Kesmas

Volume 20
Issue 2 *May 2025*

Article 8

5-30-2025

Machine Learning for Preeclampsia Prediction: Enhancing Screening in Primary Health Care

Dwirani Amelia
University of Indonesia, Jakarta, Indonesia, dwiraniamelia@gmail.com

Asri Adisasmita
Universitas Indonesia, Depok, aadisasmita@gmail.com

Kemal N. Siregar
Universitas Indonesia, Depok, nazarudin.kemal51@gmail.com

Detty Siti Nurdiati
Universitas Gadjah Mada, Yogyakarta, detty@ugm.ac.id

Recommended Citation

Amelia D , Adisasmita A , Siregar KN , et al. Machine Learning for Preeclampsia Prediction: Enhancing Screening in Primary Health Care. *Kesmas*. 2025; 20(2): 147-156
DOI: 10.7454/kesmas.v20i2.2243
Available at: https://scholarhub.ui.ac.id/kesmas/vol20/iss2/8

Amelia *et al. Kesmas: Jurnal Kesehatan Masyarakat Nasional (National Public Health Journal)*. 2025; 20 (2): 147-156
DOI: 10.7454/kesmas.v20i2.2243

Kesmas: Jurnal Kesehatan Masyarakat Nasional (National Public Health Journal)

Machine Learning for Preeclampsia Prediction: Enhancing Screening in Primary Health Care

Dwirani Amelia[1,2]*, Asri Adisasmita[3], Kemal N Siregar[4], Detty Siti Nurdiati[5]

[1]Doctoral Program, Department of Epidemiology, Faculty of Public Health, Universitas Indonesia, Depok, Indonesia
[2]Budi Kemuliaan Health Institute, Jakarta, Indonesia
[3]Department of Epidemiology, Faculty of Public Health, Universitas Indonesia, Depok, Indonesia
[4]Department of Biostatistics and Population Studies, Faculty of Public Health, Universitas Indonesia, Depok, Indonesia
[5]Department of Obstetrics and Gynecology, Faculty of Medicine, Public Health and Nursing, Universitas Gadjah Mada, Dr. Sardjito Hospital, Yogyakarta, Indonesia

Abstract

Preeclampsia is a leading cause of maternal morbidity and mortality worldwide, with early detection being critical for reducing adverse outcomes. This study aimed to develop a machine learning model for predicting the risk of preeclampsia using readily available maternal characteristics such as body mass index, mean arterial pressure, and clinical history of hypertension or diabetes mellitus. Secondary data from 2,250 pregnancies were analyzed, addressing challenges such as missing data and class imbalance through preprocessing. Various algorithms, including support vector machines, random forest, and logistic regression, were evaluated. Herein, a support vector machines model with threshold adjustment showed the best performance, with a sensitivity of 67.5%, specificity of 57.23%, and an area under the curve of 0.68. These findings indicated the promising potential of scalable and interpretable prediction models for enhancing preeclampsia screening in primary health care settings. However, further refinement and validation of the proposed model are required for broader clinical integration to improve maternal and neonatal health outcomes.

Keywords: machine learning, maternal health, preeclampsia prediction, primary care screening

Introduction

Preeclampsia (PE) is a multifactorial syndrome and a leading cause of maternal morbidity and mortality globally.[1] Pregnancy-induced hypertension causes one-fifth of maternal deaths worldwide, and PE/eclampsia alone is estimated to cause 60,000 to 80,000 maternal deaths annually.[2,3] PE is characterized by elevated blood pressure and multisystem manifestations, and its etiology is poorly understood. The prevalence and incidence of PE vary globally, with the global incidence rate of severe PE estimated between 2% and 10% of all pregnancies.[4-6] Later evidence showed an increase in the incidence of hypertension during pregnancy over time, which included PE.[7]

Early detection is critical for preventing the morbidity and mortality associated with PE. Preventive interventions, such as low-dose aspirin and calcium supplementation, have been found to reduce adverse outcomes.[6,8] As its etiology, pathogenesis, and pathophysiology of PE are not fully understood, PE is challenging to predict.[6,9,10] Previously, PE prediction based on combinations of multiple features has been actively investigated. In a meta-analysis published in 2015, models based on a combination of maternal characteristics and several biomarkers have been found to exhibit better predictive performance than models based on individual biomarkers.[11]

The advantages of biochemical and biophysical marker examination with Doppler ultrasound have been demonstrated in many studies, but with limited markers practically used in clinical practice.[12,13] However, obtaining biochemical markers is often expensive, thus limiting their applicability in primary health care (PHC), where women are first encountered for antenatal care (ANC). At the same time, models based on several simple features, such as body mass index (BMI) and other routinely collected maternal characteristics such as age, ethnicity, parity, history of hypertension, and history of PE, have shown promising predictive performance.[14-16]

***Correspondence**:* *Dwirani Amelia, Doctoral Program, Department of Epidemiology, Faculty of Public Health, Universitas Indonesia, Depok, Indonesia, Email: dwiraniamelia@gmail.com.*

