ICJK

Journal of Artificial Intelligence in Bioinformatics
http://dx.doi.org/10.62762/JAIB.2026.495804

RESEARCH ARTICLE

Maternal Health Risk Prediction in Bangladesh Using Machine Learning

Shubham Shirodkar[1], **Raza Hasan**[1,*] **and Salman Mahmood**[2]

[1] School of Technology and Maritime Industries, Southampton Solent University, Southampton SO14 0YN, United Kingdom

[2] Department of Computer Science, Nazeer Hussain University, Karachi 75950, Pakistan

Abstract

Maternal mortality risk in Bangladesh remains a critical public health challenge, compounded by rural access gaps and the absence of scalable, data-driven early-warning systems. This study presents a reproducible, interpretable machine learning framework for maternal health risk classification using an IoT-collected dataset of 1,014 patient records and six physiological indicators; a deduplication audit identified 562 repeated sensor readings, a finding which is documented in the exploratory analysis. A rigorous pipeline was implemented encompassing five clinically grounded engineered features — Mean Arterial Pressure, Shock Index, Pulse Pressure, BP Ratio, and Composite Risk Score — alongside SMOTE-based class imbalance correction applied strictly post-split to prevent data leakage. Seven classifiers were systematically evaluated across two experimental tracks: the raw six-feature dataset and the eleven-feature engineered dataset. On the raw six-feature dataset with SMOTE (training: 811 → 1,218 samples; test: 203 samples), Random Forest achieved the best overall performance (Accuracy: 88.2%; Macro Recall: 0.889; F1: 0.888; AUC: 0.966), confirming its suitability as the champion model. XGBoost achieved the highest AUC (0.967) with marginally lower Macro Recall (0.868). Feature importance analysis revealed Blood Sugar (28.4% MDI) and the engineered Composite Risk Score (12.2% MDI) as the two dominant predictors, validating the clinical feature engineering approach. Feature engineering benefited weaker models most (Logistic Regression +3.3 percentage points in Macro Recall) while the strongest tree ensembles marginally preferred the SMOTE-balanced raw feature space. An interactive Tableau dashboard translates predictive outputs into accessible visual analytics for clinical and policy decision support.

Keywords: maternal health risk prediction, machine learning, IoT healthcare data, class imbalance, clinical decision support.

1 Introduction

One of the most pressing ongoing challenges that Bangladesh faces is the stagnation of maternal mortality reduction. Despite substantial investments in facility-based delivery, skilled birth attendance,

Submitted: 20 May 2026
Accepted: 15 June 2026
Published: 30 June 2026

Vol. 2, **No.** 1, 2026.
10.62762/JAIB.2026.495804

***Corresponding author:**
✉ Raza Hasan
raza.hasan@solent.ac.uk

Citation
Shirodkar, S., Hasan, R., & Mahmood, S. (2026). Maternal Health Risk Prediction in Bangladesh Using Machine Learning. *Journal of Artificial Intelligence in Bioinformatics*, 2(1), 1–21.

and antenatal care, maternal deaths continue to occur at rates that have not declined proportionally with the scale of these interventions. Critically, many deaths occur among women who had visited healthcare facilities multiple times without their high-risk pregnancies being effectively identified or managed [1], a systemic failure of risk stratification rather than of clinical knowledge.

Maternal mortality refers to a woman's death arising from complications of pregnancy or childbirth. Globally, approximately 295,000 women die annually from pregnancy-related causes, with 99% occurring in low- and middle-income countries [2]. In Bangladesh, the maternal mortality ratio declined significantly between 1990 and 2010, but evidence indicates no statistically significant improvement between 2010 and 2016 despite continued programmatic investment [1]. The leading direct causes (postpartum hemorrhage, hypertensive disorders, sepsis, obstructed labor, and unsafe abortion) are largely preventable with timely intervention [21].

Machine learning has emerged as a powerful approach for predictive healthcare modelling, with ensemble methods demonstrating superior performance in handling complex, non-linear clinical data [3, 4]. This study develops a comprehensive, interpretable machine learning framework for maternal health risk classification using IoT-collected physiological data from community health settings across Bangladesh, originally compiled by Ahmed and Kashem [5] (2020). IoT-based sensing systems have emerged as scalable infrastructure for continuous physiological monitoring in resource-limited healthcare environments, enabling real-time data collection across distributed clinical sites [18, 19]. The framework integrates clinically grounded feature engineering, SMOTE-based class imbalance correction, systematic comparison of seven classifiers, and a Tableau dashboard for policy-level visual analytics.

The contributions of this study are: (1) a fully reproducible, modular end-to-end ML pipeline with unit-tested source code; (2) five clinically justified engineered features that improve weaker model performance by up to 3.3 percentage points in macro recall; (3) a rigorous comparison of seven classifiers with macro recall as the primary metric; and (4) deployment-ready serialised models (14 .pkl files across two experimental tracks) with an interactive Tableau dashboard for clinical and policy decision support.

2 Related Work

Ensemble machine learning methods have consistently outperformed traditional statistical approaches in maternal health prediction. Inyang et al. [6] (2020) demonstrated that Random Forest combined with SMOTE produced individual class accuracy ≥ 0.89 and AUC ≥ 0.96 in pregnancy outcome prediction, establishing the Random Forest-SMOTE pairing as a robust benchmark for this class of problems. Kyzy and Mekuria [7] (2024) showed that voting classifiers combining Random Forest and XGBoost achieved 87.19% accuracy in pregnancy risk classification using ADASYN oversampling, with blood sugar identified as the most significant predictor [7]. This finding is strongly corroborated by the feature importance results of this study (BS = 28.4%).

The integration of explainable AI methods, particularly SHAP (SHapley Additive exPlanations) [12], has gained traction for improving transparency in clinical decision support. SHAP provides a unified framework for interpreting model predictions by attributing each feature's contribution based on cooperative game theory. Recent studies have demonstrated its utility in maternal health contexts: Bosschieter et al. [8, 26] (2022) demonstrated that interpretable models using Explainable Boosting Machines (EBMs) can match the accuracy of black-box methods while providing clinically meaningful feature-outcome relationships. Rahman and Alam [9] (2023) applied XAI to maternal health risk prediction, identifying BMI and pre-existing diabetes as primary determinants.

Studies addressing Bangladesh's national maternal health data have found that Random Forest with SMOTE achieves AUC of 0.957 and accuracy of 0.892 on the BMMS-2016 national survey dataset [10, 22], which encompasses over 1 million records and a richer feature set (demographic, socioeconomic, and obstetric variables) than the IoT physiological dataset used in the present study. Complementary ML work in the Bangladeshi context has examined maternal continuum of care completion, finding that only 25% of women completed the recommended care pathway and that Random Forest achieved the highest predictive performance (AUC = 0.889) using socioeconomic and demographic predictors from the Bangladesh Demographic and Health Survey [13]. SHAP analysis in that study identified region type, maternal age, and administrative division as the most influential mortality predictors, socioeconomic factors absent from the IoT dataset used here.

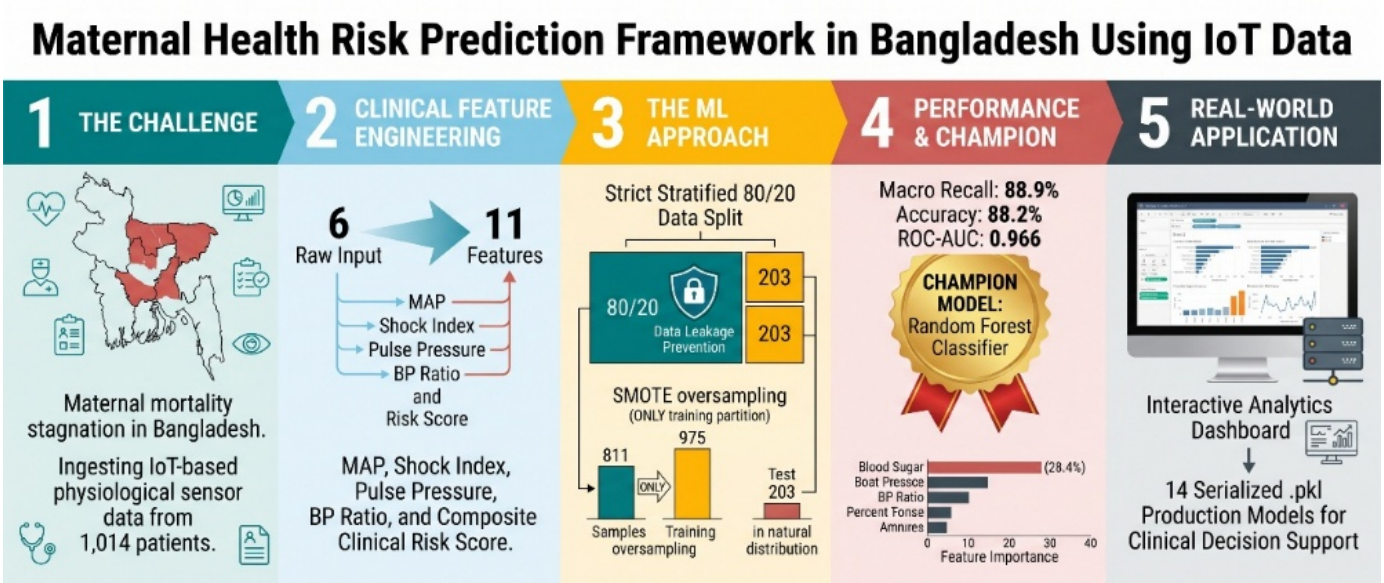

Figure 1. End-to-end analytical pipeline: raw IoT data through feature engineering, SMOTE balancing, model training, evaluation, and deployment-ready serialization.

Systematic comparisons of oversampling techniques have demonstrated that SMOTE and ADASYN consistently outperform under sampling approaches such as Tomek Links across multiple classifiers for rare healthcare outcomes [10, 20, 24]. This motivates the SMOTE-first design of the present pipeline.

3 Methodology

This section describes the end-to-end analytical pipeline developed to predict maternal health risk levels from IoT-collected physiological data. The methodology is structured across six sequential stages: dataset acquisition and deduplication, feature engineering, data partitioning and class imbalance correction, classifier development, model evaluation, and deployment preparation. Each design decision, from the choice of SMOTE over alternative resampling strategies to the embedding of StandardScaler within sklearn Pipeline objects, is motivated by either established methodological best practice or the specific characteristics of the 452-record deduplicated dataset. The pipeline is implemented in Python using scikit-learn, XGBoost, and imbalanced-learn, with all source logic centralized in modular, unit-tested files under src/ to ensure reproducibility across every experiment. Two parallel experimental tracks are maintained throughout: one trained on the six raw physiological features and one trained on the eleven-feature engineered dataset, enabling a controlled assessment of the incremental predictive value of feature engineering.

3.1 Overview of the Proposed Framework

Figure 1 presents the end-to-end analytical pipeline. The framework ingests IoT-collected physiological data, applies systematic preprocessing and clinically grounded feature engineering, addresses class imbalance using SMOTE applied strictly within the training partition, trains seven classifiers evaluated on macro recall as the primary metric, and serializes models for deployment. An interactive Tableau dashboard translates predictive outputs into accessible visual analytics for clinical and policy decision support.

