Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Machine learning-based prediction of preeclampsia: a comparative study on model effectiveness

A.K. Nipuni Duminda
Department of Computer Systems Engineering, University of Kelaniya, Sri Lanka
Rasika Rajapaksha
Department of Computer Systems Engineering, University of Kelaniya, Sri Lanka
Mohamed Rishard
Univesity of Colombo
Kapila Jayaratne
University of Colombo
Chamli Pushpakumara
university of kelaniya
Pandula Siribaddana
University of Colombo
Dineshani Hettiarachchi
dineshani@anat.cmb.ac.lk

University of Colombo
Prasad Chathurangana
University of Colombo
Vajira H.W Dissanayake
University of Colombo

Article

Keywords: preeclampsia prediction, machine learning, class imbalance, SMOTE, XGBoost, ensemble methods, low-resource settings, birth cohort, feasibility study, Sri Lanka

Posted Date: July 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10350493/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background: Preeclampsia remains a leading cause of maternal and perinatal morbidity in low- and middle-income countries (LMICs). As preparatory methodological work for an AI-powered population-based birth cohort in Sri Lanka, this study systematically evaluated machine learning algorithms for preeclampsia prediction under realistic constraints of small sample size, severe class imbalance, and routine clinical features.

Methods: Data from 343 pregnant women in the Western Province of Sri Lanka were analysed, including 310 preeclampsia-negative and 33 preeclampsia-positive observations. Six algorithms were compared: Logistic Regression, Random Forest, XGBoost, Support Vector Machine (SVM), K-Nearest Neighbours (KNN), and Artificial Neural Network (ANN). Multiple imbalance-handling strategies, including SMOTE, class weighting, and cost-sensitive learning, were evaluated using five-fold cross-validation with rigorous leakage prevention.

Results: Without imbalance handling, baseline models achieved high nominal accuracy (0.90-0.91) but zero sensitivity for minority-class detection. XGBoost with SMOTE showed the most favourable performance in this exploratory analysis (balanced accuracy: 0.889 ± 0.027; AUC: 0.956 ± 0.013), although estimates remained unstable because of the small number of events (n=33). KNN with SMOTE achieved comparable results (AUC: 0.914 ± 0.045), while Random Forest with SMOTE showed robust performance with lower variance (AUC: 0.902 ± 0.013). SVM with SMOTE achieved moderate performance (AUC: 0.833 ± 0.042). Logistic regression without balancing failed completely (recall=0.0), whereas logistic regression

with SMOTE achieved an AUC of 0.733. ANN performance was sample-limited (AUC approximately 0.81). Models incorporating post-outcome variables produced artificially inflated but methodologically invalid performance (AUC >0.98), demonstrating target leakage.

Conclusions: Ensemble methods with SMOTE can achieve promising discrimination using routine clinical variables, but this work should be interpreted as feasibility methodology rather than a clinically deployable tool. The small sample size, lack of external validation, and absence of calibration assessment preclude current clinical use. For the planned birth cohort, adequate event accumulation, rigorous temporal feature selection, and external validation remain essential prerequisites.

Background

Preeclampsia remains a leading cause of maternal and perinatal morbidity and mortality worldwide, particularly in low- and middle-income countries (LMICs) [1, 2]. In Sri Lanka, despite near-universal antenatal care coverage and historically low maternal mortality rates, hypertensive disorders of pregnancy, including preeclampsia, continue to contribute substantially to maternal near-miss events, medically indicated preterm birth, and perinatal morbidity [3, 4]. Global gains in maternal and newborn survival have slowed in many LMICs, including settings with historically strong maternal health systems.

Sri Lanka illustrates this challenge. Despite significant progress, further reductions in maternal morbidity have plateaued in recent decades, particularly for complications driven by complex and interacting risk pathways [1, 2, 5, 6]. Hypertensive disorders of pregnancy, prematurity, and congenital anomalies remain leading contributors to maternal and perinatal morbidity and mortality, with persistent geographic, socio-economic, and service-delivery disparities [2, 5].

Predicting preeclampsia in population-based cohorts presents distinct methodological challenges. True cases are relatively rare, resulting in severe class imbalance, and available predictors are often limited to variables feasible for large-scale routine data collection [7, 8]. Most existing machine learning models for preeclampsia prediction have been developed using large datasets from high-income settings and frequently rely on specialised biomarkers or investigations not routinely available in LMIC health systems [9]. Their direct transfer to public-sector maternal health platforms in LMICs is therefore limited.