Received : December 28, 2024
Accepted : May 5, 2025
Published: May 30, 2025

Machine learning (ML) is a novel approach to PE prediction by analyzing patterns in complex datasets without relying on explicit causal mechanisms.[17] Previous studies have leveraged prior work, aiming to develop prediction models that would be both effective and accessible for primary care.[14,16-18] Particularly, a study showed that a model based on biophysical (mean uterine artery pulsatility index) and biochemical markers (placental growth factor) is clinically superior to a simple feature model, but the difference is not statistically significant.[14]

This study aimed to develop an ML model for predicting PE based on readily available limited maternal clinical characteristics such as maternal history of primary hypertension and diabetes mellitus (DM), family history of hypertension and DM, primigravidity, mean arterial pressure (MAP), BMI, and history of smoking, all of which can be gathered in primary care settings. Herein, the authors evaluated the predictive performance (sensitivity, specificity, and area under the curve (AUC)) of various models and addressed the challenges associated with applying the developed prediction tools in low-resource environments. The concise and practical approach is expected to support early surveillance and prevention of PE, improving maternal and neonatal outcomes in PHC.

Method

This retrospective cohort study used secondary data on 2,250 pregnant women visiting a private Mother and Child Hospital and its branch in Jakarta from July 2012 to April 2015. The data were collected from a previous study by Savitry et al. from the Julius Center for Health Sciences and Primary Care, Julius Global Health, University Medical Center Utrecht, Utrecht, Netherlands.[19,20] Anonymous data from routine hospital checks were used; thus, this study was ethically approved as nonhuman research. All the 2,250 pregnant women whose data were collected and followed up until delivery. This analysis was a part of the Prediction Modeling of PE in Pregnant Women using an ML study conducted by the author during the doctorate program at the Faculty of Public Health, Universitas Indonesia, in 2024.

All the collected data were preprocessed, during which 584 entries were identified as null and 1,028 as duplicates. The results of Little's Missing Completely at Random (MCAR) test (p-value >0.05) showed that all missing data were missing completely at random. Thus, all the missing values and duplicates were removed from the dataset, leaving 638 data entries (Figure 1). The available features are shown in Table 1.

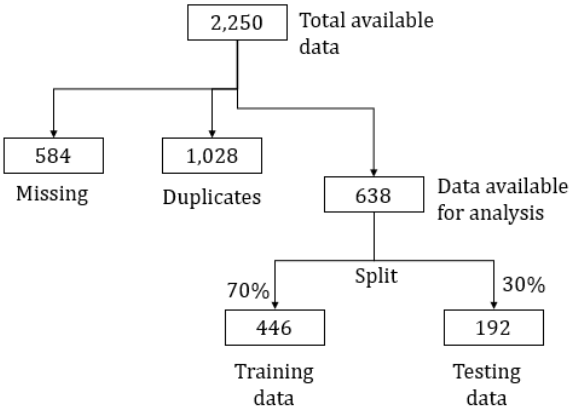

Figure 1. Flow Chart of Data Preprocessing

Data were collected during routine ANC visits by midwives, including maternal history (hypertension, DM smoking, and the family history of hypertension/DM), pre-pregnancy weight, obstetric history (gravida and miscarriage), and height (measured at the first ANC visit). Blood pressure was recorded every visit, with mean arterial pressure calculated for the first and second trimesters. Pre-pregnancy BMI was calculated from weight (kg) and height (m). PE was defined according to the International Society for the Study of Hypertension in Pregnancy (ISSHP) criteria: gestational hypertension (≥20 weeks) with ≥1 new-onset condition (proteinuria, organ dysfunction, etc.).[1]

Using Kelsey's formula, the calculated sample size for each group should be 683 subjects to achieve a statistical power of 80%.[21] In addition, the rule of 10 or 20 for the number of observations and features was considered. A common recommendation is that at least 10 events per predictor should be used to prevent overfitting and ensure robust training.[22,23] This recommendation on sample size calculation surely requires further refinement through feature selection and dimensionality reduction by identifying the best-performing features and generating synthetic data because of the lack of systematic validation.[24] Based on both considerations, all of the collected 2,250 data entries were included in the analysis.

Data analysis was conducted in Python (permissive open source license version 3.9.7) using the scikit-learn library (BSD3 license version 1.5.0) as the primary framework for ML and statistical modeling. The analysis involved several key steps, including data preprocessing, feature selection, model selection, resampling (under sampling or oversampling), hyperparameter tuning, model evaluation, model implementation, and the interpretation of results.[8,16,17] Data preprocessing included the handling of missing values, the removal of duplicates, and normalization or standardization to ensure comparability across features. Ordinal scaling and one-hot encoding were used for categorical variables. The missing mechanism was first investigated when handling the missing data. Specifically, Little's MCAR test was performed to determine whether the missing values were missing completely at random.[25] The test showed that the missing data were missing completely at random; thus, the missing data were removed.

Feature selection was performed based on a feature correlation matrix. When the correlation matrix did not adequately show a strong association (correlation coefficient <0.3), features were selected using domain knowledge. A literature review was conducted to assess the association between the features and the outcome. Several ML algorithms were considered herein: support vector machines (SVM), logistic regression, random forest (RF), decision tree, and K-nearest neighbors (KNN). The models were chosen based on their suitability for the dataset and research objectives. To optimize the model performance, hyperparameter tuning was performed using grid search with cross-validation. The dataset was split into training and testing sets (70:30 split). The model performance was evaluated on the test set, with cross-validation employed to ensure the robustness of the results. Accuracy, recall positive, specificity positive, precision positive, F-1 score weight, and the area under the receiver operating characteristic curve (AUC-ROC) were used as performance metrics to identify the best-performing model.

