Bulletin of Electrical Engineering and Informatics
Vol. 15, No. 6, December 2026, pp. 5459~5473
ISSN: 2302-9285, DOI: 10.11591/eei.v15i6.13309

Explainable machine learning-based stratification of severe obstetric complications using small-scale EMR data

I Gusti Njoman Teguh Budhi Bimantara[1], I Gusti Putu Asto Buditjahjanto[1,2], Lilik Anifah[1], Gusti Ayu Az Zahra Shabrina Putri Bukian[3]
[1]Postgraduate Program in Vocational Education, Universitas Negeri Surabaya, Surabaya, Indonesia
[2]Intelligent System Research Laboratory, Department of Electrical Engineering, Universitas Negeri Surabaya, Surabaya, Indonesia
[3]Department of Medicine, Faculty of Medicine, Universitas Muhammadiyah Surabaya, Surabaya, Indonesia

Article Info

Article history:

Received May 9, 2026
Revised Sep 11, 2026
Accepted Oct 4, 2026

Keywords:

Electronic medical records
Explainable artificial intelligence
Machine learning
Severe obstetric complications
Shapley additive explanations

ABSTRACT

Severe obstetric complications related to miscarriage remain a major contributor to maternal morbidity. This study aimed to develop and evaluate explainable machine learning (ML) models to classify patterns of severe obstetric complications using electronic medical record (EMR) data. A cross-sectional dataset of 57 patients with miscarriage-related conditions was analyzed. Severe-complication patterns were defined as incomplete, missed, or incipient miscarriage, whereas complete miscarriage and threatened/imminent miscarriage were categorized as lower-risk outcomes. Predictors included maternal age, parity, hemoglobin level, infection status, and miscarriage history. Multiple ML algorithms were evaluated using repeated stratified k-fold cross-validation and out-of-fold (OOF) analysis, with synthetic minority oversampling technique (SMOTE) applied to address class imbalance. Model interpretability was assessed using Shapley additive explanations (SHAP). Extreme gradient boosting (XGBoost) with SMOTE achieved the best performance (balanced accuracy=0.792, area under the curve (AUC)=0.806), with an accuracy of 0.877 and an F1-score of 0.926. SHAP analysis identified infection status and hemoglobin level as the leading contributors to model predictions. These exploratory findings support the feasibility of explainable ML for complication-pattern stratification using routinely available, small-scale EMR data, with potential relevance to resource-constrained maternal-health settings; however, larger multicenter datasets, calibration assessment, and external validation are required before clinical application.

This is an open access article under the CC BY-SA license.

Corresponding Author:

I Gusti Putu Asto Buditjahjanto
Postgraduate Program in Vocational Education, Universitas Negeri Surabaya
Surabaya, East Java, Indonesia
Email: asto@unesa.ac.id

1. INTRODUCTION

Miscarriage, defined as pregnancy loss before 20 weeks of gestation, remains an important global health problem associated with maternal morbidity and substantial psychological burden [1]. Pregnancy termination and related complications also contribute to the global burden of reproductive morbidity, with substantial variation across regions and populations [2]. In Indonesia, population-based evidence from Java indicates that pregnancy termination remains an important reproductive-health concern [3]. Miscarriage and related obstetric complications arise from complex interactions among biological, reproductive, clinical, and environmental factors. Previous studies have identified maternal age, parity, previous miscarriage, anemia,

infection, gestational characteristics, and maternal comorbidities as relevant determinants of adverse pregnancy outcomes [4]–[7]. Infection and inflammatory processes may be particularly relevant in complicated miscarriage presentations, while advanced maternal age is associated with biological mechanisms that increase the risk of pregnancy loss and related complications.

Despite advances in miscarriage research, clinically interpretable stratification of complication patterns remains challenging. Most existing prediction studies estimate whether miscarriage or pregnancy loss will occur, rather than distinguishing complication patterns among patients who already present with miscarriage-related conditions [8], [9]. This distinction is clinically important because patients within the same broad diagnostic category may differ in monitoring and management needs. Furthermore, many predictive models provide limited interpretability, restricting clinicians' ability to understand how individual patient characteristics influence model outputs. These limitations define the central problem addressed in this study: developing an interpretable risk-stratification approach for patients with established miscarriage-related conditions rather than another model for predicting miscarriage occurrence.

The growth of electronic medical records (EMRs) has expanded opportunities for data-driven clinical prediction [10]. Recent advances in obstetric artificial intelligence (AI) include deep learning (DL) models based on routinely collected electronic health records, such as models developed to predict clinically relevant outcomes in preeclampsia [11]. In parallel, federated-learning approaches are emerging in obstetrics and gynecology as privacy-preserving strategies that enable collaborative model development across institutions without centralizing raw patient records [12]. These developments demonstrate the increasing sophistication of data-driven obstetric prediction; however, they typically rely on larger or distributed datasets and more complex computational frameworks than the small single-center EMR cohort considered in the present study. In small clinical datasets, model interpretability remains particularly important because high-performing models may still function as "black boxes," limiting clinician confidence when predictions cannot be transparently linked to relevant patient characteristics [13], [14]. Accordingly, leakage-aware preprocessing, class-imbalance handling, reliable internal validation, and explainable model behavior are especially important when applying machine learning (ML) to small, imbalanced clinical datasets.

Explainability is therefore an important consideration in developing ML-based obstetric decision support systems (DSSs). To address this challenge, explainable artificial intelligence (XAI) has become increasingly important in clinical ML. Shapley additive explanations (SHAP) provides additive feature-attribution values based on cooperative game theory, enabling both global assessment of predictor importance and patient-level interpretation of model outputs [15]. This capability is particularly relevant in healthcare, where transparent model behavior can facilitate clinical evaluation and support the responsible development of AI-assisted decision-support systems.

Previous pregnancy-loss prediction studies have primarily estimated miscarriage risk in preconception or early-pregnancy populations. Choo *et al.* [8] developed a miscarriage risk score among preconception women, Yang *et al.* [16] investigated subsequent early pregnancy loss using preconception characteristics, and Yland *et al.* [17] evaluated miscarriage-prediction models in a preconception cohort. Collectively, these studies demonstrate the feasibility of data-driven pregnancy-loss prediction but address a different clinical question from the within-cohort complication stratification investigated here.

Recent XAI studies in obstetric prediction have used feature-attribution methods such as SHAP to interpret ML models for pregnancy-related outcomes, including preterm birth and preeclampsia [18], [19]. These studies demonstrate the value of XAI for improving model transparency and identifying clinically relevant predictors. However, their prediction targets, populations, and data scales differ from those of the present study. More importantly, previous miscarriage studies have predominantly focused on predicting the occurrence or probability of pregnancy loss rather than stratifying complication patterns among patients who already have miscarriage-related conditions [8], [16], [17]. This leaves a gap in using explainable, imbalance-aware ML for within-cohort complication stratification with small, real-world EMR datasets.

This study therefore contributes not by introducing a new ML algorithm, synthetic minority oversampling technique (SMOTE), or SHAP method individually, but by integrating them within a leakage-aware framework tailored to small-scale clinical EMR data. The prediction task is reframed from miscarriage-occurrence prediction to stratification of severe versus lower-risk miscarriage-related complication patterns; preprocessing and SMOTE are restricted to training folds; repeated stratified cross-validation is combined with patient-level out-of-fold (OOF) evaluation; and SHAP is used to examine model behavior. Accordingly, this study aimed to develop and internally evaluate an explainable ML framework for stratifying severe obstetric complication patterns using routinely available EMR variables. As a proof of concept, the framework may be particularly relevant to resource-constrained maternal-health settings where large datasets and advanced computational infrastructure are not always available, though external validation is required before clinical implementation.

