Journal of Innovations in Computer Science and Engineering
Shahid Beheshti Unversity

January 2026, Volume 3, Issue 2

Machine Learning and Hybrid Ensemble Models for Maternal Health Risk Prediction

Mohammad Reza Einollahi Asgarabad, **CODE ORCID**: 0009-0007-1118-7738
Department of Biomedical Engineering, SR.C., Islamic Azad University, Tehran, Iran, mohammadreza.einollahi@iau.ir
Mahsa Akhbari1✉, **CODE ORCID:** 0000-0002-0623-4751
Department of Biomedical Engineering, SR.C., Islamic Azad University, Tehran, Iran, akhbari.mahsa@iau.ac.ir

ABSTRACT

This study presents an applied, multi-stage machine-learning pipeline for maternal health risk classification using the Maternal Health Risk dataset (1,014 samples, six clinical features, and three risk classes). Six classifiers were evaluated across four preprocessing stages: training-fold standardization alone, Edited Nearest Neighbors (ENN)-based cleaning, Synthetic Minority Oversampling Technique (SMOTE)-based class balancing, and sequential ENN-SMOTE preprocessing. Within each fold of 10-fold cross-validation, the scaler, ENN, SMOTE, model fitting, and ensemble construction were performed using only the training fold; the corresponding test fold remained untouched until evaluation. A fixed majority-voting ensemble comprising Random Forest, Extra Trees, and XGBoost was evaluated using class-wise, macro-averaged, and weighted-averaged accuracy, precision, and F1-score. In Stage 4, the hybrid ensemble achieved a weighted accuracy of 91.5% ± 3%, precision of 88% ± 3%, and F1-score of 87.2% ± 3%. The ensemble's precision for the high-risk class exceeded 90% across all four stages. These findings indicate competitive predictive performance on this dataset; however, direct comparison with previous studies is limited by differences in preprocessing methods, validation protocols, and model configurations. External and prospective validation is required before clinical use.

CC BY

Keywords:
Maternal Health Risk, Prediction, Machine Learning, Hybrid Ensemble Model, Outlier Removal, Class Balancing

1. Introduction

In 2015, the United Nations introduced the Sustainable Development Goals (SDGs) to address global challenges, with one of the key objectives being to ensure health and well-being for all individuals at all ages, particularly pregnant women [1]. Maternal health during pregnancy, childbirth, and the postpartum period plays a critical role in women's overall health, and reducing preventable maternal and child mortality [2]. According to a United Nations report, approximately 287,000 women worldwide lost their lives during or after pregnancy in 2020, equating to an average of one maternal death every two minutes [1]. Additionally, around 6,700 newborns and 810 pregnant women die daily, underscoring the urgent need for effective interventions to mitigate pregnancy-related risks and enhance prenatal care [2]. The World Health Organization (WHO) reported that, in 2017, an average of 810 women died daily due to pregnancy or childbirth complications, with the number rising to 303,000 in 2019 [3]. Despite substantial improvements in maternal healthcare over the past decades, maternal mortality remains an important public health concern in Iran. Recent estimates indicate that Iran's maternal mortality ratio was 21 deaths per 100,000 live births in 2022 and 16 deaths per 100,000 live births in 2023, highlighting the continuing need for effective maternal health surveillance and interventions [4].

Pregnancy is a complex process accompanied by extensive physical and psychological changes. Factors such as very young or advanced maternal age, obesity, gestational diabetes, hypertension, preeclampsia, abnormal heart rate, and elevated body temperature may increase the risk of complications for the mother and fetus, making timely risk assessment important [5, 6]. Accurate assessment during pregnancy, delivery, and the postpartum period is particularly important in underserved settings with limited access to healthcare services [7, 8]. Advances in information technology and data analysis have enabled machine-learning methods to be applied to pregnancy-related data. These methods can identify statistical patterns and support risk classification, but their reported performance depends on dataset quality, class distribution, preprocessing, and validation design [6-9]. Class imbalance is a recurring concern because a classifier may favor the majority class [9]. Resampling methods such as SMOTE and SMOTE-ENN have therefore been evaluated to balance class representation and reduce the influence of inconsistent samples [9].

Submit Date: 2025-12-05
Revise Date: 2026-08-05
Accept Date: 2026-08-15
✉ Corresponding author

In recent years, several machine-learning techniques have been evaluated for maternal health risk prediction. Khadidos et al. (2024) proposed the Quad-Ensemble Machine Learning framework (QEML-MHRC), integrating bagging, boosting, stacking, and voting with Decision Tree (DT), Random Forest (RF), Gradient Boosting Tree (GBT), and K-Nearest Neighbors (KNN) models. Their study reported a weighted F1-score of 0.86 across the risk classes and up to 90% accuracy for the high-risk class. The authors emphasized the importance of preprocessing and also noted limitations related to validation across diverse populations [6].

Similarly, Setiawan et al. (2024) compared Decision Tree, Bagging with DT, AdaBoost with DT, and Random Forest using the Maternal Health Risk dataset. Random Forest with 50 decision trees produced the highest accuracy and recall in their experiments, while limitations in sample size and demographic diversity indicated the need for further validation [1]. Mauliza et al. (2024) [9] evaluated SMOTE, SMOTE-ENN, ADASYN (Adaptive Synthetic Sampling Approach for Imbalanced Learning), and ADASYN-ENN with Decision Tree, Random Forest, KNN, Naive Bayes, and Support Vector Machine (SVM). Their reported results showed that SMOTE-ENN achieved up to 96% accuracy for Decision Tree and Random Forest. The study also discussed the risks of overfitting with oversampling and the possible introduction of noise by synthetic-sampling methods [9].

In another study, Raihen and Akter (2024) compared several classification methods using an IoT-based maternal health risk dataset from rural Bangladesh. SVM achieved the highest reported accuracy (86.15%) under their evaluation protocol [2]. Wongso et al. (2023) compared SVM, Naive Bayes, and KNN with and without K-means preprocessing; KNN with K-means achieved 79.53% accuracy and an F1-score of 0.80 and was implemented in a web-based application [5]. Tiwari and Kumar (2023) evaluated multiple algorithms on the Maternal Health Risk dataset and reported the highest area under the Precision-Recall curve for XGBoost [7].

