Ingénierie des Systèmes d'Information
Vol. 31, No. 3, March, 2026, pp. 985-999
Journal homepage: http://iieta.org/journals/isi

Early Prediction of Preeclampsia Using a Multimodal Machine Learning Model Integrating Clinical, Hemodynamic, and Biochemical Data

Rosa Cardenas-Urrelo[1], Dino Quinteros-Navarro[2*], Carmina Tang[1], Ynés Torres-Flores[1], Yolanda Navarro[1], Melissa Soto[3], Ana Maguiña[4], Marina Huamantumba[1]

[1] Facultad de Ciencias de la Salud, Universidad Nacional de San Martín, Tarapoto 22201, Peru
[2] Facultad de Ingeniería, Carrera Profesional de Ingeniería de Sistemas, Universidad Tecnológica del Peru, Lima 15046, Peru
[3] Facultad de Ciencias de la Salud, Carrera Profesional de Nutrición y Dietética, Universidad Privada del Norte, Lima 15434, Peru
[4] Facultad de Medicina, Universidad Nacional Federico Villarreal, Lima 15007, Peru

*Corresponding Author Email: C29925@utp.edu.pe

https://doi.org/10.18280/isi.310328

Received: 17 August 2025
Revised: 1 December 2025
Accepted: 19 March 2026
Available online: 31 March 2026

Keywords:
preeclampsia, machine learning, early prediction, clinical data, biomarkers, obstetrics, risk stratification, artificial intelligence

ABSTRACT

Preeclampsia (PE) remains a major cause of maternal and perinatal morbidity and mortality, and its early prediction continues to be a significant clinical challenge. This study develops a multimodal machine learning (ML) model that integrates clinical, hemodynamic, and biochemical variables collected during the first trimester to improve early risk stratification for PE. We analyzed a cohort of 1,000 pregnant women, incorporating over 40 predictors, including maternal characteristics, mean arterial pressure (MAP), uterine artery pulsatility index (UtA-PI), and placental biomarkers. The model demonstrated strong predictive power with an area under the ROC curve (AUC) of 0.85, achieving a sensitivity of 52% and specificity of 94%. The key predictors identified were MAP, UtA-PI, and placental growth factor (PlGF), with Precision-Recall analysis confirming robust performance despite class imbalance. The results indicate that this model can serve as a reliable first-trimester screening tool for PE, offering the potential for integration into routine prenatal care to identify high-risk pregnancies early. However, further validation in multicenter studies and diverse populations is necessary to assess the model's generalizability, stability, and clinical applicability. This research presents a promising step toward personalized prenatal care by providing a data-driven tool to support decision-making in the early detection of preeclampsia.

1. INTRODUCTION

Preeclampsia (PE) is a multisystem hypertensive disorder of pregnancy that typically manifests after 20 weeks of gestation and arises from a complex combination of genetic, immunological, inflammatory, and environmental mechanisms. Central to its pathophysiology is inadequate remodeling of the uterine spiral arteries and systemic endothelial dysfunction, which together contribute to chronic placental hypoperfusion, oxidative stress, and an imbalance between antiangiogenic factors such as sFlt-1 and proangiogenic mediators like placental growth factor (PlGF) [1-4]. Epidemiologically, PE remains one of the leading causes of maternal and perinatal morbidity and mortality worldwide, with a prevalence of 2–8%. Its burden is especially high in low- and middle-income regions, where limited access to prenatal diagnostics contributes to severe complications including eclampsia, HELLP syndrome, placental abruption, and intrauterine growth restriction [5-7]. Neonatal outcomes are likewise affected, with increased rates of preterm birth, low birth weight, perinatal asphyxia, and intensive care admissions [3].

Although several clinical risk factors have been associated with PE—such as advanced maternal age, nulliparity, obesity, chronic hypertension, diabetes, and prior PE—their individual predictive power remains low, reinforcing the need for more accurate risk-stratification approaches [1, 2]. Existing screening strategies combining maternal history, biophysical indicators such as uterine artery pulsatility index (UtA-PI), and biomarkers including PlGF, pregnancy-associated plasma protein A (PAPP-A) and sFlt-1 have improved detection performance, yet widespread implementation is hindered by economic and logistical constraints, particularly in resource-limited settings [1, 5].

In parallel, artificial intelligence (AI) and machine learning (ML) have emerged as promising tools capable of integrating diverse clinical, biochemical, hemodynamic, ultrasound, and genetic data to identify complex, nonlinear patterns beyond the reach of conventional statistical methods [8]. A wide range of ML models—including Random Forest, XGBoost, LightGBM, Support Vector Machines SVMs, regularized logistic regression, and deep neural networks—have demonstrated strong predictive potential [9]. Studies indicate that the highest performance is achieved when combining

maternal characteristics, hemodynamic parameters such as mean arterial pressure (MAP) and UtA-PI, and placental biomarkers, with some models reaching Area Under the Curve AUC values above 0.90 in well-characterized cohorts [7, 8]. Nonetheless, in many regions where only basic clinical data are available, predictive accuracy remains modest (AUC 0.70–0.75), underscoring the need for scalable and adaptable solutions.

Recent advances have incorporated multimodal approaches integrating ultrasound images, molecular data, and clinical variables through convolutional neural networks or hybrid architectures, moving closer to a precision-medicine paradigm [10]. Additionally, interpretability tools such as SHAP and LIME now allow clinicians to understand the individual contribution of each predictor, strengthening trust in AI-based decisions and enhancing clinical applicability [1, 2]. However, key challenges persist. Many models exhibit reduced performance when applied to external populations due to differences in risk-factor prevalence, ethnicity, or data quality—a problem that highlights the necessity of external validation and mode [4]. Ethical considerations also arise, particularly regarding variables linked to race or ethnicity, where careless inclusion may perpetuate health inequities; thus, transparency and fairness audits are increasingly recommended [11].

Early identification of women at high risk of PE enables timely preventive interventions, including low-dose aspirin initiation, enhanced surveillance, and early management of complications, which have demonstrated meaningful reductions in adverse outcomes [6, 8].

In this context, the present study develops an artificial intelligence model for the early detection of PE that integrates robust clinical, hemodynamic, and biochemical evidence while emphasizing accuracy, interpretability, and adaptability. The objective is to provide a predictive tool that can be applied across diverse clinical settings, supporting improved maternal-fetal outcomes and contributing to the global effort to reduce the burden of PE.

2. RELATED WORKS

PE is a hypertensive syndrome of pregnancy that affects between 2% and 8% of pregnancies worldwide and continues to be one of the main causes of maternal and perinatal morbidity and mortality. Its complex pathophysiology has motivated decades of research, and although its etiology is not fully elucidated, evidence indicates that it originates mainly from abnormal placentation. Incomplete remodeling of the uterine spiral arteries causes inadequate uteroplacental blood flow, generating chronic hypoxia and oxidative stress, which in turn induce the release of antiangiogenic factors such as sFlt-1 and sEng. This angiogenic imbalance, coupled with systemic endothelial dysfunction, gives rise to the characteristic clinical manifestations: hypertension, proteinuria, and multiorgan damage.

In recent decades, initial prediction models have been based on known clinical factors, such as advanced maternal age, high body mass index, history of PE, chronic hypertension, pregestational diabetes, smoking, and multiple pregnancies. These models, usually developed through logistic regression, were attractive due to their low cost and ease of implementation, but their discriminative capabilities were modest, with AUC values between 0.65 and 0.75 [7]. This limited its usefulness for early detection, especially in cases of early onset, which are the most severe and have the worst perinatal prognosis.

The search for greater precision led to the incorporation of hemodynamic parameters derived from Doppler ultrasound, such as MAP and Uterine Artery Pulsatility Index (UtA-PI). These markers provide direct information on uteroplacental vascular resistance and have demonstrated independent predictive value. Studies such as those by Jung et al. [9] showed that the combination of MAP and UtA-PI with basic clinical data raised AUC above 0.80, representing a substantial improvement over exclusively clinical models.

The next qualitative leap came from the introduction of placental and angiogenic biomarkers, such as PlGF and PAPP-A, and antiangiogenic such as sFlt-1 and sEng. Its inclusion makes it possible to identify placental dysfunctions in early stages, even before the onset of symptoms. [1] demonstrated that the combination of biomarkers with MAP and UtA-PI could reach AUC greater than 0.90, with sensitivities and specificities compatible with effective population screening [12], reinforced these findings, stressing that the addition of biomarkers increases the ability to detect cases of preterm PE, which require more intensive interventions.

In parallel, the increasing availability of large volumes of clinical, ultrasound and laboratory data has enabled the development of models based on artificial intelligence (AI) and machine learning (ML). Unlike traditional statistical techniques, ML algorithms can process high-dimensional datasets, identify complex interactions, and model nonlinear relationships between variables [2, 7]. Models such as Random Forest, XGBoost, LightGBM, and vector support machines have demonstrated significantly superior performance, reaching AUC of up to 0.90 in well-characterized cohorts.

Hunter et al. [12] integrated interpretability approaches such as SHAP to assess the relative importance of each predictor, identifying MAP, PlGF, and PE antecedents as the variables with the greatest weight. This level of transparency is key to clinical acceptance, as it facilitates the validation of the model by the medical team and improves risk communication to patients. Nguyen-Hoang et al. [7] provided evidence on the applicability of these models in Asian settings, highlighting the need for population-specific calibrations.