2. MATERIALS AND METHODS

2.1. Study design and data source

This cross-sectional study analyzed secondary data obtained from the EMRs of patients diagnosed with miscarriage-related conditions at a single maternal hospital in Indonesia. A total-sampling approach yielded 57 eligible patient records. Clinical and demographic variables were extracted to develop explainable ML models for stratifying severe obstetric complication patterns. Because the cohort was small and originated from a single center, it may not adequately represent the heterogeneity of miscarriage-related presentations across institutions or populations; consequently, the resulting performance estimates should not be interpreted as externally generalizable. To reduce dependence on a single data partition and limit optimistic internal performance estimation, repeated stratified k-fold cross-validation and patient-level OOF prediction were used. OOF predictions were used to construct internally cross-validated receiver operating characteristic (ROC) curves and confusion matrices. Figure 1 illustrates the analytical pipeline. All results were therefore interpreted as exploratory internal estimates requiring confirmation in larger independent multicenter cohorts.

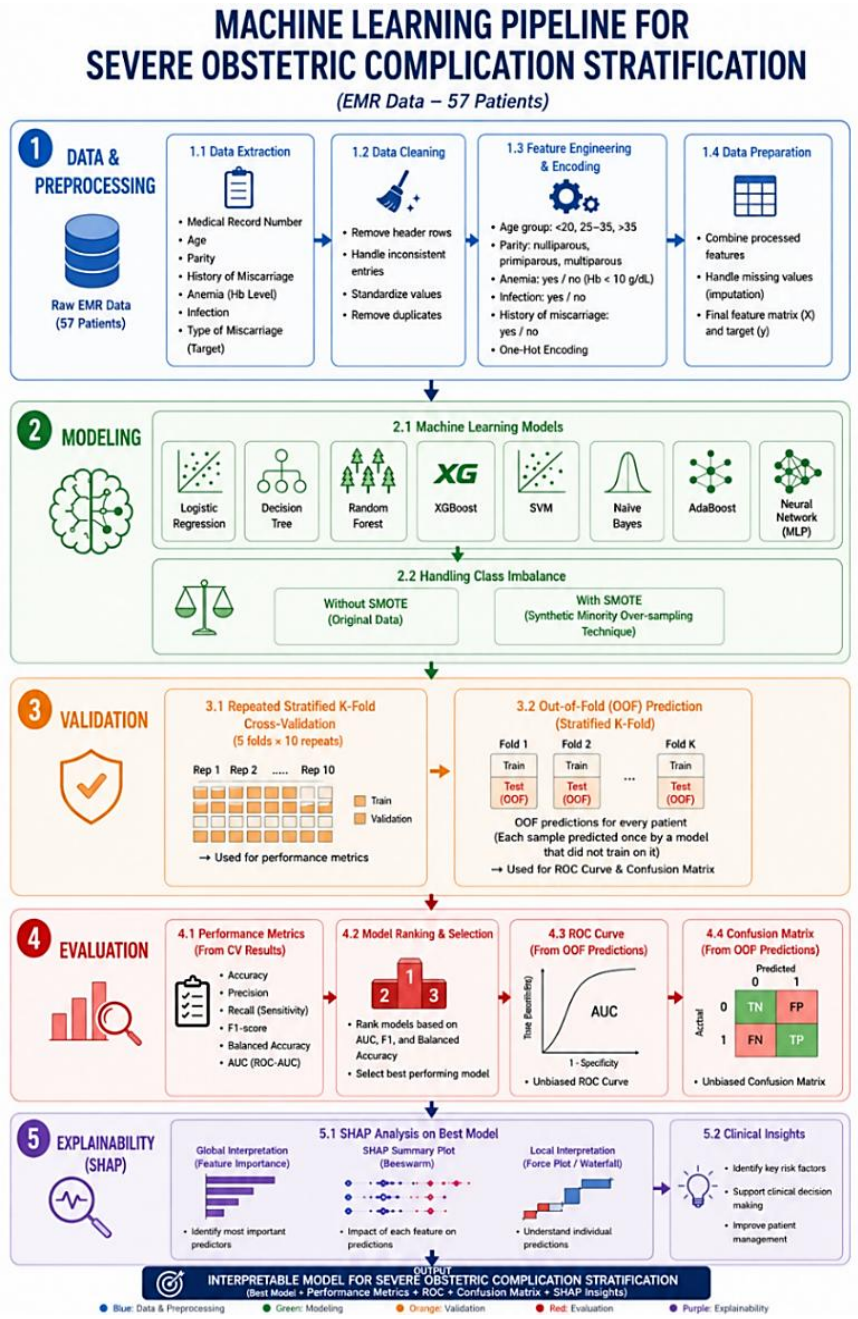


Figure 1. ML pipeline for severe obstetric complication-pattern stratification

2.2. Outcome definition

The primary outcome was a binary stratification of miscarriage-related clinical patterns according to their relative potential to require active clinical assessment or management. Incomplete, missed, and incipient miscarriage were assigned to the severe-complication pattern because these presentations may involve retained products of conception, ongoing or impending pregnancy loss, bleeding, infection, or the need for closer clinical assessment and intervention. Complete miscarriage and threatened/imminent miscarriage were grouped as lower-risk patterns for the purpose of this analytical framework because they generally represented comparatively less complicated presentations in the available records. This categorization was used as an operational modeling definition rather than as a universal clinical severity classification, and it should not be interpreted as replacing established obstetric diagnosis or individual clinical assessment. The resulting groups were:

- Severe obstetric complication pattern (positive class): incomplete miscarriage, missed miscarriage, and incipient miscarriage.
- Lower-risk miscarriage pattern (negative class): complete miscarriage and threatened/imminent miscarriage.

Following binary outcome categorization, 48 of the 57 patients (84.2%) were assigned to the severe obstetric complication class and 9 patients (15.8%) to the lower-risk class. The resulting class ratio was approximately 5.3:1, indicating substantial class imbalance. This distribution was explicitly considered during model development and evaluation.

2.3. Predictor variables

Predictor variables were selected based on their clinical relevance reported in previous obstetric and miscarriage studies. The final clinically reduced feature set included: maternal age (years), parity count, history of miscarriage, hemoglobin level (g/dL), and infection status. Maternal age was included as a continuous variable to preserve biological information related to reproductive risk. Parity count, the number of previous births, was included as an ordinal clinical feature reflecting reproductive history. The history of miscarriage was treated as a binary variable. Infection status was also represented as a binary variable based on documented infection findings in the medical records. Hemoglobin level was extracted directly from laboratory results and modeled as a continuous variable rather than a binary anemia category. This approach preserves clinically meaningful physiological variation and improves explainability through SHAP analysis.

2.4. Data preprocessing

Prior to model development, EMR variables were inspected for completeness, consistency, duplicate records, and formatting errors. Maternal age and hemoglobin level were retained as continuous numerical variables, while parity was represented by its recorded integer count. History of miscarriage and infection status were represented as binary indicator variables using 0 and 1 according to the absence or presence of the documented condition; therefore, no one-hot encoding was required for the final five-predictor dataset. Missing numerical values were imputed using the median estimated exclusively from the corresponding training fold, followed by standardization using StandardScaler. The five predictors were processed within a ColumnTransformer. Standardization was included to maintain compatibility with algorithms sensitive to feature magnitude, particularly logistic regression (LR), support vector machine (SVM), and multi-layer perceptron (MLP). All preprocessing procedures were embedded within the ML pipeline and fitted independently within each cross-validation training fold. For models evaluated with SMOTE, oversampling was applied only to the transformed training data before classifier fitting, while the corresponding validation data remained unchanged. This fold-specific procedure prevented information from validation observations from influencing imputation, scaling, or resampling, thereby minimizing information leakage [20].

