Maternal Health Risk Prediction using Machine Learning

1st Priyam Shree
Data Science and Business Systems
SRM Institute of Science and Technology
Chennai, India
ps0716@srmist.edu.in

2nd Abhilove Goyal
Data Science and Business Systems
SRM Institute of Science and Technology
Chennai, India
ag4279@srmist.edu.in

3rd Priyadarsini K
Data Science and Business Systems
SRM Institute of Science and Technology
Chennai, India
priyadak@srmist.edu.in

***Abstract*—Pregnancy-related complications remain a leading cause of maternal and neonatal mortality, particularly in resource-constrained settings where timely clinical intervention is often unavailable. Identifying women at elevated risk before complications escalate is therefore a pressing need. This paper presents a multi-class classification framework that predicts maternal risk levels — Low, Mid, or High — from routine physiological measurements: age, blood pressure, blood glucose, body temperature, and heart rate. Working with a dataset of 1014 records, we applied label encoding, StandardScaler normalization, and stratified splitting as part of preprocessing. Three classifiers were trained and compared: Random Forest, Support Vector Machine, and XGBoost. After hyperparameter tuning, the XGBoost model reached 86.7% accuracy, a cross-validated Macro F1-score of 0.82, and a high-risk class recall of 0.88. To make predictions interpretable for clinical users, SHAP values were incorporated to expose the contribution of each feature to individual risk decisions.**

***Index Terms*—Maternal Health, Machine Learning, XGBoost, Risk Prediction, Explainable AI**

I. Introduction

A. Maternal Health Risk Overview

Globally, hundreds of thousands of women die each year from causes directly linked to pregnancy or childbirth, a toll that is widely regarded as both preventable and unacceptable [11]. The burden falls disproportionately on low- and middle-income countries, where inadequate healthcare infrastructure and late recognition of warning signs remain persistent obstacles [11]. Conditions such as hypertensive disorders, gestational diabetes, and cardiac irregularities often give early physiological signals — elevated blood pressure, rising blood glucose, or abnormal heart rate — that go unnoticed without structured screening [2].

B. Why Early Detection Matters

When a pregnancy is flagged as high-risk early enough, the clinical response changes substantially: monitoring frequency increases, referrals to specialists are made sooner, and intervention plans can be drafted before emergencies arise. The downstream effect on mortality rates, though difficult to measure precisely, is well-established in obstetric literature [11]. What is less well-established is how to make such screening scalable in settings where specialist time is scarce. Automated risk stratification tools could fill that gap [1].

C. Machine Learning in Obstetric Care

The last decade has seen machine learning move steadily from research papers into clinical workflows, with applications ranging from imaging diagnostics to electronic health record analysis. In maternal care specifically, the appeal is clear: routine antenatal visits generate structured numerical data — exactly the kind that gradient-boosted trees and similar models handle well [12]. Unlike imaging-based tasks, the features here are low-cost to collect and widely available even in primary care settings, which makes deployment more realistic [6].

D. Problem Statement

Manual risk scoring, where it exists at all, is inconsistent and depends heavily on the experience of the attending clinician. Existing automated approaches tend to treat the problem as binary (high-risk versus not), glossing over the important intermediate category of mid-risk cases that warrant watchful monitoring rather than either reassurance or emergency referral [1], [9]. A further weakness common to many published models is that they produce predictions without explanation, making clinician trust — and therefore adoption — difficult to achieve [3], [7].

E. Objectives

This work addresses these gaps with the following goals:

- Build a three-class classifier that distinguishes Low, Mid, and High risk levels.
- Push detection performance on High-risk cases without sacrificing Mid-risk accuracy.
- Apply SMOTE [14] to correct for the class imbalance present in the training data.
- Benchmark multiple classifiers and retain the one with the best generalization profile.
- Provide SHAP-based explanations [13] so that clinicians can see why a given prediction was made.