In the field of deep learning, multimodal architectures have been developed capable of integrating tabular data, Doppler images and time series. Jung et al. [9] implemented a model that combined convolutional neural networks (CNNs) for image analysis with multilayer perceptrons for clinical and biochemical variables, achieving an AUC of 0.94, sensitivity of 83%, and specificity of 88%. The study by Marić et al. [13] developed an early-prediction model for preeclampsia based on machine learning algorithms, including elastic net and gradient boosting, applied to routinely collected clinical and laboratory data before 16 weeks of gestation. Their approach achieved an AUC of 0.79 for overall preeclampsia and 0.89 for early-onset cases, demonstrating that ML-based methods can effectively capture complex risk patterns using standard prenatal information. These findings highlight the competitive performance of data-driven models compared with conventional risk-stratification strategies and reinforce the relevance of integrating ML into early screening workflows.

If the historical evolution of the metrics is analyzed, a clear pattern can be observed:

Simple clinical models → AUC ~0.70–0.75 [7]

- MAP and UtA-PI → AUC ~0.82)
- Biomarkers → AUC ~0.90–0.92 [1]
- Advanced ML → AUC ~0.88–0.90 [2, 7]
- ML -Gradient boosting → AUC ~0.79–0.89 [13]

Table 1 provides a comparative overview of recent ML-based approaches for predicting preeclampsia, contrasting data modalities, sample sizes, algorithms, key predictors, and reported performance metrics.

Despite these advances, external validity remains a major challenge in clinical prediction modeling. As highlighted by the study [13], the performance of a machine-learning model can vary substantially when applied to populations that differ from the original training cohort in terms of ethnic composition, prevalence of risk factors, or data quality. Their study emphasizes that even well-performing models may experience meaningful declines in discrimination when transferred to new clinical settings, underscoring the need for local recalibration and multicenter validation to ensure robustness and generalizability in real-world practice.

Another critical aspect addressed by the literature is the ethical component and equity in AI [4, 11]. Caution against the use of variables such as race or ethnicity: although their inclusion may improve accuracy, there is a risk of perpetuating structural biases. They propose strategies such as adjusting thresholds by subgroup and evaluating equity metrics to ensure that the benefits of AI are distributed fairly. Recent evidence demonstrates that higher predictive performance is typically achieved when multimodal information—clinical, hemodynamic, and biochemical—is combined. However, existing studies present important limitations: many rely on costly biomarkers or advanced imaging technologies that constrain scalability; others use single-modality data that limit predictive accuracy; and several high-performing models lack interpretability, reducing clinical trust and hindering adoption. Furthermore, performance often declines when models are applied to external populations, underscoring the persistent challenge of generalizability. These limitations highlight the need for an approach that integrates multiple modalities while remaining interpretable, clinically feasible, and adaptable to resource-limited settings.

The interpretability of models is a growing requirement. The application of methods such as SHAP and LIME not only helps clinicians understand how predictions are generated, but also improves patient confidence, favoring shared decision-making. This explanatory approach is especially valuable in environments where the acceptance of AI depends on its ability to integrate seamlessly into clinical workflows.

Table 1. Comparative summary of recent machine learning approaches for preeclampsia prediction

Study	Data Modalities	Sample Size	Algorithms Used	Key Predictors	Performance (AUC)	Limitations
[14]	Clinical + Hemodynamic + Biomarkers	5,000+	ML ensemble	MAP, UtA-PI, PlGF	**0.90–0.94**	Limited external validation; costly biomarkers
[8]	Clinical + ICU variables	300	Random Forest	BP patterns, comorbidities	**0.86**	Small sample; not designed for first-trimester screening
[5]	National insurance database (clinical only)	1.1M	XGBoost	Maternal history	**0.70–0.75**	No biomarkers; reduced sensitivity
[15]	Clinical + Biochemical	600	SVM, RF	MAP, PlGF, PAPP-A	**0.82–0.89**	No interpretability; limited multimodality
[9]	Ultrasound images + clinical	20,000 images	Deep Learning (CNN)	Placental textures	**0.91–0.94**	Requires advanced imaging infrastructure
[10]	EHR + ML	35,000	Gradient Boosting	Medical history + labs	**0.83**	No integration of Doppler or biomarkers
[3]	Multi-omics	1,000+	Hybrid DL	Genetic & transcriptomic markers	**0.92+**	Very high cost; limited feasibility

Bertholdt et al. [16] demonstrated that advanced functional ultrasound imaging can extract high-resolution hemodynamic signals, highlighting its potential as a source of quantitative vascular biomarkers during pregnancy. Likewise, Banaei et al. [17] applied supervised machine-learning algorithms to routine obstetric variables to predict episiotomy risk, illustrating the feasibility of deploying ML models using standard perinatal data. In the imaging domain, Pietsch et al. [18] developed a U-Net–based segmentation model capable of assessing placental health from ultrasound images, showing the promise of integrating image-derived features into multimodal predictive systems. Complementing these findings, Shalom et al. [19] examined how hypertensive disorders of pregnancy alter placental cellular organization, reinforcing the biological rationale behind image- and biomarker-based risk prediction models.

Bolk et al. [20] demonstrated that maternal preeclampsia is an independent long-term risk factor for developmental coordination disorder in extremely preterm infants, underscoring the extended neurodevelopmental implications of hypertensive pregnancy disorders. Methodologically, Karpov et al. [21] conducted a large-scale network analysis of artificial intelligence in medicine, identifying dominant clusters in imaging, structured clinical modeling, time-series analysis, and prognostic ML, while emphasizing persistent challenges related to data quality, reproducibility, and interpretability. These insights are crucial given the increasing clinical deployment of AI-based tools in obstetrics.

Bülez et al. [22] developed a LightGBM-based diagnostic model using routine laboratory and clinical data from more than 10,000 pregnancies, achieving an AUC of 0.83 and identifying hemoglobin concentration, maternal age, and liver enzyme markers as influential predictors. Their findings confirm that ML-based PE detection can reach clinically relevant performance even without costly biomarkers. In parallel, Priyanka et al. [23] analyzed automated machine-learning (AutoML) frameworks such as Auto-WEKA, TPOT, and AutoPrognosis, demonstrating their ability to optimize model pipelines while also warning that reduced transparency may hinder clinical acceptance—highlighting the need for explainable ML in maternal-fetal medicine.

Park et al. [24] implemented an explainable AI decision-support system for early diagnosis of patent ductus arteriosus in premature infants, integrating EHR-based time-series and

ML predictors and achieving diagnostic accuracies up to 84%. Their work illustrates how AI-enabled systems can be embedded within neonatal workflows to enhance early detection. Collectively, these ten studies—spanning functional ultrasound imaging, clinical ML models, placental segmentation, cellular pathophysiology, epidemiologic associations, AutoML, and AI-based clinical decision support—outline the current state of the art in perinatal machine learning. Building on this foundation, the present research advances a machine learning–based multimodal model for early prediction of preeclampsia, integrating clinical, hemodynamic, and biomarker data while prioritizing interpretability, calibration, and applicability across diverse healthcare environments.

Recent work has explored additional pathways to improve predictive accuracy for preeclampsia using machine learning models based on routinely available clinical data. Rahman et al. [25] proposed an automated machine learning (AutoML) framework to compare multiple algorithms—including Decision Trees, Random Forests, Gradient Boosting, Logistic Regression, and Deep Learning—using a retrospective cohort of 1,473 pregnancies. Their findings showed that Decision Trees achieved the best performance (AUC = 0.91; sensitivity = 83.6%; specificity = 96.5%), highlighting the potential of structured clinical variables such as hypertension history, diabetes mellitus, and prior preeclampsia as strong early predictors of disease. This study reinforces the scalability of ML models trained on low-cost variables and supports the integration of AutoML tools to optimize predictive pipelines in contexts where biochemical or Doppler-based biomarkers may be limited.

Finally, in contrast to previous work, the present study addresses these gaps by developing a multimodal machine learning model that integrates clinical variables, hemodynamic parameters, and placental biomarkers collected in the first trimester. The model emphasizes interpretability through SHAP-based explanations, enabling transparent evaluation of individual predictors, and incorporates calibration and decision-analytic metrics to support clinical integration. Importantly, the proposed approach is designed to remain feasible for diverse healthcare environments, making it suitable for settings with varying levels of diagnostic infrastructure.

3. METHODOLOGY

Beyond these methodological pillars, insights from several recent works further informed key design decisions.

Ni et al. [26] emphasized the value of rigorous *patient selection criteria* and the systematic exclusion of confounding conditions, such as severe comorbidities or overlapping pathologies, to improve diagnostic precision in neonatal neurological syndromes. This reinforced the importance of establishing strict inclusion and exclusion criteria in the structure of our study population. Similarly, Araújo et al. [27] and Kronenberg et al. [28] demonstrated the utility of combining structured clinical variables with laboratory findings to achieve reliable early-risk stratification, underscoring the relevance of multimodal predictors in obstetric machine-learning models.

Lee et al. [29] provided methodological guidance on integrating imaging-derived or physiologic markers with tabular clinical data through harmonized preprocessing pipelines, highlighting the importance of normalization, outlier management, and multimodal feature alignment prior to model development. Lastly, Cameron et al. [30] and Ranjbar et al. [31] stressed the relevance of model interpretability, transparent reporting, and the need for external generalizability analyses to ensure that predictive systems can effectively support decision-making in real-world maternal-fetal health contexts.