2.5. Handling class imbalance

The binary outcome distribution was substantially imbalanced, with 48 severe-complication cases and 9 lower-risk cases. Therefore, the lower-risk group constituted the minority class in the present dataset. To investigate the effect of class balancing on model behavior, SMOTE was evaluated as an oversampling strategy within the training data only. Synthetic samples were generated exclusively within each training fold and were never introduced into the corresponding validation fold, thereby reducing the risk of information leakage. Models trained with and without SMOTE were compared to determine whether resampling improved balanced discrimination rather than merely overall accuracy. SMOTE generates synthetic minority-class observations by interpolation in feature space and is widely used for learning from imbalanced datasets [21].

The number of nearest neighbors used by SMOTE was constrained by the number of minority-class observations available within each training fold to avoid invalid oversampling in the small dataset. Under the five-fold stratified design used here, this procedure permitted the standard K_neighbors=5 setting, which was

applied consistently within the training folds. A separate sensitivity analysis across alternative K_neighbors values was not performed because the extremely small minority class would make such comparisons highly dependent on individual observations and fold composition. Consequently, the reported SMOTE results should be interpreted as conditional on this resampling specification; future validation with larger cohorts should assess the robustness of model performance across alternative neighborhood sizes. SMOTE was applied only after preprocessing within each training fold and never to the corresponding validation observations.

2.6. Machine learning models

Multiple ML algorithms from different learning paradigms were evaluated to enable a comprehensive model comparison. The evaluated models included: LR, decision tree (DT), random forest (RF), extreme gradient boosting (XGBoost), SVM, Naïve Bayes, adaptive boosting (AdaBoost), and MLP. Including diverse model families enabled comparison across linear, probabilistic, ensemble-based, margin-based, and neural network approaches for obstetric complication stratification.

2.6.1. Software environment and reproducibility

All analyses were performed in Python 3.11 using Scikit-learn, XGBoost, SHAP, NumPy, Pandas, Matplotlib, and Imbalanced-learn. A fixed random seed (random_state=42) was applied to stochastic procedures where supported. The reproducible workflow comprised loading anonymized EMR data, defining predictors and the binary outcome, creating stratified training-validation folds, fitting imputation and scaling on training data only, applying SMOTE only within training folds, fitting prespecified classifiers, generating validation and patient-level OOF predictions, computing performance metrics, selecting the final model using the predefined ranking score, and applying SHAP analysis. All preprocessing, resampling, model fitting, and evaluation steps were implemented within fold-specific pipelines to prevent information leakage from validation data.

2.6.2. Model configuration and hyperparameter settings

Model configurations were specified a priori to provide a controlled comparison across different ML families while limiting model complexity in the context of only 57 observations. Exhaustive grid search, randomized search, or automated hyperparameter optimization was intentionally not performed because extensive tuning on such a small dataset could increase selection instability and the risk of overfitting to the internal resampling structure, particularly given the limited minority class. Instead, conservative fixed configurations were applied consistently across the repeated cross-validation framework. This design prioritizes methodological stability and comparability rather than claiming that each algorithm achieved its optimal possible performance; consequently, the reported between-model comparisons should be interpreted as comparisons under the prespecified configurations used in this study.

LR was configured with a maximum of 1,000 iterations and balanced class weighting. The DT classifier used balanced class weighting with otherwise default Scikit-learn settings. RF was implemented using 300 trees with balanced class weighting. XGBoost used 200 boosting estimators, a maximum tree depth of 3, a learning rate of 0.05, and log-loss as the evaluation metric. The SVM used a radial basis function kernel, probability estimation, and balanced class weighting. Gaussian Naïve Bayes used its default configuration. AdaBoost used 100 estimators. The MLP had two hidden layers with 16 and 8 neurons, respectively, used rectified linear unit (ReLU) activation, and ran up to 1,500 training iterations.

A fixed random seed (random_state=42) was used for all stochastic algorithms where supported to improve reproducibility. These predefined configurations were evaluated using repeated stratified k-fold cross-validation rather than being selected on the basis of performance on the final OOF evaluation data. Parameters not explicitly listed in Table 1 followed the default settings of the corresponding Scikit-learn or XGBoost implementation used in the analysis. The software environment and random seed were fixed to improve computational reproducibility.

Table 1. ML model configurations used in the analysis

Model	Main configuration
LR	max_iter=1000; class_weight=balanced
DT	class_weight=balanced
RF	n_estimators=300; class_weight=balanced
XGBoost	n_estimators=200; max_depth=3; learning_rate=0.05; eval_metric=logloss
SVM	kernel=RBF; probability=True; class_weight=balanced
Gaussian Naïve Bayes	Default implementation
AdaBoost	n_estimators=100
MLP	hidden layers=(16, 8); activation=ReLU; max_iter=1500
Random seed	42, where applicable

2.7. Model development and validation

Model performance was assessed using repeated stratified k-fold cross-validation. The number of folds was determined according to the minority-class size, with a maximum of five folds. For the present dataset, five folds were used with 10 repetitions and a fixed random seed of 42. Stratification preserved the class distribution across folds. Repeated stratified cross-validation was used as an internal stability assessment across multiple data partitions. Final model-level classification metrics reported in Table 2 were subsequently calculated from the separate patient-level OOF predictions described below.

A separate non-repeated stratified five-fold cross-validation procedure with shuffling and random_state=42 was used to generate patient-level OOF class predictions and probability estimates. This distinction was necessary because OOF predictions require each patient to receive a prediction from a model trained without that patient's observation, whereas repeated cross-validation generates multiple validation appearances for each observation. The resulting OOF predictions were subsequently used to construct the confusion matrix and ROC curve. Area under the curve (AUC) was prioritized as the primary discriminative metric, while F1-score was used to evaluate classification balance between precision and recall. This evaluation design was adopted to reduce dependence on a single train-test partition and to obtain predictions for each observation from models that had not been fitted using that observation [22], [23].

Table 2. Performance comparison of ML models

Strategy	Model	Accuracy	Precision	Recall	F1	Balanced accuracy	AUC	Rank score
No SMOTE	LR	0.789	0.929	0.813	0.867	0.740	0.769	0.798
No SMOTE	DT	0.789	0.891	0.854	0.872	0.649	0.648	0.715
No SMOTE	RF	0.842	0.882	0.938	0.909	0.635	0.762	0.806
No SMOTE	**XGBoost**	**0.912**	**0.922**	**0.979**	**0.949**	**0.767**	**0.775**	**0.828**
No SMOTE	SVM	0.842	0.898	0.917	0.907	0.681	0.694	0.758
No SMOTE	Naïve Bayes	0.474	0.846	0.458	0.595	0.507	0.596	0.596
No SMOTE	AdaBoost	0.825	0.880	0.917	0.898	0.625	0.719	0.773
No SMOTE	MLP	0.877	0.918	0.938	0.928	0.747	0.785	0.828
SMOTE	LR	0.860	0.935	0.896	0.915	0.781	0.801	0.835
SMOTE	DT	0.825	0.932	0.854	0.891	0.760	0.774	0.809
SMOTE	RF	0.842	0.933	0.875	0.903	0.771	0.801	0.832
SMOTE	**XGBoost**	**0.877**	**0.936**	**0.917**	**0.926**	**0.792**	**0.806**	**0.842**
SMOTE	SVM	0.860	0.900	0.938	0.918	0.691	0.731	0.788
SMOTE	Naïve Bayes	0.491	0.880	0.458	0.603	0.563	0.622	0.616
SMOTE	AdaBoost	0.877	0.936	0.917	0.926	0.792	0.750	0.803
SMOTE	MLP	0.842	0.915	0.896	0.905	0.726	0.734	0.785

2.8. Model evaluation

Model performance was evaluated using accuracy, precision, recall (sensitivity), F1-score, balanced accuracy, and area under the receiver operating characteristic (ROC) curve (AUC). Because the outcome distribution was markedly imbalanced, balanced accuracy and AUC were emphasized alongside conventional classification metrics. Balanced accuracy represents the mean of sensitivity and specificity and therefore reduces the influence of majority-class prevalence on performance interpretation. AUC quantified each model's ability to discriminate severe from lower-risk complication patterns across classification thresholds. Confusion matrices and ROC curves were generated from patient-level OOF predictions.