Windasari (2022) evaluated a particle-swarm-based approach for improving C5.0 decision-tree classification on maternal health data and reported improved predictive accuracy under the study's experimental setting [3]. Mutlu et al. (2023) compared Decision Tree, CatBoost, Random Forest, Gradient Boosting Machines, and KNN. Decision Tree achieved the highest reported accuracy for the high-risk class, while CatBoost and Random Forest performed well for low-risk cases. The authors also noted limitations associated with dataset scope and external validity [8]. Zemariam et al. (2024) applied supervised machine-learning models, including Random Forest, CatBoost, and Light Gradient Boosting, to predict micronutrient intake among Ethiopian children and reported SMOTE-ENN as the most effective balancing method in their experiments. Although the clinical target differs from maternal health risk, the study illustrates the broader use of resampling methods in health-related classification and the need to consider sampling bias and external validation [10].

Previous studies have shown that preprocessing and ensemble methods can affect predictive performance in maternal health risk classification; however, reported results vary because of differences in validation procedures, resampling strategies, and model configurations. The present study therefore provides an applied, comparative evaluation of a multi-stage pipeline that combines ENN-based cleaning, SMOTE-based class balancing, six established classifiers, and a fixed majority-voting ensemble. The study does not introduce a new learning algorithm. Its contribution is the consistent evaluation of these components within a leakage-free, fold-wise validation procedure and the reporting of class-wise, macro-averaged, and weighted-averaged performance. The remainder of the paper is organized as follows. Section 2 describes the dataset, preprocessing stages, classification models, validation protocol, ensemble construction, and evaluation metrics. Section 3 reports the results of the individual classifiers and the hybrid ensemble. Section 4 discusses the findings, limitations, comparison with previous studies, and conclusions.

2. Materials and Methods

This section describes the dataset, data inspection, standardization, fold-wise preprocessing stages, classification models, majority-voting ensemble, validation protocol, and evaluation metrics. Every operation that learned from the data, including scaler fitting, ENN, SMOTE, model fitting, and ensemble construction, was confined to the training portion of each cross-validation fold.

2.1. Proposed method

Six classification algorithms and one fixed majority-voting ensemble were evaluated, as summarized in Table 1. Default parameter settings were used for all six classification models, and no parameter-search procedure was conducted. The ensemble contained Random Forest, Extra Trees, and XGBoost in every fold; each component was fitted only on the processed training fold, and test-fold outcomes were not used to select or modify the ensemble.

Table 1: Machine-learning algorithms and hybrid ensemble evaluated in this study

Algorithm	Rationale and key characteristics
Extra Trees Classifier (ETC)	Extremely randomized decision-tree ensemble. Random split thresholds increase diversity among trees, and predictions are combined by majority vote.
K-Nearest Neighbors (KNN)	Instance-based classifier that assigns a label from neighboring training observations after feature standardization.
Multilayer Perceptron (MLP)	Feedforward neural network trained by backpropagation to model nonlinear relationships among the clinical variables.
Random Forest Classifier (RF)	Bootstrap ensemble of decision trees with random feature selection; class predictions are aggregated by majority vote.
Support Vector Machine (SVM)	Margin-based classifier for nonlinear class separation using the library implementation applied after feature standardization.
Extreme Gradient Boosting (XGBoost)	Gradient-boosted decision-tree ensemble in which trees are added sequentially with regularization.
Hybrid Ensemble Model	Fixed majority-voting combination of RF, ETC, and XGBoost. Every component was fitted only on the processed training fold.

The performance of the proposed models was evaluated using standard classification metrics, as summarized in Table 2. In the table, each metric is presented alongside its rationale and key strengths, providing a concise assessment of model performance across maternal health risk categories.

Table 2: Evaluation metrics

Metric	Rationale and Strengths
Accuracy (ACC)	Provides a global measure of classification correctness by quantifying the proportion of correctly predicted instances across all risk classes.
Precision (PRE)	Reflects the reliability of positive predictions by measuring the proportion of true positives among predicted positives for each risk level.
F1 Score (F1)	Combines precision and sensitivity into a single metric, offering a balanced evaluation of model performance, particularly in multiclass and imbalanced settings.

The objective was to compare the six individual classifiers and the fixed hybrid ensemble across the four preprocessing stages under the same leakage-free validation protocol. Comparisons with prior studies are presented descriptively because their preprocessing methods, validation protocols, and model configurations differ.

2.2. Dataset description

The Maternal Health Risk dataset [11] contains seven columns: six input features and one class label. The input features are Age, Systolic Blood Pressure (SystolicBP), Diastolic Blood Pressure (DiastolicBP), Blood Sugar (BS), Body Temperature (BodyTemp), and Heart Rate (HeartRate). The target variable, RiskLevel, contains the original dataset labels low risk, mid risk, and high risk. Figure 1 presents the Pearson correlation matrix of the six input features

