An Explainable Machine Learning Framework for Predicting Preeclampsia and Pregnancy Outcomes Using Clinical, Doppler, and Angiogenic Biomarkers

Mohammed Bekkouche[1]*

[1]LabRI-SBA Lab., Ecole Superieure en Informatique, Sidi Bel Abbes, Algeria

*Corresponding author: m.bekkouche@esi-sba.dz

Abstract

Preeclampsia is a major pregnancy complication where accurate risk assessment is essential to improve maternal and fetal outcomes. This paper presents an interpretable machine learning framework for preeclampsia prediction using the Ljubljana dataset by integrating maternal clinical characteristics, uterine artery Doppler measurements, angiogenic biomarkers, and engineered interaction features. An ablation study evaluated the contribution of each feature group, while multiple classifiers were assessed using repeated cross-validation. Model reliability was enhanced through selective prediction, and interpretability was investigated using SHAP global feature importance, local explanations, and pairwise feature interaction analysis. In addition, gestational age at delivery was predicted using regression models. The Random Forest classifier achieved the best classification performance with a ROC-AUC of 0.880 and an F1-score of 0.819. Selective prediction further improved the F1-score to 0.962 at 44.2% coverage. SHAP analysis identified angiogenic biomarkers and engineered interaction features as the most influential predictors, while pairwise interaction analysis revealed complementary relationships between biomarkers, Doppler measurements, and engineered features learned by the model. Furthermore, the Random Forest regression model achieved an MAE of 2.63 weeks and an R^2 of 0.423 for gestational age prediction. These results demonstrate that the proposed framework provides accurate, reliable, and interpretable decision support for preeclampsia risk assessment.

Keywords

Preeclampsia, Machine Learning, Explainable Artificial Intelligence, SHAP, Uterine Artery Doppler, Angiogenic Biomarkers.

1. Introduction

Preeclampsia is a pregnancy complication characterized by new-onset hypertension and organ dysfunction after 20 weeks of gestation. According to the World Health Organization, preeclampsia affects between 2% and 10% of pregnancies and remains a leading cause of maternal and perinatal morbidity and mortality [10, 21]. Early identification of women at high risk is essential for improving pregnancy management and reducing adverse maternal and neonatal outcomes.

Preeclampsia is associated with abnormal placentation, impaired uteroplacental perfusion, and angiogenic imbalance [20]. Consequently, maternal clinical characteristics, uterine artery Doppler measurements, and angiogenic biomarkers such as placental growth factor (PlGF) and soluble fms-like tyrosine kinase-1 (sFlt-1) have become important predictors for risk assessment. Several studies have demonstrated that combining these complementary sources of information improves prediction performance compared with using clinical variables alone [23, 24]. More recently, machine learning methods have demonstrated promising results by learning complex nonlinear relationships among multimodal clinical variables, with the Ljubljana dataset [17] emerging as an important benchmark for evaluating prediction models based on clinical, Doppler, and angiogenic biomarkers [2, 11, 19, 22].

Despite these advances, several challenges remain. Most studies primarily assess predictive performance while providing limited interpretation of model decisions. Explainability is generally limited to global feature importance, whereas local explanations and interactions between clinical variables,

angiogenic biomarkers, and uterine artery Doppler measurements receive limited investigation. Furthermore, existing studies generally rely on the original feature representation without exploring clinically motivated engineered interaction features. Conventional classifiers generate predictions for every patient regardless of confidence, which may reduce their reliability in clinical practice. Finally, previous studies using the Ljubljana dataset [17] have focused on classifying preeclampsia and intrauterine growth restriction, with limited investigation of additional pregnancy outcomes.

To address these limitations, this paper proposes an explainable machine learning framework for preeclampsia prediction using the Ljubljana dataset. The framework integrates maternal clinical variables, uterine artery Doppler measurements, angiogenic biomarkers, and engineered interaction features. An ablation study evaluates the contribution of progressively enriched feature sets, while multiple machine learning classifiers are compared using repeated cross-validation. Prediction reliability is enhanced through selective prediction, whereas model transparency is achieved using SHAP global feature importance, local explanations, and pairwise feature interaction analysis. In addition, gestational age at delivery is investigated as a complementary regression task, providing a broader assessment of pregnancy outcomes.

The main contributions of this paper are as follows:

1. Development of an interpretable machine learning framework integrating clinical variables, uterine artery Doppler measurements, angiogenic biomarkers, and clinically inspired engineered interaction features.

2. Evaluation of progressively enriched feature sets and multiple machine learning classifiers through a comprehensive ablation study.

3. Integration of selective prediction to improve the reliability of automated clinical decision support.

4. Explainability analysis combining SHAP global feature importance, local explanations, and pairwise feature interaction analysis to characterize clinically meaningful relationships learned by the model.

5. Evaluation of gestational age at delivery prediction using the proposed multimodal feature representation.

The remainder of this paper is organized as follows. Section 3 describes the Ljubljana dataset and the considered variables. Section 4 presents the proposed methodology. Section 5 describes the experimental protocol. Section 6 reports the experimental results. Sections 7, 8, and 9 discuss the findings, limitations, and future research directions, while Section 10 concludes the paper.

2. Related Work

Preeclampsia prediction has evolved from conventional statistical models to advanced machine learning and deep learning techniques. Early approaches were mainly based on logistic regression and competing-risk models integrating maternal characteristics with uterine artery Doppler measurements and angiogenic biomarkers. The Fetal Medicine Foundation (FMF) screening model is among the most widely adopted clinical tools, combining maternal history, mean arterial pressure, uterine artery pulsatility index, PlGF, pregnancy-associated plasma protein-A (PAPP-A), and the sFlt-1/PlGF ratio to identify women at risk of preterm preeclampsia [3, 5].

Recent advances in machine learning have enabled more accurate prediction by modeling nonlinear relationships among heterogeneous clinical variables. Random Forest, Support Vector Machine, Gradient Boosting, XGBoost, Decision Trees, and Neural Networks are the most frequently investigated algorithms [19]. Representative studies reported competitive performance using Random Forest, Naive Bayes, Support Vector Machines, Decision Trees [1, 14], routine blood biomarkers [18], ensemble

learning strategies [9], and deep learning models trained on structured clinical data and electronic health records [3, 8].