2.9. Model selection strategy

Final model selection was based on a predefined composite ranking score combining discrimination and classification balance. AUC was assigned a weight of 0.70 and F1-score a weight of 0.30:

$$\text{Rank score} = 0.70 \times \text{AUC} + 0.30 \times \text{F1} - \text{score} \quad (1)$$

AUC received greater weight because discrimination across outcome classes was the primary model-selection objective, while F1-score was included to capture the balance between precision and recall. Only models that generated valid OOF predictions were eligible for final selection. The model with the highest resulting rank score was designated as the final model.

2.10. Explainable artificial intelligence analysis

To improve model transparency and clinical interpretability, SHAP-based explainability analysis was performed. SHAP quantifies the contribution of each predictor to a model output using Shapley values derived from cooperative game theory [15]. Both global and local interpretability analyses were performed to identify

clinically influential predictors for stratifying patients with severe obstetric complications. SHAP summary plots were generated to visualize: i) feature importance ranking; ii) direction of feature contribution; and iii) global predictor influence across all patient records. This approach enabled interpretation of how variables such as infection status, hemoglobin level, maternal age, parity, and history of miscarriage contributed to the model's predictions.

3. RESULTS

3.1. Baseline characteristics

A total of 57 patients with miscarriage-related conditions were included in the analysis. The dataset consisted of several miscarriage patterns, including incomplete miscarriage, missed miscarriage, complete miscarriage, and incipient miscarriage. For ML stratification, severe obstetric complication patterns were defined as incomplete miscarriage, missed miscarriage, and incipient miscarriage, whereas complete miscarriage and threatened/imminent miscarriage were categorized as lower-risk patterns, consistent with the binary outcome definition used for model development.

Most patients were within the reproductive age range, although maternal age and parity varied considerably across the cohort. Infection was more frequently observed in cases with severe obstetric complications, while hemoglobin levels varied substantially among patients. These variables were subsequently incorporated as predictors in the ML models.

3.2. Machine learning model performance

A total of eight ML algorithms were evaluated under two experimental conditions, without SMOTE and with SMOTE. The evaluated algorithms included LR, DT, RF, XGBoost, SVM, Naïve Bayes, AdaBoost, and MLP. Repeated stratified cross-validation (5 folds×10 repetitions) was used to assess model stability across multiple data partitions, whereas the final patient-level performance metrics reported in Table 2 were derived from a separate five-fold OOF evaluation to maintain consistency with the ROC and confusion-matrix analyses. Performance metrics included accuracy, precision, recall, F1-score, balanced accuracy, AUC, and overall rank score. In general, ensemble-based approaches such as RF and XGBoost consistently outperformed simpler probabilistic classifiers across most evaluation metrics.

Among the non-SMOTE models, XGBoost achieved the highest accuracy (0.912), F1-score (0.949), and recall (0.979), indicating high sensitivity in detecting patterns of severe obstetric complications in the present dataset. The MLP also performed competitively, with an AUC of 0.785 and an F1-score of 0.928, although it did not outperform the selected XGBoost model. After applying SMOTE, XGBoost demonstrated the best overall performance among the evaluated models, achieving accuracy=0.877, precision=0.936, Recall=0.917, F1-score=0.926, balanced accuracy=0.792, AUC=0.806, and rank score=0.842.

These findings indicate that SMOTE-enhanced XGBoost provided a favorable balance between class-specific discrimination and overall ranking performance, as reflected by its joint-highest balanced accuracy, highest AUC, and highest composite rank score. In contrast, Naïve Bayes consistently demonstrated lower predictive performance under both SMOTE and non-SMOTE conditions, which may partly reflect its conditional-independence assumption in the presence of correlated clinical predictors. Table 2 compares the performance of the evaluated ML models for severe obstetric complication-pattern stratification.

3.2.1. Effect of synthetic minority oversampling technique

Applying SMOTE improved predictive discrimination across several models, particularly LR, RF, and XGBoost, as reflected in higher AUC and balanced accuracy values. For example, XGBoost without SMOTE achieved: accuracy=0.912, F1-score=0.949, balanced accuracy=0.767, and AUC=0.775. After applying SMOTE, XGBoost demonstrated: Accuracy=0.877, F1-score=0.926, balanced accuracy=0.792, and AUC=0.806.

Although overall accuracy decreased slightly after SMOTE, balanced accuracy and AUC improved, indicating better discrimination across both outcome classes and reduced bias toward the majority class. Similarly, LR benefited substantially from SMOTE, with AUC increasing from 0.769 to 0.801 and balanced accuracy improving from 0.740 to 0.781. Overall, SMOTE improved balanced discrimination for several classifiers, particularly XGBoost, LR, and RF, although the magnitude and direction of the effect varied across algorithms. These findings indicate that oversampling benefited selected models rather than uniformly improving every classifier.

3.3. Best model selection

Among all evaluated models, XGBoost with SMOTE demonstrated the best overall performance according to the predefined ranking criterion. It achieved a joint-highest balanced accuracy of 0.792, the highest AUC (0.806), and the highest overall rank score (0.842). Although XGBoost without SMOTE achieved the

highest overall accuracy (0.912) and F1-score (0.949), the SMOTE-enhanced XGBoost model demonstrated superior balanced-classification performance and improved discrimination across both outcome classes. Consequently, XGBoost with SMOTE was selected as the final model for stratifying severe obstetric complications.

3.3.1. Out-of-fold prediction analysis

OOF predictions were generated using stratified k-fold cross-validation so that each patient's prediction was obtained from a model that had not been fitted using that observation. This approach reduced in-sample optimism and provided an internally cross-validated estimate of model performance. The resulting OOF ROC analysis demonstrated moderate-to-good internal discrimination, with an AUC of 0.806.

3.3.2. Confusion matrix analysis

The confusion matrix from OOF predictions summarized the classification performance of the final XGBoost model with SMOTE. The model correctly classified 44 severe-complication cases (true positives) and 6 lower-risk cases (true negatives), while producing 4 false negatives and 3 false positives. These findings indicate relatively high sensitivity for identifying severe-complication patterns within the present dataset. However, four severe cases were misclassified as lower-risk, which is clinically important because false-negative predictions could delay recognition of patients requiring closer assessment. Given the small number of lower-risk observations and the lack of external validation, the confusion matrix should be interpreted as an internal proof-of-concept assessment rather than as evidence of clinical screening performance. Figure 2 presents the OOF confusion matrix of the final XGBoost-SMOTE model, summarizing true-positive, true-negative, false-positive, and false-negative classifications.