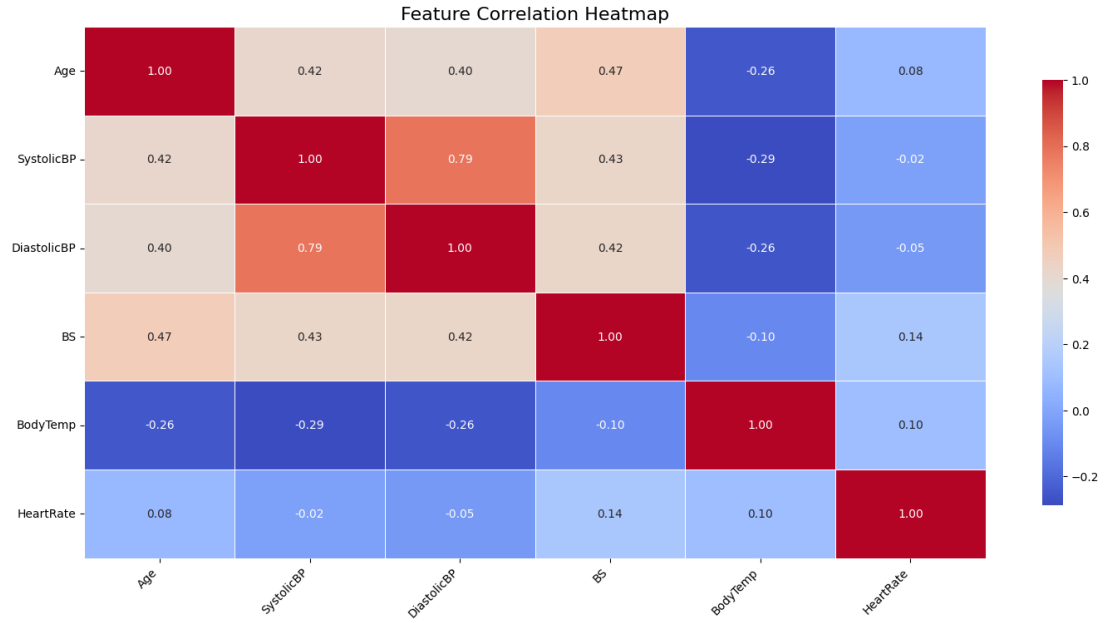


Fig. 1. Pearson correlation matrix of the six input features used for maternal health risk prediction. Red represents positive correlation, blue represents negative correlation, and values near zero indicate weak linear correlation.

The dataset comprises 1014 samples. The distribution of classes was analyzed to identify any potential imbalances in the data. The prevalence of classes is as follows: low risk (40.04%), mid risk (33.14%), and high risk (26.82%). These results indicate a slight class imbalance in the dataset. To gain a comprehensive understanding of the dataset's characteristics, a set of descriptive statistics is presented in Table 3. These statistics include measures such as the count, mean, standard deviation (std), minimum (min), 25th percentile (25%), median (50%), 75th percentile (75%), and maximum (max) for each feature in the dataset.

Table 3: Descriptive statistics of the features in the dataset samples

Feature	Mean	std	min	25%	50%	75%	max
Age	29.87	13.47	10	19	26	39	70
SystolicBP	113.2	18.4	70	100	120	120	160
DiastolicBP	76.46	13.88	49	65	80	90	100
BS	8.72	3.29	6	6.9	7.5	8	19
BodyTemp	98.66	1.37	98	98	98	98	103
HeartRate	74.3	8.09	70	70	76	80	90

2.3. Computational environment and preprocessing

All experiments were conducted in Python 3.13.0 using scikit-learn, imbalanced-learn, and XGBoost. The preprocessing and modeling workflow was implemented as a fold-wise pipeline so that information from the test fold could not influence scaler fitting, resampling, model training, or ensemble construction.

2.3.1. Data inspection and cleaning

In the first stage, the dataset was examined for missing values. Fortunately, no missing data were identified, eliminating the need for imputation techniques. This ensured the dataset's completeness for subsequent preprocessing and modeling steps.

2.3.2. Data standardization

Because the input variables are measured on different scales, StandardScaler was used within each cross-validation fold. The scaler was fitted exclusively on the training fold. The resulting training-fold mean and standard deviation were then used to transform both the training fold and the corresponding test fold according to Equation (1). The test fold was never used to estimate scaling parameters.

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

Where X is the original feature value, μ is the feature mean calculated from the training fold, σ is the feature standard deviation calculated from the training fold, and z is the standardized value.

2.3.3. Cross-validation

Stratified 10-fold cross-validation was used to evaluate all experimental stages. In each iteration, nine folds were used for training and the remaining fold for testing. StandardScaler was fitted exclusively on the training folds and then applied to the corresponding test fold. Depending on the experimental stage, ENN, SMOTE, or their combination was applied only to the training data. The classifiers and the fixed voting ensemble were subsequently trained on the resulting data, and performance was evaluated on the untouched test fold. Thus, the test data were used solely for evaluation and were not edited, resampled, included in model training, or involved in ensemble selection. Finally, the fold-level results were summarized as the mean ± standard deviation across the 10 test folds.

The leakage-free procedure followed in each cross-validation iteration is summarized in the following steps:

1. Split the original data into nine training fold and one test fold.
2. Fit StandardScaler using only the training fold.
3. Transform the training and test folds using the scaler fitted on the training fold.
4. Apply the stage-specific ENN and/or SMOTE operation only to the transformed training fold.
5. Apply each classifier using only the processed training fold.
6. Construct the fixed RF-ETC-XGBoost majority-voting ensemble from models fitted on the processed training fold.
7. Generate predictions and calculate metrics on the untouched test fold.

8. Repeat for all 10 folds and report the mean ± standard deviation of each evaluation metric.

2.3.4. Preprocessing stages

Default parameter settings were used for all six classification models. The four experimental stages differed only in the preprocessing operation applied within the training fold, as described below:

Stage 1: Baseline Preprocessing

Within each cross-validation iteration, StandardScaler was fitted on the training fold and used to transform both training and test folds. No ENN or SMOTE operation was applied. The classifier or ensemble was fitted on the standardized training fold and evaluated on the untouched test fold.

Stage 2: Outlier Removal with Edited Nearest Neighbors (ENN)

Within each cross-validation iteration, ENN was applied only to the standardized training fold to remove samples whose labels were inconsistent with their local training-data neighborhoods. The corresponding test fold was not edited or resampled. For descriptive context only, applying the same ENN procedure once to the complete dataset would remove 52 samples and leave 962 samples: 391 low-risk samples (40.64%), 320 mid-risk samples (33.26%), and 251 high-risk samples (26.09%). These complete-dataset counts are descriptive statistics only; they were not used to create any training or test fold and were not used to calculate the reported performance values.

Stage 3: Class Balancing with Synthetic Minority Oversampling Technique (SMOTE)

Within each cross-validation iteration, SMOTE was applied only to the standardized training fold to generate synthetic samples for underrepresented classes. The corresponding test fold retained its original observations and class distribution. For descriptive context only, the complete dataset contains 406 low-risk, 336 mid-risk, and 272 high-risk samples; applying SMOTE once to the complete dataset would produce 1,218 samples, with 406 samples in each class.