The analysis was implemented using scikit-learn alongside supporting libraries such as pandas for data manipulation, NumPy for numerical computations, and matplotlib/seaborn for data visualization. Custom scripts and functions were developed to streamline the analysis and ensure reproducibility. The predictive performance of the selected models was analyzed, and its implications were interpreted in the context of the research question. To improve the predictive performance of the model, the classification threshold was adjusted from the standard 0.5 to a value that balances sensitivity and specificity. This adjustment, guided by Youden's Index, helped optimize the overall performance of the model.[26] Insights derived from feature importance and model outputs were integrated into the broader discussion of the findings. In addition, the SHapley Additive exPlanation (SHAP) method was applied to some data points to determine the contribution of each studied feature.

Results

All variables representing risk factors in the study population (n = 638) are provided in Table 1. Key variables included DM, hypertension, family history, smoking status, BMI, and the prevalence of PE (16.45%). Data imbalance is observed in most of the variables in the study population, which could introduce bias. Notably, DM (1.73%), hypertension (9.88%), and obesity (11.92%) have been identified as strong predictors in many studies.[4,5,6,8,13]

The data distribution indicated a substantial imbalance between the two groups, with 533 cases of no PE and 105 cases of PE. These results suggested the need for resampling during data analysis to mitigate potential bias and ensure robust model performance. Several ML algorithms were evaluated with and without resampling techniques (oversampling and under sampling) to identify the algorithm that best addresses the research objectives.

Table 1. Variable in the Study Population

Feature	No n (%)	Yes n (%)
Type 2 diabetes mellitus	627 (98.27)	11 (1.73)
Hypertension	575 (90.12)	63 (9.88)
Family history of diabetes mellitus	458 (71.78)	180 (28.22)
Family history of hypertension	351 (55.01)	287 (44.88)
Active Smoking	355 (55.64)	283 (44.35)
Primigravidity	399 (62.54)	239 (37.46)
Mean Arterial Pressure in the 1st trimester ≥90 mmHg	409 (64.11)	229 (35.89)
Body Mass Index underweight	351 (55.02)	287 (44.98)
Body Mass Index normal weight	533 (83.54)	105 (16.45)
Body Mass Index overweight	468 (73.35)	170 (26.65)
Body Mass Index obese	562 (88.08)	76 (11.92)
Preeclampsia	533 (83.54)	105 (16.45)

Table 2. Performance Metrics of the Models Without/with Resampling

2. a. Models Without Resampling						
MODEL	**Accuracy (%)**	**Recall Positive (%)**	**Specificity Positive (%)**	**Precision Positive (%)**	**F1-score Weight (%)**	**AUC**
RF	79.17	10.00	97.37	50.00	73.21	0.62
Logistic regression	79.17	0.00	100.00	0.00	69.96	0.49
SVM	79.17	0.00	100.00	0.00	69.96	0.59
Decision tree	79.17	0.00	100.00	0.00	69.96	0.55
KNN	78.65	12.50	96.05	45.45	73.50	0.57
2. b. Models with Oversampling						
MODEL	**Accuracy (%)**	**Recall Positive (%)**	**Specificity Positive (%)**	**Precision Positive (%)**	**F1-score Weight (%)**	**AUC**
RF	67.19	35.00	75.66	27.45	68.55	0.56
Logistic regression	63.02	52.50	65.79	28.77	66.17	0.66
SVM	48.44	85.00	38.82	26.77	51.53	0.66
Decision tree	65.63	20.00	77.63	19.05	65.93	0.48
KNN	54.17	27.50	61.18	15.71	57.91	0.36
2. c. Models with Under Sampling						
MODEL	**Accuracy (%)**	**Recall Positive (%)**	**Specificity Positive (%)**	**Precision Positive (%)**	**F1-score weight (%)**	**AUC**
Random forest	58.33	47.50	61.18	24.36	62.07	0.62
Logistic regression	63.02	55.00	65.13	29.33	66.24	0.65
SVM	70.31	35.00	79.61	31.11	70.94	0.68
Decision tree	53.13	80.00	46.05	28.07	56.85	0.65
KNN	60.94	47.50	64.47	26.03	64.26	0.57

Notes: AUC = area under the curve, RF = random forest, SVM =support vector machine, KNN = K-nearest neighbor

In the initial analysis conducted on the cleaned dataset without resampling, all models showed poor performance (sensitivity and specificity) in PE prediction. In addition, all models showed low discriminatory power equivalent to random guessing (AUC <0.5). In this analysis, RF and KNN showed better sensitivity (10.00% and 12.50%, respectively) and specificity (97.37% and 96.05%, respectively) than the other models (Table 2.a).

