International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 10, NCREHAASTE - 2026

Predictive Analysis of Preeclampsia Risk in Pregnant Women Using Machine Learning Techniques

Arvind Kumar[1], Prof. (Dr.) Rahul Kumar Misra[2]
[1]Assistant Professor, Department of Computer Applications, IFTM University, Moradabad, India
Email: arvindkumar@iftmuniversity.ac.in
[2]Pro VC & Director SCSA, IFTM University, Moradabad, India
Email: rahulkumarmisra@iftmuniversity.ac.in

Abstract—Preeclampsia is a hypertensive complication of pregnancy that may develop rapidly and place both the mother and fetus at risk. Recognizing women who are more likely to develop the condition at an early stage can support closer surveillance and timely clinical evaluation. This paper presents a machine-learning framework for estimating preeclampsia risk from routinely obtainable maternal, antenatal, obstetric, and laboratory information. The framework covers ethical data acquisition, data-quality assessment, preprocessing, feature construction and selection, management of class imbalance, model training, validation, performance assessment, and explainable risk reporting. Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, and gradient-boosting approaches are considered for comparative evaluation. Performance is intended to be assessed through sensitivity, specificity, precision, F1-score, ROC-AUC, precision-recall AUC, calibration, and confusion-matrix analysis. SHAP and LIME are included to provide global and case-level explanations of model predictions. Because no research dataset or experimental outputs were supplied, this paper does not present invented numerical results. The proposed system is intended to support clinical decision-making and is not presented as an autonomous diagnostic system.

Index Terms—Preeclampsia, Machine Learning, Artificial Intelligence, Pregnancy Risk Prediction, Maternal Health, Explainable AI, SHAP, LIME, Antenatal Care.

I. IN ODUCTION

Preeclampsia is a hypertensive disorder associated with pregnancy and generally appears after the twentieth week of gestation. The World Health Organization identifies hypertensive disorders of pregnancy as an important contributor to maternal morbidity and mortality and reports that preeclampsia affects approximately 3–8% of pregnancies worldwide. These observations make timely identification and appropriate clinical management important elements of safe antenatal care [1].

In routine practice, assessment is based on information such as medical history, blood-pressure measurements, clinical examination, laboratory investigations, and recognized maternal risk factors. The growing volume of antenatal information creates an opportunity to examine relationships among several variables simultaneously. Machine learning can learn such patterns from multidimensional records and produce a predicted probability or risk category.

Recent reviews indicate that a range of machine-learning techniques—including Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, and boosting methods—have been investigated for preeclampsia prediction. Frequently reported predictors include maternal age, parity, body-mass index, diabetes, hypertension, blood-pressure measurements, and selected biomarkers. At the same time, differences in datasets, outcome definitions, validation procedures, and study populations make direct comparison difficult, while external validation remains limited in many studies [2]–[5].

The proposed research therefore emphasizes a transparent and reproducible prediction pipeline rather than accuracy alone. The framework incorporates clinically available predictors, leakage-resistant pre-processing, class-imbalance management, calibration assessment, and explain ability. The intended outcome is a model-development procedure that can subsequently be tested in diverse maternal-health populations.

II. PRO M STATEMENT

Machine-learning studies of preeclampsia have reported encouraging predictive discrimination, but performance obtained in one population may not remain reliable in another. Differences in patient characteristics, available predictors, outcome definitions, data quality, and validation design can alter model behaviour. In addition, a prediction that cannot be interpreted easily may be difficult for healthcare professionals to assess or trust.

Accordingly, there is a need for an explainable and systematically validated machine-learning framework that uses information available during antenatal care, estimates preeclampsia risk at a defined early prediction point, and communicates the important contributing factors together with appropriate uncertainty and limitations.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 10, NCREHAASTE - 2026

.

III. RESEARCH OBJECTIVES

- To determine maternal, obstetric, clinical, and laboratory variables that can contribute to early prediction of preeclampsia risk.
- To develop and compare suitable machine-learning models using appropriate validation strategies and clinically relevant performance measures.
- To incorporate explainable-AI methods for identifying the principal factors influencing overall and individual risk predictions.

IV. RELATED WORK

Amiri et al. reviewed machine-learning approaches for preeclampsia prediction and found considerable variation in algorithms, predictor sets, datasets, and study methodologies. Their findings support the need for more consistent validation and reporting practices [2].

Malik et al. examined the use of machine learning for preeclampsia prediction and highlighted the potential value of early risk estimation for supporting monitoring and preventive clinical decisions. They also discussed challenges arising from heterogeneous datasets and translation into clinical practice [3].

Dkeen et al. reviewed artificial-intelligence applications in obstetric risk prediction and reported promising findings from ensemble methods, including Random Forest and XGBoost, while emphasizing the limited availability of external validation [4]. Liu et al. similarly reported strong pooled discrimination in a recent meta-analysis but identified substantial heterogeneity and weaker performance in externally validated studies [5].

These observations directly inform the proposed framework. Model comparison should therefore include calibration and sensitivity-oriented analysis in addition to discrimination, and external or temporal validation should be used whenever the available data permit it.