Stage 4: Combined outlier removal and class balancing

Within each cross-validation iteration, ENN was first applied only to the standardized training fold, followed by SMOTE only on the ENN-processed training fold. The test fold remained unchanged apart from transformation by the training-fitted scaler. For descriptive context only, applying ENN to the complete dataset would leave 962 samples (391 low risk, 320 mid risk, and 251 high risk), and subsequently applying SMOTE would produce 1,173 samples with 391 samples in each class. These complete-dataset values are descriptive only and were not used to preprocess any test fold or to generate the reported results.

2.4. Classification models

The study evaluated six established classifiers: Extra Trees Classifier (ETC), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest (RF), Multilayer Perceptron (MLP), and Extreme Gradient Boosting (XGBoost). The descriptions below summarize their general operating principles; all classifiers were implemented with default parameter settings.

2.4.1. Random Forest (RF)

Random Forest is an ensemble of decision trees trained on bootstrap samples with random feature subsets considered at each split. For classification, the final class is obtained by aggregating the votes of the individual trees. This design can reduce variance relative to a single decision tree and is suitable for nonlinear relations in tabular data [12].

2.4.2. Extra Trees Classifier (ETC)

Extra Trees Classifier is a decision-tree ensemble that introduces additional randomization by selecting split thresholds randomly. Predictions are aggregated across the trees by majority vote. The added randomization can reduce variance and computational cost in tabular classification tasks [13].

2.4.3. K-Nearest Neighbors (KNN)

K-Nearest Neighbors is a non-parametric, instance-based classifier. A new observation is assigned the class most common among its neighboring training observations according to a distance measure. Because distances are affected by feature scale, KNN was used after training-fold standardization [14].

2.4.4. Support Vector Machine (SVM)

Support Vector Machine separates classes by identifying a decision boundary with a maximum margin between classes. For nonlinear data, kernel functions map the observations into a space in which class separation may be more effective. Multiclass prediction was handled by the library implementation used in this study [15].

2.4.5. Multilayer Perceptron (MLP)

Multilayer Perceptron is a feedforward neural network composed of an input layer, one or more hidden layers, and an output layer. The network learns nonlinear relationships through backpropagation and nonlinear activation functions [16].

2.4.6. Extreme Gradient Boosting (XGBoost)

XGBoost is a gradient-boosted decision-tree method in which trees are added sequentially to reduce errors from the preceding ensemble. The method incorporates regularization and efficient tree construction and is widely used for structured classification data [17].

2.5. Model evaluation metrics

Model performance was evaluated using accuracy, precision, and F1-score [20, 21, 22]. Because the target variable contains three classes, class-wise metrics were calculated using a one-vs-rest formulation: each risk class was treated as the positive class in turn and the other two classes were treated as negative. Macro averages assign equal weight to each class, whereas weighted averages weight each class by its support. Fold-level metric values were summarized as mean ± standard deviation across the 10 test folds.

Accuracy: The model's overall accuracy is calculated as the ratio of correct predictions to the total number of predictions. Its formula is:

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})} \quad (2)$$

Precision: Also known as positive predictive value (PPV), precision indicates the model's ability to identify positive cases correctly and is calculated as:

$$\text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})} \quad (3)$$

F1 Score: The F1 score represents the harmonic mean of precision and Sensitivity (where $\text{Sensitivity} = \frac{\text{TP}}{(\text{TP}+\text{FN})}$), summarizing both metrics into a single value. It is calculated as:

$$\text{F1} = \frac{\text{Precision} \times \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}} \times 2 \quad (4)$$

These metrics were selected because they are widely used for multiclass classification and allow both overall and class-sensitive evaluation [20, 21, 22].

3. Results

The six classifiers were evaluated across the four preprocessing stages using the leakage-free, fold-wise procedure described in Section 2.3.3. Table 4 reports results obtained exclusively from the untouched test fold in each of the 10 cross-validation iterations. Stage 1 used training-fold standardization only; Stage 2 additionally applied ENN only to the training fold; Stage 3 additionally applied SMOTE only to the training fold; and Stage 4 applied ENN followed by SMOTE only to the training fold. No result in Table 4 was obtained by preprocessing the complete dataset before cross-validation.

Table 4: Classification performance of the six algorithms across the four leakage-free experimental stages