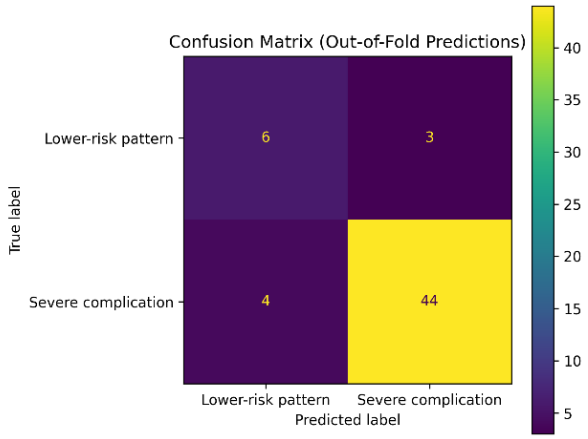


Figure 2. Confusion matrix derived from OOF predictions

3.3.3. Receiver operating characteristic curve analysis

Figure 3 shows the ROC curve derived from OOF predictions. The selected XGBoost-SMOTE model achieved an OOF AUC of 0.806, indicating moderate-to-good internal discrimination between severe and lower-risk obstetric complication patterns. The curve remained predominantly above the diagonal reference line, demonstrating discrimination beyond random classification across the evaluated thresholds. However, the stepwise appearance of the ROC curve reflects the small number of observations, particularly in the lower-risk class, and therefore individual changes in classification threshold can produce relatively large changes in sensitivity and false-positive rate.

Clinically, the AUC of 0.806 indicates that the model has moderate-to-good ranking ability within this dataset, but it should not be interpreted as evidence that the model is ready for diagnostic or treatment decisions. The OOF confusion matrix further showed that four severe-complication cases were incorrectly classified as lower-risk, while three lower-risk cases were classified as severe. In an obstetric screening context, false-negative predictions are particularly important because they could potentially delay recognition of patients requiring closer assessment. Therefore, these ROC findings support discriminatory feasibility rather than clinical deployment, and prospective use will require threshold selection, probability calibration, and external validation.

3.4. Shapley additive explanations analysis

The SHAP summary plot demonstrated that infection status had the greatest overall contribution to model output, followed by hemoglobin level, parity count, maternal age, and history of miscarriage. Positive SHAP values indicate contributions that shift the model output toward the severe-complication class, whereas negative SHAP values shift the prediction toward the lower-risk class.

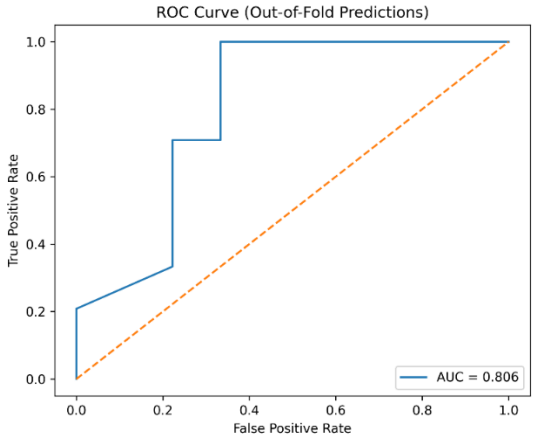


Figure 3. OOF ROC curve of the final XGBoost model with SMOTE for stratification of severe obstetric complication patterns

Infection status showed a clear directional separation. Higher feature values were associated predominantly with positive SHAP values, shifting predictions toward the severe-complication class, whereas lower values were associated with negative SHAP values. Hemoglobin level showed a more heterogeneous pattern. Higher relative hemoglobin values were predominantly associated with positive SHAP contributions, whereas several lower values produced negative contributions, although substantial variation was observed across individual patients. This model-derived direction should be interpreted cautiously and should not be considered evidence that higher hemoglobin levels are biologically associated with greater obstetric severity. Rather, it reflects the associations learned by the fitted model from this small dataset.

Parity also showed a directional pattern, with higher values generally contributing positively toward severe-complication predictions and lower values contributing negatively. Maternal age demonstrated smaller and more heterogeneous contributions distributed on both sides of zero, suggesting a less consistent relationship with the model output. History of miscarriage showed SHAP values concentrated close to zero and therefore contributed comparatively little to prediction in the present dataset. Overall, the SHAP results indicate that infection status and hemoglobin level were the dominant contributors to model behavior, but the direction and magnitude of these associations should be interpreted as model-specific and exploratory. Table 3 presents the relative feature contribution based on SHAP analysis.

Table 3. Model-derived interpretation of predictor contributions based on the SHAP summary plot

Predictor	SHAP pattern	Interpretation
Infection status	Higher values predominantly showed positive SHAP values	Strongest contribution toward severe-complication classification
Hemoglobin level	Higher values were predominantly associated with positive SHAP values, with heterogeneous individual effects	Strong contribution, but direction should be interpreted cautiously because it may reflect sample-specific model behavior
Parity count	Higher values generally showed positive SHAP contributions	Higher parity tended to shift predictions toward the severe-complication class
Maternal age	Contributions occurred on both sides of zero	Smaller and heterogeneous influence on model output
History of miscarriage	SHAP values concentrated near zero	Limited contribution in the present dataset

The SHAP distributions showed substantial between-patient heterogeneity, particularly for hemoglobin level and maternal age. Infection status exhibited the clearest directional separation, whereas history of miscarriage contributed little additional model information. These SHAP patterns should be interpreted as model-derived associations rather than causal effects. In particular, the counterintuitive hemoglobin pattern may reflect sample-specific interactions or instability associated with the limited dataset; therefore, the ranking, direction, and magnitude of feature effects should be regarded as exploratory.

Predictors are ordered according to their overall contribution to model output. Each point represents an individual patient-level SHAP value. Positive SHAP values shift the model output toward the severe-complication class, whereas negative values shift the output toward the lower-risk class. Point color represents the relative feature value, ranging from lower (blue) to higher (red/pink) values. Infection status and hemoglobin level showed the greatest overall contributions, followed by parity count and maternal age, whereas history of miscarriage contributed comparatively little as seen in Figure 4. SHAP values represent associations learned by the fitted model and should not be interpreted as causal effects.

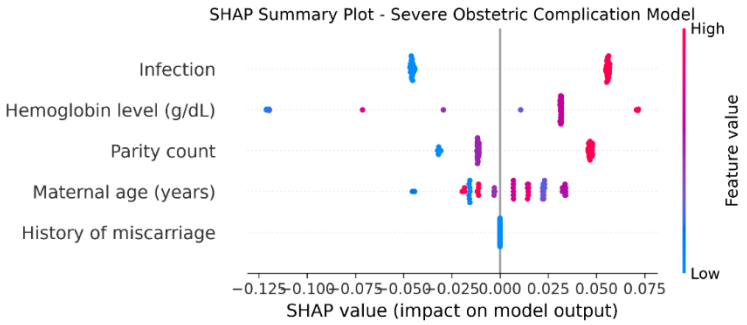


Figure 4. SHAP summary plot of the final model illustrating feature contributions

4. DISCUSSION

This study provides exploratory evidence that explainable ML can internally stratify severe versus lower-risk miscarriage-related complication patterns using structured EMR data. The analysis focuses on a clinically distinct within-cohort stratification task and evaluates both discrimination and model interpretability under a leakage-aware, imbalance-aware validation framework.