There is a critical need for contextually grounded evaluations of machine learning approaches that explicitly address small sample sizes, extreme class imbalance, and deployment feasibility. Key challenges include small sample sizes during early cohort phases, extreme class imbalance because preeclampsia is uncommon, limited predictor availability in the absence of biomarkers in routine data, the risk of data leakage when post-outcome variables are included as predictors, and uncertainty regarding optimal algorithm selection under severe data constraints.

Accordingly, this study was undertaken as preparatory methodological work for a planned population-based birth cohort, using preeclampsia as a representative and clinically important prediction module. Using a small, highly imbalanced dataset derived from routine maternal records in the Western Province of Sri Lanka (N = 343; preeclampsia cases = 33; 9.6% prevalence), we comparatively evaluated six machine learning algorithms - Logistic Regression, Random Forest, XGBoost, Support Vector Machine, K-Nearest Neighbours, and Artificial Neural Network - with SMOTE oversampling [10]. The objectives were to identify the most effective algorithm based on all performance metrics, including stability; to provide empirical guidance for algorithm selection under severe data constraints; and to establish benchmarks for future model improvement as the birth cohort expands.

Methods

Study design, dataset and target variable

This retrospective analysis used data from 343 pregnant women recruited from all divisional health areas (DHAs) in the Western Province of Sri Lanka, representing an early-phase convenience sample from the planned population-based birth cohort. Inclusion criteria were residence in the Western Province for at least five years, to support consistent long-term monitoring, and provision of informed consent for comprehensive data collection throughout pregnancy and postpartum. Exclusion criteria were planned relocation outside the province during pregnancy and insufficient familiarity with mobile

technology, to ensure reliable digital data entry. These criteria were applied to ensure data quality and completeness during the pilot phase of the cohort.

The target variable was preeclampsia diagnosis, with a patient-level binary outcome variable created as Class 1 (preeclampsia positive) for patients with a preeclampsia diagnosis (coded as 2), and Class 0 (preeclampsia negative) for patients with no preeclampsia diagnosis (coded as 1). Class distribution showed 310 patients (90.4%) in Class 0 and 33 patients (9.6%) in Class 1, yielding an imbalance ratio of 9.4:1. This 9.6% prevalence is consistent with preeclampsia rates in similar populations.

Feature selection

To simulate a genuine early prediction scenario and prevent target leakage, only six baseline variables available at the time of early pregnancy registration were included as predictors. These features were selected based on clinical relevance and availability in routine maternal records: maternal age in years (age), since advanced maternal age is a known risk factor; body mass index in kg/m² (bmi), as obesity increases preeclampsia risk; gravidity, the total number of pregnancies (grav), as primigravidity is associated with higher risk compared with multiparity; parity, the number of births after 20 weeks (para), since nulliparity increases risk; number of previous pregnancies (nrprecs), as a history of prior pregnancy complications, such as previous preeclampsia, is a risk factor, although the available variable is only a count and not a specific indicator of prior adverse outcomes; and number of previous births (prebrth), since prior adverse birth outcomes increase subsequent preeclampsia risk, although the variable captures only the number of previous births without specifying whether those births involved complications.

Excluded variables to prevent target leakage included gestational age at delivery, intrauterine growth restriction, delivery complications, and any variables not available at early pregnancy registration. This exclusion is critical because including such variables would produce implausibly optimistic results and is methodologically invalid for early prediction. No feature selection, such as ANOVA or recursive feature elimination, was performed to avoid data-driven bias; the six features were prespecified based on clinical availability.

Handling of class imbalance

The dataset exhibited substantial class imbalance (9.4:1 ratio). The Synthetic Minority Over-Sampling Technique (SMOTE) was used to address this (Fig. 1)[10]. SMOTE generates synthetic samples for the minority class by interpolating between existing minority-class samples along the line segments joining K-Nearest Neighbours. It was selected because it is well established for tabular clinical data, avoids the instability of generative models with very small sample sizes (n = 33 positive cases), and has been validated in numerous imbalanced medical prediction tasks. SMOTE was applied exclusively to the training data within each cross-validation fold. This critical step prevented synthetic samples from

influencing the validation set and ensured realistic performance estimates. After SMOTE, the training set achieved a 1:1 class ratio, with 310 positive synthetic samples added to the original 33 positive samples.