Stage	Evaluation metrics (%)	Class	SVM	KNN	RF	ETC	XGBoost	MLP
Stage 1	Accuracy Precision F1 Score %	low risk	76.6 ± 4 66.3 ± 4 74.5 ± 4	76.6 ± 5 69.8 ± 5 71.5 ± 5	89.2 ± 3 88.8 ± 5 86 ± 4	89.1 ± 3 88.3 ± 5 86 ± 4	89.5 ± 3 89.7 ± 5 86.4 ± 4	79.9 ± 3 72.6 ± 3 76 ± 4
		mid risk	73.2 ± 4 65.7 ± 10 50.1 ± 8	72.9 ± 3 59.2 ± 6 60 ± 4	87.6 ± 3 79.7 ± 8 82 ± 4	88.1 ± 3 80.7 ± 7 82.5 ± 4	88.1 ± 3 80.6 ± 8 82.7 ± 4	78.3 ± 4 69.7 ± 7 65.4 ± 6
		high risk	91 ± 3 83.1 ± 5 83.3 ± 5	90 ± 2 86.6 ± 5 79.7 ± 6	95.5 ± 2 92.8 ± 4 91.4 ± 3	95.5 ± 2 92.3 ± 5 91.5 ± 3	95.6 ± 2 91.9 ± 5 91.7 ± 4	93.1 ± 2 88.9 ± 7 87 ± 3
		Macro Average	80.3 ± 4 71.7 ± 6 69.3 ± 6	79.8 ± 3 71.9 ± 5 70.4 ± 5	90.8 ± 3 86.6 ± 5 86.5 ± 4	90.9 ± 3 87.1 ± 6 86.7 ± 4	91.1 ± 3 87.4 ± 6 86.9 ± 4	83.8 ± 3 77.1 ± 6 76.1 ± 4
		Weighted Average	79.3 ± 3 70.6 ± 7 68.8 ± 6	79 ± 4 70.8 ± 6 69.9 ± 5	**90.3 ± 3** **86.8 ± 6** **86.6 ± 4**	**90.4 ± 3** **86.8 ± 6** **86.3 ± 4**	**90.7 ± 3** **87.3 ± 6** **86.6 ± 4**	82.9 ± 3 76 ± 6 75.5 ± 5
Stage 2	Accuracy Precision F1 Score %	low risk	76.5 ± 3 66.3 ± 4 74.3 ± 4	76.8 ± 5 70.2 ± 6 71.8 ± 6	87.4 ± 4 85.8 ± 5 83.9 ± 5	86.9 ± 4 84.3 ± 5 83.6 ± 4	87.9 ± 4 86.3 ± 6 84.6 ± 5	79.5 ± 2 72.3 ± 2 75.4 ± 4
		mid risk	73.8 ± 4 65.4 ± 9 51 ± 7	73.3 ± 3 59.5 ± 5 60.5 ± 5	84.4 ± 4 75.5 ± 9 77.4 ± 6	84.7 ± 4 75.7 ± 8 77.7 ± 6	84.9 ± 4 75.9 ± 7 78 ± 5	76.8 ± 3 66.6 ± 7 63.8 ± 5
		high risk	90.8 ± 3 83.1 ± 5 82.8 ± 6	90.1 ± 3 87.2 ± 6 80.2 ± 5	94.5 ± 5 90.1 ± 5 89.4 ± 4	93.9 ± 2 91.7 ± 5 88 ± 5	94 ± 2 91.1 ± 5 88.4 ± 5	92 ± 3 88.4 ± 9 84.7 ± 5
		Macro Average	80.4 ± 3 71.6 ± 6 69.4 ± 6	80.1 ± 4 72.3 ± 6 70.8 ± 5	88.8 ± 4 83.8 ± 6 83.6 ± 5	88.5 ± 3 83.9 ± 6 83.1 ± 5	88.9 ± 3 84.4 ± 6 83.7 ± 5	82.8 ± 3 75.8 ± 6 74.6 ± 5
		Weighted Average	79.2 ± 3 70.4 ± 6 68.8 ± 5	79.1 ± 4 71.1 ± 6 70.2 ± 5	**88.2 ± 4** **83.7 ± 6** **83.1 ± 5**	**88 ± 3** **83.4 ± 6** **82.8 ± 5**	**88.5 ± 3** **84 ± 6** **83.4 ± 5**	81.9 ± 3 74.6 ± 5 74 ± 5
Stage 3	Accuracy Precision F1 Score %	low risk	75.3 ± 4 66.2 ± 4 71.7 ± 5	77.5 ± 6 73.6 ± 7 70.8 ± 8	88.6 ± 2 89.4 ± 5 85.7 ± 3	88.7 ± 3 88.4 ± 5 85.4 ± 3	89 ± 3 89.4 ± 5 85.4 ± 3	79.4 ± 3 75.9 ± 4 73.4 ± 5
		mid risk	71.2 ± 5 60.1 ± 11 49.3 ± 8	74.1 ± 5 59.9 ± 7 63.7 ± 6	87.8 ± 3 79.5 ± 6 92.2 ± 5	88.3 ± 3 80.7 ± 5 82.8 ± 4	87.9 ± 3 80 ± 6 82.7 ± 3	76.5 ± 4 64.4 ± 7 65.4 ± 5
		high risk	90.4 ± 2 81.1 ± 6 82.3 ± 5	90.6 ± 3 85.8 ± 6 81.7 ± 5	95.6 ± 2 92.2 ± 5 91.6 ± 4	95.7 ± 2 91.7 ± 5 91.9 ± 3	96 ± 2 92.3 ± 3 92.5 ± 4	91.4 ± 2 83.8 ± 8 84.4 ± 4
		Macro Average	79 ± 4 69.1 ± 7 67.8 ± 6	80.7 ± 8 73.1 ± 7 72.1 ± 6	90.8 ± 2 87 ± 5 86.6 ± 4	90.9 ± 3 86.9 ± 5 86.7 ± 4	90.9 ± 2 87.2 ± 5 86.8 ± 4	82.4 ± 3 74.7 ± 6 74.4 ± 5
		Weighted Average	77.9 ± 5 68 ± 7 67 ± 6	79.8 ± 5 72.3 ± 6 71.3 ± 6	**90.3 ± 2** **86.8 ± 5** **86.2 ± 4**	**90.3 ± 2** **86.7 ± 5** **86.2 ± 4**	**90.4 ± 3** **87 ± 5** **86.3 ± 4**	81.6 ± 3 74.1 ± 6 73.6 ± 5
Stage 4	Accuracy Precision F1 Score %	low risk	80.4 ± 3 68 ± 4 72.5 ± 5	84.5 ± 5 74 ± 7 74.8 ± 6	92 ± 2 89.8 ± 3 89.5 ± 3	91.8 ± 3 88 ± 4 89.5 ± 4	92.7 ± 3 90 ± 3 89 ± 3	84.6 ± 4 79 ± 4 78.8 ± 4
		mid risk	71.8 ± 9 60.5 ± 8 56.5 ± 8	69 ± 7 61.5 ± 6 61.5 ± 6	87.4 ± 5 81 ± 5 82.5 ± 5	88.4 ± 4 83.5 ± 4 82.5 ± 4	87.5 ± 5 83.5 ± 4 82.5 ± 4	76.6 ± 7 69 ± 5 67.1 ± 5
		high risk	91.5 ± 2 92 ± 2 84 ± 3	91 ± 6 86.5 ± 5 82.5 ± 5	96 ± 2 93.5 ± 2 92.5 ± 2	95.5 ± 3 93 ± 3 91.5 ± 3	96 ± 2 92.5 ± 3 93 ± 2	92.3 ± 4 87 ± 3 85.5 ± 4
		Macro Average	81.2 ± 5 73.5 ± 5 71 ± 5	81.5 ± 6 74 ± 6 72.9 ± 6	91.8 ± 3 88.1 ± 3 88.2 ± 3	91.9 ± 3 88.2 ± 4 87.8 ± 4	92.1 ± 3 88.7 ± 3 88.2 ± 3	84.5 ± 5 78.3 ± 4 77.1 ± 4
		Weighted Average	79.6 ± 5 71.1 ± 5 69.5 ± 5	80.2 ± 6 72.4 ± 6 71.7 ± 6	**90.6 ± 3** **87 ± 3** **87.1 ± 3**	**90.7 ± 3** **86.9 ± 4** **86.8 ± 4**	**90.9 ± 3** **87.6 ± 3** **87 ± 3**	83.1 ± 5 77 ± 4 75.9 ± 4

3.1.Hybrid ensemble model

The hybrid ensemble used the same fixed set of three component classifiers—RF, ETC, and XGBoost—in every fold. Within each fold, all three components were fitted only on the processed training fold, and the voting ensemble was constructed before evaluating the test fold. The final class was determined by majority vote. Test-fold results were not used to select, fit, or modify the ensemble.

