute segments. Its robustness to noise, imputation, and non-linear patterns makes it well-suited for this task. Feature extraction libraries like TSFEL

(time-domain) and TSFresh also provided reliable representations, highlighting the importance of statistical and time-domain features in CTG signal analysis.

Crucially, a 10-minute window achieved 91.5% accuracy—on par or better than the full 90-minute signal (88.9%) and 20-minute segments (91.3%). The last 10-minute segment, often presumed to be more informative, performed worse (88.2%), likely due to increased signal loss and imputation artifacts. This suggests that longer monitoring does not inherently improve predictive power and that early, cleaner segments may be more diagnostically valuable.

These findings have significant implications: although labor can last over 24 hours, only 10 minutes of signal data are sufficient to provide a reliable prediction of fetal distress. This enables rapid decision-making, reduced computational cost, and aligns with clinical demands for timely intervention.

However, this study is not without limitations:

- The dataset used was relatively small, with limited demographic and clinical variability.
- For the findings to be clinically validated and generalizable, larger and more diverse datasets are essential ideally incorporating a broader set of maternal-fetal variables, ethnic backgrounds, and clinical contexts.
- Currently the maximum signal length was 90 mins yet labour can last much longer than that and fetal distress could therefore occur much earlier than 90 mins. Longer segments are needed.
- Future work should also explore feature fusion strategies and integrate explainable AI to support clinician trust and adoption.

In conclusion, this study provides a strong foundation for building efficient, real time AI driven fetal monitoring tools, demonstrating that accurate predictions are possible with minimal data paving the way for scalable and practical solutions in maternal healthcare.

Declarations

Ethics Approval and Consent to Participate Not applicable.

Consent for Publication All authors have consented.

Availability of Data and Materials The dataset used in this study is publicly available from the **CTU-UHB Intrapartum Cardiotocography Database**, which can be accessed at **https://physionet.org/content/ctu-uhb-intrapartum-cardiotocography-database/**.

Competing Interests The corresponding author declares that there is **no conflict of interest** on the part of any of the authors, including themselves.

Funding This research was supported by **Innovate UK Grant Nr 10065844**.

Bibliography

[1] Aeberhard, J.L., et al.: Introducing artificial intelligence in interpretation of foetal cardiotocography: Medical dataset curation and preliminary coding—an interdisciplinary project. Scientific Reports **12**, 17632 (2022)

[2] Arun, K., Phaneesh, S., Reddy, S.P., Sreeman, S., Ghildiyal, Y.: An information system on fetal health classification based on cnn and hybrid-cnn with dimensionality reduction. In: E3S Web of Conferences. vol. 430, p. 01029. EDP Sciences (2023)

[3] Barandas, M., Ribeiro, R.P., et al.: Tsfel: Time series feature extraction library. SoftwareX **11**, 100456 (2020)

[4] Campos, D., Bernardes, J.: Cardiotocography. UCI Machine Learning Repository (2000), DOI: https://doi.org/10.24432/C51S4N

[5] Ayres-de Campos, D., Bernardes, J.: Fetal heart rate variability analysis with deep learning during acidosis: A systematic review. European Journal of Obstetrics & Gynecology and Reproductive Biology **244**, 45–51 (2020)

[6] Cesarelli, M., Romano, M., Bifulco, P., Fedele, F., Bracale, M.: An algorithm for the recovery of fetal heart rate series from ctg data. Computers in biology and medicine **37**(5), 663–669 (2007)

[7] Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research **16**, 321–357 (2002)

[8] Christ, M., Braun, N., Neuffer, J., Kempa-Liehr, A.W.: Time series feature extraction on the fly (tsfresh): A python package. Neurocomputing **307**, 72–77 (2018)

[9] Chudacek, V., Spilka, J., Jank, P., Hruban, K., Burgr, J., Lhotska, L., Huptych, M.: Automatic evaluation of intrapartum fetal heart rate recordings: a comprehensive analysis of useful features. Physiological Measurement **35**(7), 1311 (2014)

[10] Chudáček, V., Spilka, J., Burša, M., Jank, P., Hruban, L., Huptych, M., Lhotská, L.: Open access intrapartum ctg database. BMC pregnancy and childbirth **14**, 1–12 (2014)