The Ljubljana dataset has emerged as a valuable public benchmark for developing and evaluating machine learning models for placental disorders because it combines maternal clinical characteristics, uterine artery Doppler measurements, angiogenic biomarkers, and pregnancy outcomes. Previous studies using this dataset investigated the prediction of preeclampsia and intrauterine growth restriction through comparisons of conventional machine learning algorithms, feature selection methods, and multimodal data integration [2, 22]. Their findings demonstrated that combining clinical characteristics with Doppler and angiogenic biomarkers substantially improves predictive performance compared with individual feature groups, highlighting the complementary diagnostic value of these heterogeneous sources of information.

Despite these advances, several limitations remain. Most studies focus primarily on maximizing predictive performance while providing limited model interpretability. When explainability is considered, it is generally restricted to global feature importance, whereas local explanations and feature interaction analyses remain largely unexplored. Furthermore, previous investigations based on the Ljubljana dataset employ the original clinical, Doppler, and biomarker variables without introducing clinically motivated engineered interaction features designed to capture relationships between placental perfusion and angiogenic imbalance [2, 22]. In addition, prediction reliability has received limited attention, as conventional classifiers generate predictions for every patient regardless of confidence. Finally, previous studies using the Ljubljana dataset have primarily focused on the classification of preeclampsia and intrauterine growth restriction, with limited investigation of additional pregnancy outcomes, such as gestational age at delivery.

Motivated by these limitations, the proposed framework integrates maternal clinical characteristics, uterine artery Doppler measurements, angiogenic biomarkers, and engineered interaction features. It further combines an ablation study, SHAP-based global, local, and interaction analyses, selective prediction, and gestational age at delivery prediction within a unified explainable machine learning framework.

3. Materials

3.1. Ljubljana Dataset

The experiments were conducted using the publicly available Ljubljana preeclampsia dataset[1] [17], which contains maternal clinical characteristics, uterine artery Doppler measurements, angiogenic biomarkers, and pregnancy outcomes. The dataset includes pregnancies with and without preeclampsia and was collected at the University Medical Centre Ljubljana, Slovenia. The measurements were acquired during the second trimester before the clinical diagnosis of preeclampsia, enabling the evaluation of predictive models prior to disease manifestation.

After preprocessing and data cleaning, the dataset comprised 95 pregnancies, including 54 cases diagnosed with preeclampsia and 41 control pregnancies. The relatively small sample size makes robust validation procedures essential to obtain reliable estimates of model performance.

3.2. Clinical Variables

The clinical variables describe maternal demographic and obstetric characteristics routinely collected during prenatal care. These variables include maternal age, prepregnancy weight, maternal height, body mass index (BMI), and parity. Such maternal characteristics are well-established risk factors for adverse pregnancy outcomes and constitute the baseline feature set used throughout the experiments.

[1] `https://data.mendeley.com/datasets/zsjhvy9ytx/1`

3.3. Uterine Artery Doppler Variables

Uterine artery Doppler ultrasound provides a noninvasive assessment of placental perfusion and vascular resistance. The dataset includes Doppler-derived measurements from both the left and right uterine arteries, including the pulsatility index (PI), resistance index (RI), peak systolic velocity (PSV), and the presence of bilateral uterine artery notching. In addition, the mean pulsatility index (Mean PI), mean resistance index (Mean RI), and mean peak systolic velocity (Mean PSV) were computed by averaging the corresponding left and right uterine artery measurements, providing global indicators of uteroplacental blood flow. Since impaired placentation is a hallmark of preeclampsia, these Doppler variables provide valuable physiological information for risk prediction.

3.4. Angiogenic Biomarkers

The dataset contains circulating angiogenic biomarkers that reflect placental function and endothelial balance. Specifically, PlGF and soluble sFlt-1 were considered in this study. In addition to the original biomarker measurements, the clinically established sFlt-1/PlGF ratio was included because it has demonstrated strong diagnostic and prognostic value for placental dysfunction and preeclampsia in previous studies.

3.5. Outcome Variables

Two prediction tasks were investigated in this study. The primary outcome was the presence or absence of preeclampsia, which was formulated as a binary classification problem. In addition, gestational age at delivery was considered as a secondary outcome and formulated as a regression problem to evaluate the ability of the proposed framework to predict clinically relevant pregnancy outcomes beyond disease classification.

4. Methodology

4.1. Overview of the Proposed Framework

Figure 1 illustrates the overall workflow of the proposed framework. Starting from the Ljubljana dataset, the pipeline consists of data preprocessing, feature engineering, feature set construction, machine learning model development, selective prediction, and explainability analysis. The framework is designed to evaluate the contribution of heterogeneous maternal data for preeclampsia prediction while providing reliable and interpretable predictions. The same framework is subsequently extended to predict gestational age at delivery.

4.2. Data Preprocessing

The dataset was first inspected to identify missing values, inconsistent records, and duplicated observations. Missing numerical values were imputed using the median of the corresponding feature. Continuous variables were standardized when required by the learning algorithm, whereas tree-based models were trained using the original feature scales. The target variable for the classification task was encoded as a binary label, while gestational age at delivery was retained as a continuous variable for regression.

4.3. Feature Engineering

Feature engineering was performed to exploit the complementary information contained in clinical characteristics, uterine artery Doppler measurements, and angiogenic biomarkers. In addition to the original variables, clinically motivated interaction features were designed to better characterize placental dysfunction.

