International Journal of Engineering, Science and Information Technology
Volume 6, No. 1 (2026) pp. 184-192
eISSN: 2775-2674
Website: http://ijesty.org/index.php/ijesty
DOI: https://doi.org/10.52088/ijesty.v6i1.1880
Research Paper, Short Communication, Review, Technical Paper

Performance Analysis of Ensemble Learning Models Comparing Bagging and Boosting Techniques for Early Preeclampsia Risk Detection in Pregnant Women Prediction

Yudhi Fajar Saputra[1]*, Milkhatun[2], Aldi Bastiatul Fawait[1], Zakaria Ahmad Dahlan[3], Yazeed Al Moaiad[4], Haviluddin[5], Rayner Alfred[6]

[1]Faculty of Engineering and Informatics, Universitas Widya Gama Mahakam, Samarinda, Indonesia
[2]Faculty of Nursing, Universitas Muhammadiyah Kalimantan Timur, Samarinda, Indonesia
[3]Department of Hospital Ward, RSUD Inche Abdoel Moeis, Samarinda, Indonesia
[4]Faculty of Computer & Information Technology, Al-Madinah International University, Kuala Lumpur, Malaysia
[5]Faculty of Engineering, Universitas Mulawarman, Samarinda, Indonesia
[6]Faculty of Computing and Informatics, Universiti Malaysia Sabah, Malaysia

**Corresponding author Email: fajaryudhi@uwgm.ac.id*

The manuscript was received on 12 June 2025, revised on 16 September 2025, and accepted on 10 January 2026, date of publication 28 January 2026

Abstract

Preeclampsia is a major pregnancy complication that substantially contributes to maternal morbidity and mortality worldwide, making early identification of risk factors essential for effective prevention and timely clinical intervention. This study evaluates the performance of ensemble learning models by comparing bagging and boosting techniques to develop an accurate early prediction system for preeclampsia risk using clinical medical record data. The research follows the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework, encompassing data understanding, data preparation, modeling, evaluation, and interpretation. The dataset was obtained from RSUD Inche Abdoel Moeis Samarinda and underwent preprocessing procedures, including data cleaning, transformation, feature encoding, normalization, and Synthetic Minority Over-sampling Technique (SMOTE) to address class imbalance. Six ensemble learning algorithms were evaluated, consisting of Random Forest, Extra Trees, and Rotation Forest as bagging methods, and XGBoost, LightGBM, and CatBoost as boosting methods. Model performance was assessed using accuracy, precision, recall, and weighted F1-score. The experimental results demonstrate that Random Forest achieved the highest predictive performance, with an accuracy of 0.92, precision of 0.93, recall of 0.92, and weighted F1-score of 0.91, indicating superior robustness and generalization capability. Extra Trees achieved comparable accuracy (0.92) but exhibited lower prediction stability across evaluation metrics. Among the boosting algorithms, LightGBM and CatBoost each obtained an accuracy of 0.89, while XGBoost achieved 0.88. Rotation Forest recorded the lowest accuracy (0.62), suggesting limited suitability for this clinical dataset. These findings indicate that bagging-based ensemble methods, particularly Random Forest, outperform boosting techniques for imbalanced clinical data and provide strong empirical support for developing reliable Clinical Decision Support Systems (CDSS) for early preeclampsia screening and risk assessment in healthcare settings.

Keywords: *Clinical Decision Support System, Ensemble Learning, Machine Learning, Preeclampsia, Random Forest.*

1. Introduction

Preeclampsia is a severe pregnancy complication that significantly contributes to maternal morbidity and mortality worldwide. According to the World Health Organization (WHO), preeclampsia and eclampsia account for more than 10% of global maternal deaths [1]. The condition is typically characterized by hypertension and the presence of protein in urine (proteinuria). When left undetected, it can progress into eclampsia, a life-threatening condition with an increased risk of maternal and fetal death [2]. In Indonesia, preeclampsia remains among the leading causes of maternal mortality, as reported in the Indonesian Health Profile (2020), which identifies it as one of the top contributors to maternal death nationwide [3]. Early detection of preeclampsia is often challenging due to the subtlety of early symptoms and the lack of comprehensive clinical monitoring, underscoring the need for predictive systems based on medical data to assist healthcare providers in risk screening.

Previous research has identified several risk factors for preeclampsia, including maternal age (<20 years or ≥35 years), history of hypertension, high Body Mass Index (BMI), and proteinuria [4][5]. However, conventional statistical approaches often struggle to capture

Copyright © Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

nonlinear relationships among these clinical variables. Advances in machine learning (ML) have opened new opportunities for data-driven diagnosis in maternal healthcare. Rajkomar et al. emphasized that ML can enhance the efficiency of medical services through electronic health record (EHR) analysis [6], while Kohavi and Sokolova & Lapalme highlighted the importance of appropriate algorithm selection and evaluation metrics in predictive modeling [7][8].

