

Predicting Maternal Health Risks in Sierra Leone Using Machine Learning Models

Alhaji Osman Bah^{1*}; Abdul Hamid Turay²;
Iyeba Alpha Kallon³; Alhajie Mohamed Sesay⁴

^{1,2,3}Department of Electrical and Electronics Engineering, Fourah Bay College, University of Sierra Leone

⁴Department of Physics, Fourah Bay College, University of Sierra Leone

Corresponding Author: Alhaji Osman Bah^{1*}

Publication Date: 2026/10/30

Abstract: Maternal health risk prediction can support the early identification of pregnancies that may require closer clinical attention, particularly in resource-constrained settings. This study evaluates seven supervised machine-learning classifiers—Logistic Regression, Support Vector Machine (SVM) with Polynomial and Radial Basis Function (RBF) kernels, Gradient Boosting Machine (GBM), XGBoost, LightGBM, and CatBoost—for three-class maternal health-risk classification. The primary analysis uses a 500-record dataset obtained from Poly Clinic, Peninsula Road, Sierra Leone, containing eight predictors: age, systolic blood pressure, diastolic blood pressure, blood glucose, BMI, heart rate, parity, and hemoglobin. Models were evaluated using stratified 10-fold cross-validation under both the original class distribution and a SMOTE-balanced training condition, and the results were compared with the UCI Maternal Health Risk dataset and the Mendeley Maternal Health Risk Assessment Dataset. On the Poly Clinic dataset, SVM with the Polynomial kernel achieved the highest accuracy without SMOTE (85.0%), while SVM-RBF and LightGBM achieved 81.0% with SMOTE. In the cross-dataset comparison, LightGBM achieved 99.13% accuracy on Mendeley and 81.07% on UCI without SMOTE, and the highest macro-AUC values were 0.86 (Poly Clinic), 0.89 (UCI), and 0.99 (Mendeley). The findings demonstrate substantial dataset dependence and show that SMOTE does not uniformly improve overall accuracy, although it can improve recall for selected minority classes. The models should be regarded as research-level decision-support signals, not as evidence of clinical effectiveness or mortality reduction; external validation on larger multi-facility datasets and prospective clinical evaluation are required before deployment.

Keywords: Maternal Health, Machine Learning, Maternal Risk Prediction, Logistic Regression, Support Vector Machine, Gradient Boosting, XGBoost, LightGBM, CatBoost, SMOTE, Sierra Leone.

How to Cite: Alhaji Osman Bah; Abdul Hamid Turay; Iyeba Alpha Kallon; Alhajie Mohamed Sesay (2026) Predicting Maternal Health Risks in Sierra Leone Using Machine Learning Models. *International Journal of Innovative Science and Research Technology*, 11(10), 2226-2236. <https://doi.org/10.38124/ijisrt/26oct769>

I. INTRODUCTION

Maternal health remains a major public health concern worldwide, particularly in low- and middle-income countries where preventable pregnancy-related complications continue to result in high maternal mortality. According to the World Health Organization (WHO), approximately 287,000 women died from pregnancy and childbirth-related complications in 2020, with Sub-Saharan Africa accounting for the majority of these deaths [1]. Sierra Leone continues to face significant maternal health challenges owing to limited healthcare resources, shortages of skilled health professionals, delayed access to quality antenatal care, and inadequate infrastructure, which makes early identification of high-risk pregnancies a critical priority [2].

Recent advances in artificial intelligence (AI) and machine learning have created new opportunities for supporting disease diagnosis, clinical decision-making, and patient risk prediction. Unlike traditional statistical approaches, machine-learning algorithms can identify complex relationships within clinical data and generate predictions without explicitly programmed rules [3, 4]. These capabilities have led to successful applications in disease diagnosis, cardiovascular risk assessment, and personalized treatment planning [5, 6]. In maternal healthcare, predictive models can help clinicians identify women at greater risk of pregnancy complications, enabling earlier intervention and improved patient management [7].

Several supervised algorithms have shown strong performance in healthcare prediction tasks. Logistic Regression provides a simple and interpretable baseline,

while Support Vector Machines (SVMs), particularly those with Polynomial and Radial Basis Function (RBF) kernels, capture complex nonlinear relationships in clinical data [8, 9]. Ensemble methods, including Gradient Boosting Machine (GBM), Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), and CatBoost, have attracted considerable attention because of their predictive accuracy, robustness, and ability to model structured healthcare data efficiently [10–13].

A common challenge in maternal health datasets is class imbalance, in which low-risk cases significantly outnumber medium- and high-risk cases, reducing the ability of models to identify clinically important minority classes. The Synthetic Minority Oversampling Technique (SMOTE) addresses this by generating synthetic samples for underrepresented classes [14, 15]. To obtain reliable performance estimates, *k*-fold cross-validation is widely adopted because it evaluates models across multiple training and testing partitions, reducing bias in the assessment of generalization [16, 17].

Although machine learning has been widely applied to maternal healthcare prediction, relatively few studies have used data from Sierra Leone. Many existing models were developed on data from other countries, which may not represent the demographic and clinical characteristics of the Sierra Leonean population. There is therefore a need to develop and evaluate predictive models using locally relevant maternal health data.

This study evaluates seven machine-learning models—Logistic Regression, SVM with Polynomial kernel, SVM with RBF kernel, GBM, XGBoost, LightGBM, and CatBoost—for maternal health-risk classification. The models are assessed under two conditions: without SMOTE, and with SMOTE applied to the training data to address class imbalance. A 10-fold cross-validation framework is used, and accuracy, precision, recall, and F1-score are reported. In

addition to the original Poly Clinic dataset, two external datasets are used to examine how consistently the models perform across datasets with different characteristics and risk distributions: the UCI Maternal Health Risk dataset [18] and the Maternal Health Risk Assessment Dataset from Mendeley Data [19].

The evaluation therefore focuses not only on identifying the model with the highest overall accuracy, but also on how effectively each model classifies the Low-, Medium-, and High-Risk groups, and on whether class balancing improves performance or the original data distribution gives better results. Receiver operating characteristic (ROC) analysis is used as a complementary, threshold-independent measure of class discrimination.

II. METHODOLOGY

➤ Dataset

The primary dataset was obtained from Poly Clinic, Peninsula Road (2019), and contains 500 patient records. Each record represents one patient and includes eight input variables—age, systolic blood pressure, diastolic blood pressure, blood glucose, BMI, heart rate, parity, and hemoglobin—and one target variable, Risk Level. The target has three categories: Medium Risk (353 patients, 70.6%), High Risk (104 patients, 20.8%), and Low Risk (43 patients, 8.6%), as shown in Fig. 1. The class distribution is therefore substantially imbalanced, with Medium Risk as the majority class. No missing values were identified.

