CTG Triage: Interpretable Machine Learning for Fetal Health Classification

Abstract

We built an interpretable, reliable machine-learning pipeline to classify fetal cardiotocography (CTG) recordings into *Normal*, *Suspect*, and *Pathologic* categories using the public UCI CTG dataset (Ayres-de-Campos et al., 2000; UCI ML Repository, 2019). Rigorous preprocessing, medically-aware feature engineering, and stratified 5-fold cross-validation were applied. A calibrated logistic-regression model achieved high macro F1 (≈ 0.89) and balanced accuracy (≈ 0.90) with low Expected Calibration Error (ECE ≈ 0.04), providing transparent and robust triage performance.

Methods

Data cleaning: Removed duplicates, coerced numeric types, imputed medians, standardized column names, and dropped empty columns (filename, segfile).

Feature engineering: Derived seven interpretable indicators such as tachycardia > 160 bpm, bradycardia < 110 bpm, variability ratio = ALTV / (MSTV + ε), decelerations-per-contraction, and instability proxy = WIDTH / (MSTV + 1), following clinical heuristics (Ayres-de-Campos et al., 2000).

Pipeline: All transformations wrapped in a scikit-learn (Pedregosa et al., 2011) Pipeline using CTGEngineer, SimpleImputer, and StandardScaler.

Models: Dummy → Logistic Regression → SVM (RBF) → Random Forest; evaluated via stratified 5-fold CV.

Metrics: Macro F1 and Balanced Accuracy, plus calibration and robustness tests (Niculescu-Mizil & Caruana, 2005).

Results

Model	Macro F1 ± SD	Balanced Acc ± SD
Dummy (Stratified)	0.335 ± 0.012	0.340 ± 0.011
Logistic Regression (multinomial)	**0.887 ± 0.018**	**0.904 ± 0.015**
SVM (RBF, balanced)	0.874 ± 0.019	0.891 ± 0.017
Random Forest (400 trees)	0.869 ± 0.021	0.885 ± 0.018

The logistic model offered the best balance of accuracy and interpretability.
Calibration: Macro ECE ≈ 0.038.
Top features: Baseline FHR, deceleration burden (DL + DS + DP), variability ratio (ALTV / MSTV), and accelerations—consistent with clinical guidelines (Spilka et al., 2014).
Robustness: 1 % Gaussian noise changed macro F1 by −0.006.

Discussion

Linear models with domain-aware features can rival complex ensembles while remaining interpretable. The inclusion of clinically grounded ratios (variability, deceleration-to-contraction) helped the model align with real physiological meaning. Calibration reduced overconfidence, ensuring reliable probabilities for potential clinical flagging. Future extensions could analyze temporal CTG traces with deep RNN/CNN architectures, integrate maternal metadata, or deploy the calibrated model in a Streamlit dashboard for bedside decision support.

Conclusion

The CTG Triage pipeline demonstrates that transparent, statistically robust models can provide clinically useful fetal-health predictions without sacrificing interpretability. The workflow—cleaning → feature engineering → calibrated evaluation → robustness—forms a reproducible baseline for responsible healthcare AI.

References

1. Ayres-de-Campos, D., Bernardes, J., Garrido, A., Marques-de-Sá, J., & Pereira-Leite, L. (2000). *Cardiotocographic evaluation of fetal state by means of artificial neural networks.* **Biological Cybernetics, 82(2)**, 143–152.

2. Spilka, J., Chudáček, V., Janku, P., & Huptych, M. (2014). *Automatic evaluation of intrapartum fetal heart-rate recordings by machine learning.* **Physiological Measurement, 35(7)**, 1319–1334.

3. UCI Machine Learning Repository (2019). *Cardiotocography Data Set.* https://archive.ics.uci.edu/ml/datasets/Cardiotocography

4. Pedregosa, F. et al. (2011). *Scikit-learn: Machine Learning in Python.* **Journal of Machine Learning Research, 12**, 2825–2830.

5. Niculescu-Mizil, A., & Caruana, R. (2005). *Predicting good probabilities with supervised learning.* **ICML 2005**, 625–632.