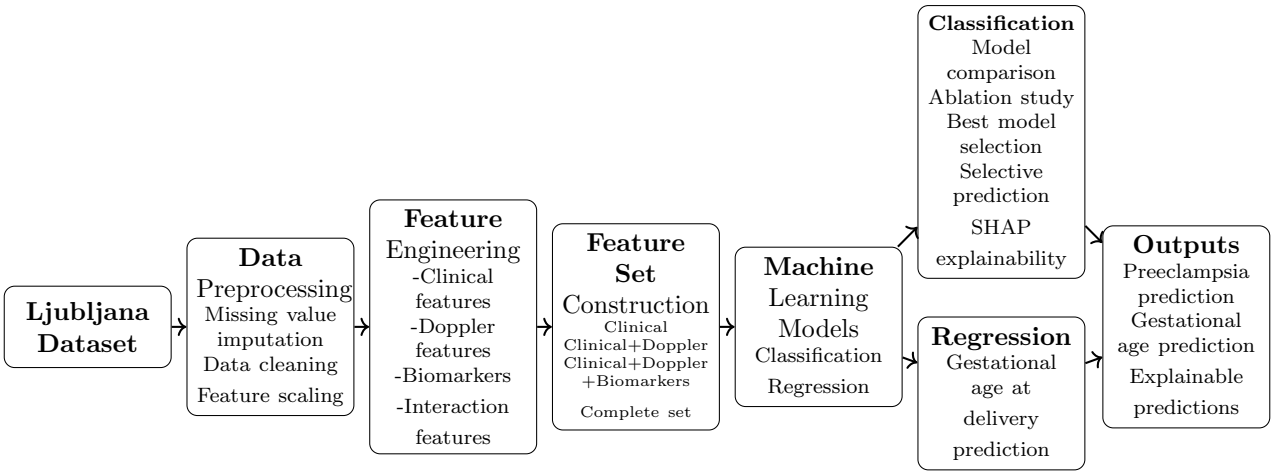


Figure 1. Overview illustrating the proposed explainable machine learning framework for preeclampsia classification and gestational age at delivery prediction. The workflow includes data preprocessing, feature engineering, feature set construction, model training, selective prediction, regression analysis, and SHAP-based explainability.

4.3.1. Clinical Features

The clinical feature group contains routinely collected maternal characteristics, including demographic information, obstetric history, blood pressure measurements, and other clinical risk factors available in the Ljubljana dataset. These variables constitute the baseline predictor set.

4.3.2. Doppler Features

The Doppler feature group includes uterine artery ultrasound measurements describing placental blood flow and vascular resistance. These variables provide complementary physiological information regarding placental perfusion and are widely used for preeclampsia risk assessment.

4.3.3. Biomarker Features

The biomarker feature group consists of angiogenic biomarkers associated with placental development and endothelial dysfunction. In addition to the individual biomarker concentrations, the sFlt-1/PlGF ratio was considered because of its established clinical relevance for preeclampsia prediction.

4.3.4. Engineered Interaction Features

To capture complementary information between placental perfusion and angiogenic imbalance, four clinically motivated interaction features were engineered from the original Doppler and biomarker variables. These features combine standardized or transformed measurements to explicitly model physiological relationships that are not represented by the original variables alone.

The angiogenic imbalance was first quantified using the logarithm of the sFlt-1/PlGF ratio:

$$\textit{log\_sFlt1\_PlGF} = \log\left(\frac{\mathrm{sFlt1} + \varepsilon}{\mathrm{PlGF} + \varepsilon}\right), \tag{1}$$

where $\varepsilon = 10^{-8}$ avoids division by zero.

The *Doppler-Angiogenic Index* combines angiogenic imbalance with uterine artery impedance:

$$\begin{gathered}\textit{Doppler\_Angiogenic\_Index} = \\ \mathrm{MeanPI} \times \textit{log\_sFlt1\_PlGF}.\end{gathered} \tag{2}$$

To summarize placental dysfunction, standardized (z-score) versions of Mean PI, sFlt-1, and PlGF were combined as

$$\begin{aligned} &\textit{Placental\_Dysfunction\_Score} = \\ &z(\text{MeanPI}) + z(\text{sFlt1}) - z(\text{PlGF}), \end{aligned} \tag{3}$$

where

$$z(x) = \frac{x - \mu_x}{\sigma_x}, \tag{4}$$

with μ_x and σ_x denoting the mean and standard deviation computed from the training data.

Finally, the *PlGF-Doppler Discordance* feature measures the disagreement between Doppler assessment and angiogenic status:

$$\textit{PlGF\_Doppler\_Discordance} = z(\text{MeanPI}) - z(\text{PlGF}). \tag{5}$$

These engineered features were designed to integrate complementary physiological information while remaining simple to compute and clinically interpretable.

4.4. Feature Set Construction

Four feature sets were evaluated through an ablation study. The first includes only clinical variables. The second combines clinical and Doppler variables. The third further incorporates angiogenic biomarkers. The final feature set augments these variables with the engineered interaction features. This progressive evaluation enables the contribution of each feature category to be quantified.

4.5. Classification Models

Several supervised machine learning algorithms were evaluated for preeclampsia prediction, including Logistic Regression, Naive Bayes, Decision Tree, Random Forest, Gradient Boosting, and XGBoost. To reduce the impact of class imbalance, algorithms supporting class weighting were trained using balanced class weights whenever applicable. Model performance was compared using an identical evaluation protocol.

4.6. Regression Models

The prediction of gestational age at delivery was formulated as a regression problem. Linear Regression, Ridge Regression, Lasso Regression, ElasticNet, Random Forest Regressor, Extra Trees Regressor, and Gradient Boosting Regressor were evaluated using the complete feature set to investigate the predictive value of the proposed framework for pregnancy outcome prediction.

4.7. Selective Prediction

To improve prediction reliability, a selective prediction strategy was applied to the best-performing classification model. Instead of producing predictions for every patient, the classifier was allowed to abstain from low-confidence predictions based on the estimated class probability. Performance was subsequently evaluated as a function of prediction coverage to assess how reliability varies with prediction availability.

4.8. Explainability

Model interpretability was investigated using SHAP (SHapley Additive exPlanations) [12], which provides consistent feature attribution scores for individual predictions and global model behavior. Several complementary analyses were performed to obtain a comprehensive understanding of the learned prediction mechanisms.

Global SHAP values were computed to quantify the overall importance of each feature, while local explanations were generated to analyze feature contributions for a representative individual patient. These analyses provide clinically interpretable explanations of the model predictions.

Pairwise SHAP interaction values were computed for the model classifier to quantify the joint contribution of feature pairs to each prediction. The average absolute interaction value was used to rank interactions according to their overall influence. This analysis complements global SHAP importance by identifying synergistic relationships learned by the model.

5. Experimental Setup

5.1. Experimental Design

The proposed framework was evaluated on two prediction tasks. The primary task consists of binary classification for preeclampsia prediction, whereas the secondary task investigates gestational age at delivery prediction as a regression problem. Four feature sets were considered in the classification experiments: (i) clinical variables, (ii) clinical and Doppler variables, (iii) clinical, Doppler, and angiogenic biomarkers, and (iv) the previous feature set augmented with engineered interaction features. The best-performing classifier was subsequently used for selective prediction and SHAP-based explainability analyses.