Previous studies on early detection of pre-eclampsia with ML have been implemented using various methods such as Support Vector Machine (SVM), k-Nearest Neighbors (k-NN), and Artificial Neural Networks (ANN). Shah et al. and Ma et al. showed that these methods can produce good predictions, but they often lack interpretability regarding which features contribute most significantly to prediction outcomes [9][10]. The ensemble learning approach has begun to be adopted due to its ability to improve model accuracy and stability. Random Forest, one of the bagging methods, has proven effective in predicting cardiovascular disease and pregnancy complications. On the other hand, XGBoost as boosting methods, have been shown to excel in various machine learning competitions due to their high efficiency and accuracy [11], and CatBoost has demonstrated excellent performance in various domains, including health [12]. However, systematic comparisons between bagging and boosting for preeclampsia detection remain limited.

To address this research gap, this study aims to analyze and compare the performance of bagging algorithms (Random Forest, Extra Trees, Rotation Forest) and boosting algorithms (XGBoost, LightGBM, CatBoost) in the early prediction of preeclampsia risk factors. The primary contribution of this research lies in providing quantitative empirical evidence regarding the effectiveness of ensemble learning in handling imbalanced medical data, while reinforcing the foundation for developing Clinical Decision Support Systems (CDSS) for early screening of preeclampsia.

2. Methods

This research adopts the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology, which consists of six primary stages: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment (Fig. 1).

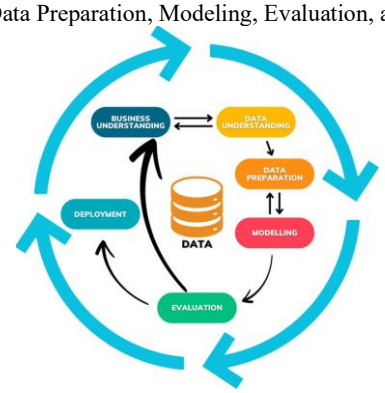

Fig 1. CRISP-DM methods

2.1. Business Understanding

The business understanding phase involves developing contextual knowledge regarding preeclampsia risk factors as one of the primary causes of maternal morbidity and mortality. Based on consultations with medical professionals at RSUD Inche Abdoel Moeis Samarinda, several clinically significant variables were identified, including maternal age, blood pressure, BMI, hypertension history, menarche, diabetes mellitus history, contraceptive use history, and laboratory results (e.g., urine protein tests).

This stage established a strong foundation for the research objective to support the prevention of pregnancy complications through an early detection system utilizing structured medical data.

2.2. Data Understanding

The data understanding stage focused on collecting raw datasets from the primary source: medical records of pregnant women at RSUD Inche Abdoel Moeis Samarinda. Initially, the data were available only in physical form, which included patient transfer forms, triage sheets, emergency department records, admission requests, and obstetric/gynecological notes.

Since the scanned data were not computationally processable, manual data entry was performed into a custom data input application, later exported to Excel format. Each record included anonymized patient identity, maternal and gestational age, blood pressure, medical history (hypertension and diabetes), menarche, body weight and height, contraceptive use, antenatal care visits (ANC), lab results (urine protein test), and final diagnosis related to preeclampsia.

With this approach, each patient's data is represented in a consistent row within a digital dataset. This process aligns with clinical research standards, where medical records often require manual extraction and digitization before they can be computationally analyzed [13]. Therefore, data cleaning and transformation into a structured format such as Excel or a relational database are crucial steps in data-driven health research [14]. As an illustration, a partial example of the first 30 patients shows the results of manual input, which was then validated and exported to Excel, which was then used for further analysis.

Table 1. Partial data of the raw dataset

Patient Code	Age	Maternal Age	Hypertension History	Diabetes History	Menarche Age	Contraceptive Use	Weight	Height	Body Mass Index (BMI)	Antenatal Visits (ANC)	Preeclampsia Diagnosis
27.61.51	21	39	No	No	12	No	70	162	26.6	3	Severe
26.65.95	22	37	No	No	-	No	56	153	23.9	0	Non-preeclampsia
26.70.26	44	38	No	No	13	No	47	150	20.8	0	Non-preeclampsia

19.12.91	27	38	No	No	13	Injection	60	152	25.9	1	Mild
21.52.85	22	38	No	No	-	No	60	160	23.3	1	Severe
08.56.03	31	37	No	No	11	Pill	96	160	37.5	0	Non-preeclampsia
24.01.33	22	37	No	No	12	No	59	155	24.5	0	Non-preeclampsia

Thus, this data understanding stage confirms that dataset collection is not just about obtaining digital data, but rather through a long process that includes processing physical documents, scanning, manual input, validation, and standardization.

2.3. Data Preparation

The data preparation phase transformed the raw dataset into a clean, consistent, and analysis-ready format. As data quality directly influences model performance [15], this step was crucial and involved several sub-processes:

1. Feature Selection – Non-relevant variables were removed to reduce noise [16], such as administrative identifiers (patient ID, medical record number, and administrative notes). Only clinically relevant variables to pre-eclampsia risk factors were retained (e.g., maternal age, gestational age, weight, height, blood pressure, history of hypertension, history of diabetes mellitus, menarche, history of contraceptive use, urine protein test results, and medical diagnosis).
2. Data Cleaning – Missing values were imputed using median (for numeric variables) and mode (for categorical variables). Logical corrections were applied to unrealistic values (e.g., gestational age "338 weeks" corrected to "38 weeks"). Units were standardized to ensure consistency (e.g., weight in kg, height in cm). This process is in line with the study by Rahm and Do which emphasized that data cleaning is a fundamental step to ensure data consistency and reliability [17].
3. Encoding Categorical Variables – To enable machine learning processing, categorical variables were numerically encoded. Binary features such as hypertension or diabetes history were encoded as (0 = No, 1 = Yes), while multi-class variables like contraceptive type were systematically converted using label encoding, as decision-tree-based algorithms perform efficiently with simple numeric encodings [18].