With oversampling, the RF classifier showed the highest accuracy (0.67) among the evaluated models (Table 2.b). This approach addressed some issues observed in the logistic regression model, particularly in classifying positive cases. Hyperparameter tuning was performed using HalvingGridSearchCV (Successive Halving Grid Search Cross-Validation) to optimize the model configuration. However, even after hyperparameter tuning, the RF classifier still showed limited sensitivity for positive cases and suboptimal overall performance, with an AUC of 0.56 and recall positive of 0.35

Under sampling was further employed to balance the class distribution and improve predictive performance. Specifically, under sampling was performed to reduce the dominance of the negative class and improve the detection of positive cases. With under sampling, SVM showed the best performance, achieving an accuracy of 70.31% (Table 2.c). The ROC curve for this SVM model exhibited the trade-off between sensitivity and specificity across various prediction thresholds. An AUC of 0.68 indicated that the model outperforms random guessing (AUC = 0.50); however, its overall performance was still suboptimal, as an ideal model would have an AUC close to 1.0.

The ROC curves and AUC values for the considered models are shown in Figure 2. The ROC curves showed the predictive performance of the five models (SVM, logistic regression, RF, decision tree, and KNN) in binary classification, with AUC values indicating predictive accuracy (compared to the diagonal line representing random guessing, AUC = 0.5). SVM showed the best predictive performance (AUC = 0.68), and KNN was the worst (AUC = 0.57).

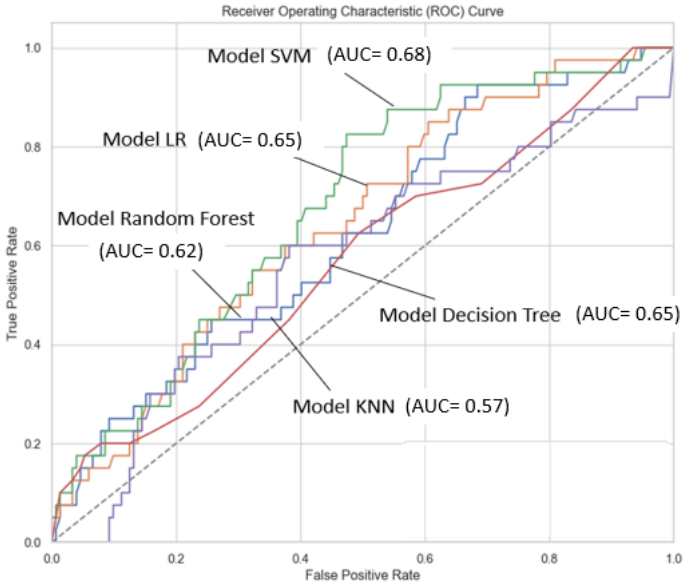

Figure 2. Comparison of Receiver Operating Characteristic Curve Among Different Algorithms
Notes: SVM = support vector machine, LR = logistic regression, KNN = K-nearest neighbor, AUC = area under the curve

Further analysis was conducted to improve the predictive performance of the SVM model. The SVM model, with an accuracy of 59.37%, showed a sensitivity of 67.5%, a specificity of 57.23%, and an AUC of 67%, indicating moderate discriminatory power compared to other algorithms (Table 3). A low precision (high number of false positives) indicated problems with the compensation data and the limitations of feature engineering.

Table 3. Performance Metrics of the Considered Models with Under Sampling and a Classification Threshold of 0.43

MODEL	Accuracy (%)	Recall Positive (%)	Specificity Positive (%)	Precision Positive (%)	AUC	Best YI
RF	48.96	75.00	42.00	25.42	0.62	0.17
Logistic regression	54.17	67.50	51.00	26.47	0.65	0.18
SVM*	59.37	67.50	57.23	29.34	0.67	0.24
Decision tree	53.13	80.00	46.05	28.07	0.65	0.26
KNN	53.64	62.50	51.31	25.25	0.56	0.13

Notes: AUC = area under the curve, YI = Youden's Index, RF = random forest, SVM = support vector machine, KNN = K-nearest neighbor
*SVM is the best-performing model on the employed dataset; furthermore, it performs better than the models from a systematic review by Ranjbar et al., who reported a recall of 0.420 and accuracy of 0.740 for RF, and a recall of 0.789 and precision of 0.447 for EGBM (Extreme gradient boosting model) using maternal characteristics similar to those employed herein and biochemical and biophysical markers.[16]

In addition, the SHAP analysis on 100 data points showed that mean arterial pressure in the first trimester has the greatest impact on outcomes, with smoking, history of hypertension and DM, and family history of hypertension and DM showing a lower impact. A case illustration was provided involving a 17 -17-year-old primigravida with a MAP of 92 mmHg and a BMI of 26. Overall, the subject had no documented history of hypertension or DM prior to pregnancy but had a family history of hypertension. The subject data were input into the best-performing classifier, the SVM model, with a sensitivity of 67.5%, a specificity of 57.23%, and a classification threshold of 0.43 (supported by Youden's Index of 0.24). The model predicted a 53.73% probability of developing PE (high-risk pregnancy).

Discussion

In this retrospective study, sparse multivariable models developed for predicting the probability of PE showed their ability to balance predictive performance with an acceptable rate of false positive predictions. Such models are crucial for enhancing the predictive capacity of existing PE screening programs outlined in the Maternal and Child Health (MCH) handbook. Although a systematic review has shown that adherence to the MCH handbook improves maternal service utilization, improved mitigation of complications such as PE could further improve outcomes.[27,28] A predictive model can assist doctors in PHC in identifying high-risk pregnant women and making informed decisions about providing preventive treatments for PE.