The imbalance is important because a classifier can obtain high overall accuracy while performing poorly on the less represented Low- and High-Risk categories. Accordingly, models are evaluated both with and without SMOTE, rather than assuming that oversampling will improve performance.

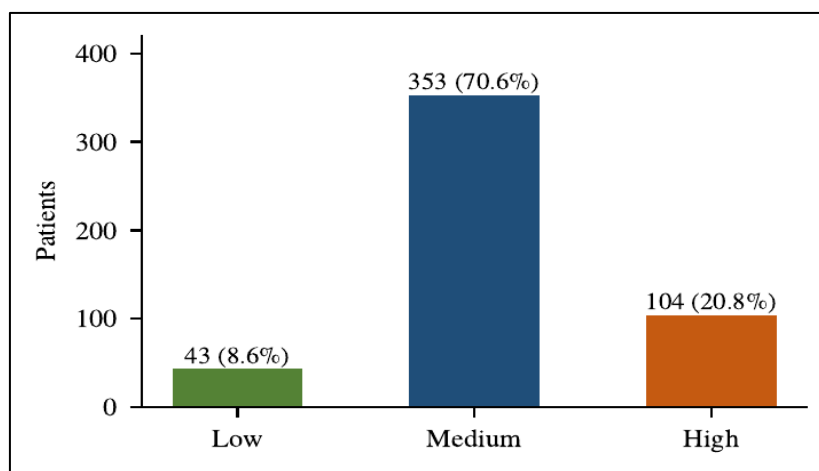


Fig 1 Distribution of Patients Across the Low-, Medium-, and High-Risk Categories in the Poly Clinic Dataset ($n = 500$).

➤ Data Preprocessing

Preprocessing included checks for missing values, duplicate records, outliers, and class distribution. As no

values were missing, no imputation was required. Duplicate observations were removed, and clinical variables were reviewed for implausible values; only observations

considered invalid were excluded, while medically plausible variation was retained. The target variable was encoded as Low Risk = 0, Medium Risk = 1, and High Risk = 2.

The eight clinical variables (Table 1) were retained as the predictor set because they are directly represented in the dataset and have plausible relevance to maternal health-risk assessment. No claim of algorithmic feature-selection superiority is made.

Predictors were standardized using StandardScaler:

$$Z = \frac{X - \mu}{\sigma},$$

Where Z is the standardized value, X the original feature value, μ the feature mean, and σ its standard deviation. This

places variables measured on different scales on a comparable footing. To avoid information leakage, μ and σ were estimated from the training portion of each cross-validation fold and then applied to the corresponding validation portion.

Models were evaluated under two conditions: (1) without class balancing, using the original training distribution; and (2) with SMOTE, applied only to the training portion of each fold and never to the validation fold. This allows the effect of oversampling on minority-class recognition to be assessed without introducing synthetic observations into the evaluation data. The overall workflow is shown in Fig. 2.

Table 1 Description of the Selected Maternal Health Features

Feature	Description	Unit	Mean $\pm$ SD
Age	Age of the pregnant patient	years	27.5 $\pm$ 6.2
Systolic BP	Pressure during heart contraction	mmHg	119.5 $\pm$ 15.8
Diastolic BP	Pressure while the heart relaxes	mmHg	79.5 $\pm$ 10.6
Blood glucose	Glucose concentration in blood	mg/dL	99.2 $\pm$ 22.4
BMI	Body weight relative to height	kg/m ²	26.4 $\pm$ 5.1
Parity	Number of previous births	births	2.1 $\pm$ 1.5
Hemoglobin	Blood hemoglobin concentration	g/dL	11.8 $\pm$ 1.6
Heart rate	Heart beats per minute	bpm	82.4 $\pm$ 12.7

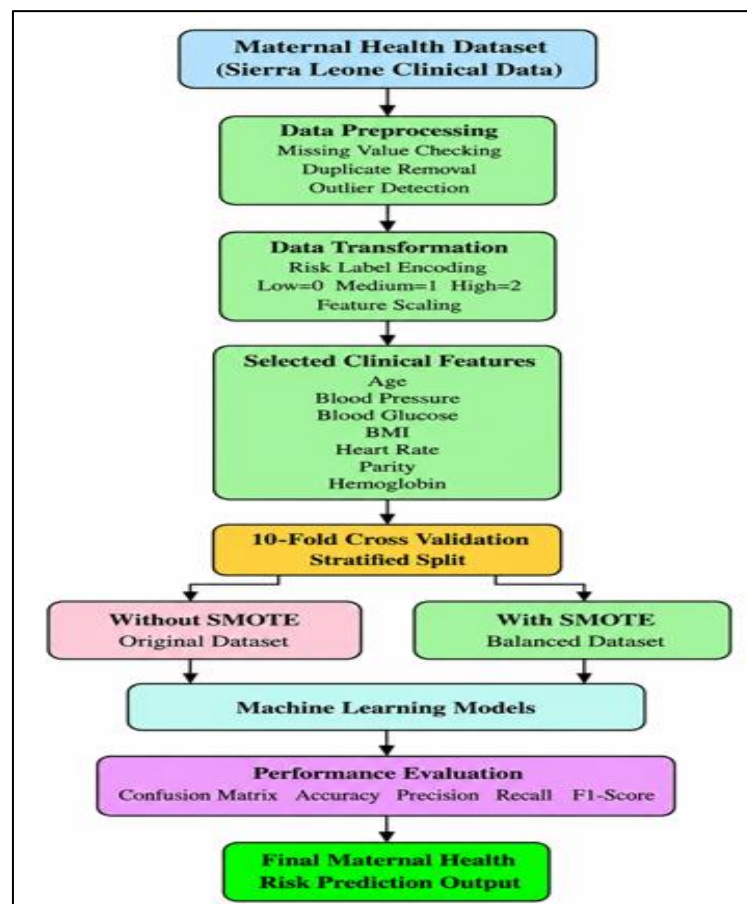


Fig 2 Conceptual Workflow: Data Preprocessing, Feature Selection, Stratified 10-Fold Cross-Validation with and Without SMOTE, Model Training, and Performance Evaluation.

➤ Experimental Setup

Stratified 10-fold cross-validation was used rather than a single fixed train/test split. In each round the data were partitioned into ten folds while preserving the class distribution as far as possible; nine folds were used for training and one for validation, until every fold had served once as the validation fold. Within each training fold, preprocessing parameters were fitted on the training data only. For the SMOTE condition, oversampling was then performed exclusively on that training fold, and the validation fold remained untouched. The two conditions used the same partitions, so that differences in performance can be attributed to SMOTE rather than to different validation samples.

