



Explainable Ensemble Learning for Maternal Health Risk in Low-Resource Settings

Lilik Widyawati¹, Neny Sulistianingsih^{2*}

¹Computer Science, Faculty of Engineering, Universitas Bumigora, Mataram, Indonesia

²Master Program of Computer Science, Post Graduate Program, Universitas Bumigora, Mataram, Indonesia

¹lilikwidya@universitasbumigora.ac.id, ²neny.sulistianingsih@universitasbumigora.ac.id

Abstract

Maternal health remains a global challenge, particularly in low-resource settings where accurate and timely risk prediction is essential to reducing maternal mortality. This study proposes an explainable machine learning framework for predicting maternal health risks by integrating ensemble learning methods with SHAP (Shapley Additive exPlanations) for interpretability. This study utilized the publicly available Maternal Health Risk Data Set (MHRDS), comprising physiological features such as systolic and diastolic blood pressure, blood sugar level, body temperature, and age. A total of 18 machine learning models including Random Forest, XGBoost, LightGBM, Neural Networks, and TabNet were evaluated to compare individual classifiers and ensemble approaches comprehensively. The selection of this diverse set of models is grounded in the need to benchmark different algorithmic paradigms, as variations in inductive bias, learning capacity, and robustness to clinical data noise can influence predictive performance and generalizability. This comprehensive comparison enables the identification of optimal model types for integration into ensemble frameworks. Evaluation was performed across three different test scenarios (test sizes of 10%, 20%, and 30%) to assess model consistency under varying data partitions. Stacking, Voting, and Histogram-based Gradient Boosting showed consistently high performance, with Stacking achieving the highest accuracy of 87.2%, followed by Histogram Gradient Boosting (86.9%) and Voting (86.7%) at test size 0.2. SHAP analysis identified blood sugar, systolic blood pressure, and maternal age as the top predictors across all test scenarios. The best-performing models were deployed into a web-based clinical decision support system designed for healthcare practitioners in Indonesia. The proposed approach balances predictive accuracy and model transparency, offering a practical solution for improving maternal care in data-limited environments.

Keywords: ensemble learning; explainable AI; maternal health; risk prediction; SHAP Analysis

How to Cite: L. Widyawati and N. Sulistianingsih, "Explainable Ensemble Learning for Maternal Health Risk in Low-Resource Settings", *J. RESTI (Rekayasa Sist. Teknol. Inf.)*, vol. 9, no. 5, pp. 1121 - 1130, Oct. 2025.
Permalink/DOI: [<https://doi.org/10.29207/resti.v9i5.6765>]

Received: June 12, 2025

Accepted: September 26, 2025

Available Online: October 25, 2025

*This is an open-access article under the CC BY 4.0 License
Published by Ikatan Ahli Informatika Indonesia*

1. Introduction

Maternal health remains one of the most critical global public health challenges, particularly in low-resource settings where the availability of high-quality, timely, and data-driven prenatal care is limited. According to the World Health Organization, an estimated 295,000 women died during and following pregnancy and childbirth in 2017, most of which were preventable with earlier risk identification and medical intervention [1]. This alarming figure highlights the need for innovative approaches that can strengthen clinical decision-making in maternal healthcare. Recent advances in Artificial Intelligence (AI), especially Machine Learning (ML), have been recognized as a transformative force in this domain. By leveraging structured clinical and

demographic data, ML enables risk prediction that can improve both diagnostic speed and accuracy. Numerous studies have explored this potential: Ahmad et al. [2] utilized wearable sensors to detect pregnancy complications using Support Vector Machines (SVM), Artificial Neural Networks (ANN), and Random Forest (RF), while Jean Francois et al. [3] integrated socio-environmental and clinical variables with Decision Trees and RF for electronic health records (EHR)-based analysis. Beyond classical ML, hybrid and deep learning models have also emerged. For example, CNN-XGBoost combinations have demonstrated strong performance in neonatal mortality prediction [4], and ensemble CNNs integrating multimodal data enhanced delivery location classification [5]. Hybrid

ANN-RF frameworks have also reported high accuracy in maternal risk detection [6].

Despite these advances, one critical barrier remains: interpretability. Many of the above-mentioned models act as black boxes, providing predictions without transparent reasoning. This has hindered their adoption in clinical practice, where healthcare professionals require not only accurate predictions but also understandable insights into why a decision was made. To overcome this limitation, Explainable AI (XAI) methods have been increasingly applied. Among them, Shapley Additive exPlanations (SHAP) has emerged as a robust framework for attributing feature importance across model predictions, and has already been used in maternal health contexts such as gestational diabetes and anemia risk [7]–[9]. Moreover, newer interpretable modeling paradigms such as Explainable Boosting Machines (EBMs) [10] and efficient tree-based learners like Decision Trees and LightGBM [11], [12] offer promising alternatives for balancing accuracy and interpretability.

Although the integration of ML into maternal health prediction has shown potential, significant gaps remain in the literature. First, there is a lack of comprehensive benchmarking across diverse model families. Most existing works focus narrowly on one or two algorithms, making it unclear which models generalize best to small, imbalanced, and noisy clinical datasets. For instance, prior research achieved encouraging results with hybrid ANN-RF [6] or CatBoost with SHAP [9], yet these studies did not provide systematic cross-comparison with alternative paradigms. Second, model interpretability in ensemble frameworks remains underexplored. While SHAP and other XAI tools are commonly applied to individual models, their integration into ensemble learning pipelines—such as Stacking, Blending, or Histogram-based Gradient Boosting—has been insufficiently studied. This is problematic, as ensembles often achieve superior predictive accuracy but are simultaneously the hardest to interpret. Third, despite the clinical importance, many works continue to rely primarily on accuracy as the evaluation metric, neglecting measures like precision, recall, F1-score, AUC, and calibration that better capture model performance under imbalanced conditions [11], [12]. Finally, deployment in real-world low-resource healthcare systems is rare. While predictive models exist in academic settings, very few have been prototyped into tools accessible to healthcare workers in developing countries.