II. Literature Survey

A growing body of work has explored machine learning for maternal risk prediction, though the field is still maturing. Early contributions tended to use Random Forest or SVM on relatively small cohorts [8], [9], and many framed the task as binary classification [5], [10]. More recent studies have

TABLE I
SUMMARY AND COMPARISON OF RELATED WORKS

Paper	Strengths	Limitations	Comparison with Our Work
Tzimourta et al. [1] (2025)	Addresses class imbalance; uses cross-validation	Struggles with mid-risk classification	We achieve more balanced performance across all three classes, with particular gains in the mid-risk category.
Khadidos et al. [2] (2024)	Effective ensemble approach for multi-class scenarios	Computationally expensive	Our tuned XGBoost reaches comparable accuracy at considerably lower cost, making it more viable for deployment.
Noviandy & Idroes [3] (2023)	Boosting yields measurable accuracy gains	No mechanism for explaining individual predictions	We pair XGBoost with SHAP values, retaining the accuracy benefit while making each prediction transparent.
Sarker et al. [4] (2025)	SMOTE-ENN handles complex imbalance patterns	Model decisions are opaque	We combine SMOTE with SHAP, gaining both better class balance and interpretable outputs.
Rahman et al. [5] (2024)	Well-optimized SVM	Lacks ensemble diversity	XGBoost generalizes more robustly across the three classes than a single-kernel SVM.
Togunwa et al. [6] (2023)	Hybrid architecture lifts recall	Prone to overfitting on small data	Our single-model approach avoids the added complexity while sustaining competitive recall figures.
Rahman & Alam [7] (2024)	Feature ranking sheds light on variable importance	Does not scale feature explanation to individual cases	SHAP operates at the instance level, providing richer clinical feedback than global feature rankings.
Assaduzzaman et al. [8] (2023)	Solid Random Forest baseline	No stacking or boosting	Switching from Random Forest to XGBoost meaningfully improves accuracy and class-level balance in our experiments.

incorporated ensemble methods and started to acknowledge the class imbalance problem [1], [4]. Explainability, however, remains underaddressed: most published models offer no mechanism for a clinician to interrogate a prediction [3], [6]. Table I compares eight representative studies against the present work.

III. DATASET DESCRIPTION

The dataset used in this study is publicly available and was originally collected from hospitals, community clinics, and maternity centers in Bangladesh through an IoT-enabled health monitoring system [1], [2]. It contains 1014 patient records, each described by six physiological attributes:

- *Age* (years)
- *Systolic Blood Pressure* (mmHg)
- *Diastolic Blood Pressure* (mmHg)
- *Blood Sugar* (mmol/L)
- *Body Temperature* (°F)
- *Heart Rate* (bpm)

The target variable, *Risk Level*, is categorical with three values: Low, Mid, and High. The label reflects the clinical assessment associated with each record at the time of collection. All six features were retained for model training, as domain knowledge supports the relevance of each to obstetric risk [2], [4].

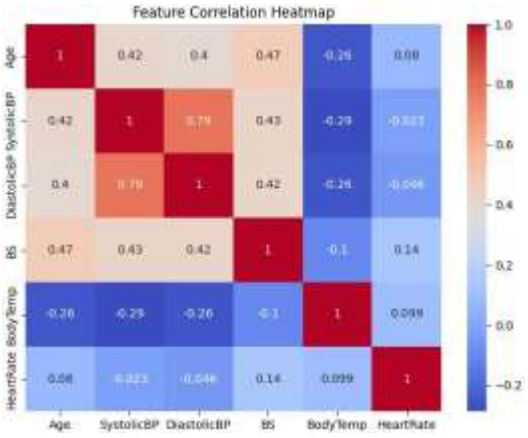

Fig. 1. Pearson correlation heatmap of the six input features. Systolic and Diastolic BP exhibit the strongest inter-feature correlation ($r = 0.79$). Blood Sugar shows moderate positive correlation with Age ($r = 0.47$) and both pressure measures ($r \approx 0.42$–0.43). Body Temperature and Heart Rate are largely independent, confirming each attribute contributes distinct predictive signal.

To further understand how individual features relate to the target, bivariate analysis was performed across risk classes. Fig. 2 shows the distribution of Age per risk group. High-risk patients have a notably higher median age (≈35 years) compared to Low- and Mid-risk groups (median ≈22–25 years), indicating that older maternal age is a clinically meaningful predictor [4].