5.2. Cross-Validation Strategy

Model performance was estimated using repeated 5-fold cross-validation with 10 repetitions, providing 50 independent train-test evaluations for each model. Mean performance and standard deviation were reported across all runs to obtain robust estimates of model generalization [7].

5.3. Evaluation Metrics

Classification models were evaluated using ROC-AUC, F1-score, accuracy, precision, and recall [16]. Regression models were assessed using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2) [4].

5.4. Implementation Details

All experiments were implemented in Python using Scikit-learn [15]. Data preprocessing was performed with Pandas [13] and NumPy [6], while SHAP [12] was used to explain model predictions. Fixed random seeds were adopted throughout the experiments to ensure reproducibility.

6. Results

6.1. Ablation Study

An ablation study was conducted to evaluate the contribution of progressively enriched feature sets to preeclampsia prediction. As summarized in Tables 1 and 2, four feature sets were constructed by incrementally incorporating clinical variables, uterine artery Doppler measurements, angiogenic biomarkers, and engineered interaction features. This progressive design enabled a systematic assessment of the contribution of each feature category to predictive performance.

Table 3 summarizes the best classification performance obtained for each feature set. Using only clinical variables resulted in the lowest predictive performance (ROC-AUC = 0.703). Adding uterine artery Doppler variables substantially improved the ROC-AUC to 0.767, demonstrating the complementary value of placental perfusion measurements. The largest improvement was achieved after incorporating angiogenic biomarkers, increasing the ROC-AUC to 0.882. Finally, introducing the engineered interaction features further improved the F1-score (0.819) and accuracy (0.789), while

Table 1. Feature sets evaluated in the ablation study.

Feature Set	Included Variables	No. Features
Clinical	Clinical	5
Clinical + Doppler	Clinical + Doppler	15
Clinical + Doppler + Biomarkers	Clinical + Doppler + Biomarkers	18
Clinical + Doppler + Biomarkers + Interaction Features	Clinical + Doppler + Biomarkers + Interaction Features	22

Table 2. Variables included in each feature group.

Feature Group	Variables
Clinical	Maternal age (years), Pre-pregnancy weight (kg), Maternal height (m), BMI (kg/m^2), Parity.
Doppler	Right uterine artery RI, Right uterine artery PI, Right uterine artery PSV, Left uterine artery RI, Left uterine artery PI, Left uterine artery PSV, Mean PI, Mean RI, Mean PSV, Bilateral notch.
Biomarkers	PlGF (μg/L), sFlt-1 (μg/L), sFlt-1/PlGF ratio.
Interaction Features	log(sFlt-1/PlGF), Doppler Angiogenic Index, Placental Dysfunction Score, PlGF Doppler Discordance.

Table 3. Best classification performance obtained for each feature set during the ablation study.

Feature Set	Best Model	ROC-AUC	F1-score	Accuracy
Clinical	Logistic Regression	0.703 ± 0.120	0.644 ± 0.081	0.642
Clinical + Doppler	Random Forest	0.767 ± 0.124	0.769 ± 0.046	0.726
Clinical + Doppler + Biomarkers	Random Forest	**0.882 ± 0.072**	0.808 ± 0.044	0.779
Clinical + Doppler + Biomarkers + Interaction Features	Random Forest	0.880 ± 0.069	**0.819 ± 0.048**	**0.789**

Table 4. Selective prediction performance obtained using the selected Random Forest classifier.

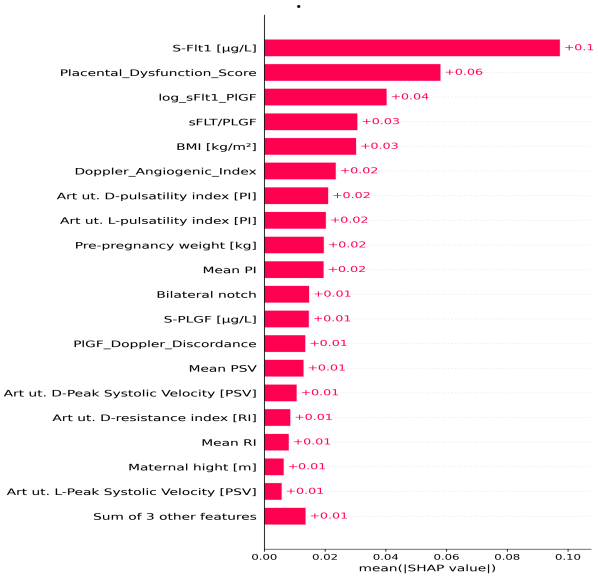
Threshold	Coverage (%)	ROC-AUC	Accuracy	F1-score
0.70	76.8	0.905	0.849	0.871
0.75	65.3	0.921	0.887	0.907
0.80	53.7	0.935	0.902	0.918
0.85	44.2	0.936	0.952	**0.962**
0.90	26.3	0.934	0.920	0.938

maintaining a comparable ROC-AUC (0.880). These results indicate that Doppler measurements and angiogenic biomarkers provide the greatest contribution to prediction performance, whereas the proposed interaction features mainly improve the balance between precision and recall.

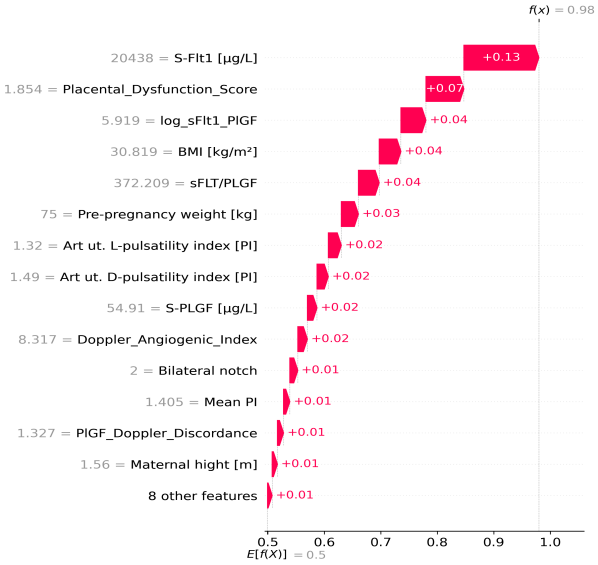
6.2. Selective Prediction Performance

Random Forest achieved the best overall classification performance (Section 4.5) and was therefore selected for the selective prediction analysis. Instead of producing predictions for every patient, the classifier was allowed to abstain from cases associated with low confidence. Performance was subsequently evaluated at different confidence thresholds to examine the relationship between prediction coverage and classification performance.

