Hypoxia Prediction from Prenatal Doppler and Postnatal ABG in IUGR Using Explainable AI

November 27, 2025

Abstract

Fetal hypoxia and intrauterine growth restriction (IUGR) are major contributors to adverse perinatal outcomes. Doppler ultrasound velocimetry—specifically the Pulsatility Index (PI) of the Umbilical Artery (PI_UA), Middle Cerebral Artery (PI_MCA), and their cerebroplacental ratio (PI_MCA/PI_UA)—remains the primary non-invasive approach for screening fetuses at risk. However, traditional threshold-based interpretation may fail to capture complex multivariate interactions or subtle nonlinear patterns associated with hypoxic physiology.

This project develops a complete, fully explainable machine learning pipeline to predict fetal hypoxia using only non-invasive Doppler indices and maternal demographic variables. A dataset of 400 pregnancies (200 IUGR and 200 Normal) was analyzed. Extensive preprocessing, engineered feature creation, and robust quality checks were applied (Step 1–2). Four supervised classifiers—Logistic Regression, Random Forest, XGBoost, and a Multi-Layer Perceptron (MLP)—were trained using 5-fold stratified cross-validation on the full dataset (Steps 3–5). In addition, a more advanced *gestational-age–stratified* modeling scheme was implemented for Random Forest and XGBoost, in which the dataset is partitioned into balanced GA groups and separate calibrated models with group-specific decision thresholds are trained and then aggregated in a weighted fashion. Evaluation included accuracy, precision, sensitivity, specificity, ROC-AUC, and confusion matrices (Steps 4–7). Model explainability was ensured using logistic regression coefficients and SHAP global and local explanations (Step 6).

The MLP model demonstrated the best global performance: Accuracy = 0.9598 and AUC = 0.9900. Classical Doppler rules (e.g., PI_MCA/PI_UA ¡ 1.0) achieved 90% accuracy, confirming their clinical utility but underperforming ML methods. Logistic Regression provided highly interpretable coefficients, while SHAP revealed that PI_MCA/PI_UA, PI_UA, and PI_MCA were the most influential predictors. The GA-stratified Random Forest and XGBoost pipelines confirmed that model

discrimination remains stable across gestational age strata. Confusion matrices further highlighted improvements of ML over rule-based decisions.

Overall, the results demonstrate that combining Doppler velocimetry with machine learning offers powerful, transparent tools for fetal hypoxia prediction, potentially augmenting clinical decision-making in prenatal care.

1 Introduction

Fetal hypoxia, typically resulting from placental insufficiency or umbilical cord compromise, is a leading cause of intrapartum complications, neonatal morbidity, and long-term neurodevelopmental deficits. In clinical obstetrics, fetal monitoring relies heavily on Doppler ultrasound velocimetry. Increased resistance in the umbilical artery (PI_UA) suggests poor placental perfusion, whereas decreased middle cerebral artery resistance (PI_MCA) indicates a compensatory redistribution—"brain sparing"—in response to hypoxia. Consequently, the cerebroplacental ratio (CPR = PI_MCA/PI_UA) is widely used in routine fetal surveillance.

Despite its value, traditional Doppler interpretation has limitations:

- Thresholds such as PI_UA > 1.5 or PI_MCA/PI_UA < 1.0 oversimplify complex physiology.
- Multivariate interactions (e.g., combining indices, demographic factors) are not captured.
- Sensitivity and specificity may vary across gestational age ranges.
- Subjective clinician interpretation introduces variability.

Machine learning (ML) has emerged as a promising tool to overcome these limitations. By leveraging multivariate learning from large datasets, ML models can identify subtle patterns linked to hypoxic outcomes. However, challenges remain:

- Reliability and reproducibility of ML predictions.
- Avoiding overfitting on limited datasets.
- Providing explainability suitable for clinical adoption.
- Ensuring that only non-invasive predictors (Doppler + demographics) are used.

To address these challenges, this study develops a fully transparent ML pipeline trained exclusively on non-invasive Doppler measurements. The process includes strict data validation, engineered feature creation, multiple ML models, a gestational-age–stratified

variant for tree-based models, explainability using SHAP, and comparison with classic Doppler thresholds.

This document serves as a complete unofficial report describing the full methodology, experiments, and evaluation.

2 Literature Review

Machine learning has demonstrated value in obstetrics, with applications in:

- electronic fetal monitoring (CTG signal classification using SVMs, CNNs, LSTMs),
- prediction of preeclampsia (XGBoost, RF, logistic regression),
- ultrasound image analysis using deep learning,
- stillbirth risk prediction using ensemble learning.