[11] Cömert, Z., Kocamaz, A.F.: Evaluation of fetal distress diagnosis during delivery stages based on linear and nonlinear features of fetal heart rate for neural network community. Int. J. Comput. Appl **156**(4), 26–31 (2016)

[12] Cömert, Z., Kocamaz, A.F., Subha, V.: Prognostic model based on image-based time-frequency features and genetic algorithm for fetal hypoxia assessment. Computers in biology and medicine **99**, 85–97 (2018)

[13] Comert, Z., Kocamaz, U.: A study of artificial neural network training algorithms for classification of cardiotocography signals. Bitlis Eren University Journal of Science and Technology **8**(1), 37–44 (2018)

[14] Daydulo, Y.D., Thamineni, B.L., Dasari, H.K., Aboye, G.T.: Deep learning based fetal distress detection from time frequency representation of car-

diotocogram signal using morse wavelet: research study. BMC Medical Informatics and Decision Making **22**(1), 329 (2022)
[15] Dempster, A., Petitjean, F., Webb, G.I.: Rocket: Exceptionally fast and accurate time series classification using random convolutional kernels. Data Mining and Knowledge Discovery **34**, 1454–1495 (2020)
[16] Georgieva, A., Payne, S.J., Moulden, M., Redman, C.W.: Phase-rectified signal averaging for intrapartum electronic fetal heart rate monitoring is related to acidosis at birth. BJOG: An International Journal of Obstetrics & Gynaecology **120**(4), 451–459 (2013)
[17] Goldberger, A.L., Amaral, L.A., Glass, L., Hausdorff, J.M., Ivanov, P.C., Mark, R.G., Mietus, J.E., Moody, G.B., Peng, C.K., Stanley, H.E.: Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. circulation **101**(23), e215–e220 (2000)
[18] Kuzu, A., Santur, Y.: Early diagnosis and classification of fetal health status from a fetal cardiotocography dataset using ensemble learning. Diagnostics **13**(15), 2471 (2023)
[19] Löning, M., Bagnall, A., et al.: Sktime: A unified interface for machine learning with time series. In: Workshop on Systems for ML at NeurIPS 2019. pp. 1–3 (2019)
[20] Lubba, C.H., Sethi, S.S., et al.: Catch22: Canonical time-series characteristics. Data Mining and Knowledge Discovery **33**, 1821–1852 (2019)
[21] Mendis, L., et al.: Rapid detection of fetal compromise using input length invariant deep learning on fetal heart rate signals. Frontiers in Physiology **13**, 1002854 (2022)
[22] Petrozziello, A., Jordanov, I., Papageorghiou, T., Redman, W., Georgieva, A.: Deep learning for continuous electronic fetal monitoring in labor. In: 2019 40th annual international conference of the IEEE engineering in medicine and biology society (EMBC). pp. 5866–5869. IEEE (2019)
[23] Romano, M., Faiella, G., Bifulco, P., D'Addio, G., Clemente, F., Cesarelli, M.: Outliers detection and processing in ctg monitoring. In: XIII Mediterranean Conference on Medical and Biological Engineering and Computing 2013: MEDICON 2013, 25-28 September 2013, Seville, Spain. pp. 651–654. Springer (2014)
[24] Salini, Y., Mohanty, S., Ramesh, J., Yang, M., Chalapathi, M.: Cardiotocography data analysis for fetal health classification using machine learning models. IEEE Access (2024)
[25] World Health Organization: Child mortality and causes of death (2025), https://www.who.int/data/gho/data/themes/topics/topic-details/GHO/child-mortality-and-causes-of-death, accessed: 2025-04-10
[26] Zhao, Z., Zhang, Y., Comert, Z., Deng, Y.: Computer-aided diagnosis system of fetal hypoxia incorporating recurrence plot with convolutional neural network. Frontiers in physiology **10**, 255 (2019)
[27] Zou, C., Zhang, Y., Yuan, Z.: An intelligent adverse delivery outcomes prediction model based on the fusion of multiple obstetric clinical data. Computer Methods in Biomechanics and Biomedical Engineering pp. 1–15 (2023)