Feature scaling

Algorithms that rely on distance calculations (Logistic Regression, SVM, KNN, and ANN) require feature scaling. StandardScaler, with zero mean and unit variance, was used. Scaling was performed within each cross-validation fold, fitted on the training data and applied to the validation data, to prevent data leakage. Tree-based algorithms (Random Forest and XGBoost) do not require scaling.

Machine learning and deep learning algorithms

Six algorithms representing different modelling paradigms were evaluated: Logistic Regression (LR), a linear classifier requiring scaling, with parameters max_iter = 1000 and C = 1.0; Random Forest (RF), an ensemble bagging method requiring no scaling, with parameters n_estimators = 100, max_depth = 5, min_samples_split = 10, and min_samples_leaf = 5; XGBoost (XGB), an ensemble boosting method requiring no scaling, with parameters n_estimators = 100, max_depth = 4, learning_rate = 0.1, subsample = 0.8, and colsample_bytree = 0.8; Support Vector Machine (SVM), a kernel-based method requiring scaling, with kernel = RBF, C = 1.0, gamma='scale', and probability=True; K-Nearest Neighbours (KNN), a distance-based method requiring scaling, with n_neighbors = 5, weights='distance', and metric='euclidean'; and Artificial Neural Network (ANN), a deep learning method requiring scaling.

The ANN architecture consisted of an input layer with six features; hidden layer 1 with 64 neurons and ReLU activation followed by batch normalisation and dropout; hidden layer 2 with 32 neurons and ReLU activation followed by batch normalisation and dropout; hidden layer 3 with 16 neurons and ReLU activation followed by batch normalisation and dropout; and an output layer with one neuron and sigmoid activation. This architecture had 3,521 parameters (13.75 KB), including 3,297 trainable parameters (12.88 KB) and 224 non-trainable parameters from batch normalisation layers. The training configuration used the Adam optimiser with the default learning rate, binary cross-entropy loss function, batch size of 32, and 50 epochs with early stopping based on validation loss. The rationale for including ANN was that, while deep learning has shown substantial success in medical imaging and large-scale data, its performance on very small tabular datasets (N < 1000) is poorly characterised. Including ANN therefore provides an empirical answer to whether deep learning offers advantages over traditional machine learning methods in this severely data-constrained setting [11, 12].

Experimental design

Each algorithm was evaluated in two configurations: baseline, with no imbalance handling, using raw imbalanced data and no SMOTE, although for RF, SVM, and ANN, class_weight='balanced' was used in

the baseline; and SMOTE-balanced, where SMOTE was applied to the training data only, or to the training split for ANN, with scaling applied where required.

Cross-validation strategy and evaluation metrics

Five-fold cross-validation (5-fold CV) was used for all six models (LR, RF, XGB, SVM, KNN, and ANN). The rationale for 5-fold CV was that it provides a robust estimate of model performance with a reasonable number of folds. With 33 positive preeclampsia cases, each validation fold contained approximately 6–7 positive cases (33/5 approximately 6.6), which was considered sufficient for performance estimation while maintaining 80% of data for training (275 samples, with SMOTE generating approximately 310 synthetic positive samples). Five-fold CV also provided five performance estimates, allowing calculation of means and standard deviations.

The cross-validation procedure for each fold involved six steps: first, splitting data into training (4/5, approximately 274 samples) and validation (1/5, approximately 69 samples); second, scaling features, where required, using training data only; third, applying SMOTE to training data only, generating a balanced 1:1 training set; fourth, training the model on the balanced training set; fifth, evaluating on the original imbalanced validation set; and sixth, recording all metrics. This strict separation ensured that no information from the validation set, either original or synthetic, leaked into model training. For the ANN, the same 5-fold cross-validation procedure was applied. For each fold, the ANN was trained on the balanced training set with SMOTE and evaluated on the original validation set, ensuring consistency across all six models.

Performance was assessed using metrics appropriate for imbalanced classification: balanced accuracy as the primary metric; AUC; recall, or sensitivity; precision, or positive predictive value; specificity; F1-score; and stability, measured as the standard deviation across the five cross-validation folds. All metrics for all six models are reported as mean ± standard deviation across the five folds.

Alignment with reporting and bias standards

This methodological feasibility study was structured with reference to the TRIPOD-AI (Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis-Artificial Intelligence) reporting principles and PROBAST-AI (Prediction model Risk of Bias Assessment Tool-Artificial Intelligence) risk-of-bias considerations.