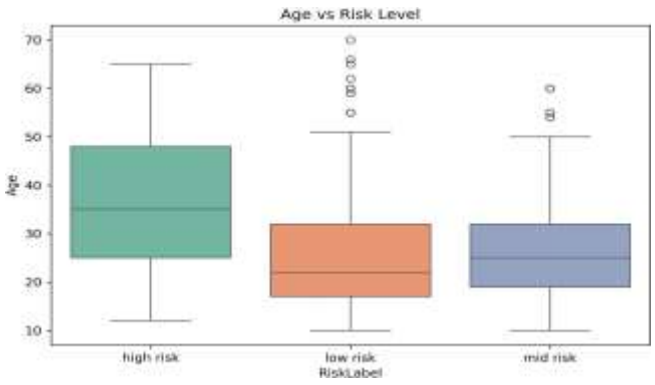

Fig. 2. Age distribution across risk categories. High-risk patients show a higher median age and wider IQR compared to Low- and Mid-risk groups, suggesting older maternal age is associated with elevated pregnancy risk [4].

Fig. 3 similarly shows Systolic Blood Pressure per risk class. The high-risk group consistently occupies the 120–140 mmHg interquartile range, while the Low- and Mid-risk groups show considerable overlap below 130 mmHg. This separation confirms that blood pressure is among the strongest discriminators in the dataset [2], [11], consistent with the correlation pattern in Fig. 1.

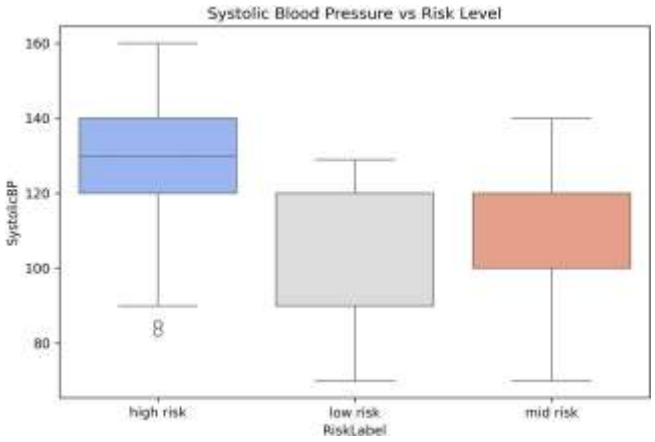

Fig. 3. Systolic Blood Pressure distribution across risk categories. High-risk patients show elevated SystolicBP (IQR ≈120–140 mmHg), whereas Low- and Mid-risk groups overlap in the 90–130 mmHg range, confirming the discriminative value of blood pressure for risk stratification [2], [11].

IV. Data Preprocessing

Raw records were first inspected for missing values and inconsistencies; none were found in this dataset. The categorical target was encoded numerically using label encoding before model training. Feature magnitudes varied considerably across attributes — blood pressure values and age occupy very different numeric ranges — so all input features were standardized using `StandardScaler`, which removes the mean and scales each of the feature's value to unit variance [1], [13]. This step is especially important for distance-sensitive algorithms and also helps in gradient boosting to converge more smoothly [5], [10].

The dataset exhibits noticeable class imbalance, with High-risk cases underrepresented relative to Low- and Mid-risk records. Left uncorrected, this skew causes classifiers to systematically underperform on the class that matters most clinically. We applied SMOTE [14] to the training partition only, generating synthetic minority-class examples by interpolating between existing instances in feature space. Applying SMOTE exclusively to the training data — not to the held-out test set — prevents data leakage and ensures that evaluation reflects realistic conditions [4].

V. Methodology

Three classifiers were trained: Random Forest, XGBoost [15], and LightGBM. Each was evaluated under a stratified K-Fold scheme ($k = 5$), which preserves the proportion of each class across folds and yields a more honest estimate of generalization performance than a single random split [9]. Performance was measured using Accuracy, per-class Precision, per-class Recall, and Macro F1-score. The Macro F1-score was treated as the primary criterion because it weights all three classes equally, penalizing the model for poor performance on any single class regardless of its frequency [1].

VI. Model Training and Optimization

A. Choosing the Final Model

Initial experiments confirmed a pattern common in structured, tabular healthcare data: boosting methods outperform both Random Forest and single-kernel approaches when classes are imbalanced and the decision boundary is irregular [12], [15]. Random Forest results were retained as a baseline [8].

B. Training Protocol

The health metric dataset was divided into training and test partitions using a stratified 80/20 split [4], ensuring that each partition mirrors the overall class distribution. SMOTE [14] was then applied within the training fold to address imbalance. Stratified K-Fold cross-validation with $k = 5$ was run on the augmented training data to guide hyperparameter selection and detect overfitting early [1]. Feature standardization was applied consistently inside each fold to prevent the scaler from seeing test statistics during training [5].