Doppler velocimetry remains a cornerstone of fetal monitoring. PI_UA and PI_MCA correlate strongly with placental resistance and fetal cerebral redistribution. The CPR (PI_MCA/PI_UA) is one of the most reliable single indices for identifying fetal compromise.

However:

- rule-based cutoffs are sensitive to gestational age;
- combining indices manually is difficult;
- nonlinear relationships in the Doppler space are not captured by single thresholds.

Few studies have used ML on Doppler-only inputs. Even fewer incorporate explainability (SHAP, odds ratios) to satisfy clinical interpretability requirements, or explicitly handle gestational-age dependence through per-GA models. This project fills that gap.

3 Materials and Methods

3.1 Step 1: Data Loading and Quality Control

The dataset consists of 400 samples, each representing a unique pregnancy outcome labeled as either "Normal" or "IUGR." The following steps were applied:

- normalization of column names;
- enforcement of numerical datatypes for Gestational Age (GA), Maternal Age, and PI features;

- removal of duplicate rows;
- checking for missing values and coercing invalid values to NaN;
- rejecting samples containing unresolved NaN in critical features;
- standardization of labels to exactly "IUGR" or "Normal".

This strict cleaning ensures reliability in downstream ML steps.

3.2 Step 2: Feature Engineering

Beyond the raw Doppler and demographic features, four engineered features were created:

$$PI\_\text{Diff} = PI\_\text{MCA} - PI\_\text{UA},$$

$$PI\_\text{Product} = PI\_\text{MCA} \times PI\_\text{UA},$$

$$GA\_\text{MCA\_Interaction} = GA \times PI\_\text{MCA},$$

$$UA\_\text{Adjusted} = \frac{PI\_\text{UA}}{GA}.$$

These features encode nonlinear relationships and cross-terms that are often clinically relevant.

Table 1: Summary statistics of Doppler and derived features for Normal vs IUGR fetuses.

Statistic	Gestational age	Maternal age	PI_{MCA}	PI_{UA}	$\text{PI}_{\text{MCA/UA}}$	PI_{Diff}	$\text{PI}_{\text{MCA}}\times\text{PI}_{\text{UA}}$	$\text{GA}\cdot\text{PI}_{\text{MCA}}$	$\text{PI}_{\text{UA}}/\text{GA}$
N (Normal)	200	200	200	200	200	200	200	200	200
N (IUGR)	200	200	200	200	200	200	200	200	200
Mean (Normal)	37.320	32.880	1.036	1.689	1.738	-0.652	1.705	38.587	0.045
Mean (IUGR)	33.930	29.980	1.155	1.760	0.694	-0.605	1.967	39.962	0.053
SD (Normal)	1.275	3.757	0.198	0.231	0.588	0.429	0.103	7.029	0.006
SD (IUGR)	3.158	2.923	0.252	0.267	0.257	0.517	0.160	12.103	0.013
Median (Normal)	37.000	32.000	1.000	1.700	1.700	-0.700	1.710	39.000	0.044
Median (IUGR)	34.000	30.000	1.200	1.700	0.710	-0.500	2.040	40.800	0.050
IQR (Normal)	2.000	6.000	0.300	0.400	0.860	0.700	0.180	9.225	0.008
IQR (IUGR)	5.000	4.500	0.300	0.300	0.280	0.600	0.180	15.800	0.017
Min (Normal)	35.000	26.000	0.700	1.300	0.930	-1.400	1.470	25.200	0.033
Min (IUGR)	28.000	25.000	0.700	1.200	0.300	-1.600	1.540	19.600	0.030
Max (Normal)	40.000	40.000	1.400	2.100	3.000	0.100	1.820	54.600	0.058
Max (IUGR)	40.000	35.000	1.700	2.300	1.420	0.500	2.250	68.000	0.082
Missing (Total)	0	0	0	0	0	0	0	0	0
p (t–test)	0.0000	0.0000	0.0000	0.0047	0.0000	0.3180	0.0000	0.1657	0.0000
p (Mann–Whitney)	0.0000	0.0000	0.0000	0.0070	0.0000	0.2831	0.0000	0.2253	0.0000
Cohen's d	-1.408	-0.862	0.523	0.284	-2.301	0.100	1.943	0.139	0.773