Table 4 shows that increasing the confidence threshold progressively reduced prediction coverage while improving classification performance. Compared with the baseline classifier without abstention



(a) Global SHAP feature importance.



(b) SHAP waterfall explanation for Patient 54.

Figure 2. Global and local SHAP explanations for the selected Random Forest classifier. The summary bar plot ranks features according to their mean absolute SHAP values, whereas the waterfall plot illustrates how individual feature contributions produce the prediction for a representative patient.

(ROC-AUC = 0.880 and F1-score = 0.819), selective prediction achieved a ROC-AUC of 0.936 and an F1-score of 0.962 at a confidence threshold of 0.85, while providing predictions for 44.2% of the patients. Increasing the threshold to 0.90 further reduced coverage to 26.3% without improving ROC-AUC or F1-score. These findings indicate that a confidence threshold of 0.85 provides the most suitable balance between prediction reliability and patient coverage, allowing highly confident predictions to be identified while referring uncertain cases for additional clinical assessment.

6.3. SHAP Analysis

SHAP was employed to interpret the predictions generated by the selected Random Forest classifier through global feature importance, local explanations for individual predictions, and pairwise feature interaction analysis.

Figure 2a presents the mean absolute SHAP values, which quantify the average contribution of each

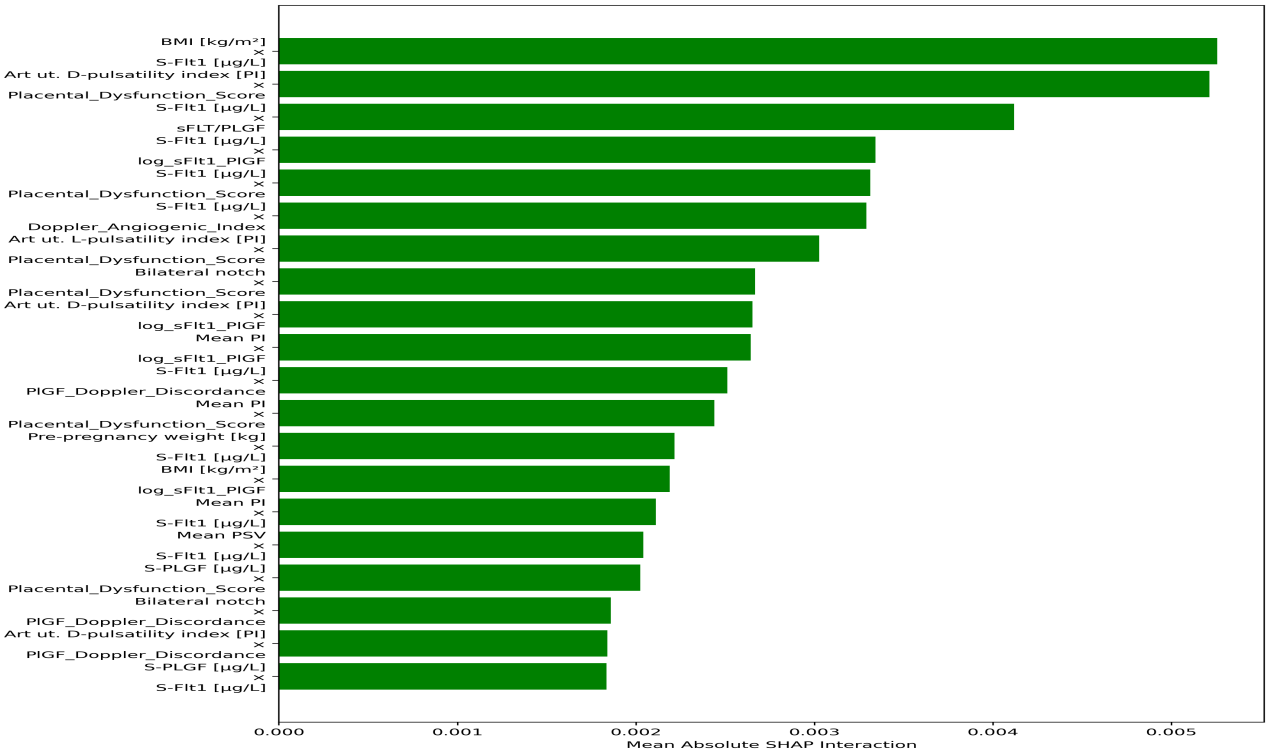


Figure 3. Top pairwise feature interactions ranked by the mean absolute SHAP interaction value for the selected Random Forest model. Higher values indicate stronger joint contributions of feature pairs to preeclampsia prediction.

feature to the model predictions. The results show that *S-Flt1* is the most influential predictor, with a mean absolute SHAP value of 0.097, followed by the engineered feature *Placental_Dysfunction_Score* (0.058), *log_sFlt1_PlGF* (0.040), the *sFlt1/PlGF* ratio (0.031), and *BMI* (0.030). Doppler-derived variables, including the uterine artery pulsatility indices and the proposed *Doppler_Angiogenic_Index*, also ranked among the most important predictors. In contrast, maternal age, parity, and uterine artery resistance indices exhibited relatively small SHAP values, indicating a limited contribution to the final predictions.

These findings demonstrate that angiogenic biomarkers constitute the primary source of predictive information, while Doppler measurements provide complementary information. Furthermore, the high ranking of the proposed engineered features confirms that combining biomarkers and Doppler measurements captures clinically meaningful information that is not fully represented by the original variables alone.

Figure 2b illustrates the SHAP waterfall explanation for a representative patient predicted as having a high risk of preeclampsia. The model assigned a prediction probability of 0.98, substantially higher than the baseline probability of 0.50. The largest positive contribution originated from the patient's elevated *S-Flt1* concentration, which increased the prediction by approximately 0.13. Additional positive contributions were provided by the *Placental_Dysfunction_Score*, *log_sFlt1_PlGF*, *sFlt1/PlGF*, elevated BMI, and several Doppler measurements. Collectively, these features shifted the prediction from the baseline value to a very high predicted risk, illustrating how the model combines multiple clinical indicators to support its decision.