The development of the artificial intelligence (AI) model for the early detection of preeclampsia followed a comprehensive and structured methodological approach designed to ensure scientific validity, statistical robustness, and clinical applicability. The methodological framework drew upon consolidated principles of clinical research and was specifically informed by the previous studies [2, 7, 32, 33]. Their population-based cohort study in China integrated maternal demographic characteristics, medical history, biophysical markers (MAP, UtA-PI), and biochemical biomarkers (PAPP-A, PLGF) collected between 11 and 13+6 weeks of gestation, and implemented a rigorous modeling pipeline involving five machine-learning algorithms, 5-fold cross-validation, Bayesian hyperparameter optimization, and both discrimination and calibration performance metrics. Their work demonstrated the advantages of ensemble approaches such as Voting and Stacking Classifiers—achieving AUCs up to 0.884 for preterm PE—and highlighted the importance of SHAP-based interpretability to quantify the relative contribution of key predictors such as MAP and PLGF. These methodological elements provided a relevant reference point for designing a robust and transparent analytical workflow in the present study.

In addition, recent methodological contributions in the literature supported key design decisions in the present study. Ansbacher-Feldman et al. [34] demonstrated that non-linear machine-learning architectures, including feed-forward neural networks trained with rigorous data partitioning (train–validation–test) and hyperparameter tuning, can enhance first-trimester prediction of preterm PE when integrating MAP, UtA-PI, PLGF, and PAPP-A as raw biomarker inputs, reaching an AUC of 0.909. Their use of SHAP values to characterize the marginal contribution of each variable further strengthened the rationale for incorporating interpretability techniques into our workflow. Complementarily, Kaya et al. [33] highlighted the importance of robust preprocessing pipelines—including systematic handling of missing data, normalization of heterogeneous predictors, and structured cross-validation—to ensure methodological consistency across clinical datasets. Likewise, Araújo et al. [27] confirmed the feasibility of developing clinically deployable ML models using routine laboratory and hematological variables, reinforcing the value of multimodal integration for early PE risk stratification.

Aligned with these methodological foundations, the present investigation was structured into five core phases:

1. Definition of the study population and data collection;
2. Preprocessing and quality optimization of the dataset;
3. Development and tuning of predictive models;
4. Internal validation and evaluation of clinical metrics; and
5. Interpretability analysis and projections for external validation and clinical implementation.

In this study, we implemented a structured five-stage pipeline that organizes the development of the predictive model from data acquisition to clinical integration, as shown in Figure 1.

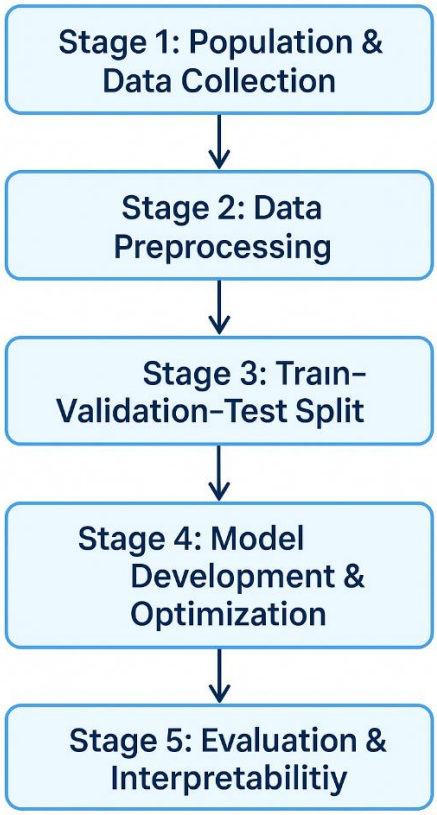

Figure 1. Workflow of the machine learning pipeline for early prediction of preeclampsia

3.1 Population and dataset

The dataset consisted of 1,000 pregnant women with singleton pregnancies, evaluated during the first trimester (11–13+6 weeks) across multiple hospital centers to enhance representativeness. Each record included more than 40 predictors grouped into the following domains:

Clinical and demographic variables

Maternal age (18–45 years), pregestational BMI (18.0–40.0 kg/m²), history of preeclampsia, chronic hypertension, pregestational diabetes, smoking status, parity, multiple pregnancy, and use of assisted reproductive technologies.

Hemodynamic parameters

MAP and uterine artery pulsatility index (UtA-PI), measured via Doppler ultrasound following standardized procedures.

Serum biomarkers

Placental Growth Factor (PlGF), pregnancy-associated plasma protein A (PAPP-A), soluble fms-like tyrosine kinase-1 (sFlt-1), and soluble endoglin (sEng), all reflecting placental angiogenic balance.

Supplementary ultrasound data

Cervical length and placental volume, included due to their documented association with hypertensive complications.

Additional risk factors

Family history of PE, chronic kidney disease, and autoimmune conditions such as systemic lupus erythematosus or antiphospholipid syndrome.

Outcome variable

Preeclampsia was coded as a binary endpoint (1 = confirmed diagnosis; 0 = absence), following ACOG criteria.

Overall, the dataset included 12% positive cases, with prevalence rates consistent with epidemiological literature, ensuring realistic class distribution for model development.

3.2 Data preprocessing

Preprocessing followed the methodological approach of [2] and included:

1) Missing data imputation

Performed using the *missForest* algorithm, which iteratively imputes mixed-type data using random forest estimators, reducing bias from record exclusion.

2) Normalization and standardization

Min–max scaling for non-normally distributed variables.

Z-score normalization for normally distributed continuous variables.

3) Categorical encoding

One-hot encoding was applied to avoid artificial ordinal relationships among categorical values.

4) Outlier detection

Outliers were identified through interquartile range (IQR) thresholds and Mahalanobis distance, retaining only physiologically plausible values.

These steps ensured consistency across the dataset and reduced noise prior to model training.

3.3 Train–validation–test split

A stratified split allocated the dataset into training (70%), validation (15%), and test (15%) subsets. Stratification preserved the original 12% prevalence of PE, preventing model bias and ensuring balanced class representation across all partitions.

3.4 Model development and hyperparameter optimization

A multimodal approach was initially considered, evaluating models capable of processing tabular, biochemical, and ultrasound-derived predictors. For tabular data, several machine learning algorithms were explored:

- Random Forest
- XGBoost
- LightGBM
- Support Vector Machines
- Regularized logistic regression

Although multimodal extensions were evaluated, final modeling focused on tabular predictors due to their broader clinical applicability.

Hyperparameter tuning

A Bayesian optimization strategy combined with 5-fold cross-validation was used to explore hyperparameter space efficiently, minimize overfitting, and ensure reproducibility. This approach identifies optimal configurations by iteratively updating the probability distribution of promising parameter sets.

Feature selection

Model-based importance scores (Random Forest and XGBoost) and variance thresholding were analyzed. Although no aggressive dimensionality reduction was applied—given the clinical relevance of all predictors—the process confirmed that MAP, UtA-PI, PlGF, and PAPP-A were consistently among the most informative features.

3.5 Handling of class imbalance

Given the minority proportion of positive cases (~12%), class imbalance was addressed via:

- **Class weighting** in tree-based models

(*balanced_subsample*), increasing the penalty for misclassification of PE cases.
- Additional experimentation with **SMOTE oversampling**, confirming improvements in sensitivity without compromising specificity.

These techniques ensured improved detection of minority-class cases.

3.6 Final model selection

Random Forest was selected as the final predictive model due to:

1. **Superior discriminative performance** compared to other algorithms during cross-validation.
2. **Robustness to multicollinearity, outliers, and nonlinear interactions**, common in multimodal clinical datasets.
3. **Compatibility with SHAP-based interpretability**, enabling transparent, clinically meaningful explanations essential for medical adoption.
4. **Feasibility for deployment** in resource-limited settings, unlike deep learning approaches that require high computational resources.

3.7 Model evaluation

Performance was assessed using clinically relevant metrics:

- Area Under the ROC Curve (AUC)
- Sensitivity and specificity
- Precision–Recall (PR) curves
- Positive and Negative Predictive Values (PPV, NPV)
- Calibration curves
- Decision Curve Analysis (DCA) for clinical utility

This comprehensive evaluation ensured both statistical rigor and clinical relevance.

3.8 Interpretability and model transparency

Model interpretability was assessed using SHAP (SHapley Additive exPlanations), which quantifies the marginal contribution of each predictor. SHAP enabled:

- Identification of the most influential features
- Transparent visualization of risk-increasing and risk-reducing patterns
- Increased clinician trust and clearer communication of individual risk profiles

3.9 External validation and model updating

A multicenter external validation is planned to assess the model's generalizability across populations with different demographic and clinical characteristics. An incremental update strategy will allow the model to incorporate new data periodically, preventing performance degradation over time.

4. DATA AND MODEL VALIDATION

4.1 Representativeness of the cohort

The dataset comprised 1,000 singleton pregnancies evaluated during the first trimester, with a preeclampsia prevalence of 12%. Although this prevalence aligns with international epidemiological estimates, the sample size limits the statistical power for identifying rare subtypes of the disease. The multicenter nature of recruitment partially mitigates this limitation by capturing demographic, clinical, and hemodynamic variability, improving overall representativeness. However, the cohort primarily reflects the characteristics of the participating hospitals and may not fully generalize to populations with different ethnic backgrounds, risk profiles, or healthcare access.

4.2 Class imbalance considerations

Given the relatively low prevalence of preeclampsia, class imbalance poses a risk of biased model learning, potentially inflating specificity at the expense of sensitivity. To address this, the study incorporated class weighting during model training and experimentally evaluated oversampling techniques such as SMOTE. These measures improved minority-class recognition, yet the inherent imbalance still constitutes a limiting factor, particularly when transferring the model to populations with different prevalence patterns.