3.2. Results of the hybrid ensemble model across four stages

Table 5 presents the hybrid ensemble results across the same four stages. These results were generated using the same leakage-free fold-wise pipeline as Table 4: the scaler and stage-specific ENN/SMOTE operations were fitted or applied only on the training fold, the component models and ensemble were fitted only on the processed training fold, and evaluation was performed on the untouched test fold. Class-wise,

macro-averaged, and weighted-averaged metrics are reported as mean ± standard deviation across the 10 folds.

Table 5: Performance of the fixed majority-voting ensemble across the four leakage-free experimental stages

Stage	Metrics (%)	Class	Ensemble
Stage 1	Accuracy Precision F1 Score %	low risk	89.5 ± 3 89.1 ± 5 86.6 ± 4
		mid risk	88 ± 3 80.6 ± 8 82.5 ± 4
		high risk	95.5 ± 2 92.5 ± 5 91.4 ± 3
		Macro Average	91 ± 3 87.4 ± 6 86.8 ± 4
		Weighted Average	90.6 ± 3 87.2 ± 6 86.5 ± 4
Stage 2	Accuracy Precision F1 Score %	low risk	87.3 ± 4 85.3 ± 6 84 ± 5
		mid risk	84.3 ± 4 75.3 ± 8 77.2 ± 6
		high risk	94.3 ± 2 91.8 ± 4 88.8 ± 5
		Macro Average	88.6 ± 3 84.1 ± 6 83.3 ± 5
		Weighted Average	88.1 ± 4 83.6 ± 6 83 ± 5
Stage 3	Accuracy Precision F1 Score %	low risk	88.6 ± 2 88.9 ± 3 85.6 ± 3
		mid risk	88.1 ± 3 80.5 ± 6 82.6 ± 4
		high risk	95.9 ± 2 92 ± 5 92.3 ± 4
		Macro Average	90.9 ± 2 87.1 ± 5 86.8 ± 4
		Weighted Average	90.4 ± 2 86.9 ± 5 86.3 ± 4
Stage 4	Accuracy Precision F1 Score %	low risk	92.4 ± 2 89.8 ± 3 87.2 ± 3
		mid risk	89.2 ± 4 83.8 ± 4 84 ± 4
		high risk	96.5 ± 2 93.8 ± 3 94.5 ± 3
		Macro Average	92.7 ± 3 89.1 ± 3 88.6 ± 3
		Weighted Average	**91.5 ± 3** **88 ± 3** **87.2 ± 3**

Because the class distribution is imbalanced, Precision-Recall (PR) curves and the corresponding PR-AUC values were also examined [19]. For each classifier in Stage 4, a PR curve and PR-AUC were calculated separately for each class using a one-vs-rest formulation. The three class-specific PR-AUC values were then combined using a weighted average, with each class weighted by its support. Figure 2 and the values discussed below were obtained from predictions generated by the leakage-free fold-wise pipeline.

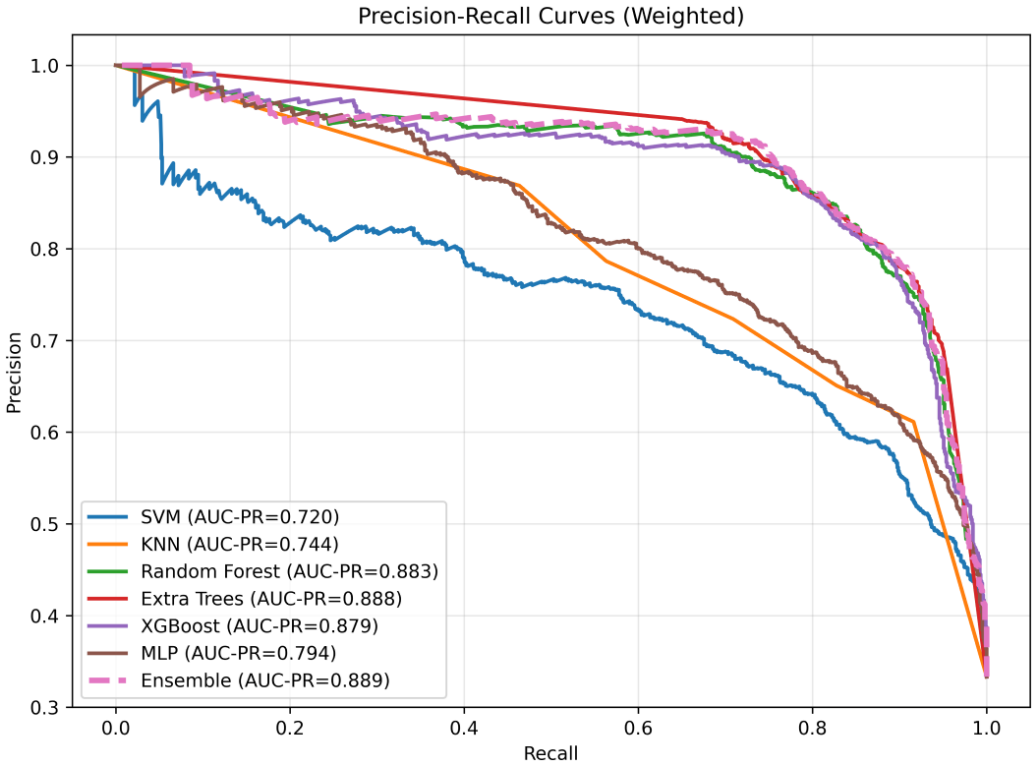


Fig. 2. Precision-Recall curves for the seven evaluated classifiers in Stage 4. Multiclass PR-AUC was calculated using a weighted-average one-vs-rest approach, with class contributions weighted by support. Curves and PR-AUC values were generated from the leakage-free fold-wise evaluation.