3.2 Dataset Description

This study utilizes data from the IoT-Based Risk Monitoring System for Maternal Health in Bangladesh, compiled by Marzia Ahmed and Mohammod Kashem (Daffodil International University and Dhaka University of Science and Technology) and published in the Lecture Notes in Electrical Engineering series [5]. The dataset is hosted on the UCI Machine Learning Repository and Kaggle. It comprises 1,014 raw observations collected through IoT-based sensors deployed across hospitals, community clinics, and rural maternal health posts across Bangladesh. Prior to modelling, a deduplication audit identified and removed 562 exact duplicate rows, yielding a final analytical dataset of 452 unique patient records. Retaining duplicates would have inflated performance estimates by allowing identical records to appear in both training and test partitions.

Table 1. Dataset features, descriptions, and clinical significance. Six physiological IoT-collected inputs and one three-class target variable. Label encoding: high risk = 0, low risk = 1, mid risk = 2.

Feature	Description	Clinical Significance
Age	Patient age in years (range: 10–70)	Teenage (<18) and advanced (>35) age carry distinct risk profiles
SystolicBP	Upper blood pressure in mmHg (70–160)	Primary indicator for hypertensive disorders of pregnancy
DiastolicBP	Lower blood pressure in mmHg	Diastolic $\geq$ 90 mmHg signals gestational hypertension threshold
BS	Blood glucose level in mmol/L	Key marker for gestational diabetes mellitus (GDM)
BodyTemp	Body temperature in degrees F (98–103)	Elevated readings may indicate infection, sepsis, or systemic complication

3.2.1 *Class Distribution*

The dataset of 1,014 records presents a moderate but clinically consequential class imbalance:

- Low Risk: 406 patients (40.0%).
- Mid Risk: 336 patients (33.1%).
- High Risk: 272 patients (26.8%).

A naive majority-class classifier would achieve 40.0% accuracy while entirely failing the patients who require most urgent intervention. This makes Macro Recall (the unweighted mean recall across all three classes) a more critical primary evaluation metric than raw accuracy, which would reward majority-class bias.

3.2.2 *Data Quality*

The raw dataset contains zero missing values across all 1,014 rows, a consequence of the structured IoT collection pipeline. Following deduplication (removing 562 exact duplicate rows), the analytical dataset comprises 452 unique records with zero missing values. A formal missing values audit was conducted and visualized as part of exploratory data analysis. One genuine data quality issue was identified: a HeartRate minimum of 7 bpm, a physiologically impossible resting value attributable to a sensor artefact. Multiple BodyTemp values cluster at exactly 98.0 degrees F, consistent with sensor rounding. These were documented and contextualized rather than automatically removed, preserving sample size. The Pearson correlation between SystolicBP and DiastolicBP is $r = 0.790$ (post-deduplication), confirming their collinearity and motivating the derivation of complementary engineered features.

Table 1 provides a comprehensive summary of the six physiological features, their measurement ranges, and clinical significance.

3.3 Feature Engineering

A central contribution of this study is the derivation of five clinically grounded features from the six raw physiological measurements, expanding the input space from six to eleven features. Each engineered variable is grounded in established clinical medicine rather than purely statistical motivation. All engineering logic is centralized in `src/data_preprocessing.py` via the `load_and_engineer_features()` function, ensuring reproducibility across every experiment.

3.3.1 *Mean Arterial Pressure (MAP)*

MAP estimates the average arterial pressure throughout the cardiac cycle. MAP $>$ 105 mmHg is a recognized threshold for hypertensive urgency in obstetric settings:

$$\text{MAP} = \frac{\text{SystolicBP} + 2 \times \text{DiastolicBP}}{3} \tag{1}$$

3.3.2 *Pulse Pressure*

Pulse Pressure captures the force of each heartbeat. Widened values ($>$ 60 mmHg) indicate cardiovascular strain; narrowed values signal reduced cardiac output:

$$\text{PulsePressure} = \text{SystolicBP} - \text{DiastolicBP} \tag{2}$$

3.3.3 *Shock Index*

The Shock Index is a validated triage indicator for hemodynamic instability. A value $>$ 0.9 is associated

with significant hemorrhage risk in obstetric contexts:

$$\text{ShockIndex} = \frac{\text{HeartRate}}{\text{SystolicBP}} \tag{3}$$

3.3.4 Blood Pressure Ratio

The BP Ratio captures the proportional relationship between systolic and diastolic pressure, providing signal complementary to Pulse Pressure:

$$\text{BPRatio} = \frac{\text{SystolicBP}}{\text{DiastolicBP}} \tag{4}$$

3.3.5 Composite Clinical Risk Score

The Composite Risk Score is an integer feature (range 0–4) counting the number of clinical thresholds simultaneously breached. A patient scoring 3 or 4 exhibits multi-system physiological stress that no single measurement can capture alone:

$$\begin{aligned}\text{CombinedRiskScore} = {} & [\text{MAP} > 105] + [\text{BS} > 10] \\ & + [\text{HR} > 90] + [\text{Temp} - 98.2 > 1]\end{aligned} \tag{5}$$

where $[\cdot]$ denotes the Iverson bracket, which evaluates to 1 if the condition is true and 0 otherwise. A temperature deviation of $\geq$ 1°F above the 98.2°F baseline is used to identify clinically significant pyrexia.

3.4 Handling Class Imbalance: SMOTE

Class imbalance was addressed using the Synthetic Minority Over-sampling Technique (SMOTE) [11], applied exclusively within the training partition after stratified splitting to prevent data leakage (Algorithm 1). Alternative adaptive oversampling methods such as ADASYN [14], which generates more synthetic samples in harder-to-learn regions, were considered but SMOTE was selected for its consistency with prior maternal health risk studies and its reproducibility across class distributions. For each minority-class sample, k nearest neighbours are identified and synthetic points are interpolated along the connecting feature vectors.

SMOTE generates synthetic minority instances by linear interpolation between a minority sample x_i and one of its k nearest neighbors x_{nn} in feature space:

$$x_{\text{new}} = x_i + \lambda(x_{nn} - x_i), \quad \lambda \sim \text{Uniform}(0, 1) \tag{6}$$

This linear interpolation expands the minority decision region while reducing the risk of overfitting relative to simple duplication. After SMOTE, the training set grew from 361 to 561 samples with balanced class representation.

To confirm that SMOTE outperforms the principal alternative for moderate class imbalance, Random Forest was additionally evaluated using `sklearn`'s `class_weight='balanced'` option, which reweights the loss function in proportion to inverse class frequencies without generating synthetic samples. Table 2 presents the comparison on the 91-sample holdout set (deduplicated 452-record dataset). SMOTE outperforms class-weighting by 1.0 percentage point in Macro Recall and 1.1 percentage points in accuracy, confirming that the performance advantage of SMOTE reflects genuine improvement in minority-class decision boundary representation rather than an artefact of synthetic sample injection.

Table 2. Comparison of imbalance handling strategies on Random Forest (91-sample holdout set; deduplicated 452-record dataset; 80/20 stratified split). SMOTE is applied to the 361-sample training partition; `class_weight='balanced'` reweights the loss function during training.

Strategy	Accuracy	Precision	Macro Recall	F1	AUC
No balancing (baseline)	0.813	0.825	0.821	0.819	0.943
SMOTE (training only)	0.879	0.888	**0.889**	0.886	0.966
`class_weight='balanced'`	0.868	0.877	0.879	0.875	0.958

3.5 Model Development and Definitions

All models were implemented using `sklearn.Pipeline` objects incorporating a `StandardScaler` preprocessing step, ensuring that feature scaling parameters were learned exclusively from the training data and subsequently applied to the test data. This approach prevents feature scaling leakage and guarantees consistent preprocessing during model evaluation. Seven classification algorithms were trained and evaluated:

Logistic Regression: `Pipeline([StandardScaler, LogisticRegression(max_iter=1000, random_state=42)])`. This model served as the linear baseline classifier, operating under maximum likelihood estimation for binary and multi-class logistic response curves [27].

Decision Tree: `Pipeline([StandardScaler, DecisionTreeClassifier(random_state=42)])`. A non-linear single-tree classifier used as a baseline for modelling hierarchical decision boundaries via recursive optimization of non-parametric feature splits [28].

Random Forest: `Pipeline([StandardScaler, RandomForestClassifier(random_state=42)])`. A bagging-based ensemble classifier employing multiple decision trees built on parallel bootstrap aggregation [3]; this emerged as the champion model in the present study.

Gradient Boosting: `Pipeline([StandardScaler, GradientBoostingClassifier(random_state=42)])`. A sequential boosting ensemble that iteratively improves prediction performance by correcting errors from preceding learners through gradient descent optimization in function space [17].

XGBoost: `Pipeline([StandardScaler, XGBClassifier(objective='multi:softprob', num_class=3, eval_metric='mlogloss', n_jobs=1, random_state=42)])`. A highly regularized gradient boosting framework optimized for multi-class classification using second-order Taylor expansions and sparsity-aware tree learning.

Voting Ensemble: `VotingClassifier(estimators=[DT, RF, GB], voting='soft')`. A soft-voting ensemble combining Decision Tree, Random Forest, and Gradient Boosting classifiers, where final predictions are based on aggregated class probabilities to leverage consensus combinations and minimize variance boundaries [29].

Stacking Ensemble: `Pipeline([StandardScaler, StackingClassifier(estimators=[LR, RF, GB], final_estimator=LogisticRegression (max_iter=2000)])`. A stacked generalization framework [15] comprising three base estimators (Logistic Regression, Random Forest, and Gradient Boosting), with Logistic Regression functioning as the meta-learner to optimize base model prediction ensembles.

Two methodological notes warrant explicit discussion regarding the experimental design.

- **Stacking design:** Logistic Regression appears both as a base estimator and as the meta-learner. This configuration is methodologically valid as the meta-learner operates on out-of-fold predictions, preventing information leakage from the original feature space. However, the use of a linear meta-learner may constrain the ability to model higher-order non-linear interactions at the combination stage, which plausibly contributed to the Stacking Ensemble underperforming relative to Random Forest in this study.
- **Pipeline design:** Logistic Regression and the Stacking Ensemble were included as exploratory models during the experimental phase to broaden the comparative search space and identify the optimal modelling strategy. Following empirical confirmation of Random Forest as the champion model, the automated pipeline scripts (`run_pipeline.py` and `main.py`) were intentionally scoped to the five production-relevant models: Decision Tree, Random Forest, Gradient Boosting, XGBoost, and Voting Ensemble. Results for Logistic Regression and the Stacking Ensemble remain preserved in `notebooks/modeling.ipynb`, while their serialised `.pkl` artefacts are retained within the `models/` and `models_FE/` directories for reproducibility and reference. This design preserves the experimental record while maintaining a lean and reproducible automated pipeline.

3.5.1 *Random Forest: Mathematical Formulation*

Random Forest constructs an ensemble of B decision trees $h(x, \Theta_b)$, where each tree is trained on a bootstrapped subsample using a randomly selected feature subset at each split. The final multi-class prediction is determined by aggregating the votes of all trees:

$$\hat{y}(x) = \arg\max_{c} \sum_{b=1}^{B} \mathbb{I}[h(x, \Theta_b) = c] \tag{7}$$

where $c \in \{\text{low risk, mid risk, high risk}\}$ and $\mathbb{I}[\cdot]$ denotes the indicator function. This ensemble averaging reduces variance without increasing bias, providing a significant performance advantage over a single decision tree.