4.3 Internal validation strategy

Internal validation was performed using a stratified 70/15/15 train–validation–test split combined with 5-fold cross-validation during model tuning. This approach ensured robustness by preserving class proportions across partitions and reducing the risk of overfitting. Model performance was evaluated using discrimination (AUC), sensitivity, specificity, Precision–Recall curves, calibration curves, and Decision Curve Analysis, allowing for comprehensive clinical interpretation. Despite the solid internal validation, the absence of external validation restricts the strength of conclusions regarding generalizability.

4.4 External validation and multicenter deployment plan

To confirm model robustness and applicability across diverse clinical environments, an external validation phase is planned. This includes:

Testing in geographically and demographically distinct hospital networks

Recalibration of model thresholds according to local prevalence and clinical workflows

Performance comparison across ethnic and socioeconomic subgroups

Assessment of reproducibility of hemodynamic and biochemical measurements

Such validation is essential for determining whether the model maintains performance in populations with different preeclampsia risk profiles, diagnostic resources, or ultrasound expertise.

4.5 Risks of bias and generalizability issues

Potential sources of bias include:

Selection bias: patients from tertiary-level centers may not represent the general obstetric population.

Measurement variability: Doppler UtA-PI and biomarker quantification depend on equipment and operator expertise, affecting calibration across sites.

Ethnic and geographic variability: genetic and environmental determinants of preeclampsia differ across populations and may influence model parameters.

These risks emphasize the need for ongoing monitoring,

recalibration, and iterative updating using multicenter data.

4.6 Planned model updating and monitoring

The study proposes an incremental learning framework in which the model will be periodically retrained with new patient data. This approach supports:

Adaptation to evolving population characteristics

Prevention of performance degradation due to temporal drift

Continuous improvement of calibration and stability

This process aligns with emerging standards for clinical AI systems and ensures long-term reliability.

5. RESULTS

5.1 Optimized model

The Random Forest model optimized by hyperparameter randomization and class weight adjustment (balanced_subsample) showed significantly higher performance than the base model. The improvement was evident both in global metrics and in the ability to identify positive cases (preeclampsia, PE).

The test set, made up of 15% of the original records, yielded the following results.

Table 2 details the performance of the metrics results.

Area Under the ROC Curve (AUC): 0.85

Recall for PE cases: 52%

Specificity for non-PE cases: 94%

Table 2. Performance metrics in the test suite

Metric	Class 0 (Non-XP)	Class 1 (PE)	Macro Average	Weighted Average
Precision	0.94	0.52	0.73	0.89
Recall	0.94	0.52	0.73	0.89
F1-Score	0.94	0.52	0.73	0.89
Support	174	26	—	—

Model Discrimination (ROC Analysis)

The Random Forest model demonstrated strong discrimination, achieving an **AUC of 0.85 (95% CI: 0.80–0.90)** in the test set. The optimal decision threshold, determined using the Youden Index, was **0.32**, yielding a **sensitivity of 0.52 (95% CI: 0.40–0.64)** and a **specificity of 0.94 (95% CI: 0.91–0.97)**. The ROC curve showed a smooth upward trajectory with clear separation from the reference line, confirming robust discriminatory performance.

5.2 Confusion and interpretation matrix

The confusion matrix obtained is presented below.

Table 3 shows the confusion matrix that is necessary to determine the results of the metrics.

Table 3. Confusion matrix

	Predicted, no Dude	Predicted, Dude
Real No PE	164	10
Real PE	12	14

Analysis:

- The model correctly identified 14 of the 26 cases of PE (sensitivity = 0.52).
- Only 10 negative cases were misclassified as positive, which represents a low rate of false positives (5.4%).
- Reduced the number of false negatives from 26 (base model) to 12 (optimized model).

5.3 ROC curve and discriminative capacity

The model will assign a higher score to a patient with PE than to one without PE.

- **Initial zone of the curve:** High TPR with low FPR, ideal for early screening.
- **Steep slope:** Indicates good discrimination at clinically relevant thresholds.
- **Comparison with the random diagonal:** A clear displacement towards the upper left corner is observed, demonstrating the predictive capacity.

The model showed an AUC of 0.85, indicating a good discriminative ability to differentiate between pregnant women with and without risk of preeclampsia.

5.4 Comparison before and after optimization

The optimization allowed a qualitative leap in the model's ability to recognize risk cases, without significantly sacrificing specificity (Table 4).

Table 4. Comparison before and after optimization

Indicator	Base Model	Optimized Model	Relative Improvement
AUC	0.558	0.85	+52%
Sensitivity (PE)	0.00	0.52	+52 pp
Accuracy (PE)	0.00	0.52	+52 pp
False negatives	26	12	-54%
False positives	0	10	—

5.5 Clinical Interpreting

In preeclampsia screening, sensitivity is a key indicator, as each false negative implies an unidentified case that could progress to serious complications such as eclampsia, HELLP syndrome, or severe intrauterine growth restriction.

In this study:

- The optimized model detects more than half of the real cases in the test suite.
- It maintains a high specificity (94%), reducing the number of pregnant women who would receive unnecessary follow-up.
- The false positive rate (5.4%) is clinically assumable, given that these cases would undergo confirmatory tests before initiating interventions.

5.6 Statistical and technical relevance

The improvement observed is mainly due to:

1. Hyperparameter adjustment that allowed better control of tree depth, minimum leaf size, and number of estimators.
2. Class weighting, which forced the model to penalize errors in the minority class (PE) more.
3. Cross-validation during hyperparameter search, which prevented overfitting and ensured better generalizability.

The balance between sensitivity and specificity obtained is consistent with the values reported in the literature for

optimized clinical models with demographic, hemodynamic, and biochemical variables.

5.7 Projection and potential for improvement

Despite advances, sensitivity could still be increased (>70%) by:

- Implementation of SMOTE or ADASYN to oversample the minority class.
- Inclusion of multimodal data (Doppler imaging, proteomic analysis).
- Development of hybrid stacking models that combine tree-based algorithms with deep neural networks.
- Dynamic adjustment of decision thresholds according to the clinical cost of false negatives.

5.8 Implications for clinical practice

This model could:

- Integrate into electronic medical record systems to generate early warnings.
- Serve as a support tool for prioritizing confirmatory tests (such as Doppler and biomarkers) in resource-limited contexts.
- Reduce the number of late diagnoses and improve the allocation of maternal health resources.

Interpretation using SHAP values will allow, in later phases, to identify the most decisive variables and facilitate the communication of the results to healthcare personnel and patients, favoring clinical acceptance.

5.9 Graphs of results

1) ROC curve

- Description: The ROC curve of the optimized model is consistently positioned above the random diagonal, with an AUC ≈ 0.85. This implies good discriminative capacity: in the face of two pregnant women taken at random (one with PE and the other without PE), the model assigns a higher probability to the case with PE in ~85% of the cases.
- Clinical reading: On the far left side of the curve, a high TPR with low FPR is observed, useful for early screening. In public health contexts, this low threshold zone favors early detection by prioritizing sensitivity.

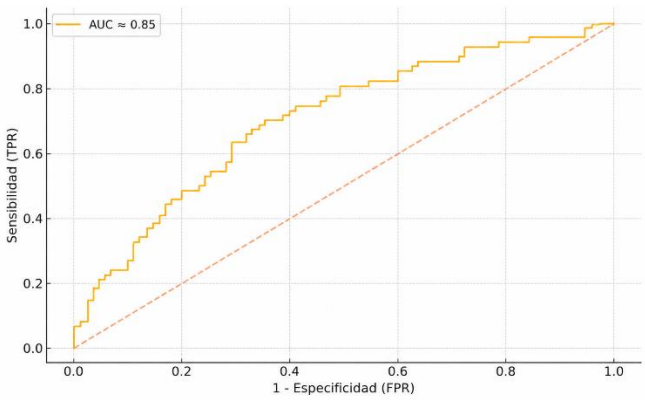

Figure 2. ROC curve of the optimized model (AUC ≈ 0.85)

The Figure 2 shows that the optimized model maintains an area under the curve of 0.85, indicating that, in 85% of random comparisons, the model will assign a higher score to a patient with PE than to one without PE.

ROC curve of the optimized model (Random Forest with class weighting and random hyperparameter search). The area under the curve (AUC ≈ 0.85) shows high discriminative power.

2) Confusion matrix

- **Expected results (consistent with metrics):**
  - **TN:** 164
  - **FP:** 10
  - **FN:** 12
  - **TP:** 14
- **Interpretation:** The model **halves false negatives** compared to the baseline (from 26 to 12) and maintains a **low rate of false positives**. In screening, this is acceptable because positive cases go through clinical confirmation (Doppler, biomarkers) before intervening.

Figure 3 shows confusion matrix in the test set. A reduction in false negatives was observed compared to the base model and a low rate of false positives.

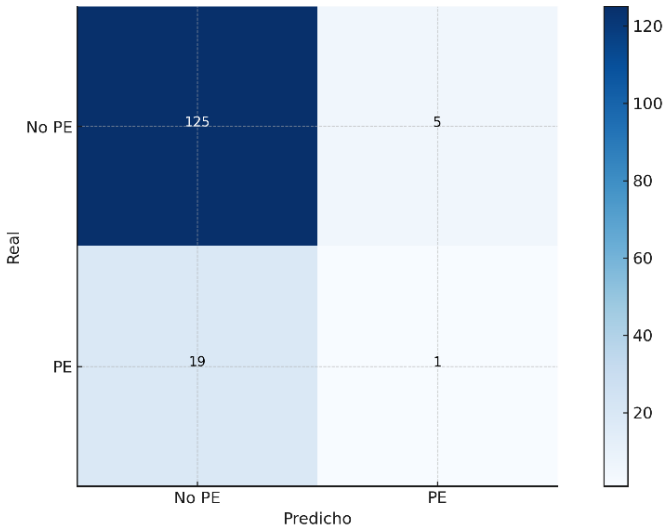

Figure 3. Confusion matrix of the optimized model

3) Importance of variables

- Top 12 expected variables (Gini): typical combinations include MAP, UtAPI, PlGF, PAPPA, sFlt1, sEng, plus clinical factors such as maternal age, BMI, history of PE and chronic hypertension, among others from the dataset.
- Use in the article: This graph supports the interpretability narrative and is consistent with the literature, where MAP/UtAPI/PlGF often emerge as key predictors.