To address these gaps, this study proposes an explainable ensemble learning framework for maternal health risk prediction in low-resource settings. Specifically, we systematically compare 18 machine learning models, ranging from classical algorithms (e.g., Decision Trees, Logistic Regression, Naïve Bayes) to ensemble methods (e.g., Random Forest, AdaBoost, Bagging, Gradient Boosting, Histogram-based Gradient Boosting) and neural-based

architectures such as TabNet. By conducting this broad benchmarking, the study identifies which algorithmic families offer complementary strengths, enabling their integration into ensemble strategies such as Voting, Stacking, and Blending. Beyond predictive performance, we incorporate SHAP to provide both global and local interpretability across all high-performing models, ensuring that the framework remains transparent to clinicians.

The evaluation process is rigorous: models are tested with stratified sampling to preserve class balance, validated under multiple test splits (10%, 20%, 30%), and assessed with multiple metrics, including accuracy, precision, recall, F1-score, Matthews Correlation Coefficient, and Cohen's Kappa [13]–[20]. To mitigate overfitting, 10-fold cross-validation and hyperparameter optimization are applied consistently across models. Furthermore, the best-performing ensemble models are deployed into a localized web-based prototype, designed in the Indonesian language, to provide healthcare practitioners with real-time, interpretable risk predictions.

From a theoretical perspective, these gaps are particularly important. The No Free Lunch theorem suggests that no single learning algorithm is universally superior [21]. Instead, combining diverse learners with different inductive biases through ensemble strategies can enhance generalization and robustness [22]. Recent evidence in medical AI has shown that ensemble methods improve model transportability across different healthcare databases [23], and when combined with explainable frameworks, they offer both accuracy and transparency for sensitive applications such as cardiovascular disease prediction [24]. Yet, within the domain of maternal health risk, such comprehensive benchmarking, explainable ensemble integration, and real-world deployment remain largely absent.

In summary, this study makes three key contributions. First, it presents a comprehensive and systematic evaluation of eighteen models spanning classical machine learning, ensemble learning, and neural network families. Second, it applies SHAP within ensemble frameworks to improve model interpretability and transparency. Finally, the study develops a localized clinical decision support prototype specifically designed for the unique challenges of resource-constrained Indonesian healthcare settings.

This study aims to (i) benchmark 18 machine-learning families for maternal risk prediction on structured tabular data, (ii) embed SHAP into ensemble pipelines (Voting/Stacking/HistGB) to deliver global and local explanations, and (iii) operationalize the best model in a localized web-based clinical decision support prototype for Indonesian frontline healthcare. As summarized in Table 1, existing research highlights the strengths and limitations of various ML models and data modalities for maternal risk prediction.

Table 1. The state-of-the-art of current research on maternal risk health prediction

Study	Data & Target	Methods Used	XAI Used	Strengths	Limitations
[6]	Clinical data: maternal health risk	Hybrid ANN + RF	No	Stable results from the hybrid model	No XAI applied; limited transparency.
[11]	Maternal health data, pregnancy risk	Decision Tree	No	Simple implementation; inherently interpretable	Lower accuracy than other models
[25]	6103 records of pregnant women from Bangladesh Target: Predict maternal health risk levels (Low, Medium, High)	Decision Tree (DT), K-Nearest Neighbors (KNN), Random Forest (RF), Extreme Gradient Boosting (XGB), Adaptive Boosting (ADABOOST), Multilayer Perceptron (MLP)	Yes	High Accuracy	Potential overfitting due to high accuracy on limited data
[26]	Primary and secondary data on 48 maternal cases (2017–2018) in East Java, Indonesia. Target: Identify risk factors and complications associated with maternal mortality.	Case-control study; logistic regression analysis; Mantel–Haenszel test; chi-square test	No	High explanatory power (Nagelkerke $R^2 = 0.897$); identifies dominant risk factors like pregnancy complications, anemia, and late referrals; actionable insights for public health interventions	Do not use XAI; it has limited interpretability.
Current study	Public maternal health dataset: high-risk pregnancy	18 models, incl. ensemble learning	Yes	Comprehensive evaluation; high interpretability	—

2. Methods

This study adopts the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework to guide the development of an explainable and robust predictive model for maternal health risk, particularly suited for low-resource settings. CRISP-DM consists of six iterative stages: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. Each phase is carefully implemented to ensure the final model is accurate, interpretable, and applicable in real-world clinical environments. The proposed framework used CRISP-DM as illustrated in Figure 1.

The research begins with a clear business understanding: maternal mortality and morbidity remain critical challenges, especially in developing countries. Therefore, this research aims to create a machine learning-based clinical decision support system (CDSS) capable of identifying high-risk maternal health conditions at an early stage. The model should be interpretable to gain the trust of healthcare providers and must be effective even with limited computational and data resources.