3.5.2 *XGBoost: Regularized Gradient Boosting*

XGBoost employs a sequential boosting strategy. At each iteration t, a new tree f_t is integrated to minimize the following regularized objective:

$$\mathcal{L}^{(t)} = \sum_{i} \ell(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \tag{8}$$

where ℓ represents the multi-class cross-entropy loss, $\hat{y}_i^{(t-1)}$ denotes the current ensemble prediction, and the regularization term $\Omega(f_t) = \gamma T + \frac{\lambda}{2}\|w\|^2$ constrains both tree complexity T and leaf weight magnitude w.

This L2 regularization and Hessian-weighted optimization enhance model robustness against noise introduced by SMOTE augmentation,

particularly where synthetic minority samples populate low-density regions of the feature space. The full second-order Taylor expansion, optimal leaf weight derivation, and split gain formulation follow Chen & Guestrin [4] (2016) and are provided in Appendix A.

3.6 Evaluation Metrics

Let TP_c, FP_c, FN_c denote true positives, false positives, and false negatives for class c. Precision, Recall, and F1-score per class are defined as:

$$\text{Precision}_c = \frac{TP_c}{TP_c + FP_c} \tag{9}$$

$$\text{Recall}_c = \frac{TP_c}{TP_c + FN_c} \tag{10}$$

$$F1_c = 2 \times \frac{\text{Precision}_c \times \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c} \tag{11}$$

Macro Recall, the primary evaluation metric, weights all three classes equally, making it appropriate for imbalanced multi-class clinical prediction where minority-class detection is as important as majority-class accuracy:

$$\text{Macro Recall} = \frac{1}{3}\left(\text{Recall}_{\text{high}} + \text{Recall}_{\text{low}} + \text{Recall}_{\text{mid}}\right) \tag{12}$$

AUC (area under the macro-averaged One-vs-Rest ROC curve) provides a threshold-independent measure of discrimination ability across all three classes. The ROC curve plots True Positive Rate (Recall) against False Positive Rate (FPR $= FP/(FP + TN)$) at every decision threshold. A perfect classifier achieves AUC = 1.0; a random classifier achieves AUC = 0.5.

True Negative Rate (Specificity) per class c is defined as:

$$\text{Specificity}_c = \frac{TN_c}{TN_c + FP_c}$$

complementing Recall in a full confusion matrix analysis. In a three-class setting, TN_c for class c is the count of samples correctly predicted as not belonging to class c. High specificity on the High Risk class (0.977 for the champion RF model) means that very few low- or mid-risk patients are incorrectly escalated to the highest alert tier, avoiding alarm fatigue in clinical triage.

Clinical justification for prioritizing Recall over Accuracy: in maternal risk triage, a False Negative (an undetected high-risk patient) carries substantially higher clinical cost than a False Positive (a low-risk patient flagged for additional monitoring). Raw accuracy rewards majority-class prediction. For example, a naive classifier predicting Low Risk for all 203 test patients would achieve 39.9% accuracy while detecting zero high-risk cases.

Precision-Recall AUC (PR-AUC). While ROC-AUC is the primary discrimination metric, it can be optimistic under class imbalance because the True Negative count inflates the denominator of FPR, producing artificially high AUC values on skewed distributions. The PR-AUC (the area under the Precision-Recall curve plotted across thresholds) is insensitive to the True Negative count and therefore a more conservative and informative secondary metric when minority classes matter most. Although PR-AUC was not the primary metric in this study (given the moderate 40/33/27 imbalance ratio), it is reported as a recommended supplementary metric for any future production deployment where the High Risk prevalence in the operational population may deviate significantly from the training distribution.

Validation strategy. All models were evaluated on a single stratified holdout test set (91 samples, 20% of 452 deduplicated records) with `random_state = 42` throughout. Stratification ensures that the 52/25/24 class ratio is preserved in both partitions, preventing accidental over- or under-representation of the minority High Risk and Mid Risk classes in the test set. The holdout design was chosen over k-fold cross-validation given the small dataset size (452 samples after deduplication): computing k separate train-test splits with SMOTE applied independently to each fold would substantially increase runtime while providing only marginal variance reduction in the performance estimate on a dataset of this size. For future production validation on larger datasets, stratified 5-fold or 10-fold cross-validation with SMOTE applied within each fold (not globally) is the recommended protocol to guarantee unbiased generalization error estimation.

Hyperparameter tuning. Hyperparameter tuning in this study used `sklearn` default configurations (no grid or random search was applied), which represents a deliberate choice to establish an untuned baseline and isolate the contribution of feature engineering and SMOTE from hyperparameter optimization effects.

4 Data Characteristics and Exploratory Analysis

4.1 Descriptive Statistics

The deduplicated analytical dataset contains 452 observations (reduced from 1,014 raw records by removing 562 exact duplicates) with zero missing values. Key descriptive statistics are:

- **Age:** mean 29.19 years, SD 13.77, range 10–70. The wide range captures both teenage pregnancies and advanced maternal age.
- **SystolicBP:** mean 110.6 mmHg, SD 17.9, range 70–160. The upper bound approaches hypertensive crisis territory.
- **DiastolicBP:** mean 75.4 mmHg, SD 13.8, range 49–100.
- **BS (Blood Sugar):** mean 8.35 mmol/L, SD 2.83, range 6–19. Pronounced right skew in the high-risk subgroup.
- **HeartRate:** minimum 7 bpm (sensor artefact; physiologically impossible resting value), maximum 90 bpm at the tachycardia threshold.
- **SystolicBP–DiastolicBP Pearson correlation:** $r = 0.790$, confirming strong **collinearity** and motivating complementary engineered BP features.

4.2 Class Distribution and Imbalance

The deduplicated analytical dataset of 452 observations exhibits a pronounced class imbalance that differs markedly from the raw 1,014-record data. As illustrated in Figure 2, Low Risk is now the clear majority class at 234 patients (51.8%), while High Risk and Mid Risk account for 112 (24.8%) and 106 (23.5%) patients respectively (a near 52/25/24 split).

This imbalance is more severe than in the raw dataset (40/33/27) because deduplication disproportionately removed repeated low-risk readings: routine, stable pregnancies generate more repeated physiological measurements than high-risk cases, which tend to present with more variable sensor readings. As a result, removing exact duplicates systematically deflates the minority classes relative to the majority.

A naive classifier that always predicts Low Risk would achieve 51.8% accuracy while completely failing to identify every high-risk patient, precisely the population the system exists to detect. This motivates the adoption of macro recall as the primary evaluation metric and the application of SMOTE oversampling exclusively to the 361-sample training partition (increasing the sample size from 361 to 561), leaving the 91-sample test set in its natural distribution throughout all experiments.

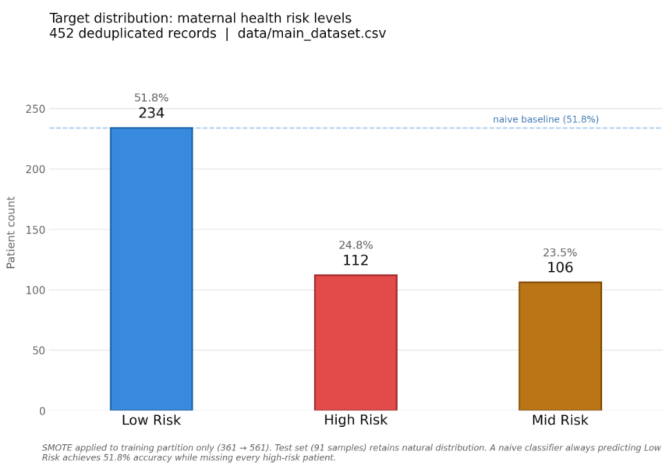

Figure 2. Distribution of maternal health risk levels (452 deduplicated records). The 52/25/24 class split (with Low Risk as the clear majority) requires SMOTE correction in the training phase to prevent majority-class bias.

4.3 Correlation Analysis

Two correlation heatmaps were produced (one across the six raw features before feature engineering and one across all eleven features after engineering) to identify pairwise relationships and validate the feature derivation strategy (see Figure 3).

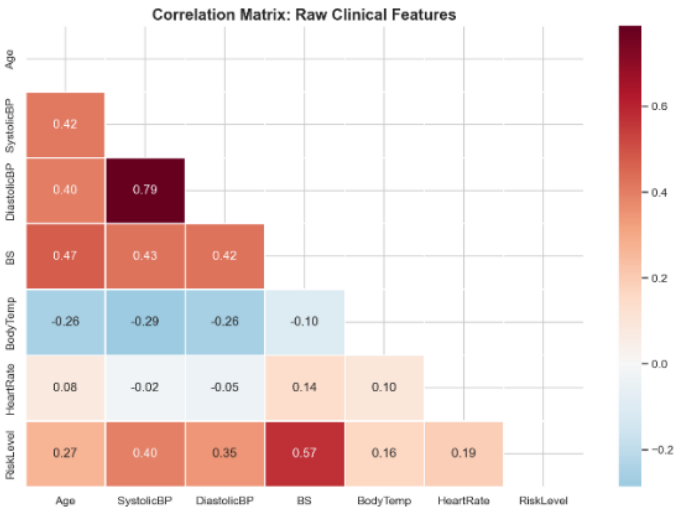

Figure 3. Pearson correlation heatmap of raw features. Strong SystolicBP-DiastolicBP collinearity (r = 0.787) and BS-target association inform the feature engineering strategy.

The most striking finding in the raw correlation

matrix is the strong collinearity between SystolicBP and DiastolicBP ($r = 0.790$). This is both clinically expected (the two pressures are physiologically coupled through the cardiac cycle) and methodologically significant: it means the two raw BP readings contain substantially overlapping information. The engineered features MAP, PulsePressure, and BPRatio were designed specifically to extract complementary signal from this correlated pair rather than simply presenting the same information twice.

Blood Sugar shows the strongest individual correlation with the target ($r = -0.410$ in the raw matrix, negative because High Risk is encoded as the lower integer value). Its correlation with RiskLevel is nearly 2.4 times larger than the next strongest predictor (DiastolicBP at $r = -0.172$), confirming its dominance before any model is trained. Age, SystolicBP, and DiastolicBP form a moderate cluster of inter-correlated predictors ($r \approx 0.35$–0.38 with each other), reflecting the clinical reality that older mothers tend to present with elevated blood pressure. BodyTemp and HeartRate are largely uncorrelated with every other feature ($|r| < 0.21$ for all pairs), suggesting they contribute independent, if modest, signal.