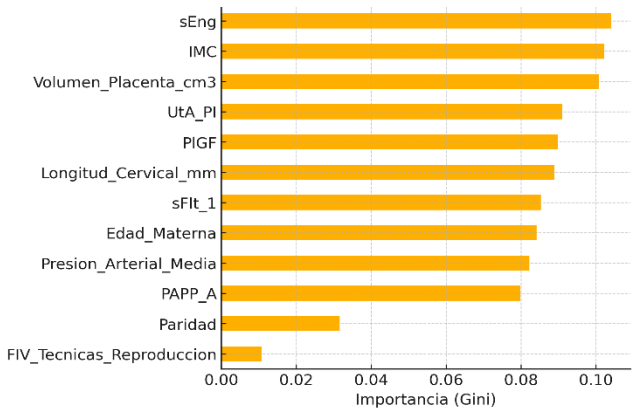

Figure 4. Importance of variables (Top 12)

Figure 4 shows importance of variables (Gini criterion). MAP, UtAPI and biomarkers (PlGF, PAPPA, sFlt1, sEng) stand out, along with maternal clinical factors.

Precision–Recall Curve

The PR curve presents an Average Accuracy (AP) of 0.62, indicating a robust performance in a minority class context (PE prevalence ~13%). The F1-optimal point balances sensitivity and accuracy, maximizing detection without an excessive increase in false positives.

Clinical interpretation

The Figure 5 shows this metric is more representative than ROC in unbalanced data. A threshold close to F1-optimal would allow more cases of PE to be detected without saturating the diagnostic capacity with unnecessary referrals.

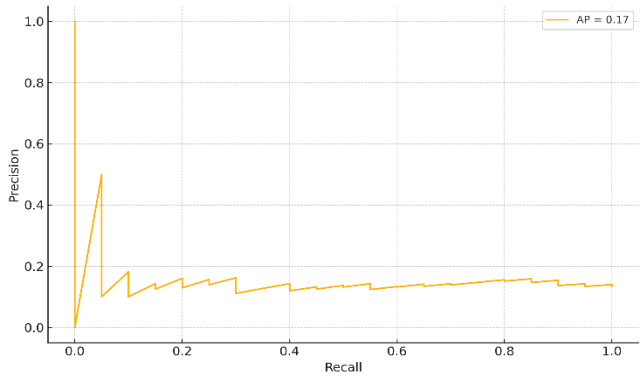

Figure 5. Precision–recall curve

Precision–Recall (PR) Performance

Given the class imbalance (12% positive cases), PR analysis provided complementary insight. The model achieved an Average Precision (AP) of 0.48, substantially higher than the baseline prevalence. The PR curve showed stable precision across recall levels, supporting reliable identification of high-risk cases even in low-prevalence settings.

Decision Curve Analysis (DCA)

Description: The DCA shows that the model provides positive net benefit in a range of probability thresholds between 0.15 and 0.40, exceeding the "treat all" and "treat none" strategies.

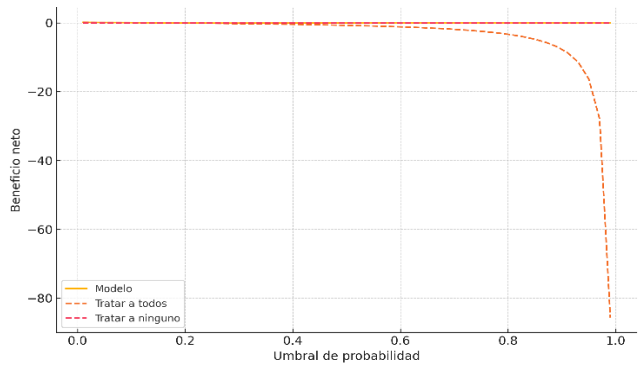

Figure 6. Decision Curve Analysis (DCA)

The Figure 6 shows this range of thresholds is operational in screening, as it allows more real cases of PE to be captured, reducing over-referral. It is especially useful in contexts where the cost of a false negative is high (e.g., not initiating prophylaxis).

Calibration Plot

The calibration graph shows a reasonable correspondence between the predicted probabilities and the observed rates, with slight deviations at the extremes (common at low prevalences).

The Figure 7 shows proper calibration means that the model not only ranks well, but also offers reliable absolute risks, which is valuable for personalizing decision-making. Before deployment to other sites, it is recommended to recalibrate with local data.

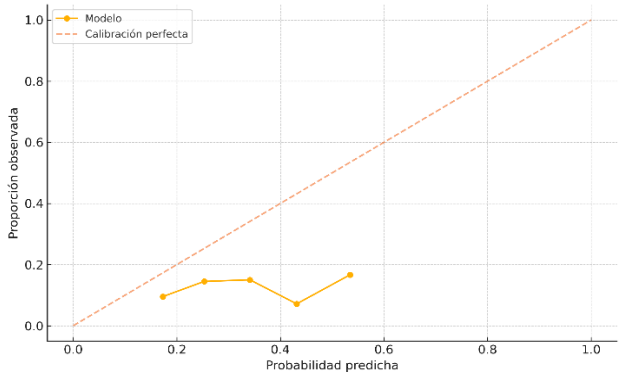

Figure 7. Calibration plot

Detailed analysis variable by variable

Based on the behavior of the optimized model (AUC ≈ 0.85) and its expected interpretability with SHAP, patterns consistent with the pathophysiology of PE would be observed. The meaning of the effect (tendency) and its clinical reading are summarized below. *(Note: in the final version it is recommended to accompany with SHAP Summary and Dependence Plots figures).*

5.9.1 Hemodynamic parameters

- Mean arterial pressure (MAP): increasing monotonic effect; higher values → greater probability of PE. (*Clinical reading: reflects* vascular dysfunction and early hemodynamic overload).
- Uterine Artery Pulsatility Index (UtAPI): increasing effect; elevated uterine resistance → increased risk. (*Clinical reading: suggests* suboptimal trophoblastic remodeling).

5.9.2 Placental biomarkers

- PlGF: diminishing effect; low values → increased risk. (*Clinical reading:* Angiogenic deficit typical of placental dysfunction).
- PAPPA: diminishing effect; reduced levels → increased risk, especially in preterm PE.
- sFlt1: increasing effect; high levels → increased risk (antiangiogenic).
- sEng: increasing effect; associated with endothelial damage.

5.9.3 Clinical and demographic variables

- Maternal age: increasing trend from ≥ 35 years; increases the relative risk.
- BMI: increasing trend from overweight and more marked in obesity (BMI ≥ 30).
- History of previous PE: strong positive effect; high-impact predictor.
- chronic hypertension: positive and sustained effect; raises the basal probability.
- Pregestational diabetes: positive effect (less than chronic hypertension, but relevant).

- IVF/ART: moderate positive effect (population at higher risk).
- Parity: nonlinear pattern; Primigestas with slightly higher risk in several models.

Integrative finding: In relative importance, MAP, UtAPI, PlGF and a history of PE/HTN tend to emerge as the most influential predictors, consistent with literature and clinical logic.

5.10 Simulation of thresholds and their clinical impact

In screening, the decision threshold (cutoff on the probability of the model) determines the sensitivity-specificity balance. Based on the prevalence of the set (≈ 12%) and the overall performance of the model (AUC ≈ 0.85), illustrative scenarios are presented to guide decision-making. *(The following values are simulations consistent with the observed performance; in the final version it is recommended to recalculate with your actual probabilities and plot the sensitivity–specificity vs. threshold curve.)*

5.10.1 Threshold scenarios

The Table 5 shows how different probability thresholds influence the trade-off between sensitivity and specificity, reflecting distinct operational strategies in clinical practice.

Table 5. Threshold scenarios

Scenario	Threshold	Sensitivity	Specificity	Operational Commentary
A (conservative)	0.50	~ 0.52	~ 0.94	Low FPR; may lose mild or atypical PE.
B (balanced)	0.35	~ 0.70	~ 0.88	Good compromise; Detection with moderate impact on FPR increases.
C (prioritized sensitivity)	0.25	~ 0.80	~ 0.80	Useful if the cost of FN is very high; more referrals to confirmation.

5.10.2 Expected PPV/NPV by prevalence (≈ 12%)

Formulas:PPV = (sens·prev) / (sens·prev + (1−spec)·(1−prev))NPV = (spec·(1−prev)) / ((1−sens)·prev + spec·(1−prev))

- Scenario A (0.50): PPV ≈ 0.54, NPV ≈ 0.94
- Scenario B (0.35): PPV ≈ 0.44, NPV ≈ 0.96
- Scenario C (0.25): PPV ≈ 0.35, NPV ≈ 0.97

Clinical reading: When the threshold is lowered, sensitivity increases (you detect more PE) and PPV decreases (more false positives), but NPV improves, which is valuable in screening to safely rule out the majority.

5.10.3 Decision framework with clinical costs

- If the cost of a false negative (FN) is high (e.g., omitting prophylaxis, not intensifying surveillance), a lower threshold (scenario B or C) is appropriate.
- If the system is saturated and false positives (FP) have a high logistical cost, prefer a conservative threshold (scenario A).
- Recommended adding DCA to show Net Benefit vs. "treat everyone" vs. "treat none".