The analysis of various ML algorithms revealed key insights into their strengths and limitations in PE prediction. Logistic regression showed the best performance in the initial analysis on the cleaned dataset without resampling, achieving an accuracy of 67%. However, its discriminatory ability, as reflected by an AUC of 0.49, was suboptimal, indicating performance slightly worse than random guessing. This limitation highlighted the challenges of the use of traditional statistical models for datasets with considerable class imbalance, as reflected in most cases (533 out of 638) being classified as no PE.

To address data imbalance, random resampling was performed. With oversampling, the RF model showed the highest accuracy (67.9%). However, despite its improved accuracy, the model showed limited sensitivity, suggesting that oversampling alone does not fully address the issue of identifying positive cases. At the same time, undersampling was performed to reduce the dominance of the majority class. The SVM emerged as the best-performing model in this scenario, with an accuracy of 63% and an AUC of 0.68, indicating moderate discriminatory power. These results underscored the importance of considering various resampling techniques to enhance model performance.

The selection between oversampling and undersampling for addressing class imbalance depends on several factors, including the dataset size, class distribution, and the potential impact on model performance, as observed herein. Oversampling, which increases the representation of the minority class by duplicating existing samples or generating synthetic ones via simple random oversampling or other advanced techniques (e.g., systematic minority oversampling technique (SMOTE)), is generally used on small datasets.[29] In such cases, oversampling is preferred because it does not discard existing data, which might be valuable for training.

Undersampling reduces the representation of the majority class by randomly selecting its subset.[29] It is used on relatively large datasets or when the majority class is excessively overrepresented. Undersampling is also preferred when the majority class contains redundant or less informative data. A combination of oversampling and undersampling (e.g., Tomek Links or SMOTE+ENN) may also be used to effectively balance the dataset while minimizing the risk of overfitting or information loss. However, no strict rule guarantees success, so it is advisable to try all the above methods and evaluate model performance using appropriate metrics such as recall, specificity, and AUC-ROC.[29]

In addition to resampling, feature selection may be used to optimize the predictive performance. A prospective cohort study in Uganda stated that maternal characteristics such as history of PE, maternal age ≥35 years, nullity, BMI, diastolic blood pressure, and simple laboratory tests can predict PE in clinical settings with 77% accuracy.[30] A systematic study of 128 citations highlighted four ML-based studies based on maternal characteristics, medical history, obstetric history, and simple laboratory and ultrasound examination results, obtaining AUC values of 0.860–0.973.[31] The best-performing models in the above two studies included Elastic net, stochastic gradient boosting, extreme gradient boosting, and RF. These findings support the methodology of the present study, albeit with limited features.[31]

One of the most important findings of this study was the impact of threshold adjustment on the model performance. A reduction in the classification threshold from the default 0.5 to 0.43 improved the balance between sensitivity and specificity for the SVM model. This adjustment, guided by Youden's Index (0.24), allowed the SVM model to achieve a sensitivity of 67.5%, specificity of 57.23%, and precision of 29.34%. Although this adjustment increased the rate of false positives, it also enhanced the detection of high-risk pregnancies, which is critical in clinical settings where missing cases of PE can have severe consequences.

Threshold adjustment is a practical strategy for addressing the inherent trade-off between sensitivity and specificity. In PE prediction, high sensitivity is particularly valuable for ensuring early PE detection and timely intervention for at-risk pregnancies.[32] However, careful evaluation beyond sensitivity and specificity needs to be considered. The increased rate of false positives necessitates careful consideration of the burden on healthcare systems, including the cost and psychological impact of additional follow-up testing. A good threshold should reflect the clinical context, and additional adjustments should be considered to develop an optimal and valid prediction model.[32] In practice, context-specific and shared decision-making among clinicians and policymakers should be considered.[33]

Performance can vary based on the specific features used, the quality of data, and the chosen model. A systematic review found that highly accurate predictions may be made based on data from routine ANC checks.[34] The AUC-ROC values of such models range from 0.64 and 0.96, sensitivity from 29% to 100%, and specificity from 26% to 96%, which are similar to those obtained herein.[34]

Another systematic review showed robust predictive performance of various algorithms using similar sets of features. The most frequently used algorithms are RF, SVM, and neural networks.[30] Nevertheless, it should be emphasized that previous studies in the systematic review in which accuracy was above 90% suffered from several limitations.[30] A Stochastic gradient boosting (SGB) model achieved an accuracy of 0,97% with a lack of trimester data and a small number of PE incidents (474 out of 10.532 cases) which was also experienced by this study.[35]

Herein, typical secondary data analysis challenges were encountered, including missing values (584 entries) and duplicates (1,028 entries). The MCAR test indicated that the missing data were entirely random, necessitating the removal of these entries to ensure the integrity of the analysis. This resulted in a reduced dataset of 638 records, with an imbalance between the no PE (533 cases) and PE (105 cases) classes.