Hyperparameters were selected using grid search and manual adjustment where applicable, always on training data within the cross-validation framework. The analysis was implemented in Python using scikit-learn and the XGBoost, LightGBM, and CatBoost libraries.

➤ Models Implemented

Seven supervised models were compared: Logistic Regression; SVM with Polynomial kernel; SVM with RBF kernel; GBM; XGBoost; LightGBM; and CatBoost. Their core formulations are summarized in Table 2.

Table 2 Mathematical Formulation of the Evaluated Models

Model	Formulation	Symbols
Logistic Regression	$P = \frac{1}{1 + e^{-(b_0 + b_1 X)}}$	P : predicted probability; b_0 : intercept; b_1 : coefficient (weight); X : input feature; e : Euler's number (≈ 2.718).
CatBoost	$\hat{y} = \sum_{t=1}^T f_t(x)$	$\hat{y}$: final prediction; x : feature vector; $f_t(x)$: prediction of tree t ; T : number of trees.
GBM	$F_m(x) = F_{m-1}(x) + \eta h_m(x)$	$F_m(x)$: model at step m ; $h_m(x)$: weak learner (tree); η : learning rate.
LightGBM	$F(x) = \sum_{m=1}^M \vartheta f_m(x)$	$F(x)$: final prediction; $f_m(x)$: decision tree; M : number of trees; ϑ : learning rate.
XGBoost	$\hat{y}_i = \sum_{k=1}^K f_k(x_i)$	K : number of trees; f_k : individual decision tree; $\hat{y}_i$: predicted output.
SVM (Polynomial)	$K(x_i, x_j) = (\gamma x_i^\top x_j + r)^d$	K : polynomial kernel similarity; x_i, x_j : input vectors; γ : kernel coefficient; r : independent term (coef0); d : polynomial degree.
SVM (RBF)	$K(x_i, x_j) = \exp(-\gamma \ x_i - x_j\ ^2)$	x_i, x_j : input vectors; γ : controls the spread of the RBF kernel.

➤ Evaluation Metrics

The confusion matrix compares predicted maternal risk categories with the actual categories. Diagonal entries are correct classifications and off-diagonal entries are misclassifications, which shows which risk categories a model predicts accurately and where it errs. For the three-class problem, the quantities in Fig. 3 are counted per class in a one-vs-rest manner, and overall accuracy is the proportion of correctly classified samples.

The following metrics are reported.

- Accuracy is the proportion of correct predictions. It is useful when classes are balanced but can be misleading on imbalanced data, because a model can score highly while performing poorly on minority classes:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- Precision is the proportion of positive predictions that are correct. High precision indicates few false positives:

$$Precision = \frac{TP}{TP + FP}$$

- Recall (sensitivity) is the proportion of actual positives that are identified. It is essential in healthcare, where failing to identify a high-risk patient may have severe consequences:

$$Recall = \frac{TP}{TP + FN}$$

- F1-score is the harmonic mean of precision and recall, and is particularly useful for imbalanced data:

	Predicted positive	Predicted negative
Actual positive	TP correctly predicted positive	FN missed positive (Type II error)
Actual negative	FP false alarm (Type I error)	TN correctly predicted negative

Fig 3 Components of a Confusion Matrix (TP, FN, FP, TN) Used to Interpret Correct and Incorrect Classifications.

$$F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}.$$

Here TP , TN , FP , and FN denote true positives, true negatives, false positives, and false negatives, respectively. Cross-dataset summaries use macro-averaged precision, recall, and F1-score, i.e., the unweighted mean over the three risk classes.

➤ ROC-AUC Analysis

The ROC curve plots the true-positive rate (sensitivity) against the false-positive rate ($1 - \text{specificity}$) over a range of decision thresholds. The area under the curve (AUC) is a threshold-independent measure of discrimination, with larger values indicating stronger separation between positive and negative cases. Because the problem has three classes, ROC-AUC is computed one-vs-rest for each class (Low, Medium, High) and summarized by the macro-average (and, where appropriate, the weighted average). For each cross-validation fold, the calculation uses the model's predicted probabilities or decision scores on the held-out fold, and the same protocol is applied to the Poly Clinic, UCI, and Mendeley datasets so that the AUC comparison remains consistent. ROC-AUC is used as a complement to, not a replacement for, accuracy, precision, recall, and F1-score.

III. EXPERIMENTAL RESULTS

➤ Confusion Matrices Without SMOTE

Fig. 4 shows confusion matrices for four of the models without SMOTE.

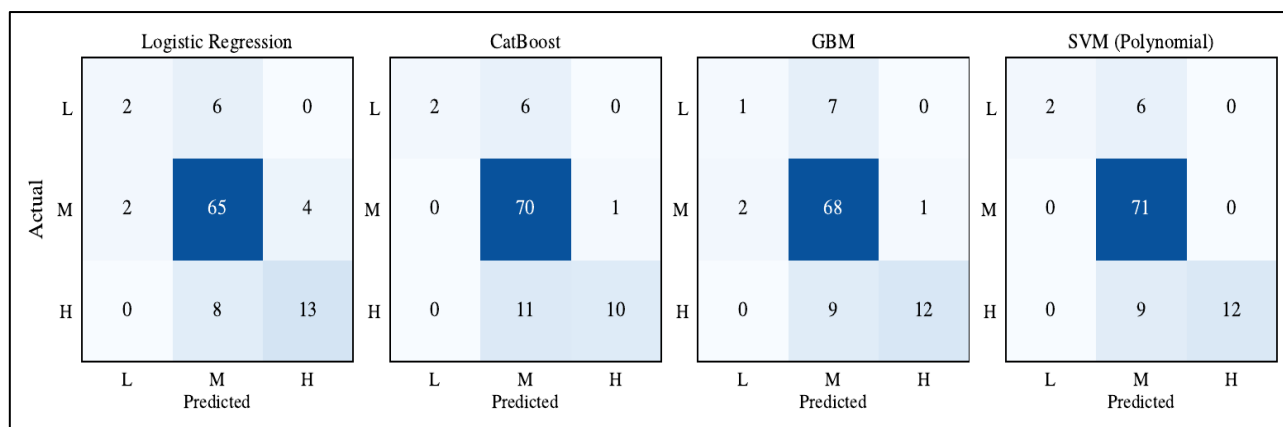


Fig 4 Confusion Matrices Without SMOTE for Logistic Regression, CatBoost, GBM, and SVM (Polynomial). L, M, and H Denote Low, Medium, and High Risk; Rows are Actual and Columns are Predicted Classes.

➤ Confusion Matrices With SMOTE

Fig. 5 shows the confusion matrices for all seven models with SMOTE.