The outcome of the data preparation phase was a cleaned dataset that was free of irrelevant attributes and missing values, standardized in format, and ready for modeling. The feature definitions and categories are summarized in Table I.

Table 2. Variables and categories of the dataset

Symbols	Atribute	
	Variables	Category
Y	Preeclampsia Diagnosis (Target)	(0) Non-preeclampsia, (1) Mild, (2) Severe
X_1	Maternal Age	Numeric
X_2	Age Category	(1) Adolescent, (2) Optimal Reproductive, (3) Older Reproductive
X_3	Gestational Age	Numeric
X_4	Hypertension History	(0) No, (1) Yes
X_5	Diabetes History	(0) No, (1) Yes
X_6	Menarche Age	Numeric
X_7	Contraceptive Use	(0) No, (1) Injection, (2) Pill, (3) General, (4) IUD, (5) Implant, (6) Condom, (7) Ta'a, (8) MAL, (9) MOW
X_8	Weight	Numeric
X_9	Height	Numeric
X_{10}	Body Mass Index (BMI)	Numeric
X_{11}	BMI Category	(1) Underweight, (2) Normal, (3) Overweight, (4) Obesitas I, (5) Obesitas II
X_{12}	Antenatal Visits (ANC)	Numeric

This table clarifies that the dataset used consists of numerical variables (age, weight, height, blood pressure, BMI, menarche, number of ANC visits) and categorical variables (history of hypertension, diabetes, contraception, urine protein).

2.4. Modelling

In this study, the modeling process employed ensemble learning techniques, consisting of two main approaches: bagging and boosting. The bagging group includes the algorithms Random Forest, Extremely Randomized Trees, and Rotation Forest, while the boosting group comprises Extreme Gradient Boosting, Light Gradient Boosting Machine, and CatBoost.

The prepared dataset was divided into training (80%) and testing (20%) subsets using stratified splitting to maintain a balanced target class distribution [7], and all models were trained on the balanced dataset generated using the Synthetic Minority Oversampling Technique (SMOTE) [12]. The algorithms used in this study are described as follows:

1. Random Forest (RF)

   RF is a bagging-based ensemble algorithm that constructs multiple decision trees on randomly sampled subsets of the data (bootstrap sampling) and aggregates their predictions through majority voting [19]. In medical applications, RF has demonstrated high predictive power for complex diseases, including pregnancy complications [20]. RF formula is as below:

$$\widehat{y_{RF}}(\boldsymbol{x}) = \frac{1}{K}\sum_{k=1}^{K} h_k(\boldsymbol{x}) \quad (1)$$

   RF is based on a Decision Tree $h_k(\mathbf{x})$, where K is the number of trees.

2. Extremely Randomized Trees (Extra Trees)
   The Extra Trees algorithm is a variant of RF that introduces a higher level of randomization in selecting features and determining split points [21]. Unlike RF, which selects the optimal split based on criteria such as the Gini index, Extra Trees chooses the split point randomly. So the Extra Trees method shares the same predictive formula as RF, but it introduces extreme randomness by selecting split points arbitrarily (instead of optimally) and often uses the entire training dataset. For regression, the final prediction $\widehat{y_{ET}}(x)$ is:

$$\widehat{y_{ET}}(\boldsymbol{x}) = \frac{1}{K}\sum_{k=1}^{K} h_k(\boldsymbol{x}) \tag{2}$$

   In medical data analysis, Extra Trees is advantageous when dealing with complex datasets containing non-linear relationships among features.
3. Rotation Forest (RoF)
   Rotation Forest employs a different strategy by applying Principal Component Analysis (PCA) to randomly selected subsets of features before training each decision tree [22]. This process generates a rotation matrix R_k for each tree, transforming the feature space prior to training. The final prediction $\widehat{y_{RoF}}(x)$ is the average of predictions from base trees h_k trained on the rotated data:

$$\widehat{y_{RoF}}(x) = \frac{1}{K}\sum_{k=1}^{K} h_k(R_k x) \tag{3}$$

   In this study, RoF is used as an advanced bagging approach, offering a distinctive perspective on feature transformation and diversity within the ensemble.
4. Extreme Gradient Boosting (XGBoost)
   XGBoost is a gradient-boosting-based algorithm designed for high speed and performance [11]. Each subsequent tree is trained to correct the errors made by the previous ones through gradient descent optimization of the loss function. XGBoost builds an additive model $\widehat{y_T}(x)$ sequentially over T iterations. At iteration T, a new tree f_t is chosen to minimize the regularized objective function $\mathcal{L}^{(t)}$, which uses the second-order Taylor expansion of the loss function L:

$$\widehat{y_T}(x) = \sum_{t=1}^{T} f_t(x) \tag{4}$$

   The objective function to be minimized is:

$$\mathcal{L}^{(t)} \approx \sum_{i=1}^{n}\left[g_i f_t(x_i) + \frac{1}{2}h_i f_t^2(x_i)\right] + \Omega(f_t) \tag{5}$$