V. PROPOSED RESEARCH FRAMEWORK

The proposed framework is organized as a sequence of seven stages: ethical data acquisition, data-quality assessment, preprocessing and feature engineering, feature selection, machine-learning model development, performance evaluation, and explainable risk reporting. The final output is intended to provide an early risk estimate that can support appropriate clinical review.

Maternal Health Data

Stage 1: Ethical Data Acquisition

Stage 2: Data Quality Assessment Missing values • Duplicates • Outliers • Leakage • Class imbalance

Stage 3: Preprocessing & Feature Engineering Imputation • Encoding • Scaling • Clinical feature construction

Stage 4: Feature Selection Statistical methods • RFE • LASSO • RF importance • SHAP

Stage 5: ML Model Development Logistic Regression • Decision Tree • Random Forest • SVM • XGBoost

Stage 6: Model Evaluation Sensitivity • Specificity • Precision • F1 • ROC-AUC • PR-AUC • Calibration

Stage 7: Explainable Risk Reporting SHAP + LIME

Early Preeclampsia Risk Prediction

Clinical Decision Support

Fig. 1. Proposed research framework for early preeclampsia risk prediction.

A. Data Acquisition

The dataset should consist of de-identified pregnancy records with a clearly specified preeclampsia outcome and a defined prediction time point. Potential predictors include maternal age, gestational age, parity, gravidity, previous preeclampsia, chronic hypertension, diabetes, body-mass index, systolic and diastolic blood pressure, mean arterial pressure, hemoglobin, clinically appropriate proteinuria measures, and available biomarkers such as placental growth factor or PAPP-A. Only information that would actually be available at the intended prediction time should be used.

B. Data Preprocessing

Before modelling, records should be checked for duplication, impossible measurements, inconsistent units, missing observations, and extreme values. Missing data should be handled according to the variable type and the underlying messiness pattern. Scaling should be applied where required by the selected algorithm. Importantly, imputation, scaling, and other learned preprocessing operations must be fitted using training data only so that information from validation or test data does not leak into model development.

C. Feature Engineering and Selection

Derived variables with clinical meaning may be created from available measurements. Examples include mean

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 10, NCREHAASTE - 2026

arterial pressure and summaries of repeated blood-pressure observations when longitudinal measurements exist. Feature selection can combine clinical knowledge with correlation analysis, regularization, recursive feature elimination, permutation importance, and model-based importance measures. Variables recorded after the intended prediction time must be excluded to preserve a realistic early-prediction setting

D. Class-Imbalance Handling

In some datasets, preeclampsia cases may form a smaller class than non-cases. The class distribution should therefore be reported before model development. Class weighting or training-only resampling methods such as SMOTE may be considered. Any resampling procedure must be performed within the training portion of each validation fold rather than before cross-validation.

E. Model Development

Logistic Regression is proposed as an interpretable baseline, while Decision Tree, Random Forest, Support Vector Machine, and gradient-boosting methods such as XGBoost can be evaluated as alternative learners. Hyper parameters should be optimized through nested cross-validation or another appropriately separated validation procedure so that model selection does not use the final test data.

F. Model Evaluation

The evaluation should include sensitivity, specificity, and precision, negative predictive value, F1-score, ROC-AUC, precision-recall AUC, calibration measures, and confusion matrices. Confidence intervals should be reported when feasible. Because failure to identify a genuinely high-risk pregnancy may have important clinical consequences, sensitivity should be considered alongside overall discrimination rather than relying on a single summary metric.

G. Explainability

HAP can be used to describe global feature importance and the contribution of individual predictors to specific predictions. LIME can provide local surrogate explanations for individual cases. These techniques explain model behavior and should not be interpreted as evidence of causal relationships. The plausibility and stability of important explanations should be reviewed across clinically relevant patient groups.

VI. MACHINE-LEARNING MODELS

Logistic Regression provides an interpretable baseline and estimates the relationship between predictors and the log-odds of the outcome. Decision Trees provide rule-based predictions but can overfit without appropriate constraints. Random Forest combines multiple decision trees and can model nonlinear relationships while reducing variance. Support Vector Machine can construct nonlinear decision boundaries using kernel functions. Gradient-boosting models sequentially improve weak learners and can capture complex interactions. Model selection should be based on validation performance, calibration, robustness, interpretability, and clinical utility rather than a single metric.

VII. PROPOSED SYSTEM ARCHITECTURE

The proposed architecture contains five logical layers: Input Layer → Preprocessing Layer → Prediction Layer → Explainability Layer → Clinical Decision-Support Output. The Input Layer receives de-identified antenatal variables. The Preprocessing Layer performs validation, imputation, encoding, and scaling. The Prediction Layer applies the selected trained model to estimate preeclampsia probability/risk category. The Explainability Layer computes feature contributions. The output layer presents the estimated risk, major contributing factors, model confidence/calibration information, and a prompt for appropriate clinical review.