Logistic Regression identified Low Risk well (11 of 13) and High Risk reasonably well (21 of 31), but only 62 of 106 Medium Risk cases were correct, with many misclassified as Low or High Risk. CatBoost classified 94 of 106 Medium, 19 of 31 High, and only 4 of 13 Low Risk cases correctly. GBM classified 91 of 106 Medium, 19 of 31 High, and 2 of 13 Low Risk cases correctly. LightGBM classified 99 of 106

Logistic Regression correctly classified 80 cases (2 Low, 65 Medium, and 13 High Risk). It performed best on Medium Risk, with some errors among all three classes, indicating difficulty in distinguishing some risk levels.

CatBoost correctly classified 82 cases (2 Low, 70 Medium, 10 High). It identified 70 of 71 Medium Risk cases; most errors occurred between the Medium and High Risk classes, while the Low-Risk class remained challenging because of its small number of samples.

GBM correctly classified 81 cases (1 Low, 68 Medium, 12 High), identifying 68 of 71 Medium Risk cases. Most misclassifications were between Low and Medium Risk and between High and Medium Risk, indicating some overlap among these categories.

SVM (Polynomial) correctly classified 85 cases (2 Low, 71 Medium, 12 High) and classified all 71 Medium Risk cases correctly. The few remaining errors were between Low and Medium and between High and Medium Risk, so these classes remain harder to separate than Medium Risk.

LightGBM and XGBoost also performed consistently under cross-validation: most Medium Risk cases were classified correctly, and the few Low- and High-Risk errors were assigned to neighboring categories. SVM (RBF) recognized the Medium Risk class well (recall 0.90) but, like the other models, found Low Risk difficult (Table 3).

Medium, 18 of 31 High, and 4 of 13 Low Risk cases correctly. XGBoost classified 95 of 106 Medium, 17 of 31 High, and 3 of 13 Low Risk cases correctly. SVM (Polynomial) classified 82 of 106 Medium, 18 of 31 High, and 7 of 13 Low Risk cases correctly. SVM (RBF) classified 94 of 106 Medium, 20 of 31 High, and 7 of 13 Low Risk cases correctly.

For all ensemble models, the Low-Risk class was the most challenging, with Low-Risk cases mostly confused with Medium Risk, suggesting overlap between Low- and Medium-Risk feature profiles.

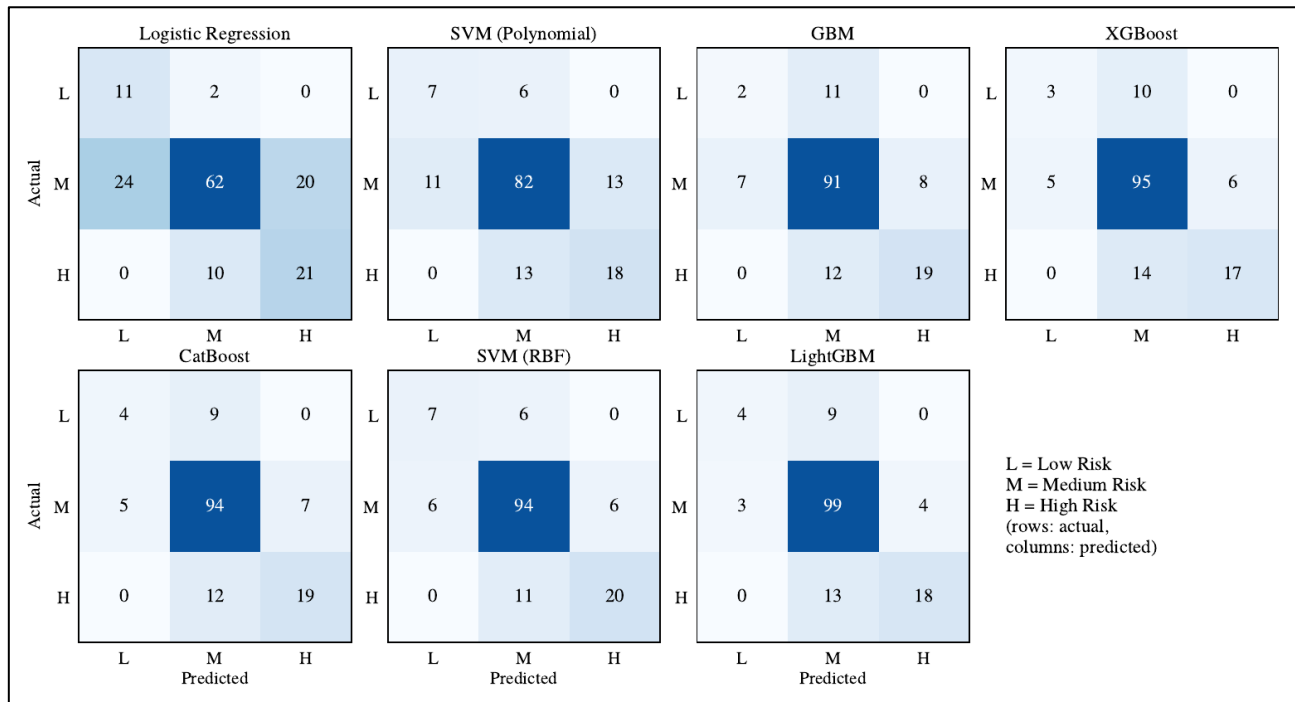


Fig 5 Confusion Matrices with SMOTE for the Seven Evaluated Models. L, M, and H Denote Low, Medium, and High Risk; Rows are Actual and Columns are Predicted Classes.

➤ Performance With and Without SMOTE

Table 3 compares accuracy, precision, recall, and F1-score for the seven classifiers under both conditions. With SMOTE, SVM-RBF and LightGBM achieved the highest accuracy (81.0%), followed by CatBoost (78.0%), XGBoost (77.0%), GBM (75.0%), SVM Polynomial (71.0%), and Logistic Regression (63.0%). Without SMOTE, SVM

Polynomial achieved the highest accuracy (85.0%), followed by CatBoost and LightGBM (82.0% each), XGBoost and GBM (81.0% each), and Logistic Regression and SVM-RBF (80.0% each). SMOTE therefore did not improve overall accuracy for most models; SVM-RBF was the exception, improving from 80.0% to 81.0% (Fig. 6).