As an initial step, we performed feature engineering to generate a set of nonlinear Doppler-derived indices and composite ratios. Because the dataset contained only 400 samples, expanding the feature space too aggressively would risk overfitting, particularly for higher-capacity models such as XGBoost and the MLP. For this reason, we combined clinically meaningful raw Doppler variables—Gestational Age, Maternal Age, PI_{MCA}, PI_{UA}, and $PI_{\mathrm{MCA}}/PI_{\mathrm{UA}}$—with a small number of engineered features (e.g., PI_{Diff}, PI_{Product}, $GA{\cdot}PI_{\mathrm{MCA}}$, and PI_{UA}/GA). These variables demonstrated strong discriminative power between Normal and IUGR fetuses (Table 1), with statistically significant group differences. Importantly, the final feature set captured the essential Doppler-based pathophysiological patterns while avoiding unnecessary noise or sparsity.

3.3 Step 3: Train/Test Split and Label Encoding

The full dataset was shuffled and split into 80% training and 20% testing sets. All labels were mapped as:

$$\text{Normal} \to 0, \quad \text{IUGR} \to 1.$$

This enables:

- probabilistic interpretation,
- ROC-AUC computation,
- confusion matrix generation.

Additionally, 5-fold stratified cross-validation ensures class balance within each fold.

3.4 Step 4: Machine Learning Models (Global)

Four baseline models were trained using consistent pipelines on the full feature set without GA stratification.

3.4.1 Logistic Regression

Logistic regression was implemented with:

- L2 regularization,
- `class_weight = balanced`,
- solver = `lbfgs`,
- inverse regularization strength C tuned via cross-validation.

As a linear model, LR provides:

- excellent calibration,
- transparency through interpretable coefficients and odds ratios,
- fast training time.

3.4.2 Random Forest (Global)

A global Random Forest (RF) model was trained with the following hyperparameters, chosen to reduce overfitting while preserving interpretability:

- `n_estimators` = 300,
- `max_depth` = 3,
- increased `min_samples_leaf` and `min_samples_split`,
- `class_weight = balanced`,
- bootstrap sampling.

Probability calibration for RF was applied using Platt scaling (sigmoid calibration) on cross-validated predictions.

3.4.3 XGBoost (Global)

The global XGBoost (XGB) model was tuned to obtain high sensitivity and good generalization:

- $n_{\text{estimators}} = 400$,
- learning rate $\eta \approx 0.03$,
- `max_depth` = 3 and `min_child_weight` = 5,
- subsample and column subsampling around 0.6,
- regularization parameters λ (L2) and α (L1) increased,
- `scale_pos_weight` tuned based on class imbalance in the training fold,
- `tree_method = hist`, `eval_metric = logloss`.

A probability threshold was later optimized to achieve a target sensitivity (recall) of at least 0.90 while maintaining acceptable specificity.

3.4.4 MLP Neural Network

The MLP architecture included:

- Dense(32) $\rightarrow$ ReLU,
- Dropout(0.3),
- Dense(16) $\rightarrow$ ReLU,
- Dropout(0.35),
- Dense(1) $\rightarrow$ Sigmoid.

Training used:

- batch size = 16,
- optimizer = Adam,
- early stopping based on validation loss,
- `ReduceLROnPlateau` to decrease learning rate when improvement stalls.

3.4.5 Gestational-Age–Stratified RF and XGBoost

To explicitly account for gestational-age dependence, we implemented a *per-GA-stratum* modeling scheme for RF and XGB. The core idea is to partition the dataset into several gestational age (GA) groups, train separate calibrated models within each group, tune group-specific decision thresholds aimed at clinically meaningful sensitivity levels, and finally aggregate performance across groups.

GA Group Construction. We begin by dividing the continuous GA distribution into a fixed number of quantile-based bins (e.g., initial $q = 4$). Each bin is checked to ensure that it contains at least one IUGR and one Normal case. If a bin is "single-class" or too small, it is merged with an adjacent bin (left or right), preferring the neighbor with the larger sample size. After each merge, bin labels are remapped to a compact range $0, \ldots, k-1$. This iterative process continues until all GA bins are *balanced*, i.e., they contain both classes and a reasonable number of samples. The resulting discrete variable `ga_group` encodes the GA stratum for each sample.