Figure 3 presents the strongest pairwise feature interactions identified using SHAP interaction values for the selected Random Forest classifier. The largest interaction was observed between BMI and S-Flt1, followed by interactions between the right uterine artery pulsatility index (PI) and the Placental Dysfunction Score, and between S-Flt1 and the sFlt-1/PlGF ratio. These results indicate that the model relies on combinations of maternal characteristics, Doppler measurements, and angiogenic biomarkers rather than isolated predictors.

Table 5. Comparison of regression models for gestational age at delivery prediction.

Model	MAE	RMSE	R^2
Linear Regression	**2.54**	3.18	0.398
Ridge	2.56	3.14	**0.426**
ElasticNet	2.57	3.16	0.417
Lasso	2.58	3.18	0.407
Random Forest	2.63	**3.12**	0.423
Extra Trees	2.64	3.18	0.405
Gradient Boosting	2.79	3.37	0.326

Several engineered interaction features also appeared among the strongest interactions. The Placental Dysfunction Score interacted prominently with S-Flt1, the right and left uterine artery PI, Mean PI, and bilateral uterine artery notch. Similarly, the Doppler Angiogenic Index showed a strong interaction with S-Flt1, while the logarithmic sFlt-1/PlGF ratio interacted with both Mean PI and the right uterine artery PI. These observations demonstrate that the engineered features remain strongly connected to their constituent variables while capturing complementary information exploited by the Random Forest classifier.

The pairwise interaction analysis complements the global SHAP importance ranking by revealing how combinations of clinical variables, Doppler measurements, angiogenic biomarkers, and engineered interaction features jointly contribute to prediction. These findings provide additional insight into the decision-making process of the proposed framework and further support the use of the proposed engineered interaction features.

6.4. Gestational Age at Delivery Prediction

In addition to preeclampsia prediction, the proposed feature set was evaluated for predicting gestational age at delivery using a regression formulation. This complementary analysis investigates whether the same clinical variables, uterine artery Doppler measurements, angiogenic biomarkers, and engineered interaction features are informative for estimating delivery timing. Performance was assessed using MAE, RMSE, and R^2.

6.4.1. Regression Model Comparison

Table 5 compares the performance of the evaluated regression models. Random Forest achieved the lowest RMSE (3.12 weeks), while Linear Regression obtained the lowest MAE (2.54 weeks) and Ridge achieved the highest R^2 (0.426). However, the differences among Random Forest, Ridge, ElasticNet, Lasso, and Linear Regression were relatively small. Considering its superior robustness to larger prediction errors, as reflected by the lowest RMSE, Random Forest was selected for further analysis.

Overall, the best-performing models explained approximately 42% of the variance in gestational age at delivery, indicating a moderate predictive capability. The average prediction error of approximately 2.6 weeks suggests that the proposed feature set captures clinically relevant information associated with delivery timing.

6.4.2. Prediction Performance

Based on the comparison presented in Table 5, Random Forest was selected as the final regression model. The model achieved an MAE of 2.63 weeks, an RMSE of 3.12 weeks, and an R^2 of 0.423.

Figure 4 compares the predicted and observed gestational ages. Most predictions follow the identity line, indicating good agreement between predicted and actual values. The largest deviations are mainly

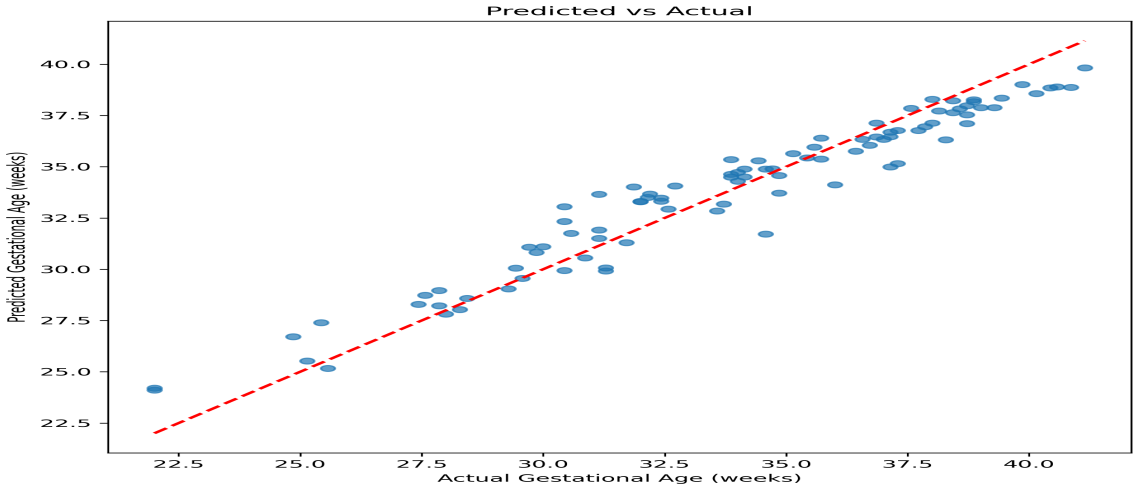


Figure 4. Predicted versus actual gestational age at delivery obtained using the selected Random Forest regression model.

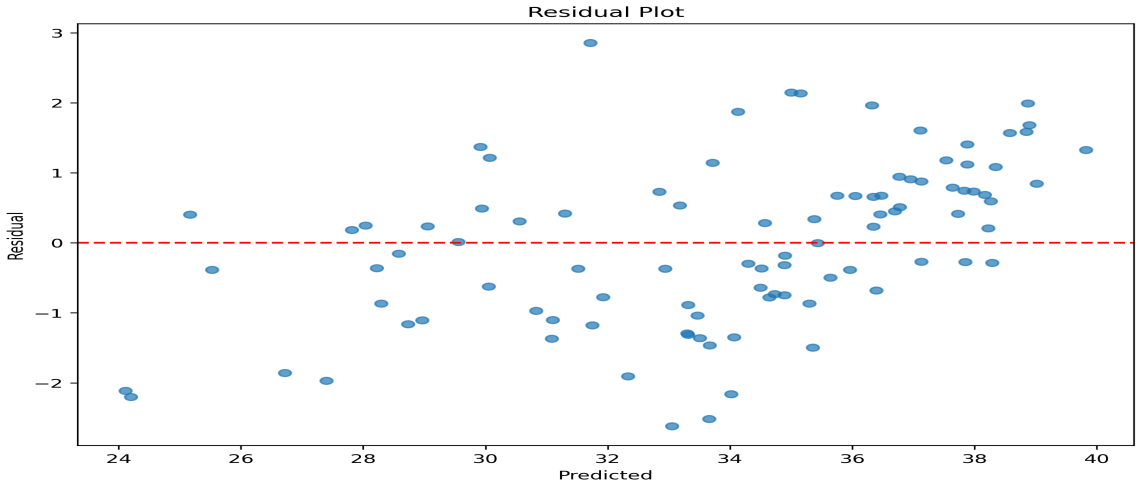


Figure 5. Residual plot obtained using the selected Random Forest regression model.