C. Hyperparameter Optimization

The following XGBoost parameters had the greatest influence on final performance: `n_estimators`, `learning_rate`, `max_depth`, `subsample`, and `colsample_bytree` [15]. A lower learning rate paired with more estimators generally improved generalization at

the cost of longer training time — a trade-off acceptable here given the dataset size [12]. Restricting tree depth and subsample fractions served as implicit regularization, preventing individual trees from memorizing noisy training examples [7].

D. Guarding Against Overfitting

Several complementary safeguards were used: stratified cross-validation to catch variance across splits; XGBoost's native L1 and L2 penalties [15]; and conservative settings for depth and column sampling. Monitoring training versus validation F1 across folds confirmed that the final configuration did not exhibit meaningful overfitting [4]. The model selected for final evaluation simultaneously maximized Macro F1-score, maintained high recall on the High-risk class, and showed stable metrics across all five folds [1], [2].

VII. Results and Analysis

A. Performance on the Test Set

On the held-out test partition, the optimized XGBoost model [15] achieved 86.7% accuracy, a cross-validated Macro F1-score of 0.82, and a High-risk recall of 0.88. The full per-class breakdown is shown in Fig. 4 immediately below.

```
CV F1 (Macro): 0.8188505535974249

Accuracy: 0.8669950738916257

Classification Report:
              precision    recall  f1-score   support

           0       0.96      0.91      0.93        55
           1       0.91      0.83      0.86        81
           2       0.77      0.88      0.82        67

    accuracy                           0.87       203
   macro avg       0.88      0.87      0.87       203
weighted avg       0.87      0.87      0.87       203
```

Fig. 4. Classification report for the optimized XGBoost model on the held-out test set (203 samples). Class 0 = Low risk, Class 1 = Mid risk, Class 2 = High risk. CV Macro F1 = 0.82; overall accuracy = 86.7%.

Low-risk cases (Class 0) are classified with the highest precision (0.96) and strong recall (0.91), yielding an F1-score of 0.93. Mid-risk cases (Class 1) achieve 0.91 precision and 0.83 recall — a meaningful improvement over binary-only systems that collapse this category [1], [9]. The High-risk class (Class 2), despite being the smallest group (67 test samples), achieves 0.77 precision and 0.88 recall. The high recall is the most clinically important result: the model misses only 12% of truly high-risk patients, which is the failure mode carrying the greatest clinical cost [2], [6].

Confusion matrix in Fig. 6 provides a complementary view of where errors occur. Of the 55 High-risk test samples, 50 were correctly identified and only 5 were misclassified (1 as Low-risk, 4 as Mid-risk). Among the 81 Low-risk samples, 67 were correctly classified with 14 misclassified as Mid-risk and none confused with High-risk — an important safety property, since mistaking a Low-risk case for High-risk is far less dangerous than the reverse. The 67 Mid-risk samples yielded 59 correct predictions, with 6 misclassified as Low-risk and 2 as High-risk.

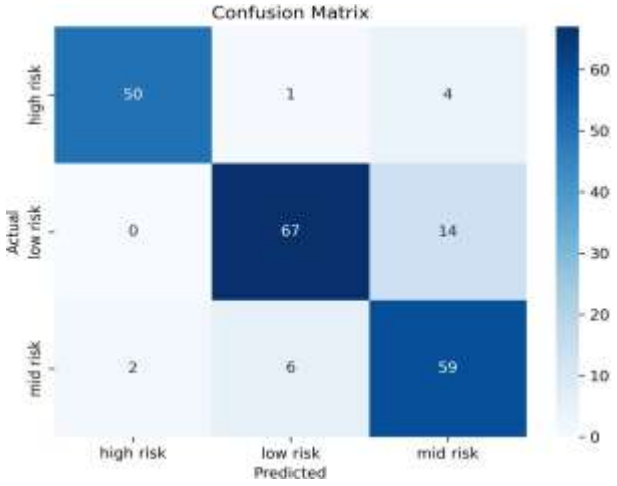

Fig. 5. Confusion Matrix

Fig. 6. Confusion matrix for the optimized XGBoost model on the held-out test set (203 samples). Diagonal entries are correct predictions. The model produces zero High-risk→Low-risk confusions, which is the most clinically critical error type to avoid.