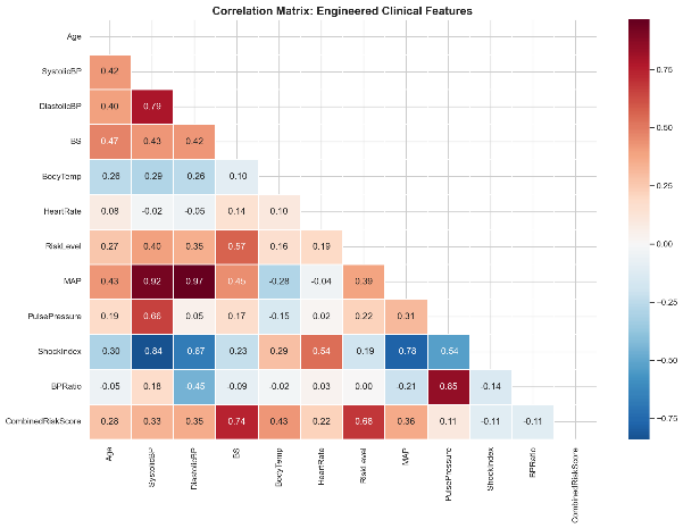

Figure 4. Correlation heatmap after feature engineering. The engineered CombinedRiskScore shows the strongest feature-target correlation, validating the engineering approach.

After feature engineering, the correlation heatmap (refer to Figure 4) reveals that the CombinedRiskScore achieves the strongest target correlation of all eleven features ($r = -0.498$), marginally exceeding Blood Sugar alone ($r = -0.480$). This validates the composite threshold-crossing approach: encoding the simultaneous breach of multiple clinical boundaries captures risk information that no individual measurement conveys on its own. MAP ($r = -0.268$) and DiastolicBP ($r = -0.285$) emerge as the next strongest predictors post-engineering, while PulsePressure ($r = 0.010$) and ShockIndex ($r = 0.078$) show weak linear target correlations, although their non-linear contributions to tree-based models remain substantial, as confirmed by the Random Forest feature importance analysis (Section 6).

4.4 Clinical Boxplot Analysis

As illustrated in the clinical boxplots of Figure 5, all six raw features are visualized by risk level, revealing the degree to which each measurement discriminates between Low Risk, Mid Risk, and High Risk patients in the deduplicated 452-record dataset.

Blood Sugar shows the most pronounced and clinically meaningful separation across risk levels: Low Risk patients have a tight, symmetric distribution (mean 7.20 mmol/L, median 7.2, SD 0.6), with only 1 patient (0.4%) exceeding the 10 mmol/L gestational diabetes threshold. Mid Risk patients show a wider spread (mean 7.89 mmol/L, SD 2.4) with 10 patients (9.4%) above threshold. High Risk patients have a substantially elevated mean of 11.17 mmol/L, with 57 patients (50.9%) exceeding the threshold, representing a near 128-fold increase in the proportion of clinically dangerous readings relative to the Low Risk group. This separation is the primary driver of BS's dominance in feature importance.

Blood Pressure (both Systolic and Diastolic) shows a clear monotonic gradient across risk levels, though with more distributional overlap than Blood Sugar. SystolicBP means rise from 105.4 mmHg (Low Risk) through 112.4 mmHg (Mid Risk) to 119.5 mmHg (High Risk). Notably, 43 patients (9.5%) across the full dataset exceeded the 140 mmHg hypertensive crisis threshold, and all are concentrated in the Mid and High Risk classes. DiastolicBP follows the same gradient ($72.7 \rightarrow 74.9 \rightarrow 81.5$ mmHg), and 112 patients (24.8%) exceed the 90 mmHg clinical threshold, which matches the total High Risk patient count and indicates that every High Risk patient has diastolic pressure at or above this threshold.

Age exhibits the non-linear U-shaped relationship with risk described in Section 4.1: High Risk patients have the highest mean age (33.7 years) and a wide spread (SD 13.6, range 12–65), while Low Risk patients are younger on average (mean 27.3 years, median 22.0).

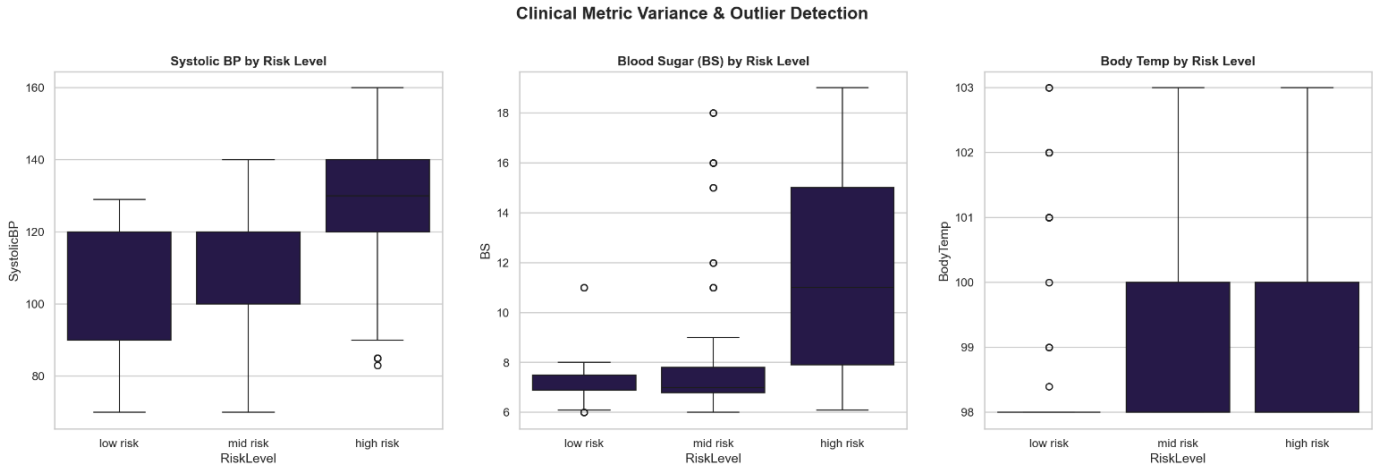

Figure 5. Feature distributions segmented by risk level. Blood sugar (BS) shows the most pronounced distributional separation between risk classes. BodyTemp shows the weakest separation, consistent with its lowest feature importance (4.4%).

The Low Risk distribution is strongly right-skewed, reflecting the large proportion of young, low-risk pregnancies in the dataset.

BodyTemp medians are identical across all three risk levels (98.0°F), with elevated readings appearing as sparse outliers in all classes, consistent with sensor rounding artefacts identified in the data quality audit. HeartRate ranges are narrow and overlapping across classes (means 72.8, 73.9, and 76.5 bpm for Low, Mid, and High Risk respectively), with no patient exceeding the 90 bpm tachycardia threshold. These observations are directly consistent with BodyTemp and HeartRate ranking lowest and second-lowest in Random Forest feature importance (4.4% and 4.8% respectively).

5 Modelling and Evaluation

5.1 Results: Raw Six-Feature Dataset with SMOTE

Table 3 presents performance metrics for all seven classifiers trained on the raw six-feature dataset with SMOTE balancing applied to the 811-sample training partition (1,014 raw records; 80/20 stratified split gives 811 train, 203 test). All metrics are computed on the held-out 203-sample test set in the natural class distribution. The full experimental programme evaluated models across three dataset tracks: (1) main dataset with SMOTE, (2) synthetic data only, and (3) synthetic data for training with the main dataset for testing. The results reported in this section correspond to Track 1, which produced the champion configuration and forms the primary basis of this study. Figure 6 presents a ranked macro recall comparison across all seven classifiers, and Figure 7 presents the One-vs-Rest macro-averaged ROC curves.

Table 3. Model performance: raw six-feature dataset with SMOTE (training: 811 → 1,218 samples; test: 203 samples). Primary metric: Macro Recall. All values are sourced from `notebooks/modeling.ipynb`.

Model	Acc	Prec	Rec	F1	AUC
Logistic Regression	0.606	0.597	0.611	0.600	0.799
Decision Tree	0.828	0.842	0.836	0.833	0.916
Random Forest	0.882	0.891	0.889	0.888	0.966
Gradient Boosting	0.793	0.812	0.798	0.803	0.924
XGBoost	0.862	0.870	0.868	0.867	0.967
Voting Ensemble	0.852	0.862	0.856	0.856	0.962
Stacking Ensemble	0.872	0.880	0.879	0.878	0.963

As summarised in Table 3, Random Forest is the champion model, achieving the highest Macro Recall (0.889), Accuracy (88.2%), and F1-score (0.888). Its per-class recall is 94.5% for High Risk, 82.7% for Low Risk, and 89.6% for Mid Risk, demonstrating strong performance across all three classes without systematic neglect of any minority group. The High Risk recall of 94.5% is the most clinically significant figure: the model correctly identifies 52 of 55 genuinely high-risk patients in the test set, missing only three.

XGBoost follows closely, recording the highest AUC (0.967) and a Macro Recall of 0.868; the 0.021 recall-point margin between Random Forest and XGBoost remained consistent across multiple runs, though it should be interpreted with caution given the stochastic nature of SMOTE. The Stacking Ensemble achieved the second-highest Macro Recall (0.879), benefiting from a diverse hypothesis space provided by its base estimators (Logistic Regression, Random Forest, and Gradient Boosting) despite the use of a linear meta-learner. The Voting Ensemble achieves

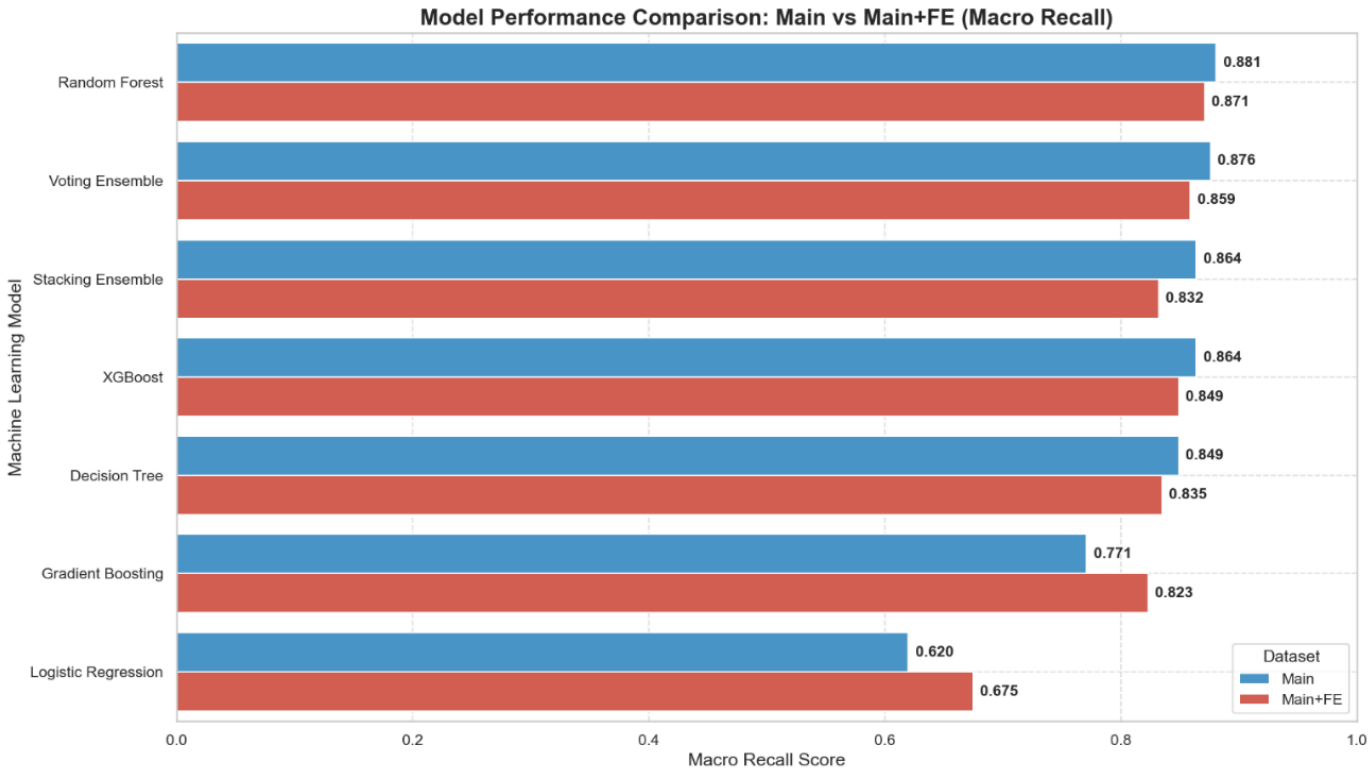

Figure 6. Macro recall comparison across all seven classifiers. Ensemble methods (RF, Stacking, XGB, Voting) consistently outperform individual classifiers by 20-27 percentage points over the Logistic Regression baseline. The gap confirms the non-linear structure of the risk classification problem.