Table 3 Performance of the Selected Models on the Poly Clinic Dataset Without and With SMOTE

Model	Risk level	Without SMOTE				With SMOTE			
		Acc.	Prec.	Rec.	F1	Acc.	Prec.	Rec.	F1
Logistic Regression	Low	80.0%	0.50	0.25	0.33	63.0%	0.31	0.85	0.46
	Medium		0.82	0.92	0.87		0.84	0.58	0.69
	High		0.76	0.62	0.68		0.51	0.68	0.58
SVM (Polynomial)	Low	85.0%	1.00	0.25	0.40	71.0%	0.39	0.54	0.45
	Medium		0.83	1.00	0.90		0.81	0.77	0.79
	High		1.00	0.57	0.73		0.58	0.58	0.58
GBM	Low	81.0%	0.33	0.12	0.18	75.0%	0.22	0.15	0.18
	Medium		0.81	0.96	0.88		0.80	0.86	0.83
	High		0.92	0.57	0.71		0.70	0.61	0.66
XGBoost	Low	81.0%	1.00	0.12	0.22	77.0%	0.38	0.23	0.29
	Medium		0.80	0.99	0.88		0.80	0.90	0.84
	High		0.91	0.48	0.62		0.74	0.55	0.63
CatBoost	Low	82.0%	1.00	0.25	0.40	78.0%	0.44	0.31	0.36
	Medium		0.80	0.99	0.89		0.82	0.89	0.85
	High		0.91	0.48	0.62		0.73	0.61	0.67
SVM (RBF)	Low	80.0%	0.50	0.50	0.50	81.0%	0.54	0.54	0.54
	Medium		0.83	0.90	0.86		0.85	0.89	0.87
	High		0.80	0.57	0.67		0.77	0.65	0.70
LightGBM	Low	82.0%	0.67	0.25	0.36	81.0%	0.57	0.31	0.40
	Medium		0.81	0.97	0.88		0.82	0.93	0.87
	High		0.92	0.52	0.67		0.82	0.58	0.68

Class-wise performance gives a more nuanced picture. Medium Risk generally had the strongest recall because it is the largest class. Under SMOTE, SVM-RBF achieved 0.85 precision, 0.89 recall, and 0.87 F1-score for Medium Risk, while LightGBM achieved 0.93 recall. Without SMOTE, SVM Polynomial achieved 1.00 recall for Medium Risk, and CatBoost and XGBoost each achieved 0.99.

The Low-Risk class was more difficult for several models. Under SMOTE, GBM and LightGBM achieved

Low-Risk recall of only 0.15 and 0.31, respectively, and without SMOTE some models produced high Low-Risk precision but low recall. Accuracy alone is therefore insufficient for this imbalanced three-class problem. For High Risk, SVM-RBF achieved the highest F1-score under SMOTE (0.70) and LightGBM the highest precision (0.82); without SMOTE, SVM Polynomial achieved a High-Risk F1-score of 0.73 and GBM 0.71. The modest High-Risk recall of several models indicates that further validation is needed before any model could be considered for clinical use.

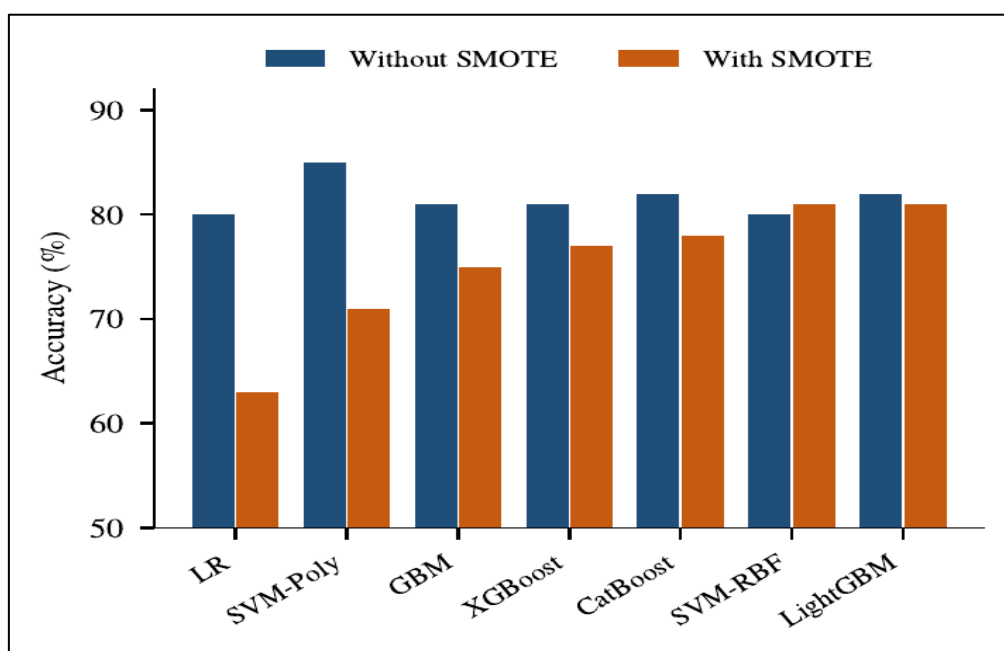


Fig 6 Model Accuracy on the Poly Clinic Dataset with and Without SMOTE.

Most models achieved higher overall accuracy without SMOTE. This does not mean that SMOTE is ineffective; rather, oversampling can trade overall accuracy for improved minority-class sensitivity. Precision, recall, and F1-score averaged over all models and classes are shown in Fig. 7. Precision was generally higher without SMOTE (approximately 0.81 versus 0.64), so the non-SMOTE models produced more precise class assignments overall, although precision alone does not indicate how many actual high-risk patients were detected.

Recall is especially important in maternal health-risk classification, because missed high-risk cases may have greater practical consequences than additional referrals for assessment. Recall varied substantially by model and class, and no single model performed uniformly well across all three risk categories. SMOTE produced a higher average

recall (approximately 0.62 versus 0.58 without SMOTE): for example, Logistic Regression achieved a Low-Risk recall of 0.85 with SMOTE versus 0.25 without, and SVM Polynomial 0.54 versus 0.25. SMOTE can thus improve sensitivity for selected minority classes even when overall accuracy decreases. The average F1-score was approximately 0.62 with SMOTE and 0.63 without, which reinforces the importance of examining individual classes rather than selecting a model on accuracy alone.

Overall, the experiments reveal a trade-off: the non-SMOTE condition generally produced higher accuracy and precision, whereas SMOTE improved recall for selected minority classes. In a future clinical implementation, the preferred operating point should be determined by explicit clinical priorities and prospective validation.