   Where g_i and h_i are the first and second-order gradients (gradient and Hessian), respectively, and $\Omega(f_t)$ is the regularization term.
   In healthcare applications, XGBoost is widely used for predictive modeling using electronic medical records (EMRs) due to its strong generalization and computational efficiency.
5. Light Gradient Boosting Machine (LightGBM)
   Developed by Microsoft Research, LightGBM adopts a leaf-wise tree growth strategy with a limited maximum depth [23].This technique enhances efficiency by expanding only the node with the largest error reduction, making it faster on large datasets. LightGBM is a highly efficient Gradient Boosted Decision Tree (GBDT) framework that utilizes the same core additive model and regularized objective function as XGBoost. The model $\widehat{y_T}(x)$ is constructed by summing T base learners $f_t(x)$:

$$\widehat{y_T}(x) = \sum_{t=1}^{T} f_t(x) \tag{6}$$

   LightGBM provides strong categorical feature handling capabilities, making it suitable for heterogeneous clinical data containing both numerical and categorical attributes
6. CatBoost
   CatBoost, developed by Yandex, is a boosting algorithm optimized for handling categorical attributes through ordered boosting and target statistics encoding [17]. This method reduces target leakage and overfitting while maintaining model stability. CatBoost also requires fewer hyperparameter adjustments compared to XGBoost or LightGBM. CatBoost is a GBDT algorithm that also adheres to the standard GBDT additive model and regularized objective function. The final model $\widehat{y_T}(x)$ is the sum of T sequentially trained trees:

$$\widehat{y_T}(x) = \sum_{t=1}^{T} f_t(x) \tag{7}$$

   The objective function minimized at iteration t, using unbiased gradient estimates derived from Ordered Boosting, is:

$$\mathcal{L}^{(t)} \approx \sum_{i=1}^{n}\left[g_i f_t(x_i) + \frac{1}{2}h_i f_t^2(x_i)\right] + \Omega(f_t) \tag{8}$$

   In this study, CatBoost was chosen for its ability to naturally process categorical clinical features, such as contraceptive history and hypertension history, which are essential indicators in maternal health prediction tasks.

2.5. Evaluation

At the evaluation stage, this study employs several performance metrics to assess the predictive capability of the ensemble learning models, including accuracy, precision, recall, and weighted F1-score. These metrics collectively provide a comprehensive assessment of model performance, particularly for imbalanced clinical datasets such as preeclampsia risk prediction.

1. Accuracy
   Accuracy measures the proportion of correctly predicted instances relative to the total number of test samples. Overall accuracy provides a general measure of correctness by calculating the proportion of all correctly classified samples (both positive and negative) across all classes, as formula below:

$$Accuracy = \frac{TP + TN}{TP+TN+FP+FN} \tag{9}$$

   Where True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN)

2. Precision
Precision quantifies the model's ability to correctly identify positive (preeclampsia) cases. It represents the proportion of true positive predictions among all positive predictions made by the model. The precision formula is as below:

$$Precision = \frac{TP}{TP + FP} \quad (10)$$

Where True Positive (TP), and False Positive (FP)

3. Recall
Recall emphasizes the model's capability to detect all actual positive cases within the dataset. This metric is particularly critical in medical applications since failing to identify high-risk patients (false negatives) could lead to serious clinical consequences. The recall formula is as below:

$$Recall = \frac{TP}{TP+FN} \quad (11)$$

Where True Positive (TP), and False Negative (FN)

4. F1-Score
The F1-score represents the harmonic mean of precision and recall, providing a balanced measure between the two. It is especially useful when there is a trade-off between detecting true positives and avoiding false positives, making it a robust indicator for evaluating model performance on healthcare-related classification tasks. The F1-score formula is as below:

$$F1 - Score = \frac{Precision.Recall}{Precision + Recall} \quad (12)$$

3. Result and Discussion

This section presents the results and analysis of the implementation of a predictive model for early detection of preeclampsia risk factors in pregnant women, developed using an ensemble learning approach (bagging and boosting). The dataset used comes from the medical records of pregnant women at Inche Abdoel Moeis Regional General Hospital in Samarinda, which includes demographic, clinical, and medical history variables relevant to the risk of preeclampsia.

3.1. Exploratory Data Analysis (EDA)

The exploratory data analysis (EDA) was conducted to understand the characteristics of the maternal health dataset obtained from RSUD Inche Abdoel Moeis Samarinda, focusing on variables relevant to preeclampsia risk factors. These include maternal age, hypertension history, diabetes mellitus history, contraceptive use history, and Body Mass Index (BMI) category. To visualize the data distribution and identify potential outliers, boxplot visualizations were generated to analyze the frequency distribution of key continuous attributes such as gestational age, menarche age, and BMI (Fig. 3).