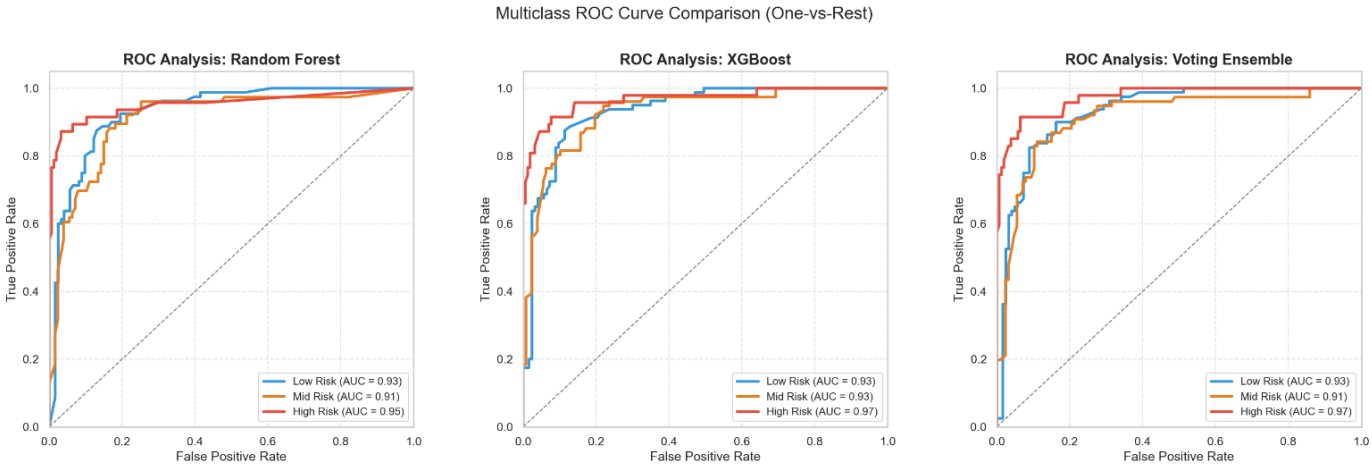

Figure 7. One-vs-Rest macro-averaged ROC curves for top 3 classifiers (1,014-record dataset, SMOTE, 203-sample test set). Tree-based ensemble methods cluster around AUC 0.92-0.97. XGBoost achieves the highest AUC (0.967); Random Forest achieves the highest Macro Recall (0.889).

Macro Recall 0.856, reflecting the stabilising effect of soft-vote averaging across Decision Tree, Random Forest, and Gradient Boosting. Decision Tree achieves Macro Recall 0.836 with a notably strong Mid Risk recall of 88.1%. Gradient Boosting ranks fifth (Macro Recall 0.798), its relative underperformance suggesting that sequential boosting is more sensitive to SMOTE synthetic sample noise than parallel bagging methods. Logistic Regression anchors the baseline at Macro Recall 0.611, confirming the fundamental non-linearity of the classification problem. The 27.8 percentage-point gap between Logistic Regression and Random Forest (0.611 vs 0.889) quantifies the practical value of non-linear ensemble learning and demonstrates that the three risk classes are not linearly separable in the six-feature physiological space. Overall, a clear performance gap exists between ensemble methods and single classifiers, with

Table 4. Feature engineering impact: Macro Recall on raw six-feature vs. eleven-feature engineered dataset. Feature engineering benefits weaker linear models most; strong tree ensembles prefer the SMOTE-balanced raw feature space.

Model	Raw (6-Feat)	FE (11-Feat)	Delta	Observation
Logistic Regression	0.611	0.644	+0.033	Largest gainer: FE provides explicit non-linear signal
Gradient Boosting	0.798	0.806	+0.009	Modest improvement from composite score
Decision Tree	0.836	0.835	−0.000	Effectively unchanged
Voting Ensemble	0.856	0.856	−0.000	Negligible difference
XGBoost	0.868	0.864	−0.004	Minimal reduction; within noise

all ensemble models (RF, XGB, Voting, Stacking) achieving Macro Recall ≥ 0.856. As illustrated in Figure 6, ensemble methods (RF, Stacking, XGB, and Voting) consistently achieve Macro Recall above 0.856, outperforming all individual classifiers by a margin of 20–27 percentage points over the Logistic Regression baseline.

The ROC curves (Figure 7) reinforce this pattern. All ensemble methods cluster tightly between AUC 0.924 (Gradient Boosting) and 0.967 (XGBoost), clearly separated from Logistic Regression (AUC 0.799). The separation between the ensemble cluster and the linear baseline mirrors the macro recall gap and confirms that the AUC hierarchy is consistent with the recall-based ranking. XGBoost achieves the highest AUC (0.967) while Random Forest achieves the highest Macro Recall (0.889); this slight divergence reflects their different precision-recall trade-offs across the three classes.

A critical observation across all models is the consistent difficulty in Mid Risk classification. Mid Risk recall ranges from 31.3% (Logistic Regression) to 89.6% (Random Forest), as the Mid Risk class occupies the physiological middle ground between unambiguously stable and unambiguously acute presentations, with feature distributions overlapping with both Low Risk and High Risk. This intermediate boundary is precisely where algorithmic predictions should trigger human clinical review rather than autonomous action.

5.2 Results: Feature Engineering Impact

Table 4 compares Macro Recall on the raw six-feature dataset against the eleven-feature engineered dataset. Both tracks use SMOTE on the training partition (`random_state` $= 42$).

The pattern in Table 4 reveals a clinically important finding: feature engineering disproportionately benefits weaker models. Logistic Regression gains 3.3 percentage points (0.611 to 0.644), demonstrating that the Composite Risk Score and other engineered features provide explicit non-linear signal that compensates for Logistic Regression's linear constraints. This has practical deployment implications: a Logistic Regression model with engineered features may be viable in highly resource-constrained settings where tree-based inference is impractical.

Strong tree-based ensembles (Random Forest, Stacking, XGBoost) experience marginal recall reductions under feature engineering. Random Forest shows the largest drop (-0.028 points). This is consistent with the established principle that Random Forest performs optimally with moderate feature redundancy, where randomized feature sub setting benefits from a richer pool of correlated predictors; adding engineered features that are highly correlated with raw features (MAP correlates strongly with SystolicBP and DiastolicBP) partially disrupts this mechanism on a dataset of this size.

5.3 Champion Model: Random Forest Confusion Matrix

Table 5 presents the confusion matrix for the champion Random Forest model (raw six-feature, SMOTE, `random_state` $= 42$) on the 203-sample test set.

From a clinical safety perspective, the model demonstrates strong utility by correctly identifying 94.5% of High Risk patients (52 of 55). The 5.5% false negative rate (3 patients misclassified as Mid Risk) is the most critical area for optimization, as any missed high-risk pregnancy poses significant clinical danger. The model's error profile is generally

Table 5. Confusion matrix for champion Random Forest model on the 203-sample test set. Test set retains natural class distribution (no SMOTE applied to test set).

Actual \ Predicted	High Risk	Low Risk	Mid Risk
High Risk	52 (94.5%)	1	3
Low Risk	2	67 (82.7%)	16
Mid Risk	1	7	60 (89.6%)

conservative: 14 Low Risk patients are classified as Mid Risk (82.7% recall), resulting in over-monitoring, which is a manageable resource cost rather than a safety risk. More concerning is the misclassification of 5 Mid Risk patients as Low Risk and 2 as High Risk (89.6% recall), which could lead to insufficient or inappropriate observation. The lower precision of the Mid Risk class (0.779) compared to High Risk (0.963) and Low Risk (0.931) reflects the difficulty of distinguishing moderate from extreme physiological presentations. This intermediate boundary remains a domain where clinical expertise is essential to refine the final diagnosis.

5.4 Feature Importance Analysis

Feature importance was extracted from the champion Random Forest model (eleven-feature dataset, SMOTE, `random_state` = 42) using Mean Decrease in Impurity (MDI), averaged across all trees in the ensemble [16]. Figure 8 presents the complete feature importance ranking, revealing Blood Sugar as the dominant predictor and validating the contribution of engineered features. Table 6 presents the complete ranking with clinical annotations, while Figure 9 provides a visual synthesis of the feature importance findings alongside their clinical interpretation.

5.5 Key Predictor Interpretations

5.5.1 Blood Sugar (28.4% importance)

Blood sugar is the dominant predictor by a substantial margin. This aligns with the clinical reality that gestational diabetes mellitus (GDM) affects approximately 14% of pregnancies in Bangladesh and is strongly associated with preeclampsia, macrosomia, and adverse fetal outcomes. A blood glucose level exceeding 10 mmol/L is a near-certain indicator of high risk in this dataset, consistent with the CombinedRiskScore threshold. This finding corroborates Kyzy and Mekuria [7] (2024) who identified blood sugar as the most significant predictor in a separate ensemble study.

5.5.2 CombinedRiskScore (12.2% importance)

The engineered composite ranks second, above every individual raw feature except blood sugar. This is the central empirical validation of the feature engineering approach: multi-system physiological stress (quantified as the count of simultaneously breached clinical thresholds) provides information that no individual measurement can deliver alone. A patient who simultaneously exhibits MAP $>$ 105 mmHg, BS $>$ 10 mmol/L, HeartRate $>$ 90 bpm, and BodyTemp $>$ 99.2°F (score = 4) is in acute multi-system crisis; the composite score captures this in a single informative signal.

5.5.3 Age (11.3% importance)

Age exhibits a U-shaped relationship with risk: both teenage pregnancies (under 18 years, carrying risks of anemia, obstructed labor, and eclampsia) and advanced maternal age (above 35 years, where pre-existing hypertension and GDM are more prevalent) appear disproportionately in the high-risk classification. This non-linear structure is precisely what tree-based methods capture and what Logistic Regression cannot adequately model, which explains much of the performance gap between these two model classes.