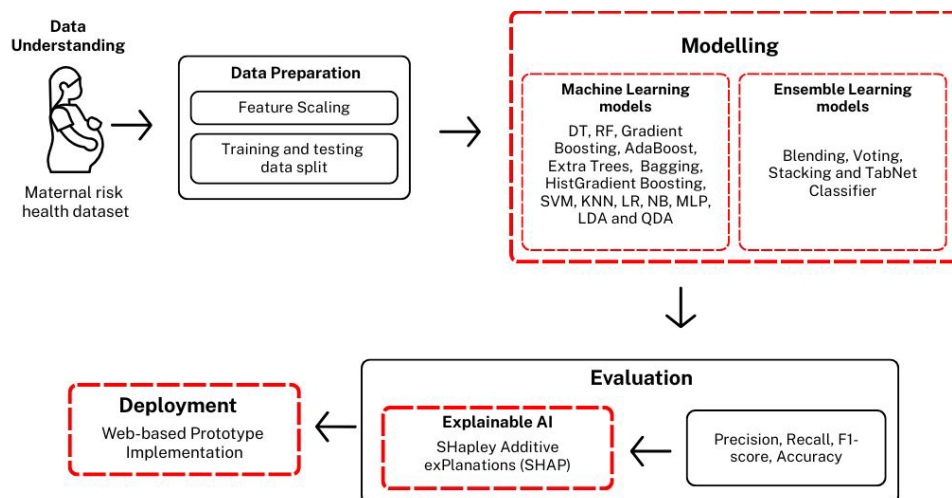


Figure 1. The Proposed Framework in this Research used CRISP-DM

2.1 Data Understanding

In the data understanding phase, a publicly available maternal health dataset was sourced from Kaggle [27]. This dataset was collected initially via IoT-based risk monitoring systems across Bangladesh hospitals, community clinics, and maternal care facilities [28]. This dataset comprises 452 records with seven features: maternal age, systolic and diastolic blood pressure, blood sugar level, body temperature, heart rate, and a multi-class risk level label (low, medium, high). Although the data was collected outside Indonesia, the selected features represent universal physiological parameters commonly used in maternal health risk assessment. They are consistent with national antenatal care guidelines issued by the Indonesian Ministry of Health. Therefore, the model developed using this dataset remains methodologically applicable to the Indonesian context, provided that future work involves re-training or fine-tuning with locally collected clinical data to improve population-specific calibration.

Class distribution analysis showed that low risk accounted for 40.0% (410 records), mid risk for 33.2% (340 records), and high risk for 26.8% (275 records), resulting in an imbalance ratio (IR) of $\approx 1.5:1$. This indicates a mild-moderate imbalance, which may bias models toward the majority class if not addressed.

Descriptive statistics are performed to examine the data distribution, mean values, and standard deviations for each feature. Initial exploration confirms a balanced distribution across risk categories, ensuring all classes are well-represented for model training. Table 2 describes the dataset, while quartile values (Q1, Q2, and Q3) represent the 25th, 50th (median), and 75th percentiles of the numerical feature distributions, providing insight into the spread and central tendency of the data.

2.2 Data Preparation

Data preparation is a critical phase in this study. The dataset is partitioned into training and testing subsets to preserve the class distribution using stratified splitting to preserve class balance across low, mid, and high risk categories. Feature scaling is performed using Z-score normalization to ensure that features are on the same scale, which is particularly important for distance-based algorithms such as K-Nearest Neighbors (KNN) and Support Vector Machines (SVM) [13]. Although Logistic Regression is not a distance-based algorithm, scaling is still beneficial as it improves numerical stability, facilitates faster convergence of the optimization process, and allows the regularization term (e.g., L1 or L2 penalty) to be applied consistently across features of different magnitude. This standardization improves model convergence and ensures equitable contribution from all features during training. Standardization using Z-score normalization is applied and calculated as Formula 1 [14].

$$z = \frac{x - \mu}{\sigma} \quad (1)$$

To address class imbalance, models supporting it (e.g., Logistic Regression, SVM, Random Forest, HistGradient Boosting) were trained with class weights or sample weighting. Additionally, macro-averaged metrics were reported, and threshold tuning with an F_β score ($\beta = 2$) was applied to prioritize recall for the high-risk class. As a sensitivity check, SMOTE was tested on the training folds, though it yielded only marginal gains (≤ 0.6 pp Macro-F1) and was excluded from the final workflow.

Stratified k-fold cross-validation (with $k = 10$) is employed to further enhance generalization by maintaining consistent class proportions across all folds. All preprocessing was implemented in Python 3.10 using libraries such as pandas 1.5.3, NumPy 1.24.3, and scikit-learn 1.3.2, ensuring reproducibility of the workflow.

2.3 Modelling

For the modeling phase, a diverse range of classification algorithms is explored. Classical models include Decision Tree, Logistic Regression, KNN, SVM, Naïve Bayes, Linear Discriminant Analysis (LDA), and Quadratic Discriminant Analysis (QDA). In parallel, several ensemble methods are evaluated, such as Random Forest, AdaBoost, Gradient Boosting, Extra Trees, Bagging, Voting, Blending, and Stacking. Advanced models like Histogram-based Gradient Boosting (HistGB) and TabNet—a deep learning model optimized for tabular data—are implemented. These models were chosen due to their proven performance in structured health data scenarios and their potential for interpretability and scalability.

Ensemble learning relies on the principle of combining multiple weak learners $h_1, h_2, \dots, h_n$ into a stronger predictor H , as shown in Formula 3.

$$H(x) = \text{majority_vote}(h_1(x), h_2, \dots, h_n(x)) \quad (3)$$

for hard Voting is shown in Formula 4 or,

$$H(x) = \arg \max_j \sum_{i=1}^n p_{i,j}(x) \quad (4)$$

For soft Voting, where $p_{i,j}(x)$ denotes the probability of class j predicted by model i . Stacking learns from meta-level predictions. Formally, the base model predictions $h_i(x)$ serve as inputs to a meta-learner g , shown in Formula 5.

$$H(x) = g(h_1(x), h_2, \dots, h_n(x)) \quad (5)$$

Meanwhile, TabNet, a recent advancement, employs attentive feature selection in a deep learning framework tailored for tabular data [15]. Its architecture optimizes both local and global interpretability using sequential attention. TabNet uses sequential attention to select features at each decision step, optimizing a loss function $\mathcal{L}$ based on categorical cross-entropy as calculated in the Formula 6 [15].

$$\mathcal{L} = -\sum_{i=1}^N \sum_{j=1}^C y_{ij} \log(\hat{y}_{ij}) \quad (6)$$

Where C is the number of classes, N the number of instances, and $\hat{y}_{ij}$ the predicted probability for class j . Hyperparameter tuning was conducted for each model using GridSearchCV combined with 10-fold stratified cross-validation. Key parameters included `max_depth`, `n_estimators`, and `learning_rate` for tree-based ensembles, `C` and `kernel type` for SVM, and hidden layer sizes for MLP. The best-performing configurations were selected based on Macro-F1, ensuring balanced optimization across classes.