4.1. Principal findings

The final XGBoost-SMOTE model achieved moderate-to-good internal discrimination, with an OOF AUC of approximately 0.81. However, these performance estimates should be interpreted as preliminary and exploratory because they were derived from only 57 patient records from a single center. The findings therefore demonstrate internal discriminatory feasibility rather than established predictive performance or clinical effectiveness. Unlike previous miscarriage-prediction studies focused primarily on miscarriage occurrence, the present study targeted clinically defined severe complication patterns, including incomplete miscarriage, missed miscarriage, and incipient miscarriage. Among the evaluated algorithms, XGBoost with SMOTE performed best overall under the predefined ranking criterion, combining a joint-highest balanced accuracy with the highest AUC and overall rank score. These findings indicate that boosting-based ensemble models, particularly XGBoost, can capture complex nonlinear relationships among obstetric clinical variables while maintaining balanced discrimination across outcome classes [24].

4.2. Model performance and algorithm comparison

The comparison across multiple ML algorithm families demonstrated that ensemble-based methods, particularly XGBoost and RF, consistently outperformed simpler statistical and probabilistic approaches. XGBoost with SMOTE achieved the strongest overall performance balance, whereas XGBoost without SMOTE produced the highest raw accuracy and F1-score. One possible explanation is that ensemble-based learning methods better capture nonlinear interactions among obstetric clinical variables. Gradient boosting algorithms such as XGBoost sequentially refine weak learners, potentially improving the detection of subtle interactions among infection status, hemoglobin level, maternal age, and parity [24]. This observation is consistent with previous healthcare ML studies showing that boosting-based ensemble models often outperform conventional regression approaches on heterogeneous clinical datasets [25].

However, direct numerical comparisons with previous miscarriage-prediction studies should be made cautiously because the target outcome differs. Previous investigations have primarily predicted the occurrence or recurrence of pregnancy loss, whereas the present study classified complication severity among patients with miscarriage-related conditions [8], [16], [17]. Differences in population, predictor availability, outcome prevalence, and validation design also limit direct comparison of AUC values across studies. Accordingly, the principal contribution of the present analysis is not superior predictive accuracy over previous models, but

demonstration of an internally evaluated and interpretable framework for a different, clinically focused stratification task using a small real-world EMR dataset.

Although the MLP performed competitively, it did not surpass XGBoost, possibly reflecting the limited sample size required for stable neural-network learning [10]. This comparison highlights the trade-off between predictive complexity and interpretability: simpler models such as LR are more directly interpretable, whereas XGBoost captures nonlinear interactions but requires explanation methods such as SHAP. In this study, SHAP provided feature-level explanations for XGBoost predictions. Moreover, XGBoost without SMOTE achieved the highest raw accuracy and F1-score, while SMOTE-enhanced XGBoost achieved better balanced accuracy and AUC. Thus, clinical model selection should consider predictive performance, class balance, and interpretability rather than accuracy alone.

4.3. Effect of synthetic minority oversampling technique and validation strategy

SMOTE improved discriminative performance across several models, particularly LR, RF, and XGBoost, as reflected in higher AUC and balanced accuracy. For example, XGBoost with SMOTE achieved higher balanced accuracy and AUC than the corresponding non-SMOTE model, despite slightly lower raw accuracy. This finding highlights an important methodological consideration in healthcare ML research. Although SMOTE improves minority-class representation, synthetic observations may not fully capture the underlying biological distribution of real clinical data. Consequently, oversampling may occasionally introduce noise or optimistic bias, especially in small datasets [21].

To minimize these risks, SMOTE was applied exclusively within the training folds during cross-validation, thereby reducing information leakage into validation data. Comparing SMOTE and non-SMOTE strategies also enabled direct assessment of the effect of oversampling on model discrimination. OOF prediction provided an additional internal assessment in which each patient's prediction was generated without using that observation for model fitting. From an engineering perspective, the integration of fold-specific preprocessing, training-only imbalance handling, repeated stratified validation, patient-level OOF evaluation, and SHAP-based explanation provides a reusable leakage-aware analytical pipeline for small and imbalanced clinical datasets. Nevertheless, neither repeated cross-validation nor OOF evaluation can establish external generalizability when the underlying dataset is small and derived from a single center. The observed performance should therefore be interpreted as evidence of internal discriminatory feasibility rather than robust generalization, and independent external validation remains necessary.

4.4. Clinical interpretation and explainability

SHAP analysis showed how the final model used individual predictors to stratify severe and lower-risk obstetric complication patterns. Infection status showed the greatest overall contribution, with higher feature values predominantly shifting predictions toward the severe-complication class. This finding is clinically plausible because infection and inflammatory processes may accompany complicated miscarriage presentations and are associated with increased maternal morbidity [1].

Hemoglobin level was the second most influential predictor; however, its direction requires more cautious interpretation. Contrary to the clinically expected association between low hemoglobin and adverse outcomes, higher relative hemoglobin values in the present model were predominantly associated with positive SHAP contributions toward the severe-complication class, while several lower values contributed toward the lower-risk class. This apparently counterintuitive pattern should not be interpreted as evidence that higher hemoglobin is a biological risk factor for severe miscarriage-related complications. Instead, it may reflect sample-specific associations, interactions with other predictors, the limited number of observations, or the structure learned by the fitted model. This finding illustrates an important advantage of explainability analysis: SHAP can reveal not only clinically plausible relationships but also unexpected model behavior that warrants further investigation rather than post-hoc clinical justification.

Parity count also showed a relatively clear model-derived pattern, with higher values generally shifting predictions toward the severe-complication class. Maternal age had a smaller and more heterogeneous contribution, with SHAP values distributed on both sides of zero, indicating that its influence depended on the individual predictor profile. History of miscarriage contributed comparatively little, with SHAP values concentrated near zero. This limited contribution does not imply that previous miscarriage is clinically unimportant; rather, it indicates that the fitted model extracted relatively little additional predictive information from this variable after accounting for the other predictors in this particular dataset.

Nevertheless, neither clinically plausible nor counterintuitive SHAP patterns should be interpreted as evidence of stable or causal predictor effects. SHAP explanations describe the behavior of a fitted prediction model and may vary when the training sample, model parameters, or correlation structure among predictors changes. This issue is particularly relevant in the present study because only 57 patient records were available. With a small dataset, relatively few observations can influence tree splits and consequently alter both model predictions and SHAP attribution patterns. Correlated clinical predictors may also redistribute attribution

among related features. Therefore, the observed importance ranking-particularly the predominance of infection status and hemoglobin level-should be regarded as hypothesis-generating and should be confirmed in larger independent datasets and through repeated explainability analyses.

4.5. Clinical implications

The present findings indicate potential clinical relevance rather than established clinical utility. The predictors identified by the model-hemoglobin level, infection status, maternal age, and parity-are routinely obtainable from EMR systems and therefore provide a practical basis for further investigation. However, the present model was developed and internally evaluated using only 57 records from a single center and has not undergone external or prospective validation. Therefore, it should not be used to guide diagnosis, treatment, or patient triage in its current form.

If subsequently validated in larger and independent populations, an explainable model of this type could potentially serve as an adjunctive risk-stratification component within an EMR-based clinical decision support system (CDSS). Rather than providing only a predicted risk estimate, SHAP-based explanations could show clinicians which routinely available patient variables contributed most strongly to an individual prediction. This transparency may facilitate clinical review of model outputs, help clinicians identify predictions that appear inconsistent with the patient context, and potentially improve confidence in and acceptance of AI-assisted decision support. Importantly, explainability should support rather than replace clinical judgment. From an engineering perspective, implementation would require reproducible preprocessing and inference pipelines, calibrated probability outputs, transparent patient-level explanations, and prospective monitoring of model performance. External discrimination, calibration, clinical utility, threshold selection, and prospective workflow performance must therefore be established before clinical deployment.