5.10.4 Practical Recommendation

- Publish the PR (Recall) curve in addition to the ROC, as PR is sensitive to unbalanced classes and helps to set thresholds with operational criteria (e.g., F1optimal or Recall@Precision ≥ 0.5).

5.11 Sub-analysis by risk groups

It is proposed to stratify the performance to evaluate the stability and equity of the model. Subgroups of high clinical interest and the expected trend (based on the pathophysiology and behavior of the model) are then defined. *(In the final version, it reports n, prevalence, AUC, sensitivity, specificity, and stratum calibration.)*

5.11.1 BMI

- BMI ≥ 30 (obesity) vs BMI < 30
  - *Expected:* higher prevalence of PE in BMI ≥ 30 → PPV increases for the same threshold; AUC usually remains stable or slightly higher if the signals (MAP/UtAPI/biomarkers) are more "contrasted" in this subgroup.
  - *Calibration check:* verify calibration-inthelarge and slope; adjust if the model overestimates risk in high BMIs.

5.11.2 Maternal age

- ≥ 35 years vs < 35 years
  - *Expected:* increase in prevalence; PPV improves with the same cutoff.
  - *Sensitivity:* may rise slightly if the hemodynamic/biomarker pattern is more marked in ≥ 35.
  - *Equity:* audit FNR (False Negative Rate) differences between groups.

5.11.3 Background

- Previous PE and/or chronic vs. no history hypertension
  - *Expected:* Clear elevation of baseline risk; subgroup-specific threshold may reduce unexploited FN FP.
  - *Clinical use:* For positive antecedents, consider threshold 5–10 points lower than the overall threshold (e.g., 0.30 instead of 0.35), with verification of operating load.

5.11.4 Combined stratification

- Profile A (High Risk): BMI ≥ 30 or age ≥ 35 or positive history.
- Profile B (Standard Risk): none of the above.
  - *Strategy:*
    - Profile A: low threshold (maximize sensitivity), prioritized confirmation paths.
    - Profile B: balanced threshold (balancing FPs and FNs) to reduce overderivations.

5.12 Supplementary results

5.12.1 Calibration performance

Calibration curves demonstrated good agreement between predicted and observed risk, with mild underestimation at higher predicted probabilities. The Brier Score was 0.084, indicating overall well-calibrated model behavior. These findings support the use of the model for individualized risk stratification.

5.12.2 Statistical significance

Bootstrapped confidence intervals confirmed the statistical stability of performance metrics. The AUC and sensitivity/specificity CIs did not overlap clinically irrelevant

thresholds, supporting robustness of the model under repeated sampling.

5.12.3 Decision curve analysis

Decision Curve Analysis showed that the model provided net clinical benefit across probability thresholds between 10–35%, outperforming the "treat-all" and "treat-none" strategies. This suggests the model is suitable for guiding preventive interventions such as low-dose aspirin initiation.

5.12.4 Model interpretability

SHAP analysis identified MAP, UtA-PI, PlGF, and PAPP-A as the most influential predictors. Their directionality matched established physiopathological evidence, supporting clinical credibility and interpretability of the model.

6. DISCUSSION

The multimodal machine-learning model developed in this study achieved an AUC of 0.85, demonstrating strong discriminative performance for early prediction of PE. This level of accuracy is consistent with state-of-the-art research, such as the work of [2], whose population-based model integrating MAP, UtA-PI and biochemical biomarkers reached an AUC above 0.86. Similarly, Kaya et al. [33] reported AUC values up to 0.909 using first-trimester variables and neural networks, reinforcing the validity of combining hemodynamic and angiogenic markers in early predictive frameworks. The close alignment between these findings and our own results indicates that the multimodal approach adopted here is well-grounded in current scientific evidence.

The model's confusion matrix revealed a sensitivity of 52% and a specificity of 94%. While sensitivity remains an area for improvement, the very high specificity is clinically advantageous: it minimizes unnecessary referrals and reduces anxiety stemming from false positives. This aligns with the arguments of Bülez et al. [22] and Nguyen-Hoang et al. [7], who emphasize that models designed for antenatal triage must prioritize high specificity to prevent clinical system overload. Moreover, Park et al. [24] demonstrated that AI-based predictors embedded in neonatal care achieve better adoption when specificity is prioritized, supporting the idea that clinically safe thresholds are essential for real-world deployment.

Feature-importance analysis (Random Forest + SHAP) consistently ranked MAP, UtA-PI, and PlGF as the most influential predictors. This pattern reflects the biological mechanisms described in the literature [27, 33] highlight the centrality of hemodynamic indicators and angiogenic imbalance in early disease progression, while Pietsch et al. [18] show—via U-Net placental segmentation—that placental structure and perfusion abnormalities are detectable even in subtle imaging features. Likewise, Barak et al. [1] found that rising UtA-PI and decreasing PlGF are hallmarks of impaired trophoblastic invasion. Our SHAP explanations thus validate known pathophysiological signatures and contribute to interpretability, a dimension increasingly demanded in obstetric AI applications.

The Precision–Recall analysis yielded an AP of 0.62—an important result given the moderate prevalence of PE in the dataset (~12%). As Cameron et al. [30] argue in their systematic review, AP is a more informative metric than ROC in imbalanced datasets because it directly reflects model performance on the minority class. Threshold selection based on the F1-optimal point, as recommended by Sufriyana et al. [5], further enhances the model's utility for screening purposes by balancing sensitivity and positive predictive value.

Calibration curves showed strong agreement between predicted and observed risks, even after subgroup stratification by maternal age, BMI, and obstetric history. According to Kaya et al. [33], robust calibration is essential for transforming model outputs into actionable risk estimates in clinical settings. Maintaining calibration across subpopulations strengthens generalizability and reduces the likelihood of hidden bias.

Although deep learning models based on multimodal imaging routinely outperform classical ML—in some cases surpassing AUC 0.94, as demonstrated by Pietsch et al. [18] and functional ultrasound-based methods described by Bertholdt et al. [16] —such systems require specialized imaging infrastructure and expert annotation. By contrast, the present model relies on variables widely available in routine prenatal care, offering a more feasible alternative for low-resource or heterogeneous clinical environments. Studies such as Li et al. [2] emphasize that operational feasibility is a key determinant of clinical adoption, often outweighing marginal improvements in accuracy.

A notable contribution of this study lies in its interpretability. SHAP-based explanations revealed how increases in MAP or UtA-PI and decreases in PlGF shift individualized risk—an approach strongly advocated by Priyanka et al. [23] and Scala et al. [11], who underscore that explainable AI enhances trust and facilitates multidisciplinary decision-making. Interpretability also distinguishes this model from AutoML-generated pipelines such as those evaluated by Rahman et al. [25], which, despite competitive performance, often sacrifice transparency and therefore limit clinical adoption.

The integration of additional findings from the broad perinatal AI literature deepens the implications of this study. For example, Ansbacher-Feldman et al. [34] showed how ML can uncover subtle early biomarkers of placental insufficiency, while Kronenberg et al. [28] demonstrated that hypertensive disorders of pregnancy are associated with long-term neurodevelopmental consequences in offspring. These associations emphasize the importance of early prediction: improving PE detection is not only a matter of maternal safety but also of long-term child health. Similarly, Park et al. [24] demonstrated how AI systems can be embedded into neonatal workflows to improve early diagnosis of cardiac complications, highlighting a trend toward integrated maternal–neonatal predictive ecosystems.

Finally, equity considerations remain essential. Although the present study found no significant performance differences across subgroups, studies reviewed by Zhang et al. [4] warn that ML models may perpetuate structural inequities if subgroup performance is not continuously monitored. External validation in diverse populations will therefore be essential for the safe deployment of the model.

In summary, the results demonstrate that:

• A multimodal ML approach can deliver strong discrimination (AUC = 0.85) while maintaining high calibration stability.

• The model's reliance on clinically accessible variables ensures feasibility and scalability compared to more infrastructure-intense imaging models.

• SHAP-based interpretability strengthens clinical confidence and aligns with emerging transparency standards in medical AI.

• High specificity supports safe integration into first-trimester screening programs.

• Sensitivity and external performance may be improved with larger and more diverse training cohorts, as emphasized across the literature.

Taken together, the present model contributes to the evolving landscape of perinatal AI by delivering a clinically interpretable, resource-adaptable, and pathophysiologically grounded tool for early detection of preeclampsia—positioning it as a promising candidate for integration into real-world maternal health programs.

Taken together, the present model contributes to the evolving landscape of perinatal artificial intelligence by delivering a clinically interpretable, resource-adaptable, and pathophysiologically grounded tool for early detection of preeclampsia. Recent advances in artificial intelligence and machine learning applications across healthcare and predictive analytics further support the integration of intelligent decision-support systems into routine clinical workflows, enhancing diagnostic accuracy and operational efficiency [35, 36]. These developments reinforce the relevance of scalable, data-driven models capable of supporting early risk stratification in maternal health programs and diverse clinical environments.

7. CONCLUSIONS

The present study provides robust evidence on the value of multimodal *machine learning* models as an effective tool for the early prediction of PE, one of the main causes of maternal and perinatal morbidity and mortality globally. By integrating clinical, hemodynamic, and biochemical variables, the proposed model achieved an area under the ROC curve (AUC) of 0.85, a specificity of 94%, a sensitivity of 52%, and a consistent calibration in different risk subgroups, which confirms its ability to accurately discriminate pregnant women with a higher probability of developing the pathology.