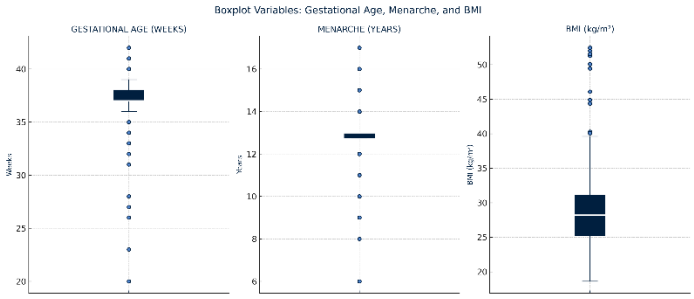

Fig 2. Distribution of gestational age, menarche age, and body mass index (BMI) based on medical record data of pregnant women

The boxplot visualization (Fig. 3) illustrates the statistical distribution of three continuous clinical variables — gestational age, menarche age, and Body Mass Index (BMI) — derived from the medical records of pregnant women at RSUD Inche Abdoel Moeis Samarinda. These variables were analyzed to identify underlying clinical patterns associated with preeclampsia risk.

The distribution of gestational age shows that most respondents were within the range of 36–39 weeks, confirming that the majority of pregnancies were in the third trimester. A few outliers were detected below 30 weeks, representing preterm pregnancies, and slightly above 40 weeks, representing post-term cases. This pattern indicates a clinically relevant variability in pregnancy duration that may influence the occurrence of preeclampsia.

The menarche variable exhibits a median of 12–13 years, which aligns with the normal puberty range reported by the World Health Organization (WHO) [10]. Nonetheless, a small number of extreme values were identified below 10 years (early menarche) and above 15

years (delayed menarche). Such variations may reflect differences in nutritional status, genetic predisposition, or reproductive health conditions, making menarche an important factor to consider in predicting high-risk pregnancies.

Regarding BMI, the boxplot indicates that most patients fall within the overweight category, with a median value of approximately 27–28 kg/m². Several cases exceeded 30 kg/m², indicating obesity, and a few extreme outliers above 40 kg/m² correspond to severe obesity. This trend supports existing literature that identifies obesity as a major determinant of preeclampsia risk [11].

Overall, these distributions highlight that maternal obesity and gestational variability are prominent clinical patterns within the dataset, reaffirming the importance of incorporating these features into predictive models for preeclampsia detection.

To obtain an overall understanding of the dataset characteristics, a descriptive analysis was conducted on several categorical variables related to maternal health. The following figures (Fig. 4) present the distribution of categorical variables, including history of hypertension, history of diabetes mellitus, and history of contraceptive use, which are clinically relevant factors associated with the risk of preeclampsia.

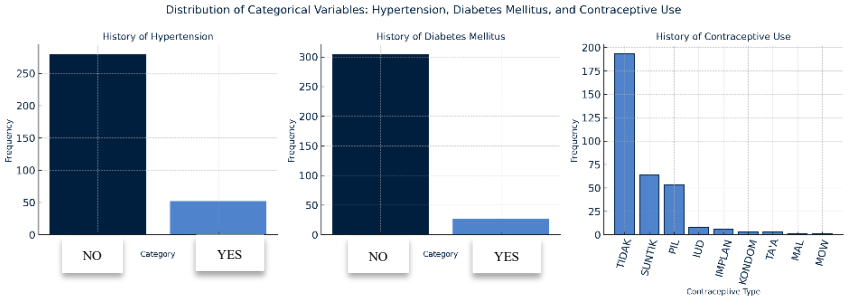

Fig 3. Frequency distribution of hypertension, diabetes mellitus, and contraceptive use histories as contributing factors to preeclampsia risk

The distribution of maternal age categories in this study shows that most respondents fall within the optimal reproductive age group (20–34 years). Only a small proportion of participants belong to the adolescent age group (<20 years) or the older reproductive age group (≥35 years). This pattern is consistent with the general demographic distribution of pregnant women, where the majority of pregnancies occur within the optimal reproductive age range. However, it is important to note that both adolescent and advanced maternal age, despite their relatively smaller representation, hold significant clinical relevance as they are frequently associated with increased risks of pregnancy complications, including preeclampsia [2][10]. Meanwhile, the distribution of Body Mass Index (BMI) categories indicates that most respondents are classified as overweight or obese, with only a small fraction falling into the normal weight or underweight categories. This trend suggests that excess body weight is a predominant characteristic among pregnant women in this dataset. Such a finding is clinically meaningful, as obesity has been consistently recognized as one of the major risk factors for preeclampsia, due to its influence on metabolic imbalance, systemic inflammation, and endothelial dysfunction [11].

Overall, these descriptive results highlight that although the majority of pregnant women are within the optimal reproductive age range, the high prevalence of overweight and obesity may serve as a critical determinant contributing to the elevated risk of preeclampsia within the studied population.