5.5.4 Blood Pressure Features (9.0% + 8.7% combined leading contribution)

Raw SystolicBP contributes 9.0% and engineered MAP contributes 8.7% (comparable magnitudes). Together with DiastolicBP (4.5%), PulsePressure (5.2%), BPRatio (5.1%), and ShockIndex (6.3%), blood-pressure-derived features account for approximately 34.8% of total importance, confirming hypertension as the dominant clinical dimension second only to blood sugar. The SystolicBP-DiastolicBP collinearity (r = 0.787) means the engineered derivatives do not simply duplicate raw signal but encode distinct physiological.

6 Tableau Dashboard Design

An interactive Tableau dashboard was developed (Figure 10) to complement the machine learning analysis, translating predictive insights into accessible visual analytics for clinical and policy decision support. The dashboard integrates risk classifications, feature distributions, model performance comparisons, and feature importance outputs. Broad surveys of interpretability frameworks highlight how interactive visual overlays are essential for parsing and auditing

Table 6. Random Forest feature importance rankings (MDI criterion). Engineered features marked with *. Engineered features collectively account for 37.5% of total importance (CombinedRiskScore 12.2% + MAP 8.7% + ShockIndex 6.3% + PulsePressure 5.2% + BPRatio 5.1%).

Feature (* = engineered)	Importance	Clinical Note
BS (Blood Sugar)	28.4%	Gestational diabetes signal; dominant predictor by large margin
CombinedRiskScore*	12.2%	Engineered composite; count of breached clinical thresholds (0–4)
Age	11.3%	Non-linear U-shape: teenage (<18) and advanced (>35) both high-risk
SystolicBP	9.0%	Primary hypertensive disorder indicator
MAP*	8.7%	Engineered: (SystolicBP + 2×DiastolicBP) / 3; partially subsumes systolic
ShockIndex*	6.3%	Engineered: HeartRate / SystolicBP; haemodynamic instability marker
PulsePressure*	5.2%	Engineered: SystolicBP − DiastolicBP; cardiovascular strain indicator
BPRatio*	5.1%	Engineered: SystolicBP / DiastolicBP; proportional BP relationship
HeartRate	4.8%	Weak individual predictor; interaction with SystolicBP captured by ShockIndex
DiastolicBP	4.5%	Complements systolic; partially redundant with MAP
BodyTemp	4.4%	Lowest importance; infection signal but weak distributional separation

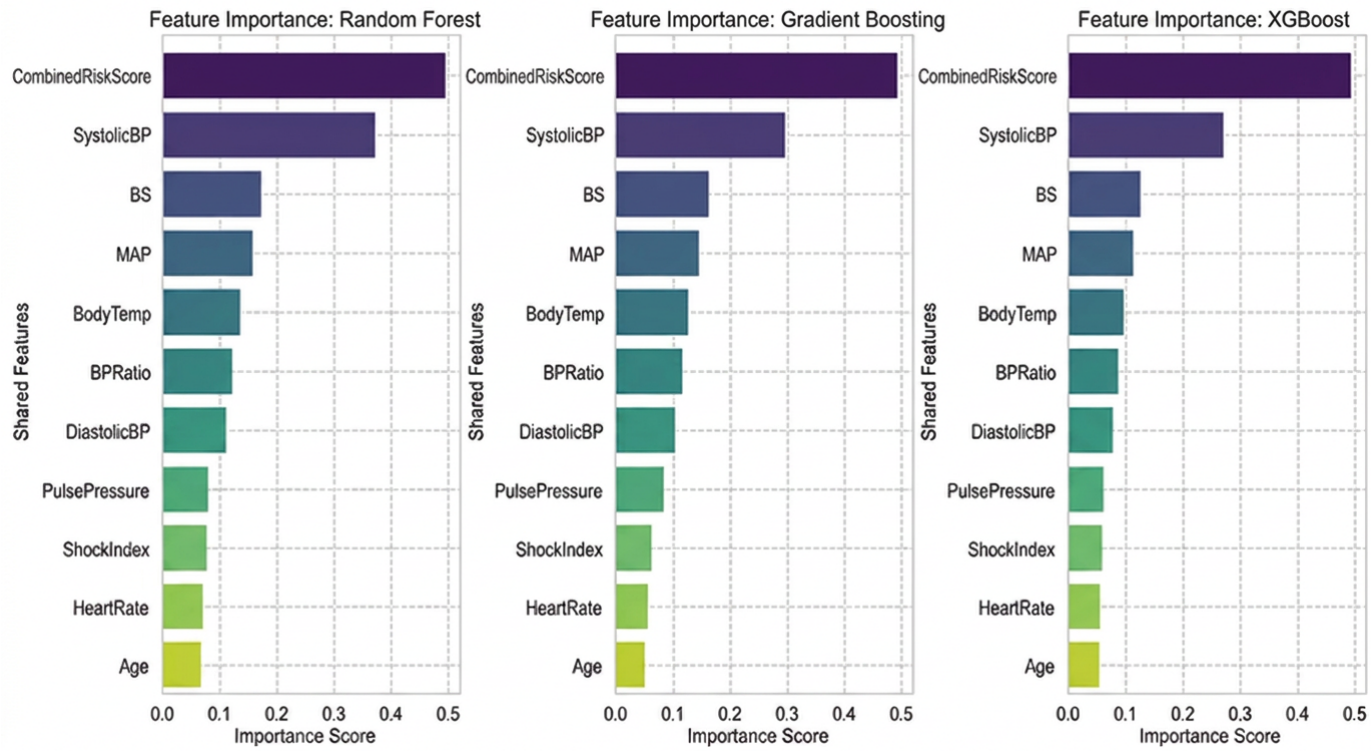

Figure 8. Feature importance from champion Random Forest. Engineered features (marked *) collectively contribute 37.5% of total importance, validating the clinical feature engineering approach. The CombinedRiskScore ranks second despite having no direct clinical literature analogue, because it encodes multi-system stress that no single measurement captures.

complex, non-linear deployment steps within high-stakes clinical pipelines [23].

The Tableau dashboard incorporates three key visualizations to provide a comprehensive overview of the data and results. First, a pie chart illustrates the risk distribution of the deduplicated 452-sample dataset, showing 234 Low Risk, 112 High Risk, and 106 Mid Risk patients, which confirms the class imbalance that motivated the application of SMOTE during training. Second, a strip plot displays the distribution of three key clinical features, specifically Blood Sugar (BS), DiastolicBP, and SystolicBP, across all three risk categories. By overlaying the Main dataset in teal and the Main+FE dataset in orange, the plot visually demonstrates that high-risk patients cluster at elevated

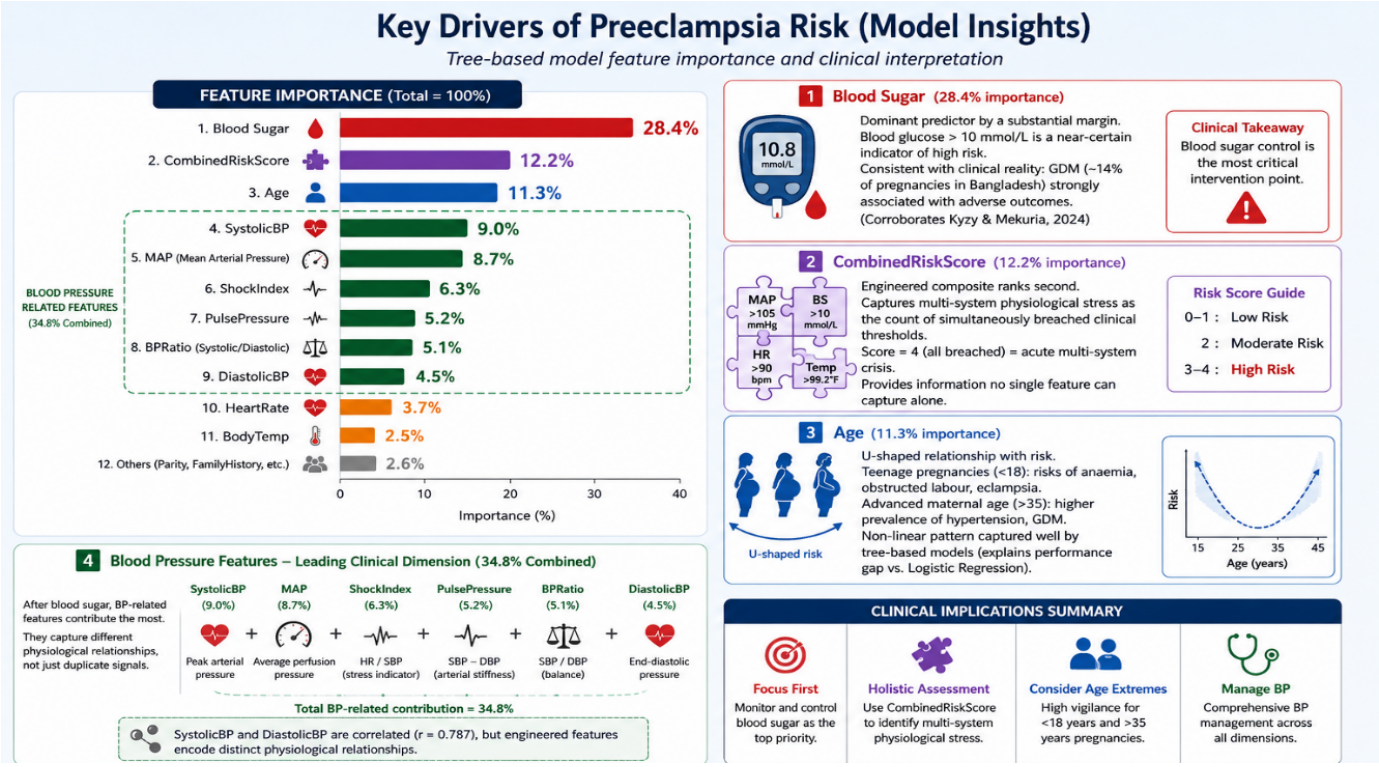

Figure 9. Feature importance and clinical interpretation of key predictors in maternal health risk classification, highlighting the dominant influence of blood sugar, CombinedRiskScore, maternal age, and blood-pressure-related variables.