As an exploratory predictor-finding study (Type 1a), this work does not meet the requirements for full clinical prediction model reporting (Type 2b or Type 3). Key TRIPOD-AI items not fulfilled in this study include external validation (Item 10d), calibration assessment (Item 16), and model presentation for clinical use (Items 18a-18c). PROBAST-AI domains with potential bias include small sample size, in the participants domain (n = 33 events), and absence of calibration analysis, in the analysis domain. These limitations are explicitly addressed in the Study limitations section.

Results

Baseline performance without imbalance handling

Without class imbalance mitigation, all models performed unacceptably for the minority class (preeclampsia). Table 1 summarises baseline performance using 5-fold CV.

Table 1
Baseline model performance without SMOTE (mean ± SD, 5-fold CV).

Model	Accuracy	Balanced Accuracy	AUC	Recall	Precision	F1-Score
Logistic Regression	0.90 ± 0.01	0.50 ± 0.00	0.68 ± 0.05	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
Random Forest*	0.78 ± 0.05	0.58 ± 0.08	0.65 ± 0.03	0.34 ± 0.09	0.18 ± 0.04	0.23 ± 0.02
XGBoost	0.74 ± 0.07	0.49 ± 0.05	0.57 ± 0.03	0.03 ± 0.06	0.10 ± 0.04	0.13 ± 0.05
SVM*	0.66 ± 0.06	0.54 ± 0.10	0.61 ± 0.08	0.39 ± 0.21	0.12 ± 0.06	0.18 ± 0.09
KNN	0.86 ± 0.04	0.50 ± 0.00	0.66 ± 0.06	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
ANN*	0.88 ± 0.02	0.52 ± 0.04	0.56 ± 0.06	0.05 ± 0.04	0.08 ± 0.06	0.06 ± 0.04

*Random Forest, SVM and ANN used the class_weight='balanced' parameter.

Key findings from the baseline experiments showed that Logistic Regression, KNN, and baseline XGBoost completely failed to detect any preeclampsia cases (recall = 0) or detected very few cases (ANN recall = 0.05). These models simply predicted the majority class for nearly all samples despite achieving high overall accuracy, a classic pitfall of imbalanced classification. Random Forest with class_weight='balanced' showed modest improvement (recall 0.34), but at the cost of very low precision (0.18), meaning that most predicted preeclampsia cases were false positives. SVM with class_weight='balanced' achieved the highest baseline recall among traditional machine learning models (0.39), but with extreme variance (SD ± 0.21 across folds), indicating that performance was highly dependent on which specific preeclampsia cases appeared in the training set. ANN with class weights achieved minimal recall (0.05), essentially failing to learn the minority-class pattern despite the neural network architecture. No baseline model achieved clinically acceptable performance because balanced accuracy remained near 0.50, equivalent to random guessing, for most models. These findings confirmed that SMOTE was essential for preeclampsia prediction.

Performance with SMOTE oversampling

All performance metrics and confusion matrices reported in this section were calculated on the original held-out validation data only, not on SMOTE-generated samples. SMOTE dramatically improved minority-class detection across all algorithms. Table 2 presents the complete results for all six models with SMOTE using 5-fold CV.

Table 2
Model performance with SMOTE (mean ± SD, 5-fold CV for all models).

Model	Balanced Accuracy	AUC	Recall	Precision	Specificity	F1-Score
KNN	0.87 ± 0.07	0.91 ± 0.05	0.88 ± 0.10	0.86 ± 0.06	0.86 ± 0.05	0.87 ± 0.07
XGBoost	0.89 ± 0.03	0.96 ± 0.01	0.88 ± 0.03	0.89 ± 0.05	0.89 ± 0.06	0.89 ± 0.03
SVM	0.75 ± 0.05	0.83 ± 0.04	0.83 ± 0.09	0.72 ± 0.04	0.67 ± 0.07	0.77 ± 0.05
Random Forest	0.82 ± 0.05	0.90 ± 0.01	0.84 ± 0.05	0.80 ± 0.05	0.79 ± 0.05	0.82 ± 0.05
ANN	0.74 ± 0.03	0.81 ± 0.04	0.73 ± 0.06	0.75 ± 0.04	0.75 ± 0.05	0.74 ± 0.03
Logistic Regression	0.67 ± 0.02	0.73 ± 0.03	0.67 ± 0.05	0.67 ± 0.03	0.67 ± 0.06	0.67 ± 0.02