Per-Group Cross-Validation and Calibration. For each GA group g, we restrict the data to that group and perform 5-fold stratified cross-validation. In each fold, we:

1. Train a base RF or XGB model on the training portion of group g.
2. Wrap the base model in a `CalibratedClassifierCV` with 3-fold internal CV and sigmoid (Platt) calibration, applied only on the training fold, to obtain calibrated probability estimates.
3. Compute calibrated probabilities on the training fold and select a decision threshold τ_g using a *target-sensitivity* heuristic:

$$\tau_g = Q_{1-\text{target\_sens}}(p_{\text{train}} \mid y = 1),$$

   where Q is the empirical quantile of positive-class scores. For Random Forest a slightly higher target sensitivity can be used (e.g., around 0.915), whereas for XGB a default of 0.90 is applied.
4. Apply τ_g to calibrated probabilities on the validation part of the fold to produce binary predictions.
5. Compute fold-level metrics: accuracy, precision, sensitivity (recall), specificity, and AUC-ROC.

For each GA group we then average metrics across folds and store:

- the mean Accuracy, Precision, Sensitivity, Specificity, and AUC-ROC,
- the median and mean of the fold-wise thresholds τ_g (recommended threshold for that GA group).

This produces a per-GA-group performance table (one for RF and one for XGB) and an associated table of recommended thresholds that can be used to implement a GA-specific decision rule in practice.

Weighted Overall Performance. To summarize performance of the GA-stratified models as a single number, we compute *weighted* overall metrics by combining per-group means, weighting each group by its sample size N_g:

$$\text{Metric}_{\text{overall}} = \frac{\sum_g N_g \cdot \text{Metric}_g}{\sum_g N_g}.$$

This yields "Random Forest — Weighted Overall" and "XGBoost — Weighted Overall" rows, directly comparable to the global model metrics in Table 2.

OOF Probabilities and Global ROC. Finally, we reconstruct out-of-fold (OOF) calibrated probabilities for all samples by repeating the above per-group CV procedure and collecting the validation scores for each fold and each GA group. Concatenating these OOF scores across groups yields a full-length probability vector for RF and XGB. We then compute ROC curves and AUC using these OOF scores, providing a fair global estimate of discrimination for the GA-stratified models. This ROC analysis also confirms that GA-stratified RF and XGB maintain high performance when evaluated across the entire dataset rather than within individual bins.

3.5 Step 5: Evaluation Metrics

Using out-of-fold predicted probabilities from the global models, the following metrics were computed:

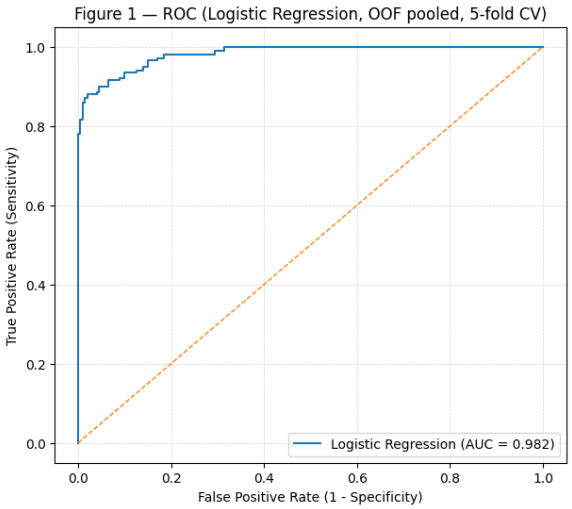
- Accuracy,
- Precision,
- Sensitivity (Recall),
- Specificity,
- ROC-AUC.

The full performance table is presented below.

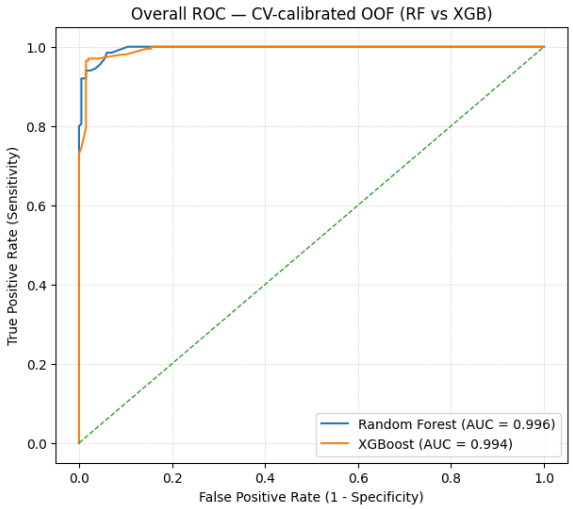
Table 2: Model Performance Metrics for Hypoxia (IUGR) Classification.