observed for pregnancies delivered at earlier gestational ages, whereas predictions for deliveries closer to term are generally more accurate. These observations are consistent with the quantitative evaluation reported in Table 5.

6.4.3. Residual Analysis

Figure 5 presents the residual distribution for the selected Random Forest model. Most residuals are centered around zero, indicating that the model does not exhibit a strong systematic tendency to overestimate or underestimate gestational age at delivery. Although some larger residuals are observed, particularly for preterm deliveries, no clear pattern is visible across the prediction range. This behavior suggests that the prediction errors are reasonably well distributed and that the Random Forest model provides stable estimates across different gestational ages.

7. Discussion

7.1. Clinical Interpretation

The experimental results demonstrate that combining clinical variables, uterine artery Doppler measurements, and angiogenic biomarkers enables accurate prediction of preeclampsia after biomarker ac-

quisition. The progressive improvement observed during the ablation study confirms that each group of variables contributes complementary clinical information. Clinical characteristics alone provided limited predictive performance, whereas Doppler measurements substantially improved discrimination by capturing impaired uteroplacental perfusion. The largest performance gain was achieved after incorporating angiogenic biomarkers, particularly sFlt-1 and PlGF, which are well-established indicators of placental dysfunction. These findings are consistent with the current understanding that preeclampsia results from the combined effects of abnormal placentation, vascular dysfunction, and angiogenic imbalance rather than a single clinical factor.

7.2. Contribution of Engineered Features

Although the inclusion of engineered interaction features produced only modest improvements in overall predictive performance, explainability analysis demonstrated that several of these features ranked among the most influential predictors. In particular, the Placental Dysfunction Score, Doppler Angiogenic Index, log_sFlt1_PlGF, and PlGF Doppler Discordance received high SHAP importance scores, indicating that they capture clinically meaningful relationships between biomarkers and Doppler measurements. These engineered variables summarize complex interactions into interpretable indicators that are easier for machine learning models to exploit than the original measurements alone. Consequently, they improve model interpretability while preserving competitive predictive performance.

7.3. Reliability Through Selective Prediction

Selective prediction further improved the reliability of the proposed framework by allowing the classifier to abstain from low-confidence predictions. As the confidence threshold increased, prediction coverage decreased while classification performance improved considerably. At a confidence threshold of 0.85, the Random Forest classifier achieved an F1-score of 0.962 with an accuracy of 95.2%, although predictions were produced for only 44.2% of the evaluated cases. Such a strategy is particularly relevant in clinical decision support, where uncertain predictions can be referred for additional examinations or specialist assessment rather than relying on potentially unreliable automated decisions. This selective prediction mechanism therefore provides a practical way to increase confidence in machine learning recommendations.

7.4. Interpretability of the Proposed Framework

Interpretability analyses demonstrate that the proposed framework provides transparent and clinically meaningful predictions. Global SHAP analysis consistently identified S-Flt1, the Placental Dysfunction Score, the sFlt-1/PlGF ratio, and Doppler-derived variables as the most influential predictors, indicating that the model primarily relies on angiogenic biomarkers and placental perfusion measurements. Local SHAP explanations complemented this analysis by illustrating how individual features contributed to each patient's prediction, allowing the factors driving high- and low-risk classifications to be identified.

Pairwise SHAP interaction analysis provided additional insight into the prediction process by revealing how combinations of clinical variables, Doppler measurements, angiogenic biomarkers, and engineered interaction features jointly influenced the Random Forest classifier. Several engineered features, including the Placental Dysfunction Score, Doppler Angiogenic Index, and log(sFlt-1/PlGF), exhibited strong interactions with their constituent variables, indicating that they capture complementary information learned by the model. Together, these explainability analyses increase confidence in the proposed framework and facilitate the clinical interpretation of its predictions.

7.5. Prediction of Pregnancy Outcomes

Beyond preeclampsia classification, the proposed feature set also demonstrated predictive capability for gestational age at delivery. The Random Forest regression model achieved an MAE of 2.63 weeks,

an RMSE of 3.12 weeks, and an R^2 of 0.423, indicating moderate agreement between predicted and observed delivery timing. Although this secondary task does not reach the predictive performance obtained for preeclampsia classification, the results suggest that the same multimodal feature representation contains information associated with pregnancy progression and delivery outcomes. This additional analysis highlights the flexibility of the proposed framework and suggests that it may be extended to support multiple prediction tasks using a unified set of clinical, Doppler, biomarker, and engineered features.

8. Limitations

Several limitations should be acknowledged. First, the proposed framework was evaluated using a single-center dataset containing a limited number of subjects, which may restrict the generalizability of the reported results. Although repeated cross-validation was employed to reduce estimation bias, external validation using independent cohorts remains necessary.

Second, the study focuses on prediction after the acquisition of angiogenic biomarkers and uterine artery Doppler measurements rather than early first-trimester screening. Consequently, the proposed framework is intended to support clinical decision-making after these examinations have been performed and should not be interpreted as an early prediction model.

Third, only gestational age at delivery was investigated as a secondary regression task. Other clinically important pregnancy outcomes, such as birth weight, neonatal complications, or maternal adverse events, were not considered.

Finally, although SHAP-based explainability provides transparent global and local explanations, these explanations should be regarded as decision-support information rather than evidence of causal relationships between predictors and preeclampsia.

9. Future Work

Future work will focus on validating the proposed framework using larger multi-center cohorts collected from different healthcare institutions to evaluate its generalizability across diverse patient populations. Additional pregnancy outcomes, including birth weight and neonatal complications, will also be investigated within a multi-task learning framework.