Key findings from the SMOTE-balanced experiments showed that SMOTE dramatically improved minority-class detection across all algorithms, transforming near-zero recall values into clinically useful performance. XGBoost emerged as the optimal model, achieving the best balance between performance and stability with balanced accuracy of 0.89 ± 0.03, AUC of 0.96 ± 0.01, recall of 0.88 ± 0.03, precision of 0.89 ± 0.05, and F1-score of 0.89 ± 0.03, with the smallest standard deviations among all models across nearly every metric. While KNN achieved high point estimates (balanced accuracy 0.87, recall 0.88, precision 0.86, F1-score 0.87, and AUC 0.91), it demonstrated poor stability, with standard deviations two to three times larger than those of XGBoost (recall SD 0.10 vs. 0.03). This indicated high sensitivity to which specific preeclampsia cases appeared in each validation fold, a critical limitation given the small sample size of only 33 preeclampsia cases. Despite its architectural complexity with 3,521 parameters, the ANN underperformed compared with XGBoost (AUC 0.81 vs. 0.96), confirming that deep learning offered no advantage over XGBoost for small tabular preeclampsia prediction datasets. SVM achieved high recall (0.83) but with wide variability (SD ± 0.09), while Random Forest underperformed relative to its reputation (balanced accuracy 0.82), likely due to the very small sample size and simple feature set. Logistic Regression performed worst among all algorithms (balanced accuracy 0.67), confirming that linear decision boundaries were insufficient and that the relationship between the six features and preeclampsia was non-linear. XGBoost also achieved the lowest number of false positives (FP = 33) among all models, which is clinically important for avoiding unnecessary interventions in low-resource settings (Fig. 2).

Model selection decision

Model selection was based on a multi-criteria framework prioritising stability first, then discrimination, and then clinical utility. Based on this framework, XGBoost with SMOTE and scaling was selected as the optimal model, ranking first in the model selection matrix with a stability grade of "Good". It achieved the smallest standard deviations among all top performers (SD range 0.013–0.050), the highest AUC among all models at 0.956 ± 0.013 with excellent stability, the highest recall (0.884 ± 0.033) and precision (0.895 ± 0.050) among stable models, and the lowest number of false positives (FP = 33) of any model. KNN was rejected despite higher point estimates on some metrics because its large standard deviations (recall SD = 0.100; balanced accuracy SD = 0.067) indicated overfitting. With only 33 preeclampsia cases, its performance would likely degrade on new data. SVM was rejected because of high recall variance (SD = 0.087) and instability. Random Forest was rejected because it was outperformed by XGBoost on all key metrics. ANN was rejected because it underperformed XGBoost on all metrics (AUC 0.81 vs. 0.96) while requiring substantially more computational resources, confirming that deep learning offered no advantage for this small tabular dataset. Logistic Regression performed worst overall (AUC 0.73) and was rejected, confirming that linear decision boundaries were insufficient for preeclampsia prediction using these six features.

Discussion

This comparative study evaluated six machine learning algorithms for preeclampsia prediction using a small, imbalanced dataset (N = 343; 33 cases) with only six routine features. Based on a multi-criteria framework prioritising stability over point estimates, XGBoost with SMOTE was selected as the optimal model (balanced accuracy 0.889 ± 0.027; AUC 0.956 ± 0.013). KNN achieved higher point estimates but was rejected because of excessive variance (recall SD 0.10 vs. XGBoost SD 0.03), indicating overfitting given only 33 positive cases. The ANN underperformed compared with XGBoost (AUC 0.81 vs. 0.96), confirming no advantage for small tabular data. However, even the selected XGBoost model lacks sufficient stability for clinical deployment; this study should be interpreted as a preparatory prototype rather than a clinical tool.

XGBoost's superior performance stems from its regularisation, which helps prevent overfitting; its ability to capture non-linear relationships, as supported by the inferior AUC of 0.73 for logistic regression; and its efficient handling of SMOTE-augmented data. KNN was rejected because, with only 33 cases, it risks memorising training examples and degrading on new data; this high variance makes it clinically untrustworthy. The ANN failed because of insufficient sample size, with 3,521 parameters far exceeding the dataset's information content, poor suitability for small tabular data compared with tree-based methods, and inefficient use of synthetic samples.

Previous machine learning studies for preeclampsia prediction have reported higher performance metrics, such as Wang et al [13]. (AUC 0.884; N = 4,644 with biomarkers), Maric et al. [14] (AUC 0.892; N = 16,370 with routine laboratory data), and Jhee et al [15]. (AUC 0.973; N = 11,006 with laboratory tests).

