Academic Report – Fetal Health Classification using Cardiotocogram (CTG) Data

Bhatia Rohit, Kapoor Arunaksh, Mohapatra Arin

# Data Overview

## The dataset used consists of 2,126 cardiotocogram (CTG) recordings, each containing 21 diagnostic features describing fetal heart rate (FHR), uterine contractions (UC), and various derived statistical measures. Each CTG was classified by expert obstetricians into one of three fetal state categories — Normal (N), Suspect (S), or Pathological (P) — and into one of ten morphological classes. During the initial inspection, the dataset showed no missing values and only a few duplicate records, which were removed. Therefore, minimal cleaning was required before analysis.

# Exploratory Data Analysis (EDA)

## Exploratory Data Analysis was performed to understand the data distributions, variability, and potential relationships between features. From the histograms (Figure 1), it was observed that several features — such as AC, FM, UC, DL, DS, and DP — were highly skewed towards zero, indicating that accelerations, fetal movements, and decelerations occurred infrequently. In contrast, continuous-valued features like LB (FHR baseline), ASTV, MLTV, Width, Min, Max, Mean, Median, and Mode exhibited near bell-shaped or multimodal distributions, suggesting natural variation across fetal states. Features such as Variance and Tendency showed long tails, indicating the presence of potential outliers. The violin plots (Figure 2) confirmed the wide variation in scales across features. Some variables, like LB and Mean, were concentrated in higher ranges (around 120–150 bpm), while others such as AC and DS had values close to zero. This variation suggested the need for standardization so that each feature contributes equally to the model. Based on these insights, we proceeded with outlier removal and standardization (z-score scaling) rather than normalization. Standardization was preferred because it centers data around zero with unit variance, preserving the shape of the distributions while accommodating features measured in different units.

# Correlation and Feature Selection

## A correlation heatmap (figure 3) was generated to study relationships between variables. However, the results showed that most features had weak linear correlations, implying that the dataset contains primarily non-linear dependencies.Higher correlations were observed among LB (baseline FHR), Mean, Median, and Mode, which is expected as they represent similar measures of central tendency. A mild negative correlation was found between NSP (fetal state) and AC (accelerations per second), suggesting that fewer accelerations tend to occur in abnormal fetal conditions. Overall, the correlation analysis offered limited insights, reinforcing the need for models capable of capturing non-linear interactions.

# Handling Class Imbalance

The dataset was imbalanced, with fewer Pathological and Suspect samples than Normal. To address this, we applied Boderline Smote to the training set. Unlike standard SMOTE, it generates synthetic samples only near the decision boundary, where minority instances are at highest misclassification risk, avoiding oversampling “safe” interior points.

In the CTG feature space, many Suspect and Pathological points (red and yellow) lie near Normal clusters (blue), making them borderline. Oversampling these regions strengthens minority representation at the edges, improving decision boundaries, especially in overlapping subspaces like ALTV–MSTV–ASTV. This enhances classifier performance (recall, F1-score) without bloating well-separated areas. Borderline-SMOTE also preserves data characteristics in high-dimensional, correlated features, maintaining local structure while being computationally efficient for moderate datasets like CTG. It was applied only to the training data to prevent test set leakage.

# Rule-Based Baseline

Before applying machine learning, we built a rule-based system inspired by how clinicians interpret CTGs. Five agents were designed, each analyzing different physiological aspects:

* Baseline Agent: Classifies baseline FHR (LB) as Normal (110–160 bpm), Suspicious (100–109 or 161–180 bpm), or Pathological (<100 or >180 bpm).
* Variability Agent: Evaluates short-term and long-term variability using MSTV and ASTV.
* Accelerations Agent: Interprets accelerations (AC) per 20 minutes.
* Decelerations Agent: Analyzes light (DL), severe (DS), and prolonged (DP) decelerations.
* Sinusoidal Agent: Detects sinusoidal patterns using variance, MSTV, and lack of accelerations.

Each agent provided a mini-diagnosis with both a label and a brief explanation. Two aggregation methods were tested:

Strict Aggregator: If any feature is Pathological → final label = Pathological. If ≥2 Suspicious → Pathological; if exactly 1 Suspicious → Suspicious; otherwise → Normal.  
Probability Aggregator: Each diagnosis contributed weighted scores (+1 Normal, +1 Suspicious, +2 Pathological), normalized into probabilities.