In a retrospective cohort study of 16,370 mothers who gave birth at Stanford, ML was used for the early prediction of PE using ML.[8] After cleaning, a dataset with 5,245 entries was obtained, among which only 561 were PE cases (10.7%). Despite the larger sample size, their predictive accuracy matches that of other studies, even though initial sample requirements were not specified. This mirrors the limitations of the present, suggesting that better data quality could improve performance.[31]

Class imbalance can cause a bias toward the majority class. Thus, both oversampling and under sampling were employed in this study. Oversampling improved accuracy but reduced the specificity of the RF model, whereas under sampling improved the sensitivity/specificity balance of the SVM model at the cost of dataset size. Therefore, a more advanced technique might be more appropriate to handle data imbalance and tiny data sets such as that used herein. This particular restriction may complicate the interpretation of the model outcomes. Class imbalance can be prevented using synthetic minority oversampling, particularly by handling the missing data based on the data distribution.[27] A previous study also addressed imbalance through oversampling, demonstrating its effectiveness alongside under sampling and cost-sensitive algorithms for rare medical outcomes.[36]

A detailed comparison of the performance metrics of different models under various resampling strategies provided valuable insights. Without resampling, logistic regression and SVM showed similar performance metrics, including low sensitivity. Oversampling improved the accuracy of RF but failed to substantially increase sensitivity, which is critical for identifying high-risk pregnancies.[31] With undersampling, the SVM model showed the highest AUC (0.68), making it the most balanced model in terms of overall predictive power. These findings underscored the importance of tailoring the resampling and modeling strategies to the specific requirements of the prediction task. For example, models with high sensitivity, such as SVM, are better suited for clinical applications where identifying positive cases is a priority, even at the cost of increased false positives. In contrast, models with higher specificity may be more appropriate in scenarios where minimizing unnecessary interventions is critical.[34]

The development of models for predicting PE is of great clinical importance, particularly in resource-constrained settings. These models may complement the existing screening protocols in the MCH handbook by providing additional risk-stratification capabilities. For instance, the SVM model with the optimized threshold could be utilized by PHC to facilitate the identification of high-risk pregnancies requiring preventive measures.

The implementation of such models in clinical practice can enable the early detection of at-risk pregnancies and, thereby, timely interventions, including the administration of low doses of aspirin and enhanced monitoring. However, the limitations of the models must also be considered, including their moderate discriminatory accuracy and potential false positive results, to prevent undue strain on healthcare infrastructure and unnecessary patient concern.

Although the developed models based on simple features derived from the MCH handbook showed decent predictive performance, feature selection and engineering could be further refined to enhance model performance. Although the SHAP analysis was employed as a promising method for feature interpretation, this study primarily focused on optimizing predictive performance. At the same time, the authors advise that future studies employ the features documented in the MCH handbook to preserve the simplicity and clinical applicability of the models.

This study has several limitations. First, reliance on secondary data restricted the inclusion of key predictors (e.g., biochemical markers and detailed clinical history), likely contributing to the moderate performance of the developed models. Second, substantial class imbalance necessitated resampling, which increased the risk of overfitting (oversampling) and reduced statistical power (undersampling). Third, the lack of external validation limited generalizability to other populations. Finally, the modest AUC scores and low statistical power (48.1% post-validation)

underscore the need for refinement before clinical implementation. Despite these limitations, the results indicated that PE can be predicted based on simple, primary care–compatible predictors. At the same time, further refinement of feature selection through advanced algorithms and external validation could enhance model performance.

Building on this study's findings, several future research directions can be proposed. First, improving the initial feature selection is critical. Although incorporating advanced predictors (biochemical or genetic markers) could enhance model performance, their implementation may be impractical for public health programs. Instead, a more systematic approach to the selection of simple features and reengineering could improve the predictive accuracy and align with findings from other studies.

Second, hybrid or ensemble methods (combining deep learning with SVM or RF) could show better forecasting accuracy and robustness. Third, prospective validation across diverse populations and clinical settings is essential for assessing the generalizability and practical utility of the developed models. Finally, the integration of these models into electronic medical record systems as decision-support tools could enable real-time risk assessments, enhancing the scalability and impact of PE screening programs.

The development of models for PE prediction is a critical step toward improving maternal and neonatal outcomes, particularly in low- and middle-income countries, where the burden of hypertensive disorders in pregnancy is disproportionately high. By enabling earlier identification of high-risk pregnancies, such models can improve the efficiency of resource allocation, reduce complications, and improve overall care delivery. Moreover, integrating ML models into maternal health research highlights the potential of data-driven approaches to address complex healthcare challenges. The increase in the amount of available data and the evolution of analytical techniques are expected to facilitate the development of robust, scalable, and clinically relevant models, driving further improvements in maternal health outcomes.

Conclusion

This study indicates the promising potential of sparse multivariable prediction models for PE prediction, thereby improving screening programs. Despite the challenges associated with data quality and class imbalance, the models developed, particularly SVM with threshold optimization, show promising performance with balanced sensitivity and specificity. However, additional refinement and validation are required to realize their potential for clinical application. Integrating these models into existing screening protocols can facilitate the detection and management of high-risk pregnancies, ultimately improving maternal and neonatal health outcomes.