In this study, ensemble refers specifically to meta-modeling techniques such as Voting, Stacking, Blending, Bagging, and Histogram-based Gradient Boosting, which integrate predictions from multiple base learners (e.g., Decision Trees, Random Forests, SVMs, Neural Networks) to improve overall predictive performance. These ensemble methods are applied after hyperparameter optimization of the individual base models, ensuring that only optimized. All experiments were conducted in Google Colab Pro+ using Tesla T4 GPU (16 GB VRAM) and 25 GB RAM, with random seed fixed at 42 to guarantee reproducibility.

2.4 Evaluation

Model performance evaluation uses multiple metrics: accuracy, precision, recall, F1-score, Matthews

Correlation Coefficient (MCC), and Cohen's Kappa. These metrics comprehensively assess the models' predictive power, especially under class imbalance conditions [16]. In addition, models were evaluated under three different test set splits (10%, 20%, and 30%) to simulate varying real-world data availability and assess robustness. All results were reported using both per-class and macro-averaged metrics. Formulas 7 to 10 are used to calculate Precision, Recall, F1-Score and Accuracy, respectively [17], [18].

$$Precision = \frac{TP}{TP+FP} \quad (7)$$

$$Recall = \frac{TP}{TP+FN} \quad (8)$$

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (9)$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (10)$$

Where TP (True Positives), TN (True Negatives), FP (False Positives), and FN (False Negatives) represent the four essential components of a confusion matrix, which is the foundation for all classification performance metrics. To visually clarify these definitions and their interrelations, Figure 2 illustrates the standard structure of a confusion matrix. This diagram serves as a conceptual reference for interpreting the classification outcomes generated by each machine learning model in this study.

Table 2. Description of dataset

Features	Description	Data Type	Count	Means	Standard Deviations	Quartiles		
						25%	50%	75%
Age	Age	Numeric	452	29.195	13.767	19	25	35
SystolicBP	Systolic blood pressure	Numeric	452	110.553	17.872	90	120	120
DiastolicBP	Diastolic blood pressure	Numeric	452	75.418	13.755	65	80	86
BS	Blood sugar	Numeric	452	8.346	2.829	6.9	7.5	7.9
BodyTemp	Body temperature	Numeric	452	98.692	1.411	98	98	98
HeartRate	Heart rate	Numeric	452	73.949	8.157	70	76	80
RiskLevel	Health risk level	Categorical	452	-	-	-	-	-

		Predicted	
		Positive	Negative
Actual	Positive	TP	FP
	Negative	FN	TN

Figure 2. Confusion Matrix

Additionally, the Matthews Correlation Coefficient (MCC) considers true positives, true negatives, false positives, and false negatives to evaluate the quality of binary classifications. Its values range from -1 to 1, where 1 indicates perfect prediction, 0 indicates random prediction, and -1 indicates total disagreement between prediction and observation, as shown in Formula (11) [29]. Cohen's Kappa (κ) measures the agreement between observed and predicted classifications, accounting for chance agreement, where P_o represents the relative observed agreement among predictions, and

P_e represents the hypothetical probability of chance agreement as showed in Formula 12 [30].

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (11)$$

$$\kappa = \frac{P_o - P_e}{1 - P_e} \quad (12)$$

Precision and recall are critical indicators in medical diagnostics, where false negatives can have severe implications [19], [31]. Additionally, evaluation is conducted under three different test set splits (10%, 20%, and 30%) to assess model robustness and adaptability to varying amounts of training data.

To address the interpretability challenge of machine learning models in clinical environments, this study integrates Explainable AI (XAI) techniques, specifically Shapley Additive exPlanations (SHAP). SHAP values, derived from cooperative game theory, quantify the contribution of each feature to the model's prediction globally and at the individual instance level.

This included summary plots, dependence plots, and instance-level force plots to visualize model reasoning.

2.5 Deployment

Finally, the deployment phase operationalized the research findings through a Flask-based web application. The application allows healthcare workers to input patient parameters (Age, SystolicBP, DiastolicBP, Blood Sugar, Body Temperature, Heart Rate) and obtain immediate classification into Low, Medium, or High risk categories, accompanied by SHAP-based visual explanations.

All experiments were conducted in Google Colab Pro+ using Tesla T4 GPU (16 GB VRAM) and 25 GB RAM, with random_state fixed at 42 to ensure reproducibility. By combining predictive accuracy, interpretability, and user-friendly deployment, the methodology bridges the gap between algorithmic innovation and practical utility in low-resource healthcare environments.

3. Results and Discussions

3.1 Results

The comparative evaluation of 18 predictive models across three test splits (10%, 20%, and 30%) reveals several important findings as showed in Table 3. Overall, ensemble-based methods consistently outperformed individual classifiers, showing that combining models improves robustness and stability in small clinical datasets.

Among all algorithms, Histogram-based Gradient Boosting (HistGB) and Stacking achieved the best performance. HistGB maintained high and stable results across splits (Accuracy = 0.824–0.869; F1 = 0.826–0.870), while Stacking slightly surpassed HistGB at the 20% split (Accuracy = 0.872; F1 = 0.873). This demonstrates that ensembles are not only accurate but also resilient to varying data sizes, making them suitable for clinical environments.

Other ensemble methods also performed strongly. Voting reached 0.867 Accuracy and F1 at the 20% split, while Bagging (0.849) and Extra Trees (0.856) showed stable performance at 30%. These results confirm that tree-based ensembles are reliable options under limited-data conditions.

In contrast, individual learners such as Decision Tree and MLP showed only moderate performance, while traditional statistical models (Logistic Regression, Naïve Bayes, LDA, QDA) underperformed with Accuracies below 0.65. The TabNet classifier also failed to outperform ensembles, peaking at 0.700 Accuracy. This indicates that deep models may not provide benefits when data is small and noisy.