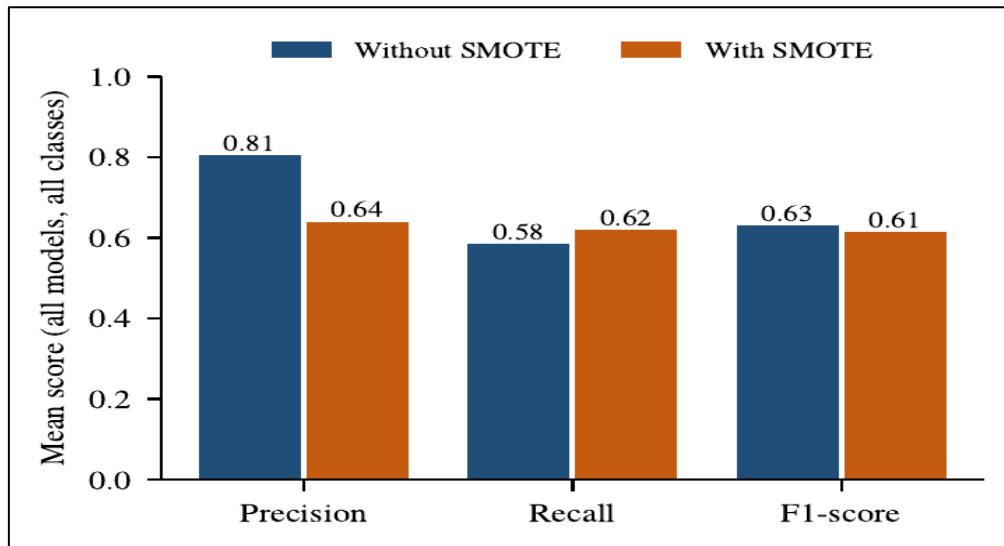


Fig 7 Mean Precision, Recall, and F1-Score Across all Models and Risk Classes on the Poly Clinic Dataset, with and Without SMOTE.

➤ Comparison Across the Three Datasets

Table 4 and Fig. 8 compare model accuracy on the Mendeley, UCI, and Poly Clinic datasets. The comparison was included to determine whether the relative performance of the seven classifiers is consistent across datasets. The strongest model differs by dataset (Table 5): LightGBM was

strongest on Mendeley (99.13%) and UCI (81.46% without SMOTE; 81.46% with SMOTE), whereas SVM Polynomial was strongest on Poly Clinic (85.00%). A model that performs strongly on one maternal-health dataset should therefore not be assumed to generalize to another.

Table 4 Comparative Accuracy of the Selected Models Across the Three Maternal Health Datasets

Model	Mendeley		UCI		Poly Clinic	
	With SMOTE	Without	With SMOTE	Without	With SMOTE	Without
Logistic Regression	56.35%	56.51%	60.94%	61.63%	63.0%	80.0%
SVM Polynomial	81.63%	81.48%	62.33%	64.21%	71.0%	85.0%
GBM	97.43%	97.39%	76.62%	74.46%	75.0%	81.0%
XGBoost	98.25%	98.23%	77.81%	77.71%	77.0%	81.0%
CatBoost	98.10%	98.12%	74.35%	76.03%	78.0%	82.0%
SVM RBF	94.79%	94.77%	70.02%	71.20%	81.0%	80.0%
LightGBM	99.12%	99.13%	81.46%	81.07%	81.0%	82.0%

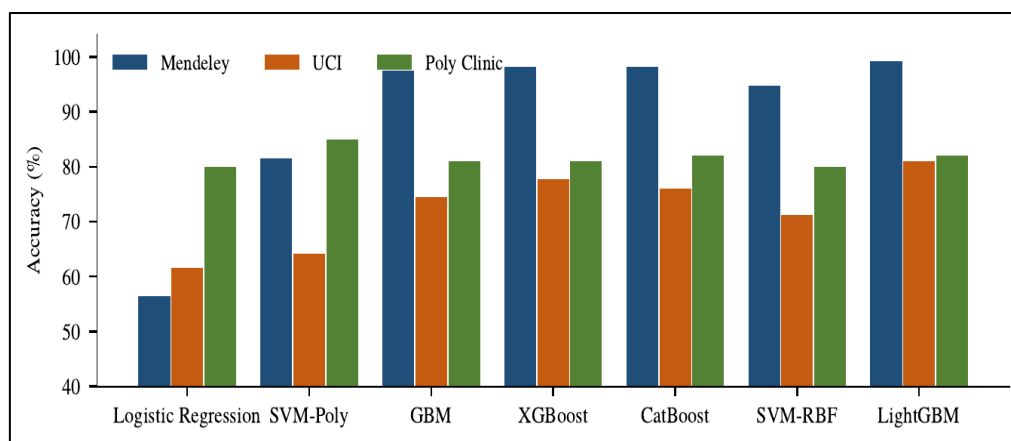


Fig 8 Accuracy of the Seven Models on the Mendeley, UCI, and Poly Clinic Datasets (Without SMOTE).

Table 5 Best-Performing Model per Dataset

Dataset	Best model	Accuracy	SMOTE
Mendeley	LightGBM	99.13%	Without
UCI	LightGBM	81.46%	With
Poly Clinic	SVM Polynomial	85.00%	Without

Table 6 adds macro-averaged precision, recall, and F1-score. As an example, the Poly Clinic Logistic Regression macro-precision is $(0.50 + 0.82 + 0.76)/3 = 0.693$, the macro-recall is $(0.25 + 0.92 + 0.62)/3 = 0.597$, and the

macro-F1 is $(0.33 + 0.87 + 0.68)/3 = 0.627$. The mean of each metric over the seven models for each dataset is shown in Fig. 9.

Table 6 Comparative Performance Evaluation of the Selected Models Across the Three Datasets (Macro-Averaged Precision, Recall, and F1-Score)

Model	Mendeley				UCI				Poly Clinic			
	Acc.(%)	Prec.	Rec.	F1	Acc.(%)	Prec.	Rec.	F1	Acc.(%)	Prec.	Rec.	F1
Logistic Regression	56.35	0.580	0.473	0.521	61.63	0.747	0.717	0.732	80.00	0.693	0.597	0.627
SVM Polynomial	81.63	0.901	0.742	0.814	64.21	0.821	0.761	0.790	85.00	0.943	0.607	0.677
GBM	97.43	0.975	0.975	0.974	74.46	0.850	0.853	0.851	81.00	0.687	0.550	0.590
XGBoost	98.25	0.983	0.983	0.982	77.71	0.867	0.864	0.866	81.00	0.903	0.530	0.573
CatBoost	98.10	0.981	0.981	0.981	76.03	0.856	0.875	0.866	82.00	0.903	0.573	0.637
SVM RBF	94.79	0.969	0.919	0.944	71.20	0.834	0.849	0.842	80.00	0.710	0.657	0.677
LightGBM	99.13	0.991	0.991	0.991	81.07	0.893	0.860	0.876	82.00	0.800	0.580	0.637