4.6. Limitations

Several limitations should be considered. First, the dataset was small and markedly imbalanced (57 patients; 48 severe-complication; and 9 lower-risk cases). Although repeated stratified cross-validation and patient-level OOF prediction reduced dependence on a single split, they cannot compensate for limited clinical information. Estimates of specificity, balanced accuracy, AUC, and feature attribution may therefore be sensitive to individual observations and fold composition. Confidence intervals and formal variance estimates across repeated runs were not reported, so performance should be interpreted as exploratory rather than as precise population-level estimates. Second, the single-center design limits transportability, and larger independent multicenter cohorts are needed to assess statistical uncertainty and external generalizability. Third, the predictor set was limited to routinely available structured variables and excluded potentially informative ultrasonographic, hormonal, longitudinal laboratory, and genomic data. Fourth, although SMOTE was restricted to training folds, synthetic oversampling does not add biological information and may alter local feature distributions; sensitivity analysis across alternative K_neighbors values was not performed, so results are conditional on the selected resampling configuration. Fifth, hyperparameters were prespecified rather than optimized through nested tuning, reducing extensive data-driven model selection but not ensuring maximal performance for each algorithm. Sixth, probability calibration was not formally assessed; future studies should evaluate calibration plots, Brier scores, calibration intercepts, and slopes. Seventh, SHAP explanations may be unstable in small samples and influenced by model specification, sampling variation, and predictor correlations, so feature rankings should be considered exploratory rather than causal. Finally, no external or prospective clinical evaluation was performed; therefore, the model should be regarded as a proof-of-concept risk-stratification framework rather than a clinically validated decision-support tool.

4.7. Future research

Future studies should prioritize larger multicenter cohorts representing different institutions, patient populations, and clinical settings to quantify statistical uncertainty, assess transportability, and perform independent external validation. With sufficient sample sizes, model development should incorporate nested hyperparameter optimization, sensitivity analyses across alternative SMOTE neighborhood settings, and confidence intervals or variance estimates for discrimination metrics. Additional clinical information, including ultrasonographic, hormonal, longitudinal laboratory, and imaging variables, may improve predictive representation. Future work should also evaluate probability calibration, threshold-specific clinical utility, and the stability of SHAP feature attribution across resamples and independent populations. Following satisfactory external validation, prospective studies should evaluate integration into EMR-based clinical workflows, including clinician interaction with model explanations, decision thresholds, usability, and monitoring of real-world model performance.

5. CONCLUSION

This exploratory pilot study provides preliminary evidence that explainable ML can internally discriminate severe from lower-risk miscarriage-related complication patterns using a small set of routinely available EMR variables. Among the evaluated approaches, XGBoost with SMOTE provided the best overall balance of performance, achieving a balanced accuracy of 0.792 and an OOF AUC of 0.806. SHAP analysis indicated that infection status and hemoglobin level contributed most strongly to model predictions, followed by parity and maternal age, although these attribution patterns should be interpreted cautiously because of the limited sample size and their model-dependent nature.

This study's contribution lies primarily in reframing the prediction task toward clinically meaningful complication stratification and integrating leakage-aware preprocessing, training-only imbalance handling, repeated cross-validation, patient-level OOF evaluation, and SHAP-based explainability within a small real-world EMR setting, rather than introducing a new ML algorithm. This integrated framework also provides reusable methodological knowledge for exploratory modeling of other small and imbalanced clinical EMR datasets. Nevertheless, the present results establish proof of concept rather than clinical effectiveness. Larger multicenter datasets, independent external validation, probability-calibration assessment, and prospective validation are required before implementation. Subsequent research should also evaluate integration into clinical workflows, including clinician interaction with model explanations, decision thresholds, usability, and prospective monitoring of model performance.

ACKNOWLEDGMENTS

The authors thank the medical and administrative staff of Rumah Sakit Ibu dan Anak Muslimat Jombang for facilitating access to the anonymized electronic medical record data. The authors also acknowledge colleagues who provided technical and data-management assistance during the study. These contributions did not meet the criteria for authorship. Generative AI disclosure: OpenAI ChatGPT was used during manuscript revision to assist with language editing, structural refinement, conciseness, and consistency checking in response to peer-review comments. No generative AI tool was used to generate or alter the study data. All scientific interpretations, statistical results, references, and final manuscript content were independently verified and approved by the authors, who take full responsibility for the submitted work.

FUNDING INFORMATION

This research received no specific research grant or contract from any funding agency in the public, commercial, or not-for-profit sectors.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
I Gusti Njoman Teguh Budhi Bimantara	✓	✓		✓	✓	✓	✓	✓	✓	✓		✓	✓	
I Gusti Putu Asto Buditjahjanto	✓	✓	✓			✓		✓	✓	✓	✓	✓		
Lilik Anifah	✓	✓	✓	✓			✓		✓	✓	✓	✓	✓	
Gusti Ayu Az Zahra Shabrina Putri Bukian	✓	✓		✓	✓	✓	✓	✓	✓	✓		✓		

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

The authors state that there is no conflict of interest.

INFORMED CONSENT

Informed consent was waived by the institutional ethics committee due to the retrospective nature of the study and the use of anonymized electronic medical record data. All data were de-identified prior to analysis, and no personally identifiable information was included.

ETHICAL APPROVAL

This study involving human participants was conducted in accordance with all relevant national regulations and institutional policies, and in compliance with the ethical principles outlined in the Declaration of Helsinki. Ethical approval for this study was obtained from the Health Research Ethics Committee of Universitas Muhammadiyah Surabaya (*Komite Etik Penelitian Kesehatan* Universitas Muhammadiyah Surabaya), under approval number 007/KET/II.3/AU/F/2023. The research protocol was reviewed and approved prior to data collection. All patient data were anonymized to ensure confidentiality and privacy protection, and no personally identifiable information was included in the analysis.

DATA AVAILABILITY

The electronic medical record dataset analyzed in the present study is not publicly available due to institutional privacy and patient confidentiality restrictions. De-identified data may be available from the corresponding author upon reasonable request and subject to institutional ethical approval.

REFERENCES

[1] S. Quenby *et al.*, "Miscarriage matters: The epidemiological, physical, psychological, and economic costs of early pregnancy loss," *The Lancet*, vol. 397, no. 10285, pp. 1658–1667, May 2021, doi: 10.1016/S0140-6736(21)00682-6.

[2] J. Bearak *et al.*, "Unintended pregnancy and abortion by income, region, and the legal status of abortion: estimates from a comprehensive model for 1990–2019," *The Lancet Global Health*, vol. 8, no. 9, pp. e1152–e1161, Sep. 2020, doi: 10.1016/S2214-109X(20)30315-6.

[3] M. M. Giorgio *et al.*, "Estimating the incidence of induced abortion in Java, Indonesia, 2018," *International Perspectives on Sexual and Reproductive Health*, vol. 46, pp. 211–222, 2020, doi: 10.1363/46e0220.

[4] Z. Al-Alami, R. Abu-Huwaij, S. Hamadneh, and E. Taybeh, "Understanding miscarriage prevalence and risk factors: Insights from Women in Jordan," *Medicina*, vol. 60, no. 7, Jun. 2024, doi: 10.3390/medicina60071044.