Figure 2 shows weighted one-vs-rest PR-AUC values of 0.889 for the hybrid ensemble, 0.883 for Random Forest, 0.888 for Extra Trees, and 0.720 for SVM. Within this experiment, the tree-based models and the hybrid ensemble produced higher PR-AUC values than SVM. These values should be interpreted only within the present dataset and validation protocol and should not be generalized to methods evaluated under different protocols.

4. Discussion and Conclusion

This study evaluated a multi-stage maternal health risk classification pipeline that combined training-fold standardization, ENN, SMOTE, six established classifiers, and a fixed majority-voting ensemble. Under the leakage-free 10-fold procedure, Stage 4 produced weighted accuracies of 90.6% ± 3% for RF, 90.7% ± 3% for ETC, and 90.9% ± 3% for XGBoost. The results indicate that the combined preprocessing stage was associated with favorable performance for these models in the present dataset; the findings do not imply that the same effect will occur in other datasets or validation settings.

The fixed majority-voting ensemble achieved a weighted accuracy of 91.5% ± 3%, precision of 88% ± 3%, and F1-score of 87.2% ± 3% in Stage 4. These values were obtained without test-fold preprocessing or test-informed model selection. The ensemble result was modestly higher than the weighted accuracy of each individual Stage 4 component in this experiment, while the magnitude and practical importance of the difference require confirmation on independent data.

At the class level, the hybrid ensemble achieved high-risk precision above 90% in each of the four stages. This result is relevant to the study objective because high-risk cases are clinically important; however, the present analysis does not determine which input features caused individual predictions, and no feature-attribution analysis was performed.

The reported results support further investigation of the pipeline as a research model for maternal health risk classification. They should not be interpreted as evidence of clinical readiness. Prospective validation, calibration assessment, subgroup analysis, and evaluation of clinical consequences are required before considering implementation in a healthcare setting.

Overall, the study demonstrates a reproducible fold-wise comparison of preprocessing strategies, individual classifiers, and a fixed voting ensemble on the Maternal Health Risk dataset. The pipeline achieved competitive performance within the present experimental setting, but direct comparison with previously

published results is limited due to differences in preprocessing, validation protocols, and model configurations.

Several limitations should be acknowledged. First, all experiments used a single public dataset with 1,014 records and six input features, which limits external validity. Second, no independent external cohort or prospective clinical evaluation was available. Third, the study used default classifier settings and did not investigate whether alternative settings would change performance. Fourth, no calibration, decision-curve, subgroup-fairness, or formal feature-attribution analysis was conducted. Fifth, the dataset may not represent demographic, socioeconomic, and healthcare-access variation across clinical populations. Future work should therefore include external and prospective validation, calibration and clinical-utility analyses, subgroup assessment, and dedicated explainability analyses conducted separately from predictive evaluation.

Table 6: Descriptive comparison with previous studies (N.A. indicates not available; direct comparison is limited by methodological differences)

Authors	Dataset	Machine-learning models				Hybrid model	Best reported model
		SVM	KNN	RF	XGBoost		
Setiawan et al., 2024 [1]	Maternal Health Risk dataset (1,014 samples) [11]	N.A.	N.A.	Overall accuracy: 81% Macro precision: 82% Weighted precision: 81% Macro F1-score: 82% Weighted F1-score: 82%	N.A.	N.A.	Random Forest with grid search (50 trees): overall accuracy 82% Bagging with DT (100 trees): overall accuracy 82%
Raihen and Akter, 2024 [2]		Accuracy: 86.13%	Accuracy (k = 5): 69.03%	Accuracy: 84.52%	N.A.	Bagging accuracy: 84.02%	SVM accuracy: 86.13%
Wongso et al., 2023 [5]		Accuracy: 74.02% Precision: 75% F1-score: 76%	Accuracy: 79.13% Precision: 80% F1-score: 80%	N.A.	N.A.	N.A.	KNN with K-means clustering Accuracy: 79.5% Precision: 80% F1-score: 80%
Mauliza et al., 2024 [9]		SMOTE-ENN Accuracy: 69% Precision: 72% F1-score: 69%	SMOTE-ENN Accuracy: 79% Precision: 81% F1-score: 80%	SMOTE-ENN Accuracy: 96% Precision: 96% F1-score: 96%	N.A.	N.A.	RF with SMOTE-ENN Accuracy: 96% Precision: 96% F1-score: 96%
Tiwari and Kumar, 2023 [7]		Precision: 58.83% F1-score: 59.66%	Precision: 71.07% F1-score: 71%	Precision: 82.18% F1-score: 82.14%	Precision: 86.54% F1-score: 86.48%	N.A.	XGBoost Precision: 86.54% F1-score: 86.48%
Mutlu et al., 2023 [8]		N.A.	Weighted accuracy: 68.47% Macro accuracy: 69.66% Weighted F1-score: 68.99% Macro F1-score: 69.64%	Weighted accuracy: 81.28% Macro accuracy: 81.85% Weighted F1-score: 81.60% Macro F1-score: 81.94%	N.A.	N.A.	Decision Tree Weighted accuracy: 89.16%
Khadidos et al., 2024 [6]		N.A.	Weighted precision: 78.09% Weighted F1-score: 78.11%	Weighted precision: 80.88% Weighted F1-score: 80.86%	N.A.	GBT + RF + DT + KNN Weighted precision: 81.18% Weighted F1-score: 80.91%	GBT + RF Weighted precision: 83.63% Weighted F1-score: 83.14%
Present study		Stage 4 weighted Accuracy: 79.6% ± 5% Precision: 71.1% ± 5% F1-score: 69.5% ± 5%	Stage 4 weighted Accuracy: 80.2% ± 6% Precision: 72.4% ± 6% F1-score: 71.7% ± 6%	Stage 4 weighted Accuracy: 90.6% ± 3% Precision: 87% ± 3% F1-score: 87.1% ± 3%	Stage 4 weighted Accuracy: 90.9% ± 3% Precision: 87.6% ± 3% F1-score: 87% ± 3%	Stage 4 hybrid ensemble Weighted accuracy: 91.5% ± 3% Weighted precision: 88% ± 3% Weighted F1-score: 87.2% ± 3%	Stage 4 hybrid ensemble Weighted accuracy: 91.5% ± 3% Weighted precision: 88% ± 3% Weighted F1-score: 87.2% ± 3%