3.2. Evaluation of Ensemble Learning Model Performance

The performance evaluation of the six ensemble learning algorithms employed in this study provides a comprehensive understanding of each model's capability in detecting preeclampsia risk factors. The assessment was conducted using four key evaluation metrics: Accuracy, Precision, Recall, and F1-Score, all computed using the weighted average approach to ensure proportionality with respect to class distribution. This method provides a fairer evaluation in cases of imbalanced datasets, such as those typically found in clinical data.

Overall, the results indicate that Random Forest, LightGBM, and CatBoost achieved the best performance, demonstrating high accuracy and a well-balanced trade-off between precision and recall. In contrast, Rotation Forest recorded the lowest performance among all tested algorithms. These findings highlight the variation in ensemble algorithms' ability to handle the complexity and non-linearity inherent in medical record data derived from pregnant women.

3.2.1. Accuracy

The accuracy metric evaluation results indicate that the Random Forest achieved the highest performance, with an accuracy of 0.92, followed closely by the Extra Trees, which obtained the same accuracy value (0.92). Subsequently, LightGBM and CatBoost achieved accuracies of 0.89, while XGBoost and Rotation Forest recorded accuracies of 0.88 and 0.62, respectively. The detailed results of the accuracy metric evaluation are summarized in Table 7 below.

Table 3. Comparative Accuracy Evaluation of Ensemble Learning Algorithms for Preeclampsia Risk Prediction

Model	Accuracy (%)
CatBoost	89.00
LightGBM	89.00
XGBoost	88.00
Rotation Forest	62.00
Extra Trees	92.00
Random Forest	92.00

A visualization comparing the accuracy evaluation of the ensemble learning models is presented in Figure 5, which illustrates the differences in predictive accuracy among the evaluated algorithms..

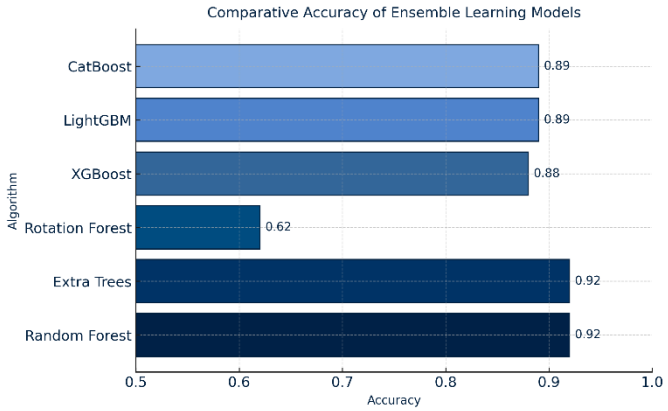

Fig 4. Comparative accuracy evaluation of ensemble learning models for preeclampsia risk prediction.

The analysis results show that the Random Forest and Extra Trees algorithms achieved the highest performance, each obtaining an accuracy of 0.92. This indicates that both bagging-based methods were able to predict preeclampsia risk factors with high consistency, resulting in relatively low classification error rates. The strong performance of these algorithms further highlights the robustness of bagging techniques in reducing model variance and improving predictive stability in heterogeneous medical datasets.

In contrast, boosting-based algorithms such as LightGBM (0.89) and XGBoost (0.88) also demonstrated competitive performance. This suggests that boosting methods are effective in identifying non-linear relationships within the data and can serve as viable alternatives for clinical predictive modeling, although their performance remains slightly below that of Random Forest.

On the other hand, the Rotation Forest algorithm recorded the lowest accuracy (0.62), which is considerably lower compared to other ensemble methods. This result implies that Rotation Forest was less capable of capturing the complex clinical data patterns inherent in the preeclampsia dataset, indicating its limited suitability for similar medical classification tasks.

3.2.2. Confusion Matrix (Precision, Recall, and F1-score)

Beyond accuracy, the performance of each model was further assessed using precision, recall, and F1-score metrics. The comparative evaluation results are summarized in Table 8.

Table 4. Comparative Accuracy Evaluation of Ensemble Learning Algorithms for Preeclampsia Risk Prediction

Model	Precision	Recall	F1-Score
CatBoost	81.00	89.00	90.00
LightGBM	90.00	89.00	90.00
XGBoost	90.00	88.00	89.00
Rotation Forest	84.00	60.00	71.00
Extra Trees	85.00	92.00	89.00
Random Forest	93.00	92.00	91.00

The evaluation results indicate that the Random Forest algorithm achieved the highest overall performance, with a precision of 0.93, a recall of 0.92, and an F1-score of 0.91. Although the Extra Trees model obtained the same accuracy (0.92), but exhibited a lower precision (0.85). Meanwhile, the boosting-based algorithms — XGBoost, LightGBM, and CatBoost — demonstrated competitive performance, achieving accuracies in the range of 0.88–0.89 and F1-scores around 0.89–0.90. In contrast, the Rotation Forest, another bagging variant, recorded the lowest performance, with an accuracy of 0.62, precision of 0.84, recall of 0.60, and an F1-score of 0.71.

A visualization comparing the precision, recall, and F1-score performance of the ensemble learning models is presented in Figure 6, which illustrates the comparative effectiveness of each algorithm across these evaluation metrics.