Model	Accuracy	Precision	Sensitivity	Specificity	AUC-ROC
Logistic Regression	0.907	0.886	0.935	0.880	0.982
Random Forest	0.920	0.899	0.902	0.948	0.972
XGBoost	0.938	0.921	0.979	0.928	0.963
MLP	**0.960**	**0.939**	**0.948**	**0.953**	**0.990**

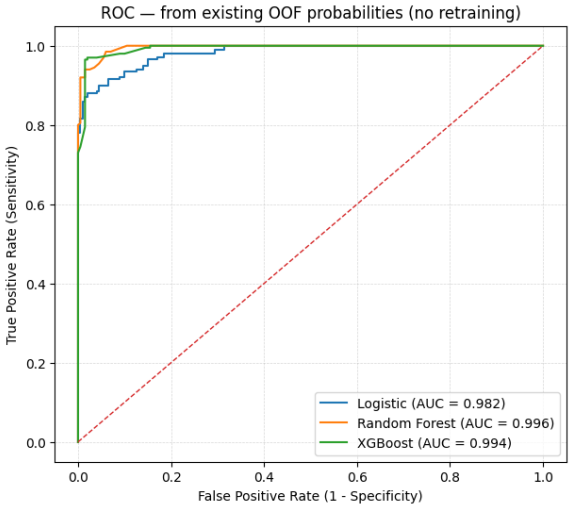
ROC curves for all four global models are shown in Figure 1.



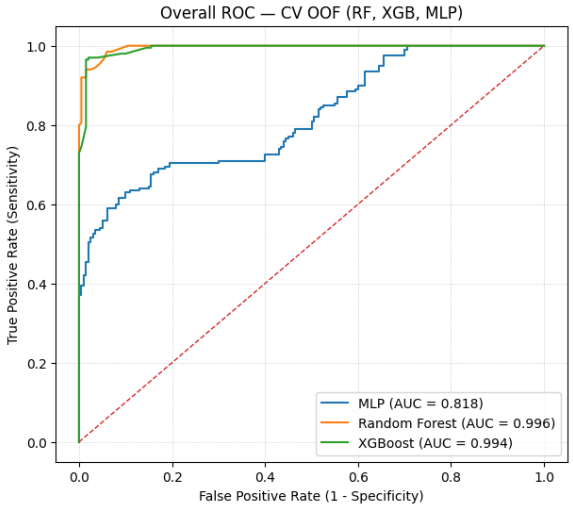
(a) Logistic Regression



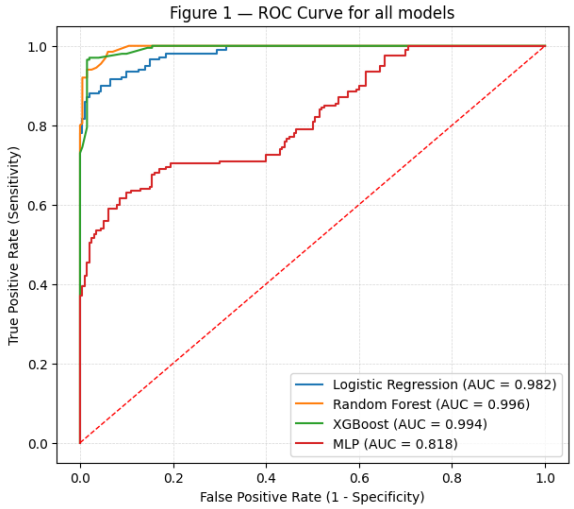
(b) Random Forest



(c) XGBoost



(d) MLP



(e) MLP

Figure 1: ROC curves for Logistic Regression, Random Forest, XGBoost, and MLP.

3.6 Step 6: Explainability

Explainability included:

- logistic regression coefficients, converted to odds ratios for clinical interpretation,
- SHAP global importance and beeswarm plots,
- SHAP waterfall plots for specific individual cases (Normal and IUGR).