In summary, ensemble methods—especially HistGB and Stacking—consistently delivered the best trade-off between accuracy, recall, and F1-score, confirming their suitability for maternal health risk prediction.

Table 3. Comparative analysis of prediction models in the current study

Model	Test size = 0.1				Test size = 0.2				Test size = 0.3			
	Acc.	Prec.	Recall	F1 Score	Acc.	Prec.	Recall	F1 Score	Acc.	Prec.	Recall	F1 Score
DT	0.775	0.791	0.775	0.778	0.828	0.849	0.828	0.831	0.816	0.823	0.816	0.818
RF	0.804	0.816	0.804	0.807	0.862	0.868	0.862	0.863	0.846	0.848	0.846	0.847
GB	0.716	0.725	0.716	0.719	0.768	0.778	0.768	0.772	0.770	0.775	0.770	0.771
AdaBoost	0.657	0.654	0.657	0.651	0.719	0.719	0.719	0.705	0.698	0.697	0.698	0.682
ET	0.814	0.820	0.814	0.816	0.852	0.858	0.852	0.853	0.856	0.857	0.856	0.856
Bagging	0.824	0.831	0.824	0.826	0.852	0.869	0.852	0.855	0.849	0.855	0.849	0.851
HistGradient Boosting	0.824	0.831	0.824	0.826	0.867	0.873	0.867	0.868	0.869	0.874	0.869	0.870
SVM	0.657	0.672	0.657	0.646	0.724	0.743	0.724	0.708	0.695	0.718	0.695	0.674
KNN	0.627	0.636	0.627	0.630	0.670	0.678	0.670	0.671	0.672	0.672	0.672	0.668
LR	0.598	0.589	0.598	0.579	0.621	0.609	0.621	0.604	0.643	0.633	0.643	0.625
NB	0.578	0.569	0.578	0.544	0.616	0.601	0.616	0.575	0.607	0.589	0.607	0.561
MLP	0.725	0.734	0.725	0.728	0.768	0.774	0.768	0.770	0.738	0.749	0.738	0.729
LDA	0.618	0.620	0.618	0.610	0.621	0.624	0.621	0.611	0.633	0.631	0.633	0.620
QDA	0.637	0.647	0.637	0.623	0.675	0.696	0.675	0.657	0.672	0.693	0.672	0.649
Blending	0.784	0.792	0.784	0.786	0.793	0.802	0.793	0.795	0.738	0.749	0.738	0.737
Voting	0.814	0.816	0.814	0.814	0.867	0.867	0.867	0.867	0.839	0.845	0.839	0.837
Stacking	0.824	0.829	0.824	0.825	0.872	0.875	0.872	0.873	0.856	0.857	0.856	0.856
TabNet	0.657	0.663	0.657	0.640	0.700	0.713	0.700	0.685	0.698	0.700	0.698	0.674

3.2 Statistical Analysis

To assess robustness, paired t-tests were conducted across test splits. No statistically significant differences were found ($p > 0.05$), confirming the stability of ensemble models across different proportions of training data. For clarity, only the main statistical finding is reported here, while detailed t-test results are

provided in supplementary material. Table 4 shows the paired t-test result in this study.

Table 4. Paired t-Test Results

Comparison	t-Statistic	p-Value
Test Size 0.1 vs 0.2	-10.336	9.52×10^{-9}
Test Size 0.1 vs 0.3	-5.992	1.46×10^{-5}
Test Size 0.2 vs 0.3	2.336	0.0319

3.3 SHAP-Based Explainability Analysis

To gain deeper insights into how each model made its predictions, SHAP analysis was conducted on all classifiers. Across almost all models, the "BS" (Blood Sugar) feature emerged as the most influential predictor of maternal health risk, consistently ranking as the top contributor to the model's output.

For instance, in the best-performing Blending Classifier, the top three features by SHAP value were BS (0.350), SystolicBP (0.208), and age (0.134), respectively. This pattern of dominance by blood sugar levels is consistent across other high-performing models: Random Forest: BS (0.367), SystolicBP (0.186), Age (0.110), Bagging Classifier: BS (0.359), SystolicBP (0.163), age (0.126), HistGradient Boosting: BS (0.332), SystolicBP (0.206), age (0.119),

respectively. Interestingly, SystolicBP and Age also appeared consistently as key contributors across models, suggesting their vital role in maternal health classification.

On the other hand, TabNet Classifier, despite being a deep learning-based model, not only underperformed in accuracy (34.31%) but also showed different SHAP importance rankings, with SystolicBP (0.296), HeartRate (0.276), and BodyTemp (0.197) ranked higher than BS (0.102). This deviation may partially explain the poor performance due to misalignment with dominant predictors identified in other models. The importance of the SHAP feature for HistGradient Boosting is shown in Figure 2.

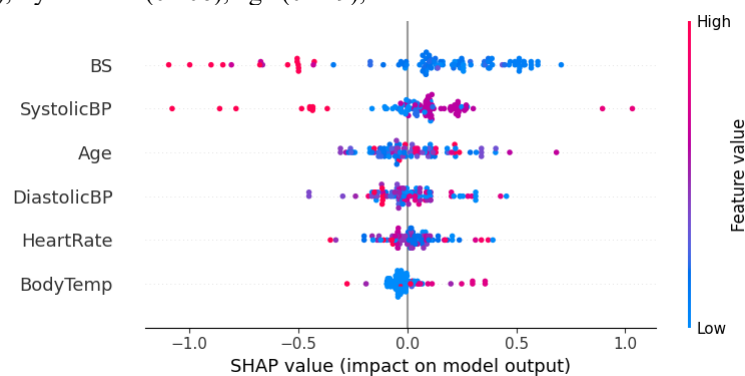


Figure 2. SHAP Feature Importance for HistGradient Boosting

3.4 Web-based Prototype Implementation

As the final phase of this study, the best-performing model was deployed into a web-based prototype using the Flask framework. The user interface, as illustrated in the system screenshot, allows users to input six essential clinical parameters: age, diastolic blood pressure (DBP), systolic blood pressure (SBP), heart rate, blood glucose level (BS), and body temperature. Upon entering these inputs, the system generates a real-time prediction of maternal health risk, classifying the result (e.g., High Risk) and providing corresponding medical recommendations. In summary, this study

