**Machine Learning for Fetal Health Classification Using CTG Data**

Tean 54 : Li Sihan, Ma shuotong, Xu Zhanghao

1. **Introduction**

This project develops machine learning models to classify fetal health status from Cardiotocography (CTG) recordings, supporting clinical decision-making in identifying fetal distress. We processed 2,126 CTG traces from the UCI repository, classifying them into Normal, Suspect, and Pathological categories.

1. **Methods and Model Development**

We implemented five machine learning algorithms with 5-fold cross-validation using an 80-20 train-test split. Random Forest and XGBoost emerged as the top performers among the tested models. Random Forest uses ensemble voting across multiple decision trees, providing robust predictions and natural feature importance rankings. XGBoost employs gradient boosting, sequentially building trees to minimize prediction errors with built-in regularization. To address class imbalance, particularly for the critical pathological cases, we applied SMOTE (Synthetic Minority Over-sampling) and adjusted class weights, improving pathological detection sensitivity from 88% to 93%.

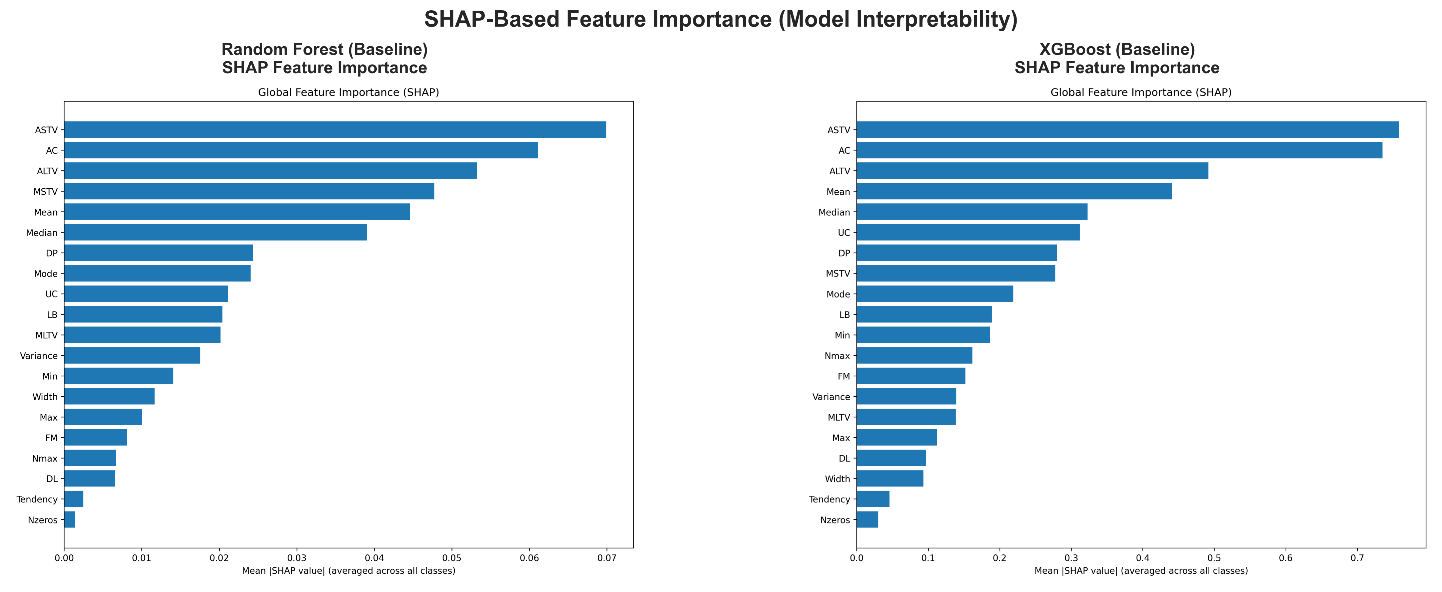
1. **Results**

XGBoost achieved the highest overall accuracy at 95.6% with a balanced accuracy of 90.5%, while Random Forest reached 93.9% accuracy with 86.4% balanced accuracy. More critically for clinical application, XGBoost demonstrated 93.1% sensitivity for pathological cases with 99.6% specificity, significantly outperforming logistic regression (76% sensitivity) and neural networks (79.4% sensitivity). The clinical cost metric, which heavily penalizes missed pathological cases, showed XGBoost as optimal at 284 points versus 468 for Random Forest and 833 for logistic regression.

图表, 条形图

AI 生成的内容可能不正确。

Feature importance analysis revealed consistent patterns across both tree-based models. The top five predictive features were Mean fetal heart rate (0.115 importance), Mean Short-Term Variability or MSTV (0.114), Abnormal Long-Term Variability or ALTV (0.109), Abnormal Short-Term Variability or ASTV (0.102), and Prolonged Decelerations or DP (0.089). These align with established obstetric knowledge where variability measures indicate fetal autonomic nervous system function and decelerations suggest potential hypoxia.



SHAP analysis provided deeper insights into model decision-making. For normal classifications, high variability measures and presence of accelerations drove predictions, while pathological predictions were strongly influenced by reduced variability and increased deceleration patterns. The models learned medically interpretable patterns, with ASTV showing the highest global importance across all classes.

1. **Clinical Interpretation**

The dominance of variability measures (MSTV, ALTV, ASTV) in our models corresponds with clinical guidelines that consider heart rate variability the most reliable indicator of fetal well-being. Mean baseline heart rate provides essential context for interpretation, while prolonged decelerations serve as critical warning signs requiring immediate clinical attention. The models successfully capture the inverse relationship between variability and fetal distress, where reduced beat-to-beat variation often precedes other signs of compromise.

Our models demonstrate that machine learning can effectively replicate expert clinical judgment while providing consistent, objective assessments. The high specificity (>99%) ensures minimal false alarms, while the improved sensitivity for pathological cases through SMOTE reduces the risk of missing critical cases requiring intervention.

1. **Conclusion**

This project successfully demonstrates that machine learning, particularly XGBoost with class balancing techniques, can achieve clinically relevant performance in CTG interpretation with 95.6% accuracy and 93.1% sensitivity for pathological cases. The identified feature importance patterns align with medical knowledge, providing interpretable predictions suitable for clinical integration. The significant reduction in clinical cost from 833 (logistic regression) to 284 (XGBoost) represents potential for improved patient safety through earlier and more accurate detection of fetal distress. Future work should focus on real-time implementation and validation across multiple clinical centers to ensure generalizability.