The comparative analysis indicates that the proposed pipeline achieved competitive performance relative to previously reported studies. However, direct numerical comparison is limited because prior studies used different preprocessing methods, validation protocols, model configurations, and reporting conventions; some studies also reported higher values for particular models or metrics. Accordingly, Table 6 is presented as a descriptive comparison rather than a cross-study ranking of methods.

References

[1] Setiawan KE, Kurniawan A, Prasetyo SY. Comparative analysis of machine learning decision tree-based models for predicting maternal health risks. Procedia Comput Sci. 2024;245:57-64.
[2] Raihen MN, Akter S. Comparative assessment of several effective machine learning classification methods for maternal health risk. Comput J Math Stat Sci. 2024;3(1):161-176.
[3] Windasari W. Optimalisasi akurasi data mining dengan menggunakan algoritma PSO. J Data Sci Theory Appl. 2022;1(2):66-71.
[4] World Health Organization, United Nations Children's Fund, United Nations Population Fund, World Bank Group, United Nations Population Division. Trends in maternal mortality 2000-2023. Geneva: World Health Organization; 2025.
[5] Wongso BP, Johan ME, Fianty MI. Empowering pregnancy risk assessment: a web-based classification framework with K-means clustering-enhanced models. J Inf Syst Inform. 2023;5(4):1221-1239.
[6] Khadidos AO, Saleem F, Selvarajan S, Ullah Z, Khadidos AO. Ensemble machine learning framework for predicting maternal health risk during pregnancy. Sci Rep. 2024;14:21483.
[7] Tiwari RG, Kumar A. Exploratory data analysis and explainable machine learning for classification of maternal health risk levels. In: 2023 IEEE Fifth International Conference on Advances in Electronics, Computers and Communications (ICAECC); 2023. p. 1-8.
[8] Mutlu HB, Durmaz F, Yücel N, Cengil E, Yıldırım M. Prediction of maternal health risk with traditional machine learning methods. Naturengs. 2023;4(1):16-23.
[9] Mauliza N, Iedwan AS, Pristyanto Y, Hartanto AD, Rohman AN. The effect of resampling techniques on model performance classification of maternal health risks. J RESTI. 2024;8(4):496-505.
[10] Zemariam AB, Adisu MA, Habesse AA, Abate BB, Bizuayehu MA, Wondie WT, et al. Employing advanced supervised machine learning approaches for predicting micronutrient intake status among children aged 6-23 months in Ethiopia. Front Nutr. 2024;11:1397399.
[11] Safrit C. Maternal Health Risk Data Set [dataset]. Kaggle; 2020. Available from: https://www.kaggle.com/datasets/csafrit2/maternal-health-risk-data
[12] Breiman L. Random forests. Mach Learn. 2001; 45(1):5–32. doi:10.1023/A:1010933404324.
[13] Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. Mach Learn. 2006;63(1):3-42.
[14] Witten IH, Frank E, Hall MA. Data mining: practical machine learning tools and techniques. 2nd ed. San Francisco: Morgan Kaufmann; 2005.
[15] Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20(3):273-297.
[16] Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature. 1986;323(6088):533-536.
[17] Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016 Aug 13-17; San Francisco, CA. New York: ACM; 2016. p. 785-794.
[18] Miller C, Portlock T, Nyaga DM, O'Sullivan JM. A review of model evaluation metrics for machine learning in genetics and genomics. Front Bioinform. 2024;4:1457619.
[19] Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One. 2015;10(3):e0118432. doi:10.1371/journal.pone.0118432.
[20] Einollahi Asgarabad M. A comprehensive comparative analysis of machine learning algorithms in heart disease prediction. Discover Artificial Intelligence. 2026 Dec;6(1):331
[21] Powers DMW. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. J Mach Learn Technol. 2011;2(1):37-63.
[22] Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. Inf Process Manag. 2009;45(4):427-437.

AUTHOR BIOGRAPHY



MohammadReza Einollahi Asgarabad, ORCID iD: https://orcid.org/0009-0007-1118-7738
Department of Biomedical Engineering, SR.C, Islamic Azad University, Tehran, Iran, mohammadreza.einollahi@iau.ir
MohammadReza Einollahi Asgarabad is a M.Sc. student in Biomedical Engineering, Bioelectric at the Islamic Azad University, Science and Research Branch, Tehran, Iran. His research focuses on Internet of Things, Machine learning, Deep learning, and Cardiac-disease image processing.



Mahsa Akhbari*, ORCID iD: https://orcid.org/0000-0002-0623-4751
Department of Biomedical Engineering, SR.C, Islamic Azad University, Tehran, Iran, akhbari.mahsa@iau.ac.ir
*Corresponding author. Tel.: +989126962697 (Mahsa Akhbari).
Mahsa Akhbari is an Assistant Professor in Biomedical Engineering, Bioelectric at the Islamic Azad University, Science and Research Branch, Tehran, Iran. She received her PhD degree from Electronique, Electrotechnique, Automatique, Traitement du Signal (EEATS), GIPSA-lab, INPG, Grenoble, France and Electrical Engineering, Sharif University of Technology, Tehran, Iran on 2016. She has also received her MSc. degree from Sharif University of Technology, Tehran, Iran in biomedical engineering on 2010. She got her BSc. degree from Electrical Engineering, University of Tehran, Tehran, Iran in Communication on 2007. Her main research focuses on Artificial Intelligence, Neural Network, Machine Learning, Deep Learning, Computer Vision, Biomedical Signal Processing, Medical Image Analysis and Processing, Medical Imaging Systems.