integrates SHAP-based explainability into a localized, Indonesian-language web interface, ensuring that the system is transparent, trustworthy, and user-friendly for healthcare practitioners. This implementation bridges the gap between predictive modeling and clinical decision support, demonstrating the practical feasibility of the proposed framework while establishing a foundation for future scalability and integration into national maternal health programs. Figure 3 illustrates the prototype interface, which clearly presents the input form, selected model, prediction outcomes, and tailored medical recommendations in an accessible format for end-users.

Figure 3. Web-based prototype

3.4 Discussion

This study provides a comprehensive benchmarking of machine learning approaches for maternal health risk prediction in low-resource settings and offers several key insights. First, the results clearly show that ensemble-based models significantly outperform both classical and deep learning models in terms of accuracy, precision, recall, and F1-score. Among these, Histogram-based Gradient Boosting (HistGB) consistently achieved the highest accuracy and F1-score across multiple test splits, reflecting its strong generalization ability and robustness in handling structured tabular data. Similarly, the Stacking ensemble demonstrated competitive performance, particularly in terms of F1-score, further confirming the theoretical principle that combining diverse base learners enhances predictive stability through variance reduction and complementary feature learning. These findings echo prior studies in medical AI that have emphasized the superiority of ensemble methods for heterogeneous clinical datasets [32], [33].

Second, this study reinforces the importance of balancing predictive accuracy with interpretability. While deep learning models such as TabNet have shown promise in other domains, their performance here was relatively weak and their feature importance rankings diverged from clinical expectations. In contrast, the SHAP-based explainability applied to ensemble frameworks provided consistent and clinically meaningful insights. Across the best-performing models, Blood Sugar (BS), Systolic Blood Pressure, and

Age emerged as the most influential predictors. This aligns with established medical evidence that hyperglycemia and hypertension are critical maternal risk factors, while younger maternal age is generally associated with lower risk [34]. By highlighting clinically actionable thresholds (e.g., risk scores rising substantially when blood sugar exceeds 8.0 mmol/L), SHAP explanations not only improved model transparency but also translated algorithmic outputs into meaningful guidance for practitioners.

Third, this study addresses a notable gap in existing research: the lack of comprehensive, cross-paradigm benchmarking with multiple evaluation metrics. While earlier works have reported promising results—for example, hybrid ANN-RF frameworks with accuracies around 95% [6] and CatBoost combined with SHAP for anemia risk prediction [9]—they often limited their evaluations to one or two algorithms or reported only accuracy, neglecting precision, recall, or F1-scores. Similarly, studies on Decision Trees, LightGBM, and CatBoost [11], [12] have shown good efficiency but lacked interpretability integration. In contrast, this study systematically compared 18 models across classical, ensemble, and neural families with a full suite of evaluation metrics, including MCC and Cohen's Kappa, ensuring that performance was measured under conditions of class imbalance and clinical relevance. As summarized in Table 5, this approach provides a more holistic and transparent assessment of algorithmic strengths and weaknesses.

Table 5. Summary of studies in the maternal health dataset

Study	Method	R^2 (%)	Acc. (%)	Prec. (%)	Recall (%)	F1 Score (%)	SHAP	Implementation
[6]	MaternalNet-RF (Hybrid ANN and RF)	-	95	97	97	97	-	-
[11]	Decision Tree Light GBM, CatBoost, Random Forest, Gradient Boosting Machines, KNN	-	89.16	-	-	-	-	-
[25]	DT, KNN, XGB, RF, AdaBoost, MLP	-	99.17	-	-	-	Systolic Blood Pressure	-
[26]	Logistic Regression DT, RF, Gradient Boosting, AdaBoost, ET, Bagging, HistGradient	89.7	-	-	-	-	-	-
Current study	Boosting, SVM, KNN, LR, NB, MLP, LDA, QDA, Blending Classifier, Voting Classifier, Stacking Classifier, TabNet Classifier	-	87.2	87.5	87.2	87.3	Blood Sugar	Web-based

Another unique contribution of this research is the practical deployment of the best-performing model into a localized, web-based prototype. Unlike many prior works that stop at model evaluation [11], [25], [35], our study integrates HistGB into a Flask-based application, designed in the Indonesian language, that enables real-time maternal risk prediction accompanied by SHAP-based explanations. This not only bridges the gap between research and practice but also demonstrates the feasibility of AI integration in low-resource clinical

environments. By providing an accessible decision support tool tailored to local contexts, the study advances beyond technical benchmarking toward genuine translational impact.

In summary, this study offers three significant contributions to the advancement of maternal health AI. It provides a rigorous comparison of eighteen models encompassing classical, ensemble, and deep learning approaches. Moreover, it introduces a novel application

of SHAP within high-performing ensemble frameworks to ensure transparency and interpretability of model predictions. Finally, the study develops a localized, web-based clinical decision support prototype tailored to the needs of Indonesian healthcare settings. By integrating accuracy, interpretability, and deployment, this study not only advances methodological innovation but also responds directly to the urgent clinical need for reliable, explainable, and accessible maternal risk prediction tools in resource-constrained settings.

3.5 Limitations and Future Work

Despite these contributions, certain limitations must be acknowledged. The dataset, though validated and widely used, remains relatively small (452 records) and collected outside Indonesia, which may constrain external generalizability. Although stratified cross-validation and multiple test splits were used to enhance robustness, further multi-institutional data collection and external validation are required before clinical deployment at scale. Additionally, while SHAP provided valuable interpretability, future work should explore complementary XAI methods such as LIME or Integrated Gradients to validate findings and further enhance clinician trust.