Abbreviations

PE: preeclampsia; PHC: primary health care; ANC: antenatal care; BMI: body mass index; ML: machine learning, DM: diabetes mellitus; MAP: mean arterial pressure; AUC: area under the curve; MCAR: Missing Completely at Random; SVM: support vector machines; RF: random forest; KNN: K-nearest neighbors; AUC-ROC: area under the receiver operating characteristic curve; SHAP: SHapley Additive exPlanation; MCH: Maternal and Child Health;

Ethics Approval and Consent to Participate

Data was collected from the database of a previous study from hospital routine service data, with proper anonymization; thus, ethical approval was provided as a nonhuman research (no. 013-1/DIN/KEP.RSBK/LKBK/III/2024).

Competing Interest

All authors do not have any competing interest in the research.

Availability of Data and Materials

The data is available upon request to the corresponding author.

Authors' Contribution

DA conceived of the presented idea, developed the overall methods and concepts, and analysis, AAS and KNS provided expert guidance on the theory and evaluation of the computations and supervised the research, AAS, KNS, and DSN evaluated the main conceptual ideas and proof outline, thus contributed to the design and implementation of the research, to the analysis of the results and the writing of the manuscript, DA wrote the manuscript with support from AAS, KNS, DSN.

Acknowledgment

Great appreciation to The Indonesian Medical Education and Research Institute Faculty of Medicine Universitas Indonesia (IMERI FKUI) for many insights and supporting analysis and database hosting while analyzing using machine learning. Special thanks to Achzab Asharudin and Abdul Hamid from aiseeyou.tech for their technical expertise with machine learning analysis.

References

1. Brown MA, Magee LA, Kenny LC, et al. Hypertensive disorders of pregnancy: ISSHP classification, diagnosis, and management: Recommendations for international practice. Hypertension. 2018; 72 (1): 24-43. DOI: 10.1161/HYPERTENSIONAHA.117.10803
2. Boushra M, Natesan SM, Koyfman A, et al. High risk and low prevalence diseases: Eclampsia. Am J Emerg Med. 2022; 58: 223-228. DOI: 10.1016/j.ajem.2022.06.004.
3. Teka H, Yemane A, Abraha HE, et al. Clinical presentation, maternal-fetal, and neonatal outcomes of early-onset versus late onset preeclampsia-eclampsia syndrome in a teaching hospital in a low-resource setting: A retrospective cohort study. PLoS ONE. 2023; 18 (2): e0281952. DOI: 10.1371/journal.pone.0281952.
4. Goel A, Maski MR, Bajracharya S, et al. Epidemiology and mechanisms of de novo and persistent hypertension in the postpartum period. Circulation. 2015; 132 (18): 1726–1733. DOI: 10.1161/CIRCULATIONAHA.115.015721.
5. Perry H, Khalil A, Thilaganathan B. Preeclampsia and the cardiovascular system: An update. Trends Cardiovasc Med. 2018; 28 (8): 505-513. DOI: 10.1016/j.tcm.2018.04.009.
6. Ives CW, Sinkey R, Rajapreyar I, et al. Preeclampsia—pathophysiology and clinical presentations: JACC state-of-the-art review. J Am Coll Cardiol. 2020; 76 (14): 1690-1702. DOI: 10.1016/j.jacc.2020.08.014.
7. Garovic VD, Dechend R, Easterling T, et al. Hypertension in pregnancy: Diagnosis, blood pressure goals, and pharmacotherapy: A scientific statement from the American Heart Association. Hypertension. 2022; 79 (2): e21-e41. DOI: 10.1161/HYP.0000000000000208.
8. Maric I, Tsur A, Aghaeepour N, et al. Early prediction of preeclampsia via machine learning. Am J Obstet Gynecol MFM. 2020; 2 (2): 100100. DOI: 10.1016/j.ajogmf.2020.100100.
9. Gathiram P, Moodley J. Pre-eclampsia: Its pathogenesis and pathophysiolgy. Cardiovasc J Afr. 2016; 27 (2): 71-78. DOI: 10.5830/CVJA-2016-009.
10. Benny PA, Alakwaa FM, Schlueter RJ, et al. A review of omics approaches to study preeclampsia. Placenta. 2020; 92: 17-27. DOI: 10.1016/j.placenta.2020.01.008.
11. Wu P, Van den Berg C, Alfirevic Z, et al. Early pregnancy biomarkers in pre-eclampsia: A systematic Review and meta-analysis. Int J Mol Sci. 2015; 16 (9): 23035-23056. DOI: 10.3390/ijms160923035.
12. Townsend R, Khalil A, Premakumar Y, et al. Prediction of pre-eclampsia: Review of reviews. Ultrasound Obstetr Gynecol. 2019; 54 (1): 16-27. DOI: 10.1002/uog.20117.
13. Al-Rubaie ZT, Hudson HM, Jenkins G, et al. Prediction of preeclampsia in nulliparous women using routinely collected maternal characteristics: A model development and validation study. BMC Pregnancy Childbirth. 2020; 20: 23. DOI: 10.1186/s12884-019-2712-x.
14. Kusuma RA, Nurdiati DS, Wilopo SA. Alternatives of risk prediction models for preeclampsia in a low middle-income setting. Open Access Maced J Med Sci. 2022; 10 (B): 1745-1750. DOI: 10.3889/oamjms.2022.9030.
15. Tiruneh SA, Moran LJ, Callander EJ, et al. Externally validated prediction models for preeclampsia: Systematic review and meta-analysis. Ultrasound Obstet Gynecol. 2024; 63 (5): 592-604. DOI: 10.1002/uog.27490.
16. Ranjbar A, Montazeri F, Ghamsari SR, et al. Machine learning models for predicting preeclampsia: A systematic review. BMC Pregnancy Childbirth. 2024; 24: 6. DOI: 10.1186/s12884-023-06220-1.
17. Bzdok D, Altman N, Krzywinski M. Statistics versus machine learning. Nat methods. 2018; 15: 233-234. DOI: 10.1038/nmeth.4642.
18. Sufriyana H, Wu YW, Su EC. Artificial intelligence-assisted prediction of preeclampsia: Development and external validation of a nationwide health insurance dataset of the BPJS Kesehatan in Indonesia. EbioMedicine. 2020; 54: 102710. DOI: 10.1016/j.ebiom.2020.102710.
19. Savitri AI, Zuithoff P, Browne JL, et al. Does pre-pregnancy BMI determine blood pressure during pregnancy? A prospective cohort study. BMJ Open. 2016; 6: e011626. DOI: 10.1136/bmjopen-2016-011626.
20. Mocking M, Savitri AI, Cuno SPM, et al. Does body mass index early in pregnancy influence the risk of maternal anaemia? An observational study in Indonesian and Ghanaian women. BMC Public Health. 2018; 18: 873. DOI: 10.1186/s12889-018-5704-2.
21. Kelsey JL, Whittemore AS, Evans AS, et.al. Methods of sampling and estimation of sample size. In: Kelsey JL, Whittemore AS, Evans AS, et al., eds., Methods in Observational Epidemiology. New York: Oxford University Press; 1996.
22. Balki I, Amirabadi A, Levman J, et al. Sample-size determination methodologies for machine learning in medical imaging research: A systematic review. Can Assoc Radiol J. 2019; 70 (4): 344–353. DOI: 10.1016/j.carj.2019.06.002.
23. Riley RD, Ensor J, Snell KI, et al. Calculating the sample size required for developing a clinical prediction model. BMJ. 2020; 368. DOI: 10.1136/bmj.m441.
24. Wagner R, Grimm MS. Empirical validation of the 10-times rule for SEM. In: Radomir L, Ciornea R, Wang H, et al., eds. State of the Art in Partial Least Squares Structural Equation Modeling (PLS-SEM). Berlin: Springer; 2023.
25. Little RJA, Rubin DB. Statistical analysis with missing data. 3rd ed. New York: Wiley; 2019.
26. Hassanzad M, Hajian-Tilaki K. Methods of determining optimal cut-point of diagnostic biomarkers with application of clinical data in ROC analysis: an update review. BMC Med Res Methodol. 2024; 24: 84. DOI: 10.1186/s12874-024-02198-2.
27. Nishimura E, Rahman MO, Ota E, et al. Role of maternal and child health handbook on improving maternal, newborn, and child health outcomes: A systematic review and meta-analysis. Children. 2023; 10 (3): 435. DOI: 10.3390/children10030435.
28. Osaki K, Hattori T, Toda A, et al. Maternal and child health handbook use for maternal and child care: A cluster randomized controlled study in rural Java, Indonesia. J Public Health (Oxf). 2019; 41: 170-182. DOI: 10.1093/pubmed/fdx175.