[5] H. S. Sonu, S. K. Das, R. Tony, and V. S. Binu, "Risk and protective factors of miscarriage: Evidence from a nationally representative sample of women in India," *Journal of Family Medicine and Primary Care*, vol. 13, no. 9, pp. 3879–3886, Sep. 2024, doi: 10.4103/jfmpc.jfmpc_329_24.

[6] R. Rosnani *et al.*, "Pregnancy termination among women of reproductive age: Evidence from the Indonesian demographic and health survey," *International Journal of Environmental Research and Public Health*, vol. 22, no. 4, Apr. 2025, doi: 10.3390/ijerph22040564.

[7] I. M. W. Supanji, A. A. R. Budayasa, P. A. N. K. Permatananda, P. N. Cahyawati, and A. A. S. A. Aryastut, "Relationship between maternal age, education, and parity in the incidence of spontaneous abortion in Bali," *Science Midwifery*, vol. 10, no. 4, pp. 3091–3095, Oct. 2022, doi: 10.35335/midwifery.v10i4.710.

[8] X. H. Choo *et al.*, "Risk score to stratify miscarriage risk levels in preconception women," *Scientific Reports*, vol. 11, no. 1, Jun. 2021, doi: 10.1038/s41598-021-91567-8.

[9] Y. Gong *et al.*, "Can single progesterone concentration predict miscarriage in early pregnant women with threatened miscarriage: a systematic review and meta-analysis," *BMC Pregnancy and Childbirth*, vol. 24, no. 1, Feb. 2024, doi: 10.1186/s12884-024-06303-7.

[10] B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi, "Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 5, pp. 1589–1604, Sep. 2018, doi: 10.1109/JBHI.2017.2767063.

[11] X. Yang *et al.*, "Predicting interval from diagnosis to delivery in preeclampsia using electronic health records," *Nature Communications*, vol. 16, no. 1, Apr. 2025, doi: 10.1038/s41467-025-58437-7.

[12] B. Sowan and B. Massad, "Federated learning in obstetrics and gynaecology: A systematic review of applications, architectures, privacy frameworks, and clinical translation challenges," *Informatics in Medicine Unlocked*, vol. 65, Sep. 2026, doi: 10.1016/j.imu.2026.101794.

[13] D. Muhammad and M. Bendechache, "Unveiling the black box: A systematic review of explainable artificial intelligence in medical image analysis," *Computational and Structural Biotechnology Journal*, vol. 24, pp. 542–560, Dec. 2024, doi: 10.1016/j.csbj.2024.08.005.

[14] H. Müller, A. Holzinger, M. Plass, L. Brcic, C. Stumptner, and K. Zatloukal, "Explainability and causability for artificial intelligence-supported medical image analysis in the context of the European In Vitro Diagnostic Regulation," *New Biotechnology*, vol. 70, pp. 67–72, Sep. 2022, doi: 10.1016/j.nbt.2022.05.002.

[15] S. M. Lundberg *et al.*, "From local explanations to global understanding with explainable AI for trees," *Nature Machine Intelligence*, vol. 2, no. 1, pp. 56–67, Jan. 2020, doi: 10.1038/s42256-019-0138-9.

[16] X. Yang, R. Wang, W. Zhang, Y. Yang, and F. Wang, "Predicting risk of the subsequent early pregnancy loss in women with recurrent pregnancy loss based on preconception data," *BMC Women's Health*, vol. 24, no. 1, Jul. 2024, doi: 10.1186/s12905-024-03206-9.

[17] J. J. Yland *et al.*, "Predictive models of miscarriage on the basis of data from a preconception cohort study," *Fertility and Sterility*, vol. 122, no. 1, pp. 140–149, Jul. 2024, doi: 10.1016/j.fertnstert.2024.04.007.

[18] W. Khan *et al.*, "Predicting preterm birth using explainable machine learning in a prospective cohort of nulliparous and multiparous pregnant women," *PLOS ONE*, vol. 18, no. 12, Dec. 2023, doi: 10.1371/journal.pone.0293925.

[19] S. A. Tiruneh, D. L. Rolnik, H. J. Teede, and J. Enticott, "Prediction of pre-eclampsia with machine learning approaches: Leveraging

important information from routinely collected data," *International Journal of Medical Informatics*, vol. 192, Dec. 2024, doi: 10.1016/j.ijmedinf.2024.105645.
[20] A. Pfob, S.-C. Lu, and C. Sidey-Gibbons, "Machine learning in medicine: a practical introduction to techniques for data pre-processing, hyperparameter tuning, and model comparison," *BMC Medical Research Methodology*, vol. 22, no. 1, Nov. 2022, doi: 10.1186/s12874-022-01758-8.
[21] A. Fernandez, S. Garcia, F. Herrera, and N. V. Chawla, "SMOTE for learning from imbalanced Data: Progress and challenges, marking the 15-year anniversary," *Journal of Artificial Intelligence Research*, vol. 61, pp. 863–905, Apr. 2018, doi: 10.1613/jair.1.11192.
[22] S. Varma and R. Simon, "Bias in error estimation when using cross-validation for model selection," *BMC Bioinformatics*, vol. 7, no. 1, Feb. 2006, doi: 10.1186/1471-2105-7-91.
[23] G. S. Collins *et al*., "TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods," *BMJ*, vol. 385, Apr. 2024, doi: 10.1136/bmj-2023-078378.
[24] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 2016, pp. 785–794, doi: 10.1145/2939672.2939785.
[25] A. Rajkomar, J. Dean, and I. Kohane, "Machine learning in medicine," *New England Journal of Medicine*, vol. 380, no. 14, pp. 1347–1358, Apr. 2019, doi: 10.1056/NEJMra1814259.

BIOGRAPHIES OF AUTHORS



I Gusti Njoman Teguh Budhi Bimantara is an obstetrician-gynecologist and a postgraduate student in vocational education with a focus on medical education. He received his medical degree from Universitas Airlangga, Surabaya, Indonesia, and completed specialist training in obstetrics and gynecology at Universitas Sebelas Maret, Surakarta, Indonesia. His research interests include evidence-based medicine, obstetric care, and medical education. He is involved in hospital-based medical education and health professions education. He can be contacted at email: igntbb@gmail.com.



I Gusti Putu Asto Buditjahjanto is a professor in the Department of Electrical Engineering at Universitas Negeri Surabaya, Indonesia. He received his bachelor of Engineering and doctoral degrees in Electrical Engineering from Institut Teknologi Sepuluh Nopember, Indonesia, in 1998 and 2011, respectively, and his master of Engineering degree in Industrial Engineering from the same institution in 2003. His research interests include fuzzy multi-attribute group decision making, decision-support systems, learning technologies, and intelligent systems. He can be contacted at email: asto@unesa.ac.id.



Lilik Anifah is a professor in the Department of Electrical Engineering at Universitas Negeri Surabaya, Indonesia. She received her bachelor's and master's degrees in Electrical Engineering, and her Ph.D. in Engineering Physics from Institut Teknologi Sepuluh Nopember, Indonesia. Her research interests include artificial intelligence, biomedical signal and image processing, decision support systems, and internet of things (IoT) applications. She has published numerous scientific papers in international journals and conferences and has been actively involved in research projects related to intelligent systems and healthcare applications. She is a member of the International Association of Engineers. She can be contacted at email: lilikanifah@unesa.ac.id.



Gusti Ayu Az Zahra Shabrina Putri Bukian is a medical doctor who graduated from Universitas Muhammadiyah Surabaya, Indonesia. Her professional interests include maternal and reproductive health. She currently works as an intern doctor at the Indonesian Red Cross. She can be contacted at email: asazzahra24@gmail.com.