Datathon Academic Report

TM-37 : Ricard Josse Meyer, Charles Lukas Chairos Yo, Poon Wei Lok

Model Design and Rationale

Cardiotocography (CTG) interpretation is highly subjective, with only ~60% inter-observer agreement [1], leading to inconsistent diagnoses in high-risk pregnancies. To address this, we developed a "glass-box" hybrid model that combines clinical feature engineering with a high-performance LightGBM classifier. Our approach leverages domain knowledge to create new, interpretable features (e.g., clinical risk score, weighted deceleration score, decel/accel ratio) from the 24 raw CTG parameters, reflecting the way clinicians assess fetal well-being. These features are then used as inputs to a LightGBM model, which captures complex, non-linear relationships while remaining fully explainable via SHAP analysis. This balances the need for a low false negative rate (high pathological recall) and interpretability, directly addressing clinical priorities.

Data Preprocessing Pipeline

The CTG dataset was first loaded and rigorously cleaned to ensure data integrity: duplicate records were removed, any traces with missing target labels were dropped, and physiologically implausible outliers—such as baseline heart rates below 50 bpm or above 200 bpm and negative acceleration counts—were excluded. We then applied six domain-informed feature transformations to capture clinically meaningful patterns. Short-term and long-term variability ratios (STV_ratio, LTV_ratio) quantify the relationship between rapid and gradual heart-rate fluctuations. Deceleration metrics aggregate all deceleration events (total_decelerations), weight severe types more heavily (weighted_deceleration_score), flag the presence of severe decelerations (has_severe_decelerations), and compare decelerations to accelerations (decel_accel_ratio). Binary indicators for bradycardia and tachycardia (LB_bradycardia, LB_tachycardia) capture critical baseline heart-rate extremes. Finally, a composite clinical risk score synthesizes these signals—abnormal variability, severe decelerations, bradycardia/tachycardia, and absent accelerations—into a single, interpretable index. After assembling 33 total features, we performed an 80/20 stratified train–test split and used Borderline-SMOTE on the training set to address class imbalance [2] and safeguard against false negatives. This preprocessing framework yields a robust, clinically grounded feature set ready for model training.

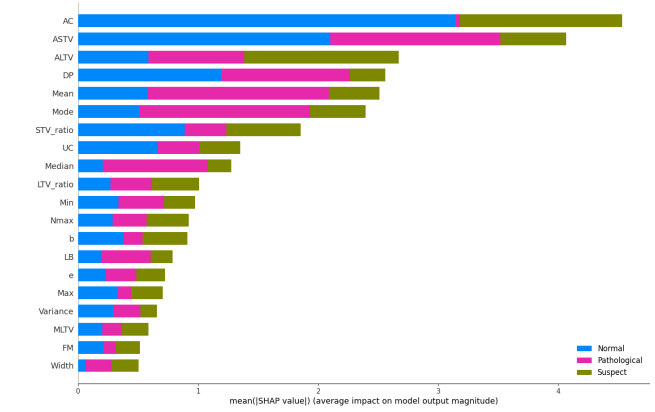
Results and Performance

Table 1: Classification Performance Metrics

Model	Accuracy	Macro-F1	Pathological Recall	Key Takeaway
Hybrid Glass-Box (LGBM)	**0.9693**	**0.9452**	**0.9714**	**Final solution: peak performance + explainable**
Superblend	**0.9693**	**0.9450**	**0.9714**	**Highest accuracy, but black-box**
Feature-Engineering RF	**0.9481**	**0.9149**	**0.9714**	**Most interpretable, strong recall**

	Predicted: Normal	Predicted: Suspect	Predicted: Pathological
True: Normal	329	1	0
True: Suspect	10	48	1
True: Pathological	0	1	34

Figure 1: Confusion Matrix. Performance on the test set (n=424).



Our final Hybrid Glass-Box (LGBM) model achieved a test set accuracy of 96.9% and a Macro-F1 score of 0.945. Critically for clinical safety, it attained a Pathological Recall of 0.971, correctly identifying 34 out of 35 high-risk cases. This performance matched that of a complex "black-box" superblend while retaining full interpretability. SHAP analysis confirms the model's decisions are driven by clinically relevant factors; the most impactful features include raw signals like AC (accelerations) and ASTV (abnormal short-term variability), alongside our engineered STV_ratio. This transparent reliance on understandable metrics validates our hybrid approach and enhances clinical trust.

Conclusion

Our work successfully demonstrates that a hybrid "glass-box" model can achieve the performance of a complex black-box ensemble without sacrificing the interpretability essential for clinical trust. By integrating domain knowledge directly into the modeling pipeline, this project provides a robust and transparent framework for AI-driven decision support in critical care settings.

References

[1] Bernardes, J., Moura, C., Marques-de-Sá, J., & Pereira-Leite, L. (1998). The roles of cardiotocography and fetal blood sampling in the management of intrapartum fetal distress. Current Opinion in Obstetrics and Gynecology, 10(2), 125-132.

[2] Han, H., Wang, W. Y., & Mao, B. H. (2005). Borderline-SMOTE: A New Over-sampling Method in Imbalanced Data Sets Learning. In Advances in Intelligent Computing (pp. 878-887). Springer Berlin Heidelberg.