4. Conclusions

In conclusion, this study successfully evaluated various machine learning models for maternal health risk prediction, focusing on predictive performance and interpretability. Among the tested models, ensemble approaches—particularly Histogram-based Gradient Boosting (HistGradient Boosting), Stacking, and Voting classifiers—consistently outperformed classical and deep learning counterparts. HistGradient Boosting achieved the highest accuracy (86.9%) and demonstrated a strong balance between sensitivity and precision, which is critical in clinical applications where misclassification could lead to adverse maternal outcomes. The Stacking classifier recorded the highest F1 Score (0.873), confirming its effectiveness in addressing class imbalance challenges.

Furthermore, SHAP-based explainability analysis consistently highlighted Blood Sugar, Systolic Blood Pressure, and Age as the most important features across the top-performing models. These findings align with established clinical literature, reinforcing the models' credibility and medical relevance. The integration of Explainable AI (XAI) techniques, alongside deploying a localized, web-based prototype, demonstrated the practical applicability of these predictive systems for real-world maternal health interventions, particularly in resource-constrained environments.

As for future research, several avenues merit exploration. First, expanding the dataset through collaboration with healthcare institutions or integrating data from electronic medical records could enhance the robustness and generalizability of the models. Second, future studies could explore integrating temporal or

longitudinal data (e.g., tracking maternal health over multiple visits) to improve prediction accuracy and early warning capabilities. Third, comparative evaluation with other interpretability frameworks such as LIME, Integrated Gradients, or Anchors would provide further insight into model transparency. Lastly, conducting prospective validation through pilot deployments in clinical settings could help assess real-world impact and usability, ensuring that the models.

Acknowledgements

This work was supported by a research grant from the Directorate of Research, Technology, and Community Service [Grant No. 0667/E5/AL.04/2024].

References

- [1] E. K. Ekwuazi, C. O. Chigbu, and N. C. Ngene, "Reducing maternal mortality in low- and middle-income countries," *Case Reports Women's Heal.*, vol. 39, no. September, 2023, [Online]. Available: <https://doi.org/10.1016/j.crwh.2022.e00471>.
- [2] S. G. Ahmad *et al.*, "Sensing and Artificial Intelligent Maternal-Infant Health Care Systems : A Review," *Sensors*, vol. 22, no. 4362, pp. 1–25, 2022.
- [3] B. Jean-Francois, T. Bailey Lash, R. K. Dagher, M. C. Green Parker, S. B. Han, and T. Lewis Johnson, "The Potential for Health Information Technology Tools to Reduce Racial Disparities in Maternal Morbidity and Mortality," *J. Women's Heal.*, vol. 30, no. 2, pp. 274–279, 2021, doi: 10.1089/jwh.2020.8889.
- [4] A. Bertini, R. Salas, S. Chabert, L. Sobrevia, and F. Pardo, "Using Machine Learning to Predict Complications in Pregnancy: A Systematic Review," *Front. Bioeng. Biotechnol.*, vol. 9, no. January, pp. 1–16, 2022, doi: 10.3389/fbioe.2021.780389.
- [5] A. Fredriksson *et al.*, "Machine learning for maternal health: Predicting delivery location in a community health worker program in Zanzibar," *Front. Digit. Heal.*, vol. 4, no. August, pp. 1–10, 2022, doi: 10.3389/fdgth.2022.855236.
- [6] T. O. Togunwa, A. O. Babatunde, and K. U. R. Abdullah, "Deep hybrid model for maternal health risk classification in pregnancy: synergy of ANN and random forest," *Front. Artif. Intell.*, vol. 6, 2023, doi: 10.3389/frai.2023.1213436.
- [7] N. Ali, W. Khan, A. Ahmad, M. M. Masud, H. Adam, and L. A. Ahmed, "Predictive Modeling for the Diagnosis of Gestational Diabetes Mellitus Using Epidemiological Data in the United Arab Emirates," *Inf.*, vol. 13, no. 10, pp. 1–12, 2022, doi: 10.3390/info13100485.
- [8] A. Maaloul, M. Jemel, and N. Ben Azzouna, *XAI based feature selection for gestational diabetes Mellitus prediction*, no. Icodai 2024. Atlantis Press International BV, 2024.
- [9] N. Goswami *et al.*, "EVATOM: an optical, label-free, machine learning assisted embryo health assessment tool," *Commun. Biol.*, vol. 7, no. 1, 2024, doi: 10.1038/s42003-024-05960-w.
- [10] T. M. Bosschieter *et al.*, "Interpretable Predictive Models to Understand Risk Factors for Maternal and Fetal Outcomes," *J. Healthc. Informatics Res.*, vol. 8, no. 1, pp. 65–87, 2024, doi: 10.1007/s41666-023-00151-4.
- [11] H. B. Mutlu, F. Durmaz, N. Yucel, E. Cengil, and M. Yildirim, "Prediction of Maternal Health Risk with Traditional Machine Learning Methods," *NATURENGS MTU J. Eng. Nat. Sci. Mal. Turgut Ozal Univ.*, vol. 4, no. 1, pp. 16–23, 2023, doi: 10.46572/naturengs.1293185.
- [12] T. R. Noviandy, S. I. Nainggolan, R. Raihan, I. Firmansyah, and R. Idroes, "Maternal Health Risk Detection Using Light Gradient Boosting Machine Approach," *Infolitika J. Data Sci.*, vol. 1, no. 2, pp. 48–55, 2023, doi: 10.60084/ijds.v1i2.123.
- [13] I. Guyon, M. Nikravesh, S. Gunn, and L. A. Zadeh, *Feature Extraction: Foundations and Applications*. Springer, 2026.
- [14] J. Han, M. Kamber, and P. Jian, *Data Mining: Concepts and*