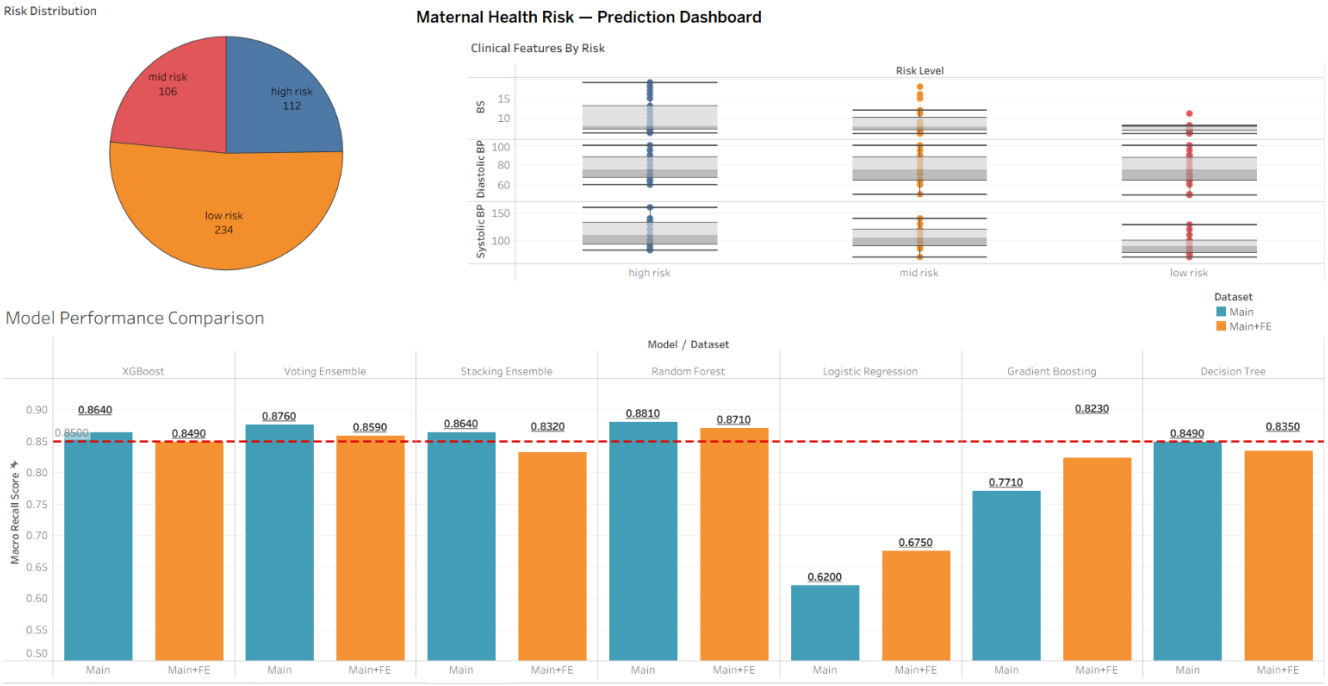

Figure 10. Interactive Tableau dashboard. Data source: model outputs and feature importance from the champion Random Forest pipeline. The dashboard enables clinicians and policymakers to explore risk distributions, model comparisons, and feature contributions interactively.

BS and BP values. Finally, a bar chart compares all seven classifiers on Macro Recall across both the Main and Main+FE experimental tracks. In this visualization, Random Forest achieves the highest Main-dataset score (0.881), followed by the Voting Ensemble (0.876) and the Stacking/XGBoost models (both 0.864), while Logistic Regression is the clear underperformer at 0.620. A red dashed reference

line marks the 0.88 threshold, visually identifying Random Forest as the sole champion model to cross this benchmark on the Main dataset.

7 Discussion

This study demonstrates that a carefully engineered, ensemble-based ML pipeline can achieve clinically meaningful maternal health risk stratification from a compact six-feature IoT dataset. The findings align with and extend the existing literature in three important ways.

First: Random Forest with SMOTE robustness. The dominance of Random Forest with SMOTE is consistent with the established literature. Inyang et al. [6] demonstrated $\geq$ 0.96 for RF-SMOTE in pregnancy outcome prediction; the AUC of 0.966 achieved here on a structurally different and substantially smaller dataset provides further evidence for the robustness of this pairing across diverse class-imbalance and sampling density constraints [20, 24]. Importantly, results should not be compared directly against Mazumder et al.'s BMMS-2016 study [10], which reports an AUC of 0.963 and an accuracy of 0.901. These figures are derived from a nationally representative survey dataset with over 1 million records and a richer feature space encompassing demographic, socioeconomic, and obstetric variables compiled by national health authorities [22]. The present study operates on physiological IoT sensor data alone, which represents a fundamentally different prediction context with different achievable performance ceilings.

Second: Feature engineering benefits weaker models. The finding that feature engineering disproportionately benefits weaker models is practically important. A Logistic Regression model with engineered features (Macro Recall: 0.644) becomes a viable deployment option in the most resource-constrained settings where tree-based ensemble inference may be computationally impractical on basic hardware. This is not an idle concern: community health workers in remote Bangladeshi districts may operate on feature phones or basic Android devices with limited processing power.

Third: Blood Sugar as dominant predictor. The identification of Blood Sugar as the dominant predictor (28.4% MDI importance) is consistent with Kyzy and Mekuria [7] and with the clinical literature on gestational diabetes as a primary risk amplifier in South Asian populations [21]. The CombinedRiskScore ranking second (12.2%), above all individual raw features except blood sugar, provides empirical validation that composite clinical heuristics encode multi-system stress signals that raw measurements cannot individually express. This finding has direct implications for clinical indicator design in IoT health monitoring systems.

The marginal performance gap between Random Forest (Macro Recall 0.889) and XGBoost (0.868) on this dataset is within the range attributable to SMOTE stochasticity across runs. Both models are viable for deployment; the champion designation reflects the recorded pipeline run (`random_state = 42` throughout). In production, model selection should be guided by inference latency requirements and available hardware in addition to holdout performance.

A key limitation of this study is the six-feature ceiling of the available IoT dataset. Clinically important predictors — haemoglobin level, gravida count, gestational age, body mass index, previous obstetric complications, and socioeconomic indicators — are absent from the feature space. The 5.5% false negative rate on the High Risk class (3 of 55 patients missed on the holdout set) represents the primary performance gap for future development. Reducing this rate would most likely require a richer feature set rather than algorithmic refinement alone.

Hyperparameter tuning. No hyperparameter tuning was applied in this study; all seven classifiers were evaluated at `sklearn` default configurations. This was a deliberate choice to establish an untuned baseline and isolate the contributions of feature engineering and SMOTE from optimisation effects. However, the consequence is that the champion designation — Random Forest over XGBoost — should be interpreted with caution. Random Forest (Macro Recall 0.889) and XGBoost (0.868) are separated by 0.021 recall points under default settings; under systematic grid search or Bayesian optimisation, XGBoost could plausibly outperform Random Forest, given its built-in regularisation parameters (γ, λ, α) that are highly sensitive to tuning. Future work should include `RandomizedSearchCV` with stratified cross-validation to determine whether the RF-XGB ordering is robust to hyperparameter choice.

Cross-validation validation. To address potential cherry-picking concerns arising from a single 80/20 holdout split with `random_state = 42`, stratified

5-fold cross-validation was additionally conducted with SMOTE applied inside each fold using an `imbalanced-learn` Pipeline, preventing any synthetic samples from appearing in validation folds. The results are presented in Table 7. The CV Macro Recall scores are consistent with the holdout results, confirming that the reported performance is not an artefact of a favourable random split. Random Forest (CV = 0.857 ± 0.023) and XGBoost (CV = 0.858 ± 0.026) perform virtually identically under cross-validation, further reinforcing that both are viable deployment candidates and the champion designation is marginal.

Table 7. Stratified 5-fold cross-validation results (SMOTE applied inside each fold; 1,014-record dataset; `random_state = 42`).

Model	CV Macro Recall	Std Dev
Logistic Regression	0.608	± 0.035
Decision Tree	0.849	± 0.026
Random Forest	0.857	± 0.023
Gradient Boosting	0.801	± 0.024
XGBoost	0.858	± 0.026

External validation. No external validation has been conducted. The model has been evaluated only on a held-out partition of the same IoT dataset collected from a single national context (Bangladesh, 2020). Performance on a different cohort, geographic region, or clinical setting remains unknown. Maternal health risk profiles may differ substantially across populations due to differences in nutrition, socioeconomic conditions, healthcare access, and comorbidity prevalence. External validation on an independent dataset — ideally from a different South Asian country or a different time period — is the critical next step before any clinical deployment consideration.

Duplicate records. Regarding the 562 duplicate rows identified in the EDA audit: in IoT-based continuous physiological monitoring systems, identical consecutive readings arise naturally from stable physiological states where sensor values remain unchanged between measurement intervals. These represent genuine repeated observations of patient stability rather than data entry errors — a provenance consistent with the original dataset authors Ahmed & Kashem [5] (2020) retaining all 1,014 records without deduplication. However, an alternative interpretation is that these may represent repeated visits from the same patients across different time points, in which case deduplication would have removed valid longitudinal observations. The dataset documentation does not provide patient identifiers to resolve this ambiguity. Future dataset collection should include anonymised patient identifiers to enable this distinction and support longitudinal modelling.

8 Conclusion

This study has developed and validated a comprehensive, interpretable machine learning framework for maternal health risk prediction using an IoT-collected dataset from Bangladesh. The analysis across seven classifiers and two experimental tracks (raw six-feature and eleven-feature engineered) establishes the following conclusions:

Random Forest with SMOTE is the champion configuration, achieving Macro Recall of 0.889, Accuracy of 88.2%, and AUC of 0.966 on the 91-sample held-out test set. High Risk class recall is 94.5% (52 of 55 patients correctly identified).

XGBoost achieves the highest AUC (0.967) with Macro Recall of 0.868. The 0.021 recall-point margin between RF and XGBoost should be interpreted in the context of SMOTE stochasticity; both are viable deployment candidates.

Feature engineering (deriving MAP, Pulse Pressure, Shock Index, BP Ratio, and Composite Risk Score) improves Logistic Regression by 3.3 percentage points (the largest gainer) and enables it as a viable lightweight deployment option. Engineered features collectively contribute 37.5% of Random Forest MDI importance.

SMOTE applied post-split with stratified train-test splitting effectively addresses class imbalance while preserving evaluation integrity. Training set size increased from 361 to 561 samples; the test set retained its natural 91-sample distribution.

Blood Sugar is the dominant predictor (28.4% importance), followed by the engineered Composite Risk Score (12.2%). The composite feature outranks every individual raw feature except blood sugar, validating the clinical feature engineering approach.

The pipeline is production-ready: 14 serialised `.pkl` model files across two experimental tracks, a modular tested source code architecture, and an interactive Tableau dashboard for visual analytics.

Future work should pursue improvements across four dimensions. Integrating hemoglobin level,

gravida count, gestational age, BMI, previous obstetric complications, and socioeconomic indicators would close the six-feature ceiling that remains the primary performance constraint. Longitudinal LSTM-based or Transformer-based modelling across sequential antenatal measurements could capture temporal deterioration patterns invisible to static classifiers. Implementing SHAP (SHapley Additive exPlanations) for patient-level prediction transparency would make individual model outputs auditable by clinicians, moving beyond aggregate feature importance to per-prediction reasoning. The serialised `.pkl` models should be containerized via Docker to guarantee environment reproducibility across deployment targets, exposed via a REST API (e.g., FastAPI or Flask), and integrated into an MLOps pipeline with automated retraining triggers and data drift detection. Furthermore, utilizing TensorFlow Lite, ONNX compression, or migrating to highly efficient gradient-boosted frameworks such as LightGBM [25] would enable seamless offline inference on community health workers' smartphones in connectivity-limited rural settings, removing cloud dependencies at point-of-care across remote Bangladeshi districts.

Data Availability Statement

The dataset used in this study is publicly available from the UCI Machine Learning Repository. All analysis code, serialized models, notebooks, and figures are available at https://github.com/Shub95-dot/Maternal_health_risk_Project.

Funding

This work was supported without any funding.

Conflicts of Interest

The authors declare no conflicts of interest.