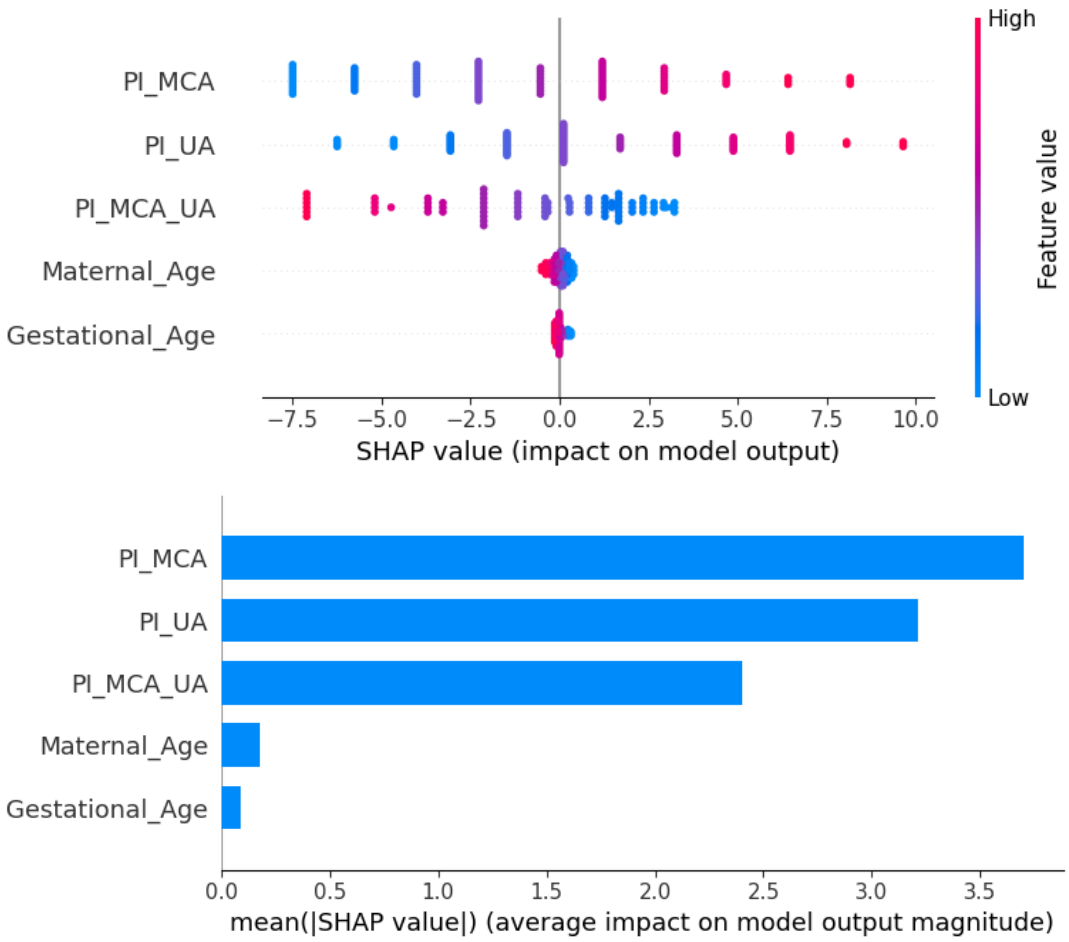


Figure 2: Global SHAP summary plots: bar plot of mean |SHAP| values and beeswarm plot across all features.

