This report outlines our approach to classifying fetal cardiotocography (CTG) data into Normal, Suspect, and Pathological (NSP) classes using interpretable machine learning. Our objective was to build a medically reliable and balanced model that supports clinical decision-making, particularly in identifying high-risk pregnancies.

Data Cleaning. Empty rows, columns, and duplicates were removed. Derived label columns (*CLASS*, *A–E*, *AD*, *DE*, *FS*, *SUSP*) were excluded to prevent target leakage, and rows with missing targets were dropped. Each feature was checked for physiologically impossible or sensor-error values; none were found, so no further corrections were made to preserve data authenticity.

Data Engineering. All 23 features were initially included in a LightGBM model, and **mean absolute SHAP values** were computed to rank global importance. A **K-feature selection test** was conducted, progressively trimming features until the optimal balanced accuracy was reached with the top eight:
Final Feature Set (8): ASTV, DP, ALTV, Median, Variance, AC, UC, Mode.
This reduced feature set improved generalisation and interpretability. Class imbalance (NSP_1: 78%, NSP_2: 14%, NSP_3: 8%) was addressed using **SMOTE** oversampling within each cross-validation fold.

Model Selection and Rationale. We compared Logistic Regression, Decision Tree, Random Forest, and MLP models, eventually selecting **LightGBM** for its superior balanced accuracy, computational efficiency, and explainability. Hyperparameters were tuned using **Optuna** with 5-fold **Stratified Cross-Validation**, optimising for balanced accuracy. We achieved Best CV balanced accuracy of 91.64%.

Metric Literacy. Metrics such as **Balanced Accuracy**, **Macro F1-score**, and **Per-Class Recall** were critical to ensure fairness and medical safety in an imbalanced dataset. Balanced Accuracy equally weights performance across all NSP classes, preventing the model from favouring the dominant Normal class. Macro F1 balances precision and recall, rewarding consistent performance across classes. Per-class recall is especially important clinically—low recall for Suspect or Pathological cases represents **false negatives**, where an unhealthy fetus may be misclassified as normal. In medical settings, such errors could lead to delayed intervention and potential loss of life. Hence, the evaluation prioritised sensitivity to minority classes, serving as a **clinical guardrail** against life-critical misdiagnoses.

Model Interpretability. Global interpretability via SHAP identified *ASTV*, *DP*, and *ALTV* as strong indicators of fetal distress, while *AC* and *Mode* stabilised normal predictions. For individual cases, **SHAP waterfall plots** were used to visualise each feature's contribution to a specific prediction. This patient-level explainability allows doctors to understand the reasoning behind the model's output, cross-check it with clinical judgment, and build trust in AI-assisted diagnosis.

Conclusion. Through systematic cleaning, SHAP-guided feature selection, SMOTE balancing, and interpretable LightGBM modelling, our pipeline achieved high balanced accuracy while maintaining clinical transparency.