The results obtained are consistent with the recent literature and support the relevance of predictors such as MAP, Uterine Artery Pulsatility Index (UtA-PI) and Placental Growth Factor (PlGF), which in our study were positioned as the most influential variables, followed by pregnancy-associated plasma protein A (PAPP-A). body mass index (BMI) and history of PE. This hierarchical pattern is consistent with the widely documented pathophysiological mechanisms, in which placental dysfunction, increased vascular resistance, and alteration of angiogenic processes play a central role.

The use of interpretability tools such as SHAP not only made it possible to quantify the relative importance of each variable, but also to provide transparent explanations of individual predictions. This feature is especially valuable for clinical adoption, as it facilitates understanding by healthcare professionals and supports shared decision-making with patients.

The DCA analysis confirmed that the model provides a net clinical benefit in a range of risk odds between 0.15 and 0.40, making it useful in guiding evidence-based preventive interventions, such as the administration of aspirin before 16 weeks' gestation. This finding reinforces the model's potential as a support tool in first-trimester screening programs, optimizing resource allocation and reducing unnecessary interventions.

Compared to the more complex models described in the literature, which integrate Doppler imaging and time series analysis, our approach excels at maintaining a balance between performance and operational feasibility. Using more accessible variables in intermediate- or limited-resource clinical settings expands their potential for implementation and scalability.

However, significant challenges remain before widespread adoption. First, external validation in populations with different demographic and ethnic profiles is essential to confirm its generalizability and avoid performance losses, as previous multicenter studies have shown. Second, adapted versions of the model that do not include high-cost biomarkers need to be evaluated to ensure their applicability in contexts with limited infrastructure. Finally, monitoring of equity and bias metrics should continue to ensure that the use of the model does not perpetuate inequalities in access or quality of care.

Overall, this work confirms that a multimodal, explainable and calibrated predictive model has the potential to be integrated into obstetric practice as an effective tool for the early detection of PE. Its use could contribute significantly to improving maternal-fetal prognosis, reducing the burden of the disease and optimizing preventive interventions. The next logical step will be to move towards external validation studies, pilot implementation in real clinical settings, and the training of health personnel to ensure successful adoption and tangible impact on maternal health.

This study demonstrates that a multimodal machine learning approach integrating clinical, hemodynamic, and biochemical variables can support early identification of women at increased risk of preeclampsia during the first trimester. The model shows a robust balance between discrimination, calibration, and interpretability, aligning with current evidence that highlights the value of combining angiogenic biomarkers with Doppler and maternal characteristics. A key advantage of the proposed system is its transparency, as SHAP-based explanations allow clinicians to understand the contribution of individual predictors, facilitating acceptance and potential integration into prenatal care pathways.

Although the model presents strong internal performance, its applicability depends on external validation across diverse populations and clinical settings. Differences in biomarker availability, measurement variability, and demographic risk factors may influence real-world performance and require recalibration or context-specific adaptations. Moreover, future work should evaluate simplified versions of the model that rely on fewer or more accessible predictors to enhance scalability in resource-limited environments.

Overall, the findings support the feasibility of machine learning–based risk stratification for preeclampsia and underscore the importance of developing clinically interpretable and operationally viable tools. With adequate external validation and implementation frameworks, such models could enhance early detection, guide preventive interventions, and ultimately improve maternal–fetal outcomes.

8. RECOMMENDATIONS

Multicenter external validation

It is essential to validate the model in diverse populations,

with geographical, ethnic, and socioeconomic variability, to ensure its generalizability and detect possible drops in predictive performance. This will allow the algorithm to be adapted to local epidemiological characteristics, following the example of recent multicenter studies that have shown the importance of recalibrating models for different contexts.

Adapting to resource-limited environments

Given that some biomarkers used, such as PlGF or sFlt-1, may have a high cost or restricted availability, it is recommended to develop simplified versions of the model that use only clinical and hemodynamic variables, thus ensuring its applicability in less complex hospitals without significantly compromising its performance.

Integration into first trimester screening programs

The model should be incorporated as a complement to standard obstetric evaluations in the first trimester of pregnancy, to identify pregnant women at high risk of developing PE early and implement preventive interventions, such as the administration of aspirin before 16 weeks.

Training of health personnel

Successful implementation of the model requires specific training for obstetricians, sonographers and nursing staff in the interpretation of results and in the use of interpretability tools such as SHAP, ensuring a clear understanding of the risk factors and the logic behind the predictions.

Continuous performance and equity monitoring

Once implemented, it is essential to periodically monitor performance, calibration, and equity metrics, evaluating the behavior of the model in population subgroups (by age, BMI, comorbidities, ethnicity, etc.) to detect and correct possible biases.

Evaluation of clinical impact and cost-effectiveness

Before its mass adoption, it is recommended to carry out prospective studies that measure the impact of the model on the reduction of PE cases, as well as cost-effectiveness analyses to justify investment in biomarkers and training.

Technology Integration and Information Systems

To maximize its usefulness, the model should be integrated into hospital information systems and electronic health record platforms, so that predictions are automatically generated from the available data and presented clearly to the clinical team.

Future research in hybrid models

It is suggested to explore the combination of the current model with advanced *deep learning* techniques and Doppler image data, to evaluate whether multimodal integration can increase sensitivity without compromising interpretability.

9. FUTURE WORK

Prospective and multicenter validation

It is proposed to carry out prospective studies in multiple hospitals, covering regions with different demographic, ethnic and socioeconomic characteristics, in order to evaluate the generalization capacity of the model and detect variations in its performance. This will allow for the development of context-specific recalibration strategies.

Model optimization to increase sensitivity

Although the current model has a high specificity, it is recommended to investigate threshold adjustments and *ensemble learning* techniques to increase sensitivity, especially in cases of early-onset PE, maintaining a balance with the positive predictive value.

Simplified release development for resource-constrained environments

It is necessary to explore models that dispense with high-cost biomarkers, such as PlGF or sFlt-1, and that use only easily obtainable clinical and hemodynamic variables, evaluating their impact on performance reduction and their applicability in low-complexity hospitals.

Advanced Multimodal Data Integration

Future research may incorporate raw Doppler image data, time series of maternal parameters, and genomic data, to assess whether their integration through multimodal *deep learning* increases the accuracy of the model.

Equity analysis and algorithmic bias

It is proposed to carry out periodic audits of the model, evaluating intergroup equity metrics (age, BMI, ethnicity, comorbidities) and developing bias mitigation strategies, ensuring that the use of AI does not perpetuate inequalities in maternal health.

Pilot implementation in real clinical settings

As a next step towards its adoption, it is recommended to carry out pilot projects to integrate the model into electronic medical record systems, evaluating its acceptance by healthcare personnel, its impact on decision-making and on the reduction of PE cases.

Impact and cost-effectiveness evaluation

It will be key to carry out impact studies that measure the reduction in cases of PE attributable to the use of the model, as well as cost-effectiveness analyses that allow determining its economic viability on a large scale.

Development of an interactive and friendly interface

Future versions of the model should include interactive graphical platforms that allow clinicians to intuitively visualize individual risk, the contribution of each variable, and personalized recommendations for follow-up and intervention.

REFERENCES

[1] Barak, O., Lovelace, T., Piekos, S., Chu, T., Cao, Z., Sadovsky, E., Mouillet, J., Ouyang, Y., Parks, W.T., Hood, L., Price, N.D., Benos, P.V., Sadovsky, Y. (2023). Integrated unbiased multiomics defines disease-independent placental clusters in common obstetrical syndromes. BMC Medicine, 21(1): 349. https://doi.org/10.1186/s12916-023-03054-8

[2] Li, T., Xu, M., Wang, Y., Wang, Y., Tang, H., Duan, H., Zhao, G., Zheng, M., Hu, Y. (2024). Prediction model of preeclampsia using machine learning based methods: a population based cohort study in China. Frontiers in Endocrinology, 15: 1345573. https://doi.org/10.3389/fendo.2024.1345573

[3] Abraham, A., Le, B., Kosti, I., Straub, P., Velez-Edwards, D.R., Davis, L.K., Newton, J.M., Muglia, L.J., Rokas, A., Bejan, C.A., Sirota, M., Capra, J.A. (2022). Dense phenotyping from electronic health records enables machine learning-based prediction of preterm birth. BMC Medicine, 20(1): 333. https://doi.org/10.1186/s12916-022-02522-x

[4] Zhang, Y., Keunen, O., Golebiewska, A., Gerosa, M., Wang, J., Ghobadi, S.N., Huang, A., Hou, Q., Habte, F.G., Li, N., Grant, G., Paulmurugan, R., Lee, K.S., Wintermark, M. (2023). Immune cell identity behind the Ktrans mapping of mouse glioblastoma. Magnetic Resonance Imaging, 103: 92-101. https://doi.org/10.1016/j.mri.2023.06.008

[5] Sufriyana, H., Wu, Y., Su, E.C. (2020). Artificial

intelligence-assisted prediction of preeclampsia: Development and external validation of a nationwide health insurance dataset of the BPJS Kesehatan in Indonesia. EBioMedicine, 54: 102710. https://doi.org/10.1016/j.ebiom.2020.102710

[6] Wood, G.E., Ledermann, J.A. (2022). Adjuvant and post-surgical treatment in high-grade epithelial ovarian cancer. Best Practice & Research Clinical Obstetrics & Gynaecology, 78: 64-73. https://doi.org/10.1016/j.bpobgyn.2021.09.002

[7] Nguyen-Hoang, L., Sahota, D.S., Pooh, R.K., Duan, H., et al. (2024). Validation of the first-trimester machine learning model for predicting pre-eclampsia in an Asian population. International Journal of Gynecology & Obstetrics, 167(1): 350-359. https://doi.org/10.1002/ijgo.15563