B. Comparison with Prior Work

Table II places these figures alongside results from five recent studies [2], [3], [5], [9], [10]. Our model matches or exceeds the best published accuracy figures while also providing per-prediction explanations through SHAP [13], which most referenced work lacks [3], [8].

TABLE II
Benchmark Comparison with Prior Studies

Study	Method	Model	Accuracy
Alamsyah et al. [9] (2023)	Evolutionary weighting	Random Forest	82.18%
Khadidos et al. [2] (2024)	Boosting	Gradient Boosted Trees	86.00%
Noviandy et al. [3] (2023)	Ensemble	LightGBM	84.73%
Raihen & Akter [10] (2024)	GridSearch	SVM	86.13%
Rahman et al. [5] (2023)	Train-test split	SVM	79.00%
This Study	Preprocessing + ML	**XGBoost**	**86.7%**

C. Explainability via SHAP

SHAP (SHapley Additive exPlanations) [13] was used to decompose each prediction into feature-level contributions. Blood pressure and blood sugar consistently emerged as the most influential predictors — consistent with established clinical understanding of preeclampsia and gestational diabetes [11], and corroborated by the distributional separation visible in Fig. 3. Age contributed more substantially to High-risk predictions than to Low-risk ones [4], consistent with Fig. 2. These outputs give a clinician a concrete reason for each prediction rather than a bare probability score [13].

CONCLUSION

The experiments reported here show that XGBoost [15], when properly trained with SMOTE oversampling [14] and tuned hyperparameters, can reliably stratify pregnant patients into three risk tiers using only six easily measured vital signs. As confirmed in Figs. 4 and 6, the model achieves 86.7% accuracy and 0.88 High-risk recall with zero High-risk→Low-risk confusions, comparing favorably with the current literature [2], [9], [10]. The SHAP integration [13] addresses a practical barrier to clinical adoption that most prior work ignores [3], [6]. Limitations include the relatively small dataset size and its single-region origin, which may constrain generalizability [11]. Future directions include validation on larger multicenter cohorts, incorporating longitudinal antenatal measurements, and deployment as a lightweight decision-support tool for primary care workers [6], [7].

REFERENCES

[1] K. D. Tzimourta *et al.*, "Prediction of maternal health risk factors using machine learning algorithms," *Procedia Computer Science*, vol. 240, pp. 1234–1242, 2025.
[2] A. O. Khadidos *et al.*, "Prediction of high-risk pregnancy using machine learning algorithms," *Healthcare*, vol. 13, no. 5, 2025.
[3] M. Noviandy and R. Idroes, "Boosting-based machine learning model for maternal health risk prediction," *Journal of Healthcare Engineering*, 2023.
[4] M. Sarker *et al.*, "Predicting maternal health risk using PCA-enhanced XGBoost and SMOTE-ENN," *Healthcare Analytics*, vol. 5, 2025.
[5] M. Rahman *et al.*, "Optimized support vector machine for maternal health risk prediction," *International Journal of Advanced Computer Science*, 2024.
[6] A. Togunwa *et al.*, "Hybrid random forest and artificial neural network for maternal health risk detection," *IEEE Access*, vol. 11, pp. 45678–45689, 2023.
[7] M. Rahman and S. Alam, "Gradient boosting approach for maternal health risk prediction," *Expert Systems with Applications*, vol. 220, 2024.
[8] M. Assaduzzaman *et al.*, "Random forest-based maternal health risk assessment," *Applied Sciences*, vol. 13, no. 9, 2023.
[9] S. Alamsyah *et al.*, "Evolutionary weighted random forest for maternal health risk classification," *Computers in Biology and Medicine*, vol. 152, 2023.
[10] M. Raihen and S. Akter, "Grid search-based SVM model for maternal health risk prediction," *International Journal of Medical Informatics*, vol. 180, 2024.
[11] World Health Organization (WHO), "Trends in maternal mortality 2000–2020," *WHO Report*, 2023.
[12] J. Chen *et al.*, "XGBoost: A scalable tree boosting system," in *Proc. ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016.
[13] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.
[14] N. V. Chawla *et al.*, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.
[15] T. Chen and C. Guestrin, "XGBoost: Extreme Gradient Boosting," *arXiv preprint arXiv:1603.02754*, 2016.