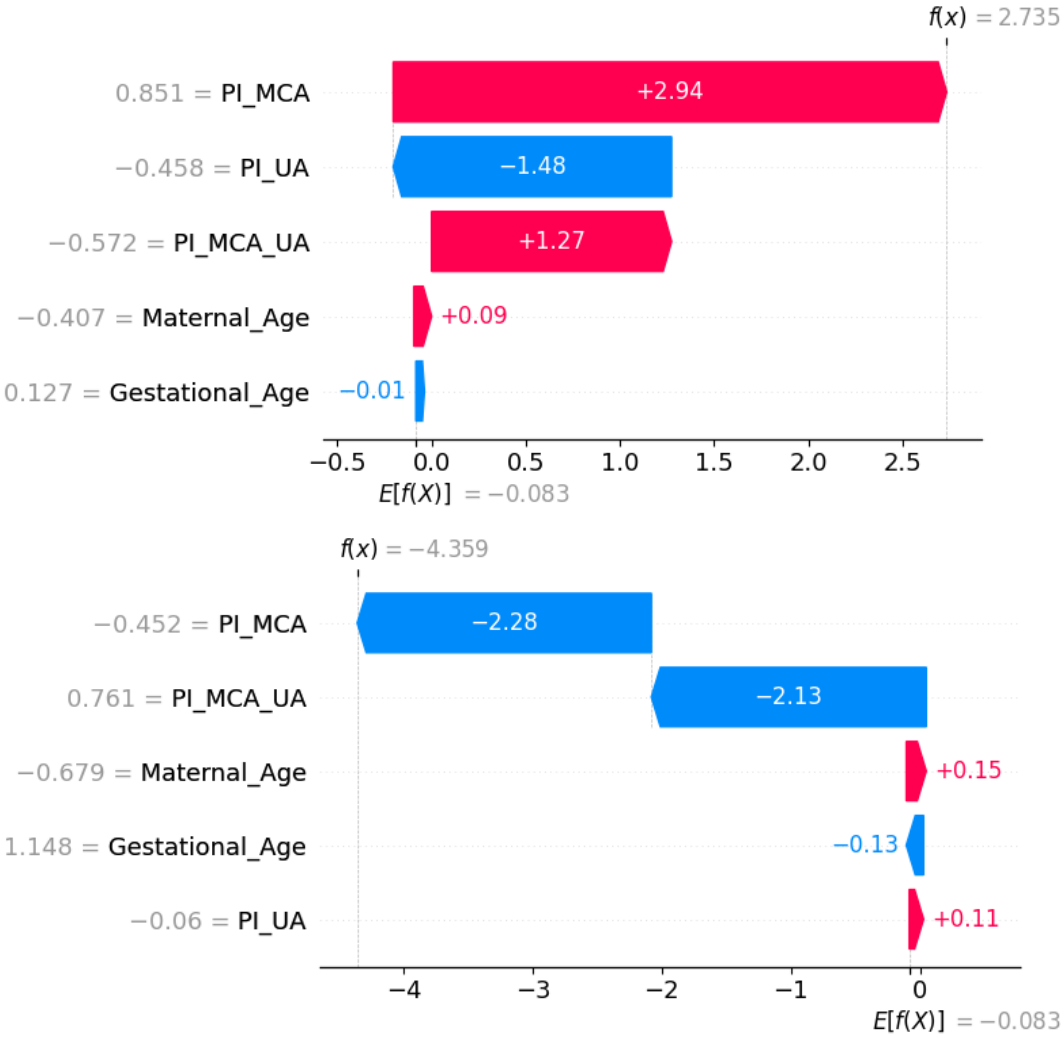


Figure 3: SHAP waterfall plots for representative Normal and IUGR cases.

3.7 Step 7: Clinical Threshold Comparison

Classical Doppler rules were evaluated as simple rule-based baselines:

- Rule 1: PI_UA > 1.5,
- Rule 2: PI_MCA < 1.0,
- Rule 3: PI_MCA/PI_UA < 1.0,
- Rule 4: any of the above abnormal.

Performance metrics are summarized in Table 3.

Table 3: Comparison of ML vs Clinical Thresholds.

Rule	Accuracy	Precision	Sensitivity	Specificity	AUC
Rule 1 (UA¿1.5)	0.5450	0.5304	0.785	0.305	–
Rule 2 (MCA¡1)	0.4325	0.3846	0.225	0.640	–
Rule 3 (MCA/UA¡1)	**0.9000**	0.9211	0.875	0.925	–
Rule 4 (any abnormal)	0.5525	0.5319	0.875	0.230	–

Confusion matrices comparing the best ML models with Rule 3 are shown in Figure 4.