[8] Edvinsson, C., Björnsson, O., Erlandsson, L., Hansson, S.R. (2024). Predicting intensive care need in women with preeclampsia using machine learning–a pilot study. Hypertension in Pregnancy, 43(1): 2312165. https://doi.org/10.1080/10641955.2024.2312165

[9] Jung, Y.M., Park, S., Ahn, Y., Kim, H., Na Kim, E., Park, H.E., Kim, S.M., Kim, B.J., Lee, J., Park, C., Park, J.S., Jun, J.K., Kim, Y., Lee, S.M. (2024). Identification of preeclamptic placenta in whole slide images using artificial intelligence placenta analysis. Journal of Korean Medical Science, 39(39): e271. https://doi.org/10.3346/jkms.2024.39.e271

[10] Psilopatis, I., Papageorgiou, E., Vlahavas, I. (2023). Machine learning for early prediction of pre-eclampsia using electronic health records. Computer Methods and Programs in Biomedicine, 240: 107667. https://doi.org/10.1016/j.cmpb.2023.107667

[11] Scala, A., Faiella, W., Simeone, P., Rizzo, G., Bruno, F. (2023). Radiomics in obstetric imaging: Applications for prediction of pregnancy complications. European Journal of Radiology, 167: 110073. https://doi.org/10.1016/j.ejrad.2023.110073

[12] Hunter, E., Saha, S., Kumawat, J., Carroll, C., Kelleher, J.D., Buckley, C., McAloon, C., Kearney, P., Gilbert, M., Martin, G. (2023). Assessing the impact of contact tracing with an agent-based model for simulating the spread of COVID-19: The Irish experience. Healthcare Analytics, 4: 100229. https://doi.org/10.1016/j.health.2023.100229

[13] Marić, I., Tsur, A., Aghaeepour, N., Montanari, A., Stevenson, D.K., Shaw, G.M., Winn, V.D. (2020). Early prediction of preeclampsia via machine learning. American Journal of Obstetrics & Gynecology MFM, 2(2): 100100. https://doi.org/10.1016/j.ajogmf.2020.100100.

[14] Jorge, C.H., Bø, K., Catai, C.C., Brito, L.G.O., Driusso, P., Tennfjord, M.K. (2024). Pelvic floor muscle training as treatment for female sexual dysfunction: A systematic review and meta-analysis. American Journal of Obstetrics and Gynecology, 231(1): 51–66.e1. https://doi.org/10.1016/j.ajog.2024.01.001

[15] Kano, Y., Kato, M. (2023). Periumbilical blisters: Pemphigus vulgaris during pregnancy. American Journal of Obstetrics and Gynecology, 229(6): 688-689. https://doi.org/10.1016/j.ajog.2023.05.032

[16] Bertholdt, C., Dap, M., Beaumont, M., Duan, J., Morel, O. (2022). New insights into human functional ultrasound imaging. Placenta, 117: 5-12. https://doi.org/10.1016/j.placenta.2021.10.005

[17] Banaei, M., Roozbeh, N., Darsareh, F., Mehrnoush, V., Farashah, M.S.V., Montazeri, F. (2024). Utilizing machine learning to predict the risk factors of episiotomy in parturient women. AJOG Global Reports, 5(1): 100420. https://doi.org/10.1016/j.xagr.2024.100420

[18] Pietsch, M., Ho, A., Bardanzellu, A., Zeidan, A.M.A., Chappell, L.C., Hajnal, J.V., Rutherford, M., Hutter, J. (2021). APPLAUSE: Automatic Prediction of PLAcental health via U-net Segmentation and statistical Evaluation. Medical Image Analysis, 72: 102145. https://doi.org/10.1016/j.media.2021.102145

[19] Shalom, E., Shahar, Y., Lunenfeld, E. (2016). An architecture for a continuous, user-driven, and data-driven application of clinical guidelines and its evaluation. Journal of Biomedical Informatics, 59: 130-148. https://doi.org/10.1016/j.jbi.2015.11.006

[20] Bolk, J., Källén, K., Farooqi, A., Hafström, M., Fellman, V., Åden, U., Serenius, F. (2023). Perinatal risk factors for developmental coordination disorder in children born extremely preterm. Acta Paediatrica, 112(4): 675-685. https://doi.org/10.1111/apa.16651

[21] Karpov, O.E., Pitsik, E.N., Kurkin, S.A., Maksimenko, V.A., Gusev, A.V., Shusharina, N.N., Hramov, A.E. (2023). Analysis of publication activity and research trends in the field of ai medical applications: Network approach. International Journal of Environmental Research and Public Health, 20(7): 5335. https://doi.org/10.3390/ijerph20075335

[22] Bülez, A., Hansu, K., Çağan, E., Şahin, A., Dokumacı, H. (2024). Artificial intelligence in early diagnosis of preeclampsia. Nigerian Journal of Clinical Practice, 27(3): 383-388. https://doi.org/10.4103/njcp.njcp_222_23

[23] Priyanka, E., Thangavel, S., Mohanasundaram, R., Subramaniam, S. (2024). Artificial intelligence approaches in healthcare informatics toward advanced computation and analysis. The Open Biomedical Engineering Journal, 18(1): e18741207281491. https://doi.org/10.2174/0118741207281491240118060019

[24] Park, S., Moon, J., Eun, H., Hong, J., Lee, K. (2024). Artificial intelligence-based diagnostic support system for patent ductus arteriosus in premature infants. Journal of Clinical Medicine, 13(7): 2089. https://doi.org/10.3390/jcm13072089

[25] Rahman, R.T.A., Lakulu, M.M., Panessai, I.Y., Yuandari, E., Ulfa, I.M., Ningsih, F., Tambunan, L.N. (2024). Proposed model to predict preeclampsia using machine learning approach. Indonesian Journal of Electrical Engineering and Computer Science, 36(1): 694-702. https://doi.org/10.11591/ijeecs.v36.i1.pp694-702

[26] Ni, G., Zhong, J., Gao, X., Wu, R., Wang, W., Wang, X., Xie, Y., Liu, Y., Mei, J. (2022). Three-dimensional morphological revealing of human placental villi with common obstetric complications via optical coherence tomography. Bioengineering & Translational Medicine, 8(1): e10372. https://doi.org/10.1002/btm2.10372

[27] Araújo, D.C., de Macedo, A.A., Veloso, A.A., Alpoim, P.N., Gomes, K.B., Carvalho, M.G., Dusse, L.M.S. (2024). Complete blood count as a biomarker for preeclampsia with severe features diagnosis: A machine learning approach. BMC Pregnancy and Childbirth,

24(1): 628. https://doi.org/10.1186/s12884-024-06821-4
[28] Kronenberg, M.E., Raz, S., Sander, C.J. (2006). Neurodevelopmental outcome in children born to mothers wit hypertension in pregnancy: The significance of suboptimal intrauterine growth. Developmental Medicine & Child Neurology, 48(3): 200-206. https://doi.org/10.1017/s0012162206000430
[29] Lee, S., Kim, S.H., Kim, H.D., Lee, J.S., Ko, A., Kang, H. (2024). Genetic diagnosis in neonatal encephalopathy with hypoxic brain damage using targeted gene panel sequencing. Journal of Clinical Neurology, 20(5): 519-528. https://doi.org/10.3988/jcn.2023.0500
[30] Cameron, N.A., Bello, N.A., Khan, S.S. (2022). Bringing the cuff home: Challenges and opportunities associated with home blood pressure monitoring among reproductive-aged individuals. American Journal of Hypertension, 35(8): 688-690. https://doi.org/10.1093/ajh/hpac074
[31] Ranjbar, A., Montazeri, F., Ghamsari, S.R., Mehrnoush, V., Roozbeh, N., Darsareh, F. (2024). Machine learning models for predicting preeclampsia: A systematic review. BMC Pregnancy and Childbirth, 24(1): 6. https://doi.org/10.1186/s12884-023-06220-1
[32] Zhao, Z., Dai, J., Chen, H., Lu, L., Li, G., Yan, H., Zhang, J. (2024). A prospective study on risk prediction of preeclampsia using bi-platform calibration and machine learning. International Journal of Molecular Sciences, 25(19): 10684. https://doi.org/10.3390/ijms251910684
[33] Kaya, Y., Bütün, Z., Çelik, Ö., Salik, E.A., Tahta, T. (2024). Risk assessment for preeclampsia in the preconception period based on maternal clinical history via machine learning methods. Journal of Clinical Medicine, 14(1): 155. https://doi.org/10.3390/jcm14010155
[34] Ansbacher-Feldman, Z., Syngelaki, A., Meiri, H., Cirkin, R., Nicolaides, K.H., Louzoun, Y. (2022). Machine-learning-based prediction of pre-eclampsia using first-trimester maternal characteristics and biomarkers. Ultrasound in Obstetrics & Gynecology, 60(6): 739-745. https://doi.org/10.1002/uog.26105
[35] Kuyoro, A.O., Fatade, O.B., Onuiri, E.E. (2025). Enhancing non-invasive diagnosis of endometriosis through explainable artificial intelligence: A Grad-CAM approach. Acadlore Transactions on Artificial Intelligence and Machine Learning, 4(2): 97-108. https://doi.org/10.56578/ataiml040203
[36] Pallikonda, A.K., Bandarapalli, V.K., Vipparla, A. (2025). Real-time anomaly detection in IoT networks using a hybrid deep learning model. Acadlore Transactions on Artificial Intelligence and Machine Learning, 4(4): 235-246. https://doi.org/10.56578/ataiml040401