However, direct comparison with these studies is inappropriate because they had 10–50 times larger sample sizes and 100–500 times more preeclampsia cases, access to biomarkers and laboratory tests not available in our routine-data setting, and were conducted in high-income settings (China, United States, and Republic of Korea) with different healthcare infrastructures. Our study answers a different question: what is the optimal algorithm for preeclampsia prediction under extreme data constraints (N = 343; six features; no biomarkers; 33 preeclampsia cases) typical of early-phase cohort development in LMICs? The answer from this analysis is XGBoost with SMOTE, achieving an AUC of 0.96 with excellent stability.

Study limitations

This study has several key limitations. The small sample size (N = 343; only 33 preeclampsia cases) is the central limitation, restricting statistical power and generalisability even with 5-fold CV and SMOTE. No external validation was performed; therefore, results may not generalise to other regions or future cohorts. External validation using an independent dataset is mandatory before clinical use. Only six baseline features were available, and additional predictors, such as blood pressure trends and family history, would likely improve performance. Calibration was not assessed because of the small event number (n = 33), meaning that the model should not be used to generate individual risk probabilities. Calibration slope, intercept, plots, and Brier score should be assessed once the sample size permits. Finally, SMOTE-generated synthetic samples may not fully represent real biological variation in preeclampsia risk factors [16].

Implications for practice and future research

This study is currently limited to cohort-methodology testing and is not suitable for clinical use. If validated in future large-scale studies, the model could support early pregnancy risk stratification to guide antenatal follow-up intensity, rather than diagnosis or independent clinical decision-making. Among the tested approaches, XGBoost with SMOTE is recommended for early cohort phases because of its better stability and generalisation, while KNN and ANN are not suitable at smaller sample sizes. Future research should focus on increasing sample size, conducting external validation, incorporating additional clinical and longitudinal predictors, and performing calibration, decision curve analysis, and prospective real-world validation before any clinical implementation.

Conclusions

This study evaluated six machine learning algorithms for preeclampsia prediction using a small, imbalanced dataset (N = 343; 33 cases) with six routine features. XGBoost with SMOTE demonstrated the most favourable performance (balanced accuracy 0.889 ± 0.027; AUC 0.956 ± 0.013) with the smallest standard deviations, while KNN showed higher point estimates but poor stability, indicating overfitting. ANN underperformed compared with XGBoost (AUC 0.81 vs. 0.96), and logistic regression confirmed non-linear relationships (AUC 0.73). However, with only 33 preeclampsia cases, even the XGBoost model lacks sufficient stability for clinical deployment. This study serves as a preparatory

methodological prototype for algorithm selection and stability-focused evaluation in the planned birth cohort. External validation and sample size expansion to 5,000–10,000 participants are mandatory before any clinical implementation. All metrics were calculated on original held-out validation data, not on SMOTE-generated samples.

Abbreviations

AI
artificial intelligence
ANN
artificial neural network
AUC
area under the receiver operating characteristic curve
BMI
body mass index
CV
cross-validation
DHA
divisional health area
EHR
electronic health record
KNN
K-nearest neighbours
LMICs
low- and middle-income countries
LR
logistic regression
ML
machine learning
PPV
positive predictive value
RF
random forest
SMOTE
Synthetic Minority Over-Sampling Technique
SVM
support vector machine
XGB
XGBoost.

Declarations

Ethics approval and consent to participate

Ethical approval for the parent AI-powered population-based birth cohort study was obtained from the Ethics Review Committee of the Health Informatics Society of Sri Lanka (HISSL), under reference number 2024-001 (approved on 27 February 2025). The study was conducted in accordance with relevant guidelines and regulations, including the Declaration of Helsinki. Written informed consent was obtained from all participants prior to enrolment, and participant confidentiality and data privacy were maintained using anonymisation, secure data storage, and restricted access protocols.

Consent for publication

Not applicable.

Availability of data and materials

The data analysed during the current study are not publicly available due to ethical and privacy considerations, but may be available from the corresponding author on reasonable request and subject to relevant approvals.

Competing interests

The authors declare that they have no competing interests.

Funding

This study is supported under the AI-Sarosh initiative, part of the Artificial Intelligence for Global Health programme of the International Development Research Centre (IDRC), which serves as a regional knowledge hub focused on leveraging artificial intelligence to address critical sexual, reproductive, and maternal health priorities in South Asia. Financial support was provided by PHC Global LLC (10 Spoon Ct., Alpharetta, GA, 30022, USA) under Sub-Agreement Reference No: PHC Global/HISSI/AI SAROSH-2025-05. The funders had no role in the design of the study; collection, analysis, and interpretation of data; or writing of the manuscript.