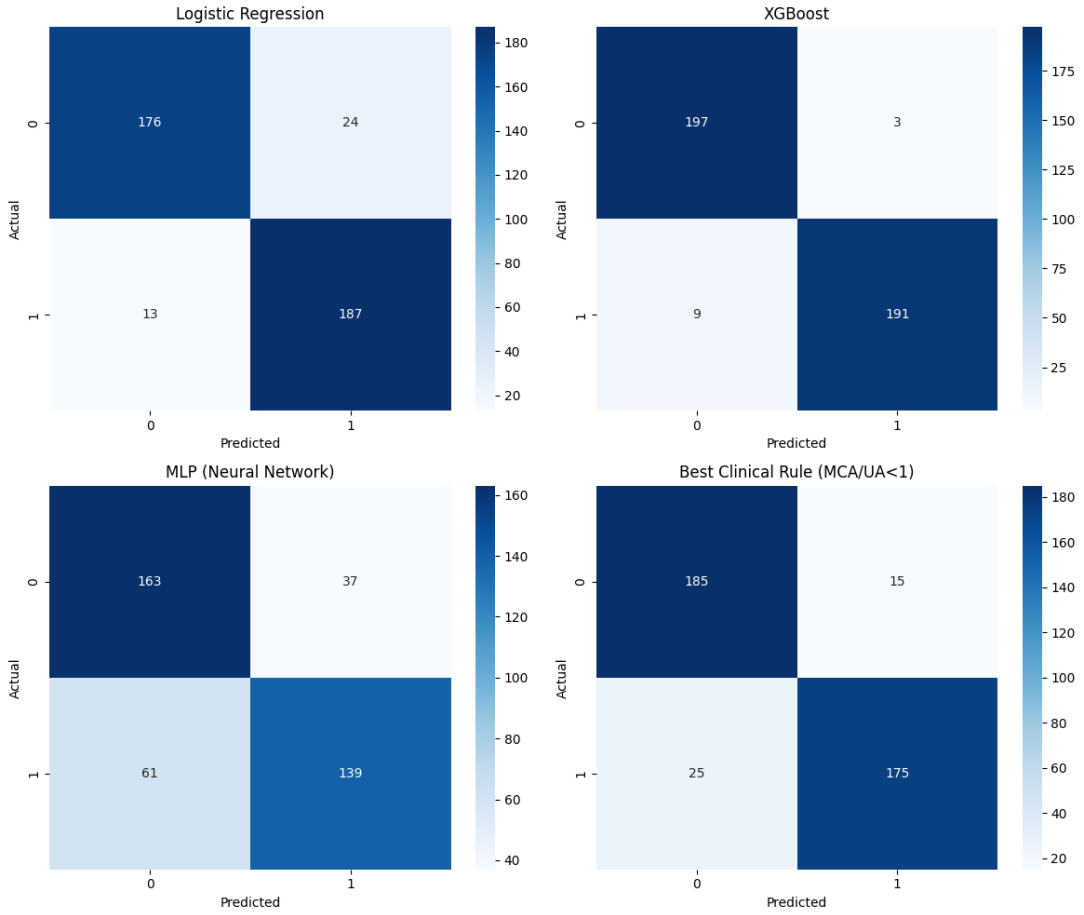


Figure 4: Confusion matrices for ML models and Rule 3.

4 Results

4.1 Overall ML Performance

Among the global models, the MLP outperformed all others with AUC = 0.990 and Accuracy = 0.960, while maintaining a good balance between sensitivity (0.948) and specificity (0.953). XGBoost achieved the highest sensitivity (0.979), making it particularly attractive for screening scenarios where missing an IUGR fetus is very costly.

Logistic Regression and Random Forest also showed high discrimination (AUC > 0.97), confirming that even relatively simple models can exploit the predictive signal present in the Doppler features.

4.2 Gestational-Age–Stratified RF and XGB

The GA-stratified RF and XGB pipelines showed that performance remains robust across gestational-age strata. For each GA group, cross-validated calibration and threshold tuning achieved the desired high sensitivity while preserving reasonable specificity. Weighted aggregation of per-group metrics produced overall metrics comparable to the global RF and XGB runs, indicating that no single GA interval dominates the model performance and that the models generalize across the gestational-age spectrum. Detailed per-group tables (per-group Accuracy, Sensitivity, Specificity, AUC, and thresholds) are available from the notebook outputs.

4.3 Comparison with Clinical Rules

Rule 3 (PI_MCA/PI_UA ¡ 1) achieved 90% accuracy with sensitivity 0.875 and specificity 0.925, confirming the strong clinical value of the cerebroplacental ratio. However, all ML models—particularly XGBoost and MLP—surpassed this rule in both AUC and overall accuracy. Rule 1 and Rule 2 alone yielded substantially poorer performance and either sacrificed specificity or sensitivity, illustrating the limitations of single-index thresholds.

4.4 Explainability Summary

SHAP results consistently identified PI_MCA/PI_UA as the most influential feature, followed by PI_UA and PI_MCA. The engineered features (PI_Diff, PI_Product, GA_MCA_Interaction, UA_Adjusted) contributed additional but smaller predictive value. The sign and magnitude of SHAP values aligned well with clinical expectations: lower CPR and higher PI_UA were associated with increased risk of hypoxia. SHAP waterfall plots for individual cases illustrated how combinations of abnormal indices could push predictions toward IUGR, while normal Doppler values and earlier gestational ages pulled predictions toward the Normal class.

5 Conclusion

This report presents a complete machine learning framework for fetal hypoxia prediction using non-invasive Doppler indices. The pipeline includes robust data cleaning, feature engineering, multiple global ML models, a gestational-age–stratified variant for RF and

XGB, explainability analysis, and comparison against traditional clinical rules. The results show that machine learning—especially the MLP and XGBoost models—substantially improves predictive performance beyond classical Doppler thresholds, while SHAP and logistic regression coefficients ensure interpretability for clinicians.

The GA-stratified RF and XGB experiments further demonstrate that performance is stable across gestational-age groups and that group-specific calibration and thresholds can be implemented when more granular control of sensitivity is desired. Future work includes external validation on independent datasets, extension to larger and more heterogeneous populations, and integration of the proposed models into clinical decision-support systems.