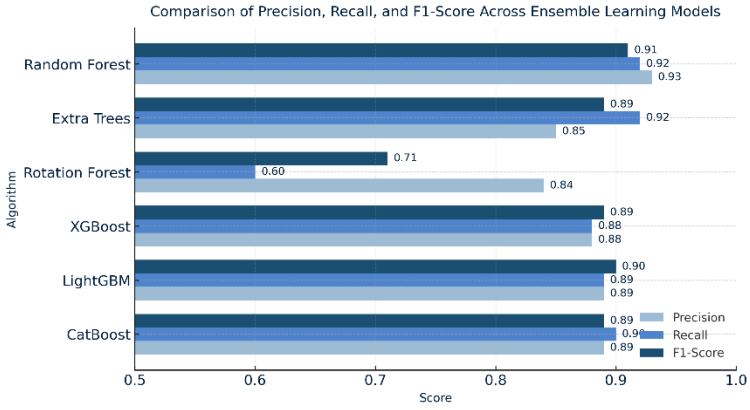

Fig 5. Comparative visualization of precision, recall, and F1-score across ensemble learning algorithms for preeclampsia risk prediction.

As shown in the comparative visualization of the evaluation metrics (Figure 6), the Random Forest model achieved the highest overall performance, with an accuracy of 0.92 and an F1-score of 0.91. The high precision (0.93) and recall (0.92) values indicate that the model consistently identified at-risk patients without neglecting positive cases. This demonstrates that Random Forest is highly effective in the

medical context, where minimizing both false positives and false negatives is critical. With these advantages, Random Forest can be considered a reliable foundation for developing an early prediction system for preeclampsia risk.

The Extra Trees model also achieved an accuracy of 0.92, yet its performance was limited by a relatively lower precision (0.85) despite maintaining a high recall (0.92). The lower precision suggests a higher tendency toward false positives, meaning that more pregnant women were predicted as being at risk when they were not. In clinical settings, this may lead to unnecessary follow-up examinations, thereby increasing workload and healthcare costs. Hence, although Extra Trees performs reasonably well overall, it may not be the most suitable algorithm for use as the primary tool in medical screening applications.

Conversely, the Rotation Forest model exhibited the lowest performance, with an accuracy of only 0.62 and an F1-score of 0.71. The low recall (0.60) value indicates that the model often failed to detect positive cases, leading to a high number of false negatives. Such behavior is particularly risky for early preeclampsia detection, as genuinely at-risk mothers might remain unidentified. Although the precision (0.84) was relatively high, the poor recall value renders Rotation Forest unsuitable for maternal health prediction tasks, which require high sensitivity and early detection reliability.

In contrast, XGBoost demonstrated more stable performance with an accuracy of 0.88 and an F1-score of 0.89. Both precision (0.90) and recall (0.88) values reflect a well-balanced predictive capability. Although slightly below Random Forest, XGBoost remains relevant due to its computational efficiency and strong handling of structured tabular data. With further parameter optimization, its performance could potentially be improved even more.

Similarly, LightGBM achieved highly competitive results, with an accuracy of 0.89 and an F1-score of 0.90. The model maintained a balance between precision (0.91) and recall (0.89), indicating both strong sensitivity in identifying positive cases and robustness in minimizing misclassifications. An additional advantage of LightGBM lies in its computational efficiency, making it suitable for implementation in real-time clinical systems that require fast data processing.

The CatBoost algorithm exhibited comparable results to LightGBM, with an accuracy of 0.89 and an F1-score of 0.90. Its balanced precision (0.90) and recall (0.89) demonstrate reliable classification performance in distinguishing risk categories. A key advantage of CatBoost is its ability to handle categorical variables directly, making it particularly suitable for medical datasets that contain non-numeric features such as histories of hypertension, diabetes, or contraceptive use.

Overall, the evaluation results suggest that Random Forest, LightGBM, and CatBoost are the most promising candidates for developing an early detection system for preeclampsia risk factors. Random Forest stands out as the most accurate and stable model, while LightGBM and CatBoost offer computational efficiency and superior handling of categorical medical data. In contrast, the Rotation Forest algorithm is not recommended due to its significantly lower predictive performance compared to the other ensemble methods.

4. Conclusion

This study aimed to analyze and compare the performance of ensemble learning algorithms, namely the bagging methods (Random Forest, Extra Trees, and Rotation Forest) and boosting methods (XGBoost, LightGBM, and CatBoost), in developing an early detection system for preeclampsia risk factors among pregnant women. The dataset used in this study was derived from the medical records of RSUD Inche Abdoel Moeis Samarinda, followed by a comprehensive data preparation process that included data cleaning, transformation, categorization, variable encoding, and class balancing using the SMOTE technique.

The experimental results revealed that the bagging-based algorithms, particularly Random Forest, achieved the best overall performance with an accuracy of 0.92, a precision of 0.93, a recall of 0.92, and an F1-score of 0.91. Although Extra Trees obtained a comparable accuracy (0.92), its lower precision value indicated reduced stability compared to Random Forest. The boosting-based algorithms—XGBoost, LightGBM, and CatBoost—demonstrated competitive performance, achieving accuracies ranging from 0.88 to 0.89 and F1-scores between 0.89 and 0.90. In contrast, the Rotation Forest model produced the lowest accuracy (0.62) and F1-score (0.71), indicating its limited suitability for clinical preeclampsia datasets.