Results:  
The strict aggregator predicted almost all samples as Pathological (≈8% accuracy) due to its rigidity. The probability-based method improved performance to about 62% accuracy, achieving high recall for the Pathological class but low precision due to many false alarms. The rule-based system failed primarily because CTG signals often contain borderline values that made deterministic rules escalate too easily.

# Machine Learning Models

Given the limitations of rule-based logic, we transitioned to data-driven machine learning models capable of capturing non-linear patterns. We experimented with three models: Extra Trees, XGBoost, and SVM with PCA.

* Extra Trees Classifier:  
   Chosen as our primary model due to its robustness and ability to handle complex, non-linear data. Compared to Random Forest, Extra Trees introduces additional randomness when choosing split thresholds, which helps reduce overfitting and improves generalization. It is also fast to train and provides interpretable feature importance scores.
* XGBoost:  
   A gradient boosting algorithm that builds trees sequentially, correcting previous errors. It performs well on tabular data and can capture subtle relationships, though it requires more fine-tuning.
* SVM with PCA:  
   PCA (Principal Component Analysis) was applied before training the SVM to reduce dimensionality and remove correlations between features. This step helps SVMs, which are sensitive to feature scaling, perform better in a lower-dimensional space.

Why PCA was not used for Extra Trees and XGBoost:  
Tree-based models like Extra Trees and XGBoost do not require feature scaling or decorrelation since they split data based on thresholds rather than distance. Applying PCA would make their outputs harder to interpret and could remove meaningful feature information. Therefore, PCA was used only for the SVM model.

# Final Model and SHAP Analysis

## After training and cross-validation, Extra Trees achieved the best overall performance in terms of f1 scores, recall, and interpretability. It also demonstrated strong stability across folds and resistance to overfitting, making it a reliable choice for medical applications. While XGBoost showed competitive accuracy, its feature importance patterns were less consistent, and the interpretability of its outputs was more complex compared to Extra Trees. To further validate the model, we applied SHAP (SHapley Additive Explanations) to both classifiers. SHAP values confirmed that clinically relevant features such as Baseline FHR (LB), ASTV, MSTV, MLTV, and histogram statistics (Mean, Median, Mode) were among the most influential in predicting fetal state. The Extra Trees SHAP beeswarm plot(Figure 6) showed stable and balanced contributions from these features, closely matching clinical reasoning. In contrast, the XGBoost SHAP plot(Figure 7) revealed more dispersed and extreme SHAP values, suggesting higher sensitivity but less interpretability. In summary, Extra Trees(table 1) was selected as the final model because it offered the best balance of predictive accuracy with best precision, robustness, and clinical transparency, with SHAP analysis providing additional confidence that the model’s decisions aligned with established medical knowledge.

Tables

Table 1

Model Comparison

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Macro Precision** | **Macro Recall** | **Macro F1** | **Weighted Precision** | **Weighted Recall** |
| **SVM + PCA** | 0.9310 | 0.8364 | 0.8659 | 0.9384 | 0.9435 |
| **Extra Trees** | 0.8541 | 0.8714 | 0.8413 | 0.9610 | 0.9609 |
| **XGBoost** | 0.8270 | 0.9048 | 0.8301 | 0.9786 | 0.9739 |

**Figure 1**

Histogram Plots

A collage of graphs

AI-generated content may be incorrect.

**Figure 2**

Feature Distribution Using Violin Plots

A graph of different colored shapes

AI-generated content may be incorrect.

**Figure 3**

Correlation Matrix

A screenshot of a diagram

AI-generated content may be incorrect.

**Figure 4**

Mutual Information Scores

A graph with red and white stripes

AI-generated content may be incorrect.

**Figure 5**

Feature Importance Using ExtraTrees

A graph showing a number of different colored bars

AI-generated content may be incorrect.

**Figure 6**

SHAP Plot for ExtraTrees

A group of blue and pink dots

AI-generated content may be incorrect.

**Figure 7**

SHAP Plot for XGBOOST

A graph of different colors

AI-generated content may be incorrect.