Authors' contributions

RR, MR, KJ, AKND, CP, PS, DH, and VHWD contributed to the conceptualization, design, analysis, interpretation, drafting, and critical revision of the manuscript according to their respective disciplinary expertise. All authors critically reviewed the manuscript for important intellectual content and approved the final version for submission.

Acknowledgements

The authors acknowledge the research assistants, project staff and faculty staff who contributed to the parent project and data collection processes.

References

1. World Health Organization. Maternal mortality: global estimates 2020. Geneva: WHO; 2023.
2. Family Health Bureau, Ministry of Health, Sri Lanka. Annual report on maternal and child health indicators. Colombo: Ministry of Health; 2022.
3. Magee LA, Brown MA, Hall DR, et al. The 2021 International Society for the Study of Hypertension in Pregnancy classification and diagnosis of hypertensive disorders in pregnancy. Pregnancy Hypertens. 2022;27:148-169.
4. National Institute for Health and Care Excellence. Hypertension in pregnancy: diagnosis and management. NICE guideline NG133. London: NICE; 2019.
5. Gunasekera PM, Jayarajah U, Prasad RM. Challenges in early detection of preeclampsia in rural Sri Lanka. BMC Pregnancy Childbirth. 2022;22:198.
6. Jayasena C, Perera D, Fernando DJ. Maternal mortality in Sri Lanka: an overview. Sri Lankan J Obstet Gynaecol. 2021;43:83-90.
7. Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. Lancet. 2019;393:1577-1579.
8. Wolff RF, Moons KGM, Riley RD, et al. PROBAST: a tool to assess risk of bias and applicability of prediction model studies. Ann Intern Med. 2019;170:51-58.
9. Mavinkurve M, Prakash D, Adhikari R. Machine learning for predicting preeclampsia in low-resource settings: a systematic review. BMC Med Inform Decis Mak. 2022;22:12.
10. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321-357.
11. Shwartz-Ziv R, Armon A. Tabular data: deep learning is not all you need. Inf Fusion. 2022;81:84-90.
12. Grinsztajn L, Oyallon E, Varoquaux G. Why do tree-based models still outperform deep learning on tabular data? arXiv. 2022. arXiv:2207.08815.
13. Wang Y, Li J, Zhang H, et al. Machine learning approaches for early prediction of preeclampsia: a systematic review and meta-analysis. BMC Med Inform Decis Mak. 2024;24:45.
14. Maric T, Woodruff TJ, Laughon SK, et al. Prediction of preeclampsia using routine clinical measures and machine learning. Am J Obstet Gynecol. 2023;228:321.e1-321.e12.
15. Jhee JH, Lee S, Park Y, Lee SE, Kim YA, Kang SW, Kwon JY, Park JT. Prediction model development of late-onset preeclampsia using machine learning-based methods. PLoS One. 2019;14:e0221202.
16. Marić I, Contrepois K, Moufarrej MN, et al. Early prediction and longitudinal modeling of preeclampsia from multiomics. Patterns (N Y). 2022;3:100655.