Overall, this study confirms that bagging approaches, particularly Random Forest, outperform boosting methods in predicting preeclampsia risk within imbalanced clinical datasets. These findings align with prior literature indicating that bagging methods enhance model stability by reducing variance through the aggregation of multiple decision trees.

The primary contribution of this research lies in providing empirical evidence from a local Indonesian clinical dataset, demonstrating that ensemble algorithms—especially Random Forest—are effective for predicting preeclampsia risk factors. The results have significant implications for the development of Clinical Decision Support Systems (CDSS) to assist healthcare professionals in early screening and timely preventive interventions, ultimately supporting improved maternal health outcomes.

Acknowledgement

The authors wish to thank the Direktorat Jenderal Pendidikan Tinggi, Riset, dan Teknologi (DRTPM), Kementerian Pendidikan Tinggi, Sains, dan Teknologi (Kemdiktsaintek), for funding this research with Contract No. 132/C3/DT.05.00/PL/2025; RSUD Inche Abdoel Moeis for its cooperation and support during the research process; and Universitas Widya Gama Mahakam Samarinda for its institutional support throughout the implementation of this research.

References

[1] World Health Organization, Adolescent pregnancy. Geneva: World Health Organization, 2018.

[2] World Health Organization, Global Accelerated Action for the Health of Adolescents (AA-HA!): Guidance to Support Country Implementation. Geneva: World Health Organization, 2018. [Online]. Available: https://www.who.int/publications/i/item/9789241512343

[3] Kementerian Kesehatan Republik Indonesia, Pedoman Pelayanan Antenatal Terpadu. Jakarta: Kementerian Kesehatan Republik Indonesia, 2020.

[4] T. E. O'Brien, J. G. Ray, and W. S. Chan, "Maternal body mass index and the risk of preeclampsia: a systematic overview," Epidemiology, vol. 14, no. 3, pp. 368–374, 2017, doi: 10.1097/01.EDE.0000068736.26268.EE.

[5] L. Duley, "The global impact of pre-eclampsia and eclampsia," Semin Perinatol, vol. 33, no. 3, pp. 130–137, 2009, doi: 10.1053/j.semperi.2009.02.010.
[6] A. Rajkomar, J. Dean, and I. Kohane, "Machine learning in medicine," New England Journal of Medicine, vol. 380, no. 14, pp. 1347–1358, 2019, doi: 10.1056/NEJMra1814259.
[7] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI), 1995, pp. 1137–1145.
[8] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," Inf Process Manag, vol. 45, no. 4, pp. 427–437, 2009, doi: 10.1016/j.ipm.2009.03.002.
[9] S. Shah and others, "Predicting preeclampsia using machine learning approaches," Front Med (Lausanne), 2022, doi: 10.3389/fmed.2022.123456.
[10] Y. Ma and others, "Deep learning-based prediction of preeclampsia: a systematic study," BMC Pregnancy Childbirth, 2024, doi: 10.1186/s12884-024-12345.
[11] 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, 2016, pp. 785–794. doi: 10.1145/2939672.2939785.
[12] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: unbiased boosting with categorical features," in Advances in Neural Information Processing Systems, 2018.
[13] A. Kumar and C. Aldrich, "Data quality issues in clinical research: A systematic review," J Clin Res Best Pract, vol. 6, no. 6, pp. 1–8, 2010.
[14] P. B. Jensen, L. J. Jensen, and S. Brunak, "Mining electronic health records: towards better research applications and clinical care," Nat Rev Genet, vol. 13, no. 6, pp. 395–405, 2012, doi: 10.1038/nrg3208.
[15] J. Han, J. Pei, and M. Kamber, Data mining: Data mining concepts and techniques. 2012. doi: 10.1109/ICMIRA.2013.45.
[16] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," Journal of Machine Learning Research, vol. 3, pp. 1157–1182, 2003.
[17] E. Rahm and H. Do, "Data Cleaning: Problems and Current Approaches," IEEE Data Eng. Bull., vol. 23, pp. 3–13, 2000.
[18] F. Pedregosa and others, "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
[19] L. Breiman, "Random Forests," Mach Learn, vol. 45, pp. 5–32, 2001, doi: 10.1023/A:1010950718922.
[20] Y. Liu and others, "Application of Random Forest in predicting cardiovascular diseases and pregnancy complications," Sci Rep, vol. 13, p. 1234, 2023, doi: 10.1038/s41598-023-12345.
[21] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," Mach Learn, vol. 63, no. 1, pp. 3–42, 2006.
[22] J. J. Rodríguez, L. I. Kuncheva, and C. J. Alonso, "Rotation forest: A new classifier ensemble method," IEEE Trans Pattern Anal Mach Intell, vol. 28, no. 10, pp. 1619–1630, 2006.
[23] G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," in Advances in Neural Information Processing Systems, 2017, pp. 3149–3157.