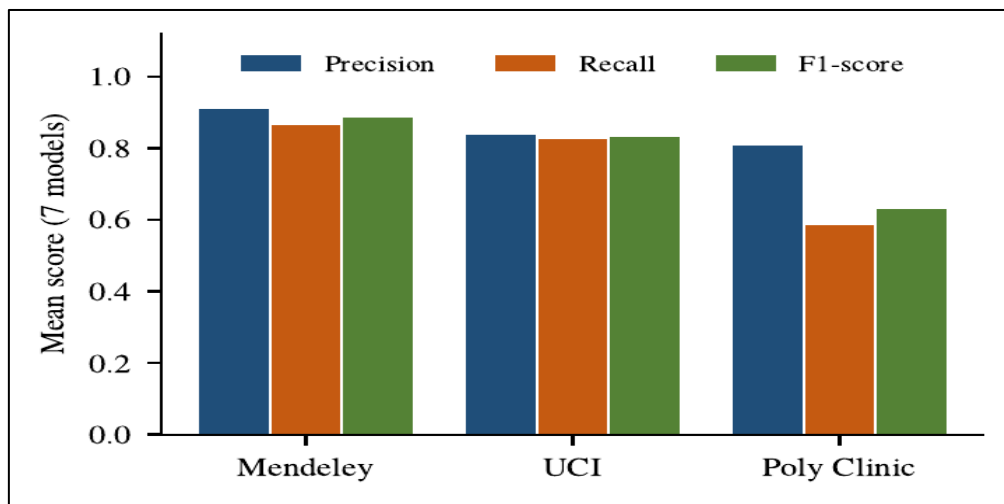


Fig 9 Mean Precision, Recall, and F1-Score Over the Seven Models for Each Dataset.

➤ ROC-AUC Results

Table 7 and Fig. 10 report the macro-AUC of each model on the three datasets. Unlike accuracy, which counts correct classifications at a particular decision rule, AUC measures how well a model separates the classes across thresholds. A macro-average is used because the problem is multiclass and gives equal importance to each risk category. Values near 1.00 indicate strong discrimination, and values near 0.50 indicate limited ability to separate the classes.

• Poly Clinic

Macro-AUC ranged from 0.72 to 0.86. Logistic Regression was lowest (0.72), followed by GBM (0.76), XGBoost (0.78), and CatBoost (0.80); SVM-RBF reached 0.82 and LightGBM 0.83. SVM Polynomial recorded the highest value (0.86), together with the highest accuracy (85.00%), suggesting relatively strong separation of the three risk categories.

• UCI

AUC values were generally higher than on Poly Clinic for several models: Logistic Regression 0.76, SVM Polynomial 0.80, GBM and SVM-RBF 0.85, CatBoost 0.87, XGBoost 0.88, and LightGBM 0.89 (accuracy 81.07%).

• Mendeley

The Mendeley dataset produced the highest macro-AUC values: Logistic Regression 0.83, SVM Polynomial 0.93, SVM-RBF 0.95, GBM 0.96, XGBoost and CatBoost 0.98, and LightGBM 0.99 (accuracy 99.13%). Such high performance should be interpreted carefully, because differences in dataset size, feature distributions, collection procedures, and class characteristics can substantially affect model performance.

Table 7 Macro-AUC Across Datasets and Models

Model	Poly Clinic	UCI	Mendeley
Logistic Regression	0.72	0.76	0.83
SVM Polynomial	0.86	0.80	0.93
GBM	0.76	0.85	0.96

XGBoost	0.78	0.88	0.98
CatBoost	0.80	0.87	0.98
SVM RBF	0.82	0.85	0.95
LightGBM	0.83	0.89	0.99

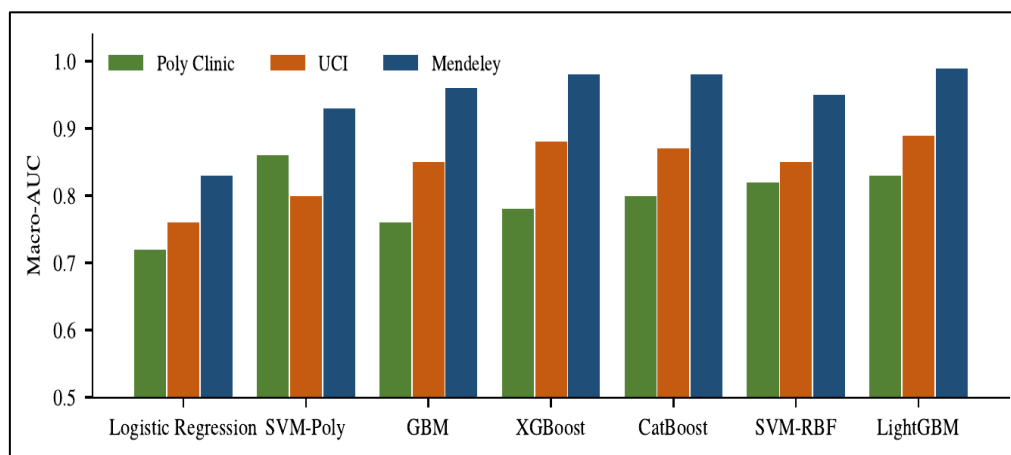


Fig 10 Macro-AUC of the Seven Evaluated Models on the Poly Clinic, UCI, and Mendeley Datasets.

IV. DISCUSSIONS

The comparison of the Mendeley, UCI, and Poly Clinic datasets shows that dataset characteristics strongly influence the performance of machine-learning models for maternal health-risk prediction: although the same algorithms were applied, performance varied considerably. Three levels of behavior can be distinguished: Mendeley gave the strongest overall performance, particularly for ensemble models; UCI gave moderate performance, with LightGBM remaining the strongest selected model; and Poly Clinic gave the highest accuracy for SVM Polynomial but with more variation between risk categories.

On Mendeley, LightGBM achieved approximately 99% accuracy and XGBoost and CatBoost approximately 98%, with precision, recall, and F1-score close to 1.00, which suggests that the relationships between inputs and risk categories in that dataset are well suited to nonlinear tree-based models. In contrast, Logistic Regression achieved only about 56%, showing that a simple linear decision boundary was insufficient there. On UCI, performance was considerably lower: LightGBM achieved 81.07% (81.46% with SMOTE), XGBoost about 78%, and GBM and CatBoost remained in the mid-70% range. High performance on one dataset therefore does not necessarily transfer to another, even when the same risk categories and similar clinical variables are used.