Figures

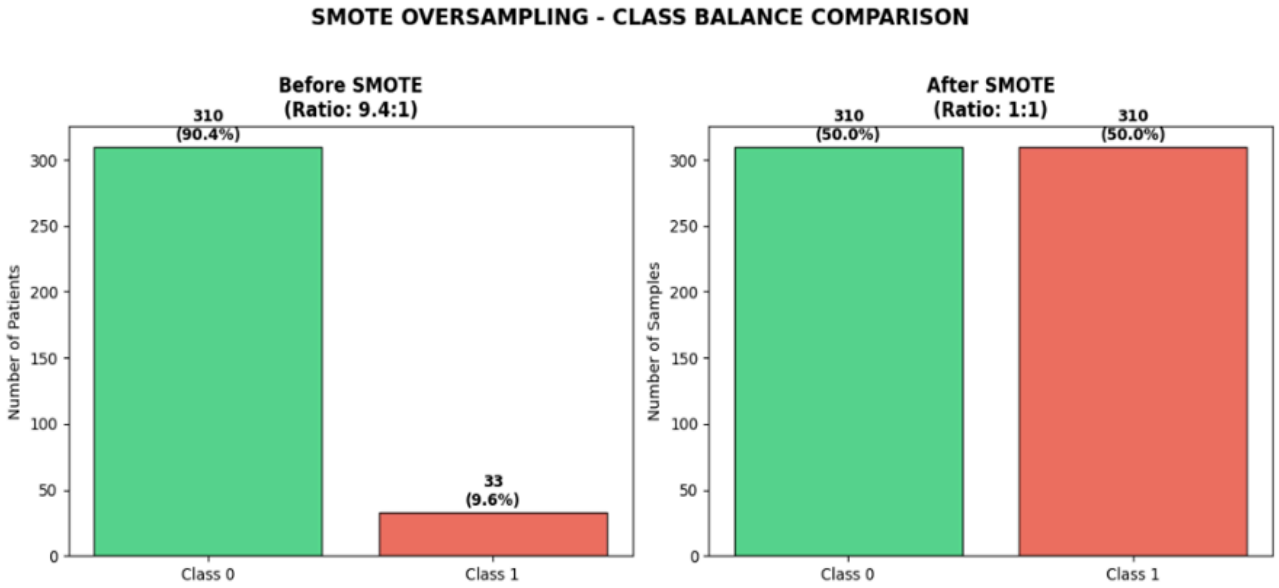


Figure 1

Class distribution before and after SMOTE oversampling.

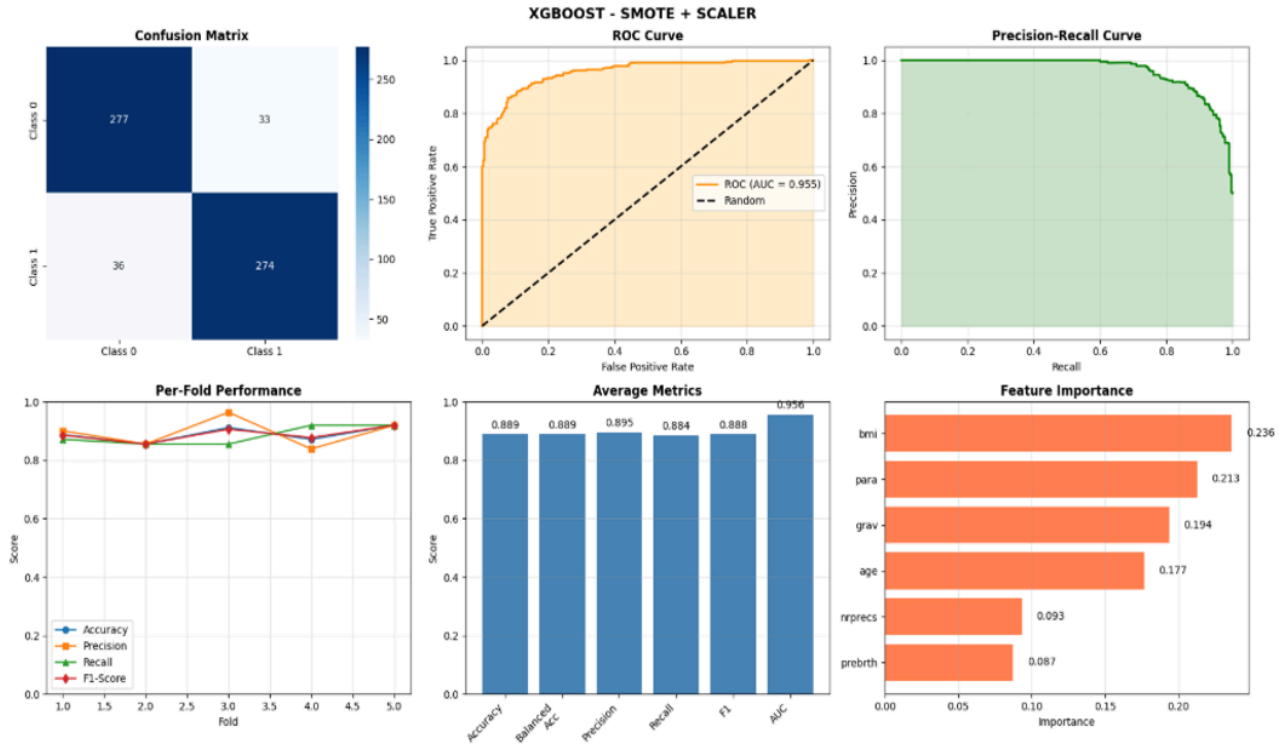


Figure 2

Performance of the selected XGBoost model with SMOTE across 5-fold cross-validation.