29. Fernández A, Garcia S, Herrera F, et al. SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. J Artif Intell Res. 2018; 61: 863-905. DOI: 10.1613/jair.1.11192.
30. Awor S, Abola B, Byanyima R, et al. Prediction of pre-eclampsia at St. Mary's hospital lacor, a low-resource setting in northern Uganda, a prospective cohort study. BMC Pregnancy Childbirth. 2023; 23: 101. DOI: 10.1186/s12884-023-05420-z.
31. Ranjbar A, Taeidi E, Mehrnoush V, et al. Machine learning models for predicting pre-eclampsia: A systematic review protocol. BMJ Open. 2023; 13: e074705. DOI: 10.1136/bmjopen-2023-074705.
32. Wynants L, van Smeden M, McLernon DJ, et al. Three myths about risk thresholds for prediction models. BMC Med. 2019; 17: 192. DOI: 10.1186/s12916-019-1425-3.
33. Zheng D, Hao X, Khan M, et al. Comparison of machine learning and logistic regression as predictive models for adverse maternal and neonatal outcomes of preeclampsia: A retrospective study. Front Cardiovasc Med. 2022; 9: 959649. DOI: 10.3389/fcvm.2022.959649.
34. Aljameel SS, Alzahranin M, Almusharraf R, et al. Prediction of preeclampsia using machine learning and deep learning models: A review. Big Data Cogn. Comput. 2023; 7 (1): 32. DOI: 10.3390/bdcc7010032.
35. Jhee JH, Lee S, Park Y, et al. Prediction model development of late-onset preeclampsia using machine learning-based methods. PLoS ONE. 2019; 14 (8): e0221202. DOI: 10.1371/journal.pone.0221202.
36. Parrales-Bravo F, Caicedo-Quiroz R, Tolozano-Benitez E, et al. OUCH: Oversampling and Undersampling Cannot Help Improve Accuracy in Our Bayesian Classifiers That Predict Preeclampsia. Mathematics. 2024; 12: 3351. DOI: 10.3390/math12213351.