Another promising direction consists of integrating longitudinal clinical observations collected throughout pregnancy, enabling dynamic risk assessment as new information becomes available. Finally, prospective clinical evaluation will be necessary to assess the practical utility of the proposed selective prediction strategy and explainability framework in routine obstetric practice.

10. Conclusion

This paper presented an interpretable machine learning framework for preeclampsia prediction using clinical variables, uterine artery Doppler measurements, angiogenic biomarkers, and engineered interaction features derived from the Ljubljana dataset. An extensive ablation study demonstrated the complementary contribution of each feature group, while engineered interaction features further improved model interpretability and produced competitive predictive performance. Among the evaluated classifiers, Random Forest achieved the best overall results, and selective prediction further increased prediction reliability by allowing the model to abstain from low-confidence decisions.

Explainability analyses, including SHAP global feature importance, local explanations, and pairwise interaction analysis, provided additional insight into the prediction process. These analyses highlighted the dominant contribution of angiogenic biomarkers and Doppler-derived indicators and showed how combinations of these variables jointly influenced the predictions generated by the Random Forest classifier.

Beyond preeclampsia classification, the proposed feature representation also demonstrated moderate capability for predicting gestational age at delivery, suggesting that the same multimodal information can support additional pregnancy outcome prediction tasks.

Overall, the proposed framework combines predictive performance, reliability, and interpretability, making it a promising decision-support approach for assisting clinicians in the assessment of pregnancies at risk of preeclampsia while providing transparent explanations that facilitate clinical interpretation.

Data Availability

The source code implementing the proposed framework, including data preprocessing, feature engineering, model training, evaluation, selective prediction, and SHAP-based explainability, is publicly available at: `https://github.com/mdbekkouche/preeclampsia-prediction`

The repository provides all scripts necessary to reproduce the experimental results reported in this study.

AI Tools Disclosure

Artificial intelligence tools were used in the preparation of this article. Specifically, OpenAI?s ChatGPT (GPT-5.5) was employed to assist in language refinement, restructuring certain paragraphs for clarity. All final content was reviewed and approved by the authors, who take full responsibility for the accuracy and integrity of the work.

Funding sources

This work was supported by the Directorate General for Scientific Research and Technological Development (DGRSDT), Algeria.

Declaration of competing interest

The author declares that there are no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Sumayh S Aljameel, Manar Alzahrani, Reem Almusharraf, Majd Altukhais, Sadeem Alshaia, Hanan Sahlouli, Nida Aslam, Irfan Ullah Khan, Dina A Alabbad, and Albandari Alsumayt. Prediction of preeclampsia using machine learning and deep learning models: a review. *Big Data and Cognitive Computing*, 7(1):32, 2023.

[2] Pakiza Amin, Saima Gulzar Ahmad, Hikmat Ullah Khan, Ehsan Ullah Munir, and Naeem Ramzan. Predictive analytics in maternal health: a machine learning approach for classification of preeclampsia. *International Journal of Computational Intelligence Systems*, 18(1):115, 2025.

[3] Zuriya Ansbacher-Feldman, Argyro Syngelaki, Hamutal Meiri, Roy Cirkin, Kypros H Nicolaides, and Yoram Louzoun. Machine-learning-based prediction of pre-eclampsia using first-trimester maternal characteristics and biomarkers. *Ultrasound in Obstetrics & Gynecology*, 60(6):739–745, 2022.

[4] Davide Chicco, Matthijs J Warrens, and Giuseppe Jurman. The coefficient of determination r-squared is more informative than smape, mae, mape, mse and rmse in regression analysis evaluation. *Peerj computer science*, 7:e623, 2021.

[5] Max Hackelöer, Leon Schmidt, and Stefan Verlohren. New advances in prediction and surveillance of preeclampsia: role of machine learning approaches and remote monitoring. *Archives of gynecology and obstetrics*, 308(6):1663–1677, 2023.

[6] Charles R Harris, K Jarrod Millman, Stéfan J Van Der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J Smith, et al. Array programming with numpy. *nature*, 585(7825):357–362, 2020.

[7] Ron Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Montreal, Canada, 1995.

[8] Haoyang Li, Yaxin Li, Chengxi Zang, Weishen Pan, He S Yang, Tracy B Grossman, Zhen Zhao, and Fei Wang. Machine learning for dynamic and short-term prediction of preeclampsia using routine clinical data. *JAMA Network Open*, 9(3):e260359, 2026.

[9] Taishun Li, Mingyang Xu, Yuan Wang, Ya Wang, Huirong Tang, Honglei Duan, Guangfeng Zhao, Mingming Zheng, and Yali Hu. Prediction model of preeclampsia using machine learning based methods: a population based cohort study in china. *Frontiers in Endocrinology*, 15:1345573, 2024.

[10] M Lin, X He, H Guo, M He, L Zhang, J Xian, T Lei, Q Xu, J Zheng, J Feng, et al. Use of real-time artificial intelligence in detection of abnormal image patterns in standard sonographic reference planes in screening for fetal intracranial malformations. *Ultrasound in Obstetrics & Gynecology*, 59(3):304–316, 2022.

[11] Lu Liu, Qixuan Zhu, Yichi Zong, Xueyuan Chen, Wei Zhang, and Jun Wang. Machine learning prediction models for preeclampsia: Systematic review and meta-analysis. *Journal of Medical Internet Research*, 28:e78714, 2026.

[12] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[13] Wes McKinney et al. Data structures for statistical computing in python. *scipy*, 445(1):51–56, 2010.

[14] Marian Melinte-Popescu, Ingrid-Andrada Vasilache, Demetra Socolov, and Alina-Sînziana Melinte-Popescu. Prediction of hellp syndrome severity using machine learning algorithms—results from a retrospective study. *Diagnostics*, 13(2):287, 2023.

[15] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[16] David MW Powers. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061*, 2020.

[17] T Premru-Srsen and TT Premru-Srsen. Uterine arteries doppler and sflt-1/plgf ratio in hypertensive disorders during pregnancy. mendeley data 2018. *Mendeley Data, V1.*

[18] Longwei Qiao, Yanqiu Zhang, Juan Wang, Xiaojuan Wu, Chunhua Zhang, Ying