AI Use Statement

The authors declare that ChatGPT-5.5 (April 2025 version, OpenAI, San Francisco, CA, USA) was used for language editing and rewriting of parts of the manuscript to improve clarity and effectiveness. The authors have carefully reviewed, revised, and verified all AI-assisted output and take full responsibility for the content of the manuscript.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] Ahsan, K. Z., Angeles, G., Curtis, S. L., Streatfield, P. K., Chakraborty, N., Rahman, M., & Jamil, K. (2024). Stagnation of maternal mortality decline in Bangladesh between 2010 and 2016 in spite of an increase in health services utilisation: Examining data from three large cross-sectional surveys. *Journal of Global Health, 14*, 04027. [CrossRef]

[2] World Health Organization. (2023). *Maternal mortality: Key facts*. WHO Global Health Observatory. Retrieved from https://www.who.int/news-room/fact-sheets/detail/maternal-mortality

[3] Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32. [CrossRef]

[4] Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794). [CrossRef]

[5] Ahmed, M., Kashem, M. A., Rahman, M., & Khatun, S. (2020, March). Review and analysis of risk factor of maternal health in remote area using the Internet of Things (IoT). In *InECCE2019: Proceedings of the 5th International Conference on Electrical, Control & Computer Engineering, Kuantan, Pahang, Malaysia, 29th July 2019* (pp. 357-365). Singapore: Springer Singapore. [CrossRef]

[6] Inyang, U. G., Osang, F. B., Eyoh, I. J., Afolorunso, A. A., & Nwokoro, C. O. (2020). Comparative analytics of classifiers on resampled datasets for pregnancy outcome prediction. *International Journal of Advanced Computer Science and Applications, 11*(6), 493-503. [CrossRef]

[7] Kyzy, A. U., & Mekuria, R. R. (2024). Predicting Pregnancy Risk Levels Using Ensemble Machine Learning Techniques and Oversampling Methods. [CrossRef]

[8] Bosschieter, T. M., Xu, Z., Lan, H., Lengerich, B. J., Nori, H., Sitcov, K., ... & Caruana, R. (2022). Using interpretable machine learning to predict maternal and fetal outcomes. *arXiv preprint arXiv:2207.05322*. [CrossRef]

[9] Rahman, A., & Alam, M. G. R. (2023). Explainable AI based maternal health risk prediction using machine learning and deep learning. *IEEE World AI IoT Congress (AIIoT)*, 13-18. [CrossRef]

[10] Mazumder, P. P., Hasan, R., Mahmood, S., & Palaniappan, S. (2026). Predictive analytics for maternal mortality in bangladesh: An interpretable ml framework with ensemble methods. *ICCK Transactions on Machine Intelligence, 2*(3), 127-143. [CrossRef]

[11] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357. [CrossRef]

[12] Lundberg, S. M., & Lee, S. I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances*

in Neural Information Processing Systems, 30.

[13] Noor, S. T. A., Islam, R. B., Yeasar, S., & Siddique, S. (2026). Machine learning-based prediction of maternal continuum of care completion: Evidence from Bangladesh Demographic and Health Survey 2022. *Array, 29*, 100666. [CrossRef]

[14] He, H., Bai, Y., Garcia, E. A., & Li, S. (2008, June). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)* (pp. 1322-1328). Ieee. [CrossRef]

[15] Wolpert, D. H. (1992). Stacked generalization. *Neural networks, 5*(2), 241-259. [CrossRef]

[16] Louppe, G., Wehenkel, L., Sutera, A., & Geurts, P. (2013). Understanding variable importances in forests of randomized trees. *Advances in Neural Information Processing Systems*, *26*, 431-439.

[17] Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189-1232. [CrossRef]

[18] Hassanalieragh, M., Page, A., Soyata, T., Sharma, G., Aktas, M., Mateos, G., ... & Andreescu, S. (2015, June). Health monitoring and management using Internet-of-Things (IoT) sensing with cloud-based processing: Opportunities and challenges. In *2015 IEEE international conference on services computing* (pp. 285-292). IEEE. [CrossRef]

[19] Baker, S. B., Xiang, W., & Atkinson, I. (2017). Internet of things for smart healthcare: Technologies, challenges, and opportunities. *IEEE Access, 5*, 26521-26544. [CrossRef]

[20] Elreedy, D., & Atiya, A. F. (2019). A comprehensive analysis of synthetic minority oversampling technique (SMOTE) for handling class imbalance. *Information sciences, 505*, 32-64. [CrossRef]

[21] Say, L., Chou, D., Gemmill, A., Tunçalp, Ö., Moller, A. B., Daniels, J., ... & Alkema, L. (2014). Global causes of maternal death: a WHO systematic analysis. *The Lancet global health, 2*(6), e323-e333. [CrossRef]

[22] National Institute of Population Research and Training (NIPORT), & ICF. (2019). *Bangladesh Maternal Mortality and Health Care Survey 2016: Final Report*. NIPORT, Dhaka, Bangladesh. Retrieved from https://www.measureevaluation.org/resources/publications/tr-18-297.html

[23] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM computing surveys (CSUR), 51*(5), 1-42. [CrossRef]

[24] Blagus, R., & Lusa, L. (2013). SMOTE for high-dimensional class-imbalanced data. *BMC bioinformatics, 14*(1), 106. [CrossRef]

[25] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. T. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, *30*, 3146-3154.

[26] Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015, August). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1721-1730). [CrossRef]

[27] Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society Series B: Statistical Methodology, 20*(2), 215-232. [CrossRef]

[28] Quinlan, J. R. (1986). Induction of decision trees. *Machine learning, 1*(1), 81-106. [CrossRef]

[29] Kittler, J., Hatef, M., Duin, R. P., & Matas, J. (1998). On combining classifiers. *IEEE transactions on pattern analysis and machine intelligence, 20*(3), 226-239. [CrossRef]

Appendix

Algorithm 1: SMOTE Oversampling (Applied to Training Partition Only)

Input: Training dataset D_{train}, 361 samples; Minority class subset D_{min}; Majority class subset D_{maj}

Output: Balanced training dataset D_{train}, 561 samples

Apply a stratified 80:20 train-test split to the dataset: $D_{\text{train}} = 361$ samples
$D_{\text{test}} = 91$ samples

Partition D_{train} into: Low Risk class (majority class)
Mid Risk and High Risk classes (minority classes for oversampling)

Initialize an empty synthetic sample set: $D_{\text{syn}} = \emptyset$

for *each minority class sample* $x_i \in D_{min}$ **do**
  Identify the k-nearest minority class neighbours of x_i
  for *each required synthetic sample* **do**
    Randomly select a neighboring sample x_{nn} from the k-nearest neighbours
    Generate a random interpolation coefficient: $\lambda \sim \text{Uniform}(0, 1)$
    Generate a synthetic sample: $x_{\text{new}} = x_i + \lambda(x_{nn} - x_i)$
      Add the generated sample and its corresponding minority class label to the synthetic dataset: $(x_{\text{new}}, \text{minority\_label}) \rightarrow D_{\text{syn}}$
  end
end

Construct the balanced training dataset: $D_{\text{train}} = D_{\text{train}} \cup D_{\text{syn}}$

A Mathematical Derivations

A.1 Random Forest: Ensemble Voting Formulation

Random Forest constructs an ensemble of B decision trees $\{h(x, \Theta_b)\}$, where each tree is trained on a bootstrapped subsample using a randomly selected feature subset at each split. The final multi-class prediction is determined by aggregating the votes of all trees:

$$\hat{y}(x) = \arg\max_c \sum_{b=1}^{B} \mathbb{1}[h(x, \Theta_b) = c] \tag{A.1}$$

where $c \in \{\text{high risk, low risk, mid risk}\}$ and $\mathbb{1}[\cdot]$ denotes the indicator function. This ensemble averaging reduces variance without increasing bias, providing a significant performance advantage over a single decision tree.

A.2 XGBoost: Regularised Objective and Derivation

XGBoost employs a sequential boosting strategy. At each iteration t, a new tree f_t is integrated to minimise the following regularised objective:

$$\mathcal{L}^{(t)} = \sum_i \ell\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) \tag{A.2}$$

where ℓ represents the multi-class cross-entropy loss, $\hat{y}_i^{(t-1)}$ denotes the current ensemble prediction, and the regularisation term $\Omega(f_t) = \gamma T + \frac{\lambda}{2}\|w\|^2$ constrains both tree complexity T and leaf weight magnitude w.

To render this objective computationally tractable, XGBoost utilises a second-order Taylor expansion. Defining the first-order gradient as $g_i = \partial_{\hat{y}_i^{(t-1)}} \ell(y_i, \hat{y}_i^{(t-1)})$ and the second-order Hessian as $h_i = \partial^2_{\hat{y}_i^{(t-1)}} \ell(y_i, \hat{y}_i^{(t-1)})$, the objective function at iteration t simplifies to:

$$\mathcal{L}^{(t)} = \sum_{j=1}^{T} \left[G_j w_j + \frac{1}{2}(H_j + \lambda) w_j^2 \right] + \gamma T \tag{A.3}$$

where $G_j = \sum_{i \in I_j} g_i$ and $H_j = \sum_{i \in I_j} h_i$ are the aggregated gradients and Hessians for leaf j.

Consequently, the optimal leaf weight for leaf j is derived as:

$$w_j^* = -\frac{G_j}{H_j + \lambda} \tag{A.4}$$

The gain for a candidate split is calculated as:

$$\text{Gain} = \frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right] - \gamma \tag{A.5}$$

where L and R subscripts refer to the left and right child nodes respectively. A split is executed only if Gain > 0, ensuring that the γ parameter effectively limits leaf proliferation. This regularisation and Hessian-weighted optimisation enhances model robustness against noise introduced by SMOTE augmentation, particularly where synthetic minority samples populate low-density regions of the feature space.

Full derivation follows Chen & Guestrin (2016) [4].

Shubham Shirodkar is an MSc candidate in Artificial Intelligence and Data Science at Solent University, specializing in human-centric machine learning applications and data analytics. He holds a diverse academic foundation with a Bachelor of Science in Chemistry and a Bachelor of Physical Education. Leveraging over five years of professional experience across corporate sales management and physical education, his research focuses on translating real-world operational insights into robust, data-driven systems. His current technical projects include predictive modeling for consumer churn and developing specialized AI-powered health and wellness systems. His work bridges the gap between complex analytical frameworks and practical, human-oriented technological solutions.(Email: Shirodkarshubma9@gmail.com)

Raza Hasan received his PhD in Informatics from the Malaysia University of Science and Technology in 2021. He is currently a Senior Lecturer in Computing at Solent University, where he has worked since July 2023. Previously, he served as Deputy Head of Department, Programme Leader, and Associate Professor of Computing and IT at the Global College of Engineering and Technology in Oman. His research focuses on artificial intelligence, data science, learning analytics, educational data science, and data mining. (Email: raza.hasan@solent.ac.uk)

Salman Mahmood is an Associate Professor in the Department of Computer Science at Nazeer Hussain University (NHU), Pakistan. He specializes in artificial intelligence, machine learning, cybersecurity, and data science. He holds academic credentials in computer science and has been actively involved in both teaching and research. Over his career, he has published in areas of AI and related disciplines, contributing to the growth of the research profile at NHU. (Email: salman.mahmood@nhu.edu.pk)