VIII. EXPERIMENTAL DESIGN

A reproducible experiment should divide the data into development and independent test components. Where sufficient data exist, stratified k-fold cross-validation can be used during training and hyperparameter tuning. A final untouched test set should be used once for unbiased performance estimation. If data are collected across time, temporal validation is preferable to random splitting when it reflects the intended deployment scenario. If multiple hospitals are available, hospital-level external validation should be considered.

The primary analysis should compare all candidate models on the same outcome definition and evaluation splits. Secondary analyses should examine performance by clinically relevant subgroups, missing-data patterns, and different prediction horizons. Calibration plots and decision-curve analysis can provide information beyond discrimination. Model reporting should follow current

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 10, NCREHAASTE - 2026

TRIPOD+AI guidance for prediction models using regression or machine learning. [6]

IX. EXPECTED RESULTS AND DISCUSSION

The proposed study is expected to identify a model that provides useful discrimination while maintaining acceptable calibration and sensitivity. Ensemble methods may perform strongly because preeclampsia risk can involve nonlinear interactions among clinical variables; however, this should be established empirically rather than assumed. A simpler model may be preferred if its performance is comparable and its interpretation is substantially clearer.

The explainability stage is expected to identify variables that consistently influence predictions. For example, blood-pressure measures, prior hypertensive disorders, diabetes-related indicators, maternal characteristics, and selected biomarkers may contribute substantially. These are expected predictors based on previous literature, but their importance in the proposed dataset must be determined experimentally. [2]–[5]

No numerical accuracy, AUC, sensitivity, or other performance value is claimed in this paper because no research dataset and experimental output were supplied. This is intentional: inserting invented results would compromise scientific integrity. Once the study is conducted, the Results section should be replaced or expanded with actual measurements, confidence intervals, statistical comparisons, calibration plots, and explain ability results.

X. ETHICAL, PRIVACY, AND CLINICAL CONSIDERATIONS

Maternal-health prediction involves sensitive clinical information. Any primary-data study should obtain approval from the relevant institutional ethics committee and follow applicable privacy and data-governance requirements. Records should be de-identified or appropriately pseudonymized before analysis. Access controls, secure storage, and audit procedures should be established.

The model should be positioned as decision support rather than autonomous diagnosis. A high-risk prediction should prompt appropriate clinical assessment, while a low-risk prediction should not be interpreted as a guarantee of safety. Before clinical use, the model requires external validation, calibration assessment, subgroup/fairness analysis, prospective evaluation, and assessment of clinical utility.

XI. LIMITATIONS

The proposed framework has several limitations. Model performance depends on dataset quality, prevalence, predictor availability, outcome definitions, and population characteristics. Models trained in one hospital or geographic setting may not generalize to other settings. Missing data and measurement differences can introduce bias. Explainability methods describe model behavior but do not establish causation. Finally, the paper presents a research framework rather than a completed clinical validation study.

XII. CONCLUSION AND FUTURE WORK

This paper presents an explainable machine-learning framework for early prediction of preeclampsia risk using maternal, antenatal, obstetric, and laboratory information. The framework combines controlled pre-processing, feature selection, comparative model development, sensitivity-aware evaluation, calibration, and explainable AI. Its emphasis on transparent methodology addresses important limitations identified in recent literature, particularly variation between study populations and limited external validation [4], [5].

Future work will involve obtaining an ethically approved dataset, defining the intended early prediction window, training and tuning the candidate models, performing independent and external validation, assessing calibration and clinical utility, and implementing an interpretable decision-support dashboard. Prospective multicentre validation should remain a long-term requirement before the framework is considered for routine maternal-health use

REFERENCES

[1] World Health Organization, “Pre-eclampsia,” WHO Fact Sheet, Dec. 10, 2025.

[2] F. Amiri et al., “Machine learning models for predicting preeclampsia: a systematic review,” BMC Pregnancy and Childbirth, vol. 23, 2023, Art. no. 845, doi: 10.1186/s12884-023-06220-1.

[3] V. Malik, N. Agrawal, S. Prasad, et al., “Prediction of Preeclampsia Using Machine Learning: A Systematic Review,” Cureus, vol. 16, no. 12, 2024, Art. no. e76095, doi: 10.7759/cureus.76095.

[4] M. Dkeen et al., “Artificial Intelligence Applications in Obstetric Risk Prediction: A Systematic Review of Machine Learning Models for Preeclampsia,” Cureus, 2025, doi: 10.7759/cureus.83961.

[5] L. Liu, Q. Zhu, Y. Zong, et al., “Machine Learning Prediction Models for Preeclampsia: Systematic Review and Meta-Analysis,” Journal of Medical Internet Research, 2026.

[6] G. S. Collins, K. G. M. Moons, P. Dhiman, R. D. Riley, A. L. Beam, et al., “TRIPOD+AI statement:

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 10, NCREHAASTE - 2026

updated guidance for clinical prediction models," BMJ, vol. 385, 2024, e078378.

[7] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Proc. 31st Int. Conf. Neural Information Processing Systems (NeurIPS), 2017, pp. 4765–4774.

[8] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.

[9] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 785–794.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)