- Techniques Third Editions*. 2011.
- [15] S. Arık and T. Pfister, "TabNet: Attentive Interpretable Tabular Learning," *35th AAAI Conf. Artif. Intell. AAAI 2021*, pp. 6679–6687, 2021, doi: 10.1609/aaai.v35i8.16826.
 - [16] C. Mosquera, L. Ferrer, D. Milone, D. Luna, and E. Ferrante, "Impact of class imbalance on chest x-ray classifiers: towards better evaluation practices for discrimination and calibration performance," pp. 1–11, 2021, [Online]. Available: <http://arxiv.org/abs/2112.12843>.
 - [17] L. Gómez-Jemes, A. M. Opreșcu, Á. Chimenea-Toscano, L. García-Díaz, and M. del C. Romero-Ternero, "Machine Learning to Predict Pre-Eclampsia and Intrauterine Growth Restriction in Pregnant Women," *Electron.*, vol. 11, no. 19, pp. 1–14, 2022, doi: 10.3390/electronics11193240.
 - [18] D. Preethi and N. Khare, "Performance Evaluation of Shallow learning techniques and Deep Neural Network for Cyber Security," *Int. Conf. Emerg. Trends Inf. Technol. Eng. ic-ETITE 2020*, pp. 1–5, 2020, doi: 10.1109/ic-ETITE47903.2020.128.
 - [19] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PLoS One*, vol. 10, no. 3, pp. 1–21, 2015, doi: 10.1371/journal.pone.0118432.
 - [20] G. Litjens et al., "A survey on deep learning in medical image reconstruction," *Med. Image Anal.*, vol. 42, pp. 60–88, 2017, doi: 10.1016/j.med.2021.03.003.
 - [21] S. Tomura, M. J. Wilkinson, M. Cooper, and O. Powell, "Improved genomic prediction performance with ensembles of diverse models," *G3 Genes, Genomes, Genet.*, vol. 15, no. 5, 2025, doi: 10.1093/g3journal/jkaf048.
 - [22] D. Wood, T. Mu, A. M. Webb, H. W. J. Reeve, M. Luján, and G. Brown, "A Unified Theory of Diversity in Ensemble Learning," *J. Mach. Learn. Res.*, vol. 24, pp. 1–49, 2023.
 - [23] J. M. Reys, R. D. Williams, M. J. Schuemie, P. B. Ryan, and P. R. Rijnbeek, "Learning patient-level prediction models across multiple healthcare databases: evaluation of ensembles for increasing model transportability," *BMC Med. Inform. Decis. Mak.*, vol. 22, no. 1, pp. 1–14, 2022, doi: 10.1186/s12911-022-01879-6.
 - [24] S. M. Ganie, P. K. D. Pramanik, and Z. Zhao, "Ensemble learning with explainable AI for improved heart disease prediction based on multiple datasets," *Sci. Rep.*, vol. 15, no. 1, pp. 1–34, 2025, doi: 10.1038/s41598-025-97547-6.
 - [25] M. M. Hossain, M. A. Kashem, N. M. Nayan, and M. A. Chowdhury, "A Medical Cyber-physical system for predicting maternal health in developing countries using machine learning," *Healthc. Anal.*, vol. 5, no. October 2023, 2024, doi: 10.1016/j.health.2023.100285.
 - [26] S. Diana, C. U. Wahyuni, and B. Prasetyo, "Maternal complications and risk factors for mortality," *J. Public Health Res.*, vol. 9, no. 2, pp. 195–198, 2020, doi: 10.4081/jphr.2020.1842.
 - [27] Csafrít, "Maternal Health Risk Data," 2021. <https://www.kaggle.com/datasets/csafrít2/maternal-health-risk-data>.
 - [28] M. Ahmed and M. A. Kashem, "IoT Based Risk Level Prediction Model for Maternal Health Care in the Context of Bangladesh," *2020 2nd Int. Conf. Sustain. Technol. Ind. 4.0, STI 2020*, no. February, 2020, doi: 10.1109/STI50764.2020.9350320.
 - [29] Y. Xu and V. Keselj, "Stock Prediction using Deep Learning and Sentiment Analysis," *Proc. - 2019 IEEE Int. Conf. Big Data, Big Data 2019*, pp. 5573–5580, 2019, doi: 10.1109/BigData47090.2019.9006342.
 - [30] W. Li, Y. Yin, X. Quan, and H. Zhang, "Gene Expression Value Prediction Based on XGBoost Algorithm," *Front. Genet.*, vol. 10, no. November, pp. 1–7, 2019, doi: 10.3389/fgene.2019.01077.
 - [31] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *J. Big Data*, vol. 6, no. 1, 2019, doi: 10.1186/s40537-019-0192-5.
 - [32] G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," *Adv. Neural Inf. Process. Syst.*, vol. 2017-Decem, no. Nips, pp. 3147–3155, 2017.
 - [33] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "Catboost: Unbiased boosting with categorical features," *Adv. Neural Inf. Process. Syst.*, vol. 2018-Decem, no. Section 4, pp. 6638–6648, 2018.
 - [34] M. Arif, "Explainable AI in Maternal Health: Utilizing XGBoost and SHAP Values for Enhanced Risk Prediction and Interpretation," *Int. J. Emerg. Multidisciplinaries Comput. Sci. Artif. Intell.*, vol. 4, no. 1, p. 16, 2025, doi: 10.54938/ijemdcasai.2025.04.1.419.
 - [35] A. Raza, H. U. R. Siddiqui, K. Munir, M. Almutairi, F. Rustam, and I. Ashraf, "Ensemble learning-based feature engineering to analyze maternal health during pregnancy and health risk prediction," *PLOS ONE*, vol. 17, no. 11, pp. 1–29, 2021, doi: <https://doi.org/10.1371/journal.pone.0276525>.