The Poly Clinic dataset produced an intermediate result. It did not yield extremely high accuracies for the ensemble models, but several models performed better on it than on UCI. SVM Polynomial achieved 85% without SMOTE, compared with 64.21% on UCI and 81.48% on Mendeley, and Logistic Regression improved from 56.51% (Mendeley) and 61.63% (UCI) to 80%. This suggests that the relationship between clinical variables and risk categories may be closer to linearly separable in the Poly Clinic data, favouring relatively simple classifiers. The local dataset is particularly

relevant to this study: although LightGBM performs very well on Mendeley and UCI, its 82% accuracy on Poly Clinic is slightly below the 85% of SVM Polynomial without SMOTE.

The Poly Clinic dataset contains 500 records and is strongly imbalanced (70.6% Medium, 20.8% High, 8.6% Low Risk), so overall accuracy can remain relatively high even when recognition of the minority Low-Risk class is weak. Possible explanations for the differences between datasets include sample size, predictor definitions, class distributions, data-collection context, feature ranges, and the degree of class overlap; these are plausible dataset-level explanations, not proof of a specific causal mechanism.

The ROC-AUC results complement the accuracy comparison. Poly Clinic generally produced lower AUC values (maximum 0.86) than UCI (0.89) and Mendeley (0.99), consistent with weaker class separation and greater overlap. AUC and accuracy do not necessarily increase together: LightGBM had a higher AUC than XGBoost on UCI, but this does not mean that every individual prediction was more accurate, since AUC reflects ranking across thresholds whereas accuracy depends on a single decision rule.

Finally, the SMOTE comparison demonstrates a trade-off rather than a universally superior preprocessing strategy. Model selection should consider class-specific clinical priorities, in particular recall for the High-Risk class, rather than accuracy alone.

V. CONCLUSION

This study evaluated seven machine-learning models for three-class maternal health-risk classification and extended the analysis by comparing performance across the Poly Clinic dataset from Sierra Leone, the Mendeley maternal-health dataset, and the UCI Maternal Health Risk

dataset. Performance was strongly influenced by the dataset and the experimental condition. On Poly Clinic, SVM with a Polynomial kernel achieved the highest accuracy without SMOTE (85.0%), while SVM-RBF and LightGBM achieved 81.0% under SMOTE. In the cross-dataset comparison, LightGBM achieved the highest accuracy on Mendeley (99.13%) and UCI (81.07%), whereas SVM Polynomial was best on Poly Clinic (85.00%).

The ROC-AUC analysis showed that macro-AUC ranged from 0.72 to 0.86 on Poly Clinic (best: SVM Polynomial), 0.76 to 0.89 on UCI (best: LightGBM), and 0.83 to 0.99 on Mendeley (best: LightGBM). Accuracy alone does not fully describe model performance, as ROC-AUC provides additional information about how well a model separates the risk categories across thresholds.

SMOTE should not be regarded as a universally beneficial preprocessing technique. On Poly Clinic it reduced overall accuracy for most models, although recall improved for selected minority classes. This trade-off highlights the importance of evaluating accuracy together with class-specific precision, recall, F1-score, and ROC-AUC on imbalanced maternal health data.

Overall, the results show that machine-learning techniques can identify meaningful patterns in maternal health data and may support research into decision-support systems in resource-constrained settings. The Poly Clinic results provide an initial assessment using data from a Sierra Leonean healthcare setting. However, the models should be considered research prototypes rather than autonomous clinical systems: the primary dataset contains approximately 500 observations from a single healthcare setting, and the study includes neither independent external validation nor prospective clinical evaluation.

Future work should develop larger and more diverse multi-facility Sierra Leonean datasets, harmonize feature definitions across cohorts, and conduct independent external validation. It should also investigate probability calibration, model explainability, cost-sensitive evaluation, and clinically meaningful performance measures. Prospective evaluation involving qualified healthcare professionals is essential to determine whether the models provide useful and reliable information in real clinical environments before any model is considered for routine clinical decision support.

REFERENCES

- [1.] World Health Organization (WHO), *Trends in Maternal Mortality 2000 to 2020: Estimates by WHO, UNICEF, UNFPA, World Bank Group and UNDESA/Population Division*. Geneva, Switzerland: WHO, 2023.
- [2.] R. Carshon-Marsh *et al.*, “Child, maternal, and adult mortality in Sierra Leone: Nationally representative mortality survey 2018–20,” *The Lancet Global Health*, vol. 10, no. 1, pp. e114–e123, Jan. 2022.
- [3.] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 4th ed. Hoboken, NJ, USA: Pearson, 2021.
- [4.] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.
- [5.] K. G. Dinesh, K. Arumugaraj, K. D. Santhosh, and V. Mareeswari, “Prediction of cardiovascular disease using machine learning algorithms,” *Procedia Computer Science*, vol. 165, pp. 570–576, 2019.
- [6.] K. Shameer, K. W. Johnson, B. S. Glicksberg, J. T. Dudley, and P. P. Sengupta, “Machine learning in cardiovascular medicine: Are we there yet?” *Heart*, vol. 104, no. 14, pp. 1156–1164, 2018.
- [7.] World Health Organization, “Improving national and district capacity to tackle maternal mortality in Sierra Leone,” WHO Regional Office for Africa, 2020.
- [8.] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*, 3rd ed. Hoboken, NJ, USA: Wiley, 2013.
- [9.] C.-C. Chang and C.-J. Lin, “LIBSVM: A library for support vector machines,” *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–27, 2011.
- [10.] J. H. Friedman, “Greedy function approximation: A gradient boosting machine,” *Ann. Statist.*, vol. 29, no. 5, pp. 1189–1232, 2001.
- [11.] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD)*, 2016, pp. 785–794.
- [12.] G. Ke *et al.*, “LightGBM: A highly efficient gradient boosting decision tree,” in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 3146–3154.
- [13.] L. Prokhorenkova, G. Gusev, A. Vorobev, A. Dorogush, and A. Gulin, “CatBoost: Unbiased boosting with categorical features,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2018, pp. 6638–6648.
- [14.] H. He and E. A. Garcia, “Learning from imbalanced data,” *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sept. 2009.
- [15.] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique,” *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.
- [16.] G. Douzas and F. Bacao, “Effective data generation for imbalanced learning using conditional generative adversarial networks,” *Expert Syst. Appl.*, vol. 91, pp. 464–471, 2018.
- [17.] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed. New York, NY, USA: Springer, 2009.
- [18.] M. Ahmed, M. A. Kashem, M. Rahman, and S. Khatun, “Maternal Health Risk,” UCI Machine Learning Repository, 2020, doi: 10.24432/C5DP5D.
- [19.] M. U. Mojumdar, M. Assaduzzaman, D. Sarker, H. A. Shifa, M. A. H. Sajeeb, S. Bari, N. R. Chakraborty, and M. J. Alam, “Maternal Health Risk Assessment Dataset,” Mendeley Data, 2024, doi: 10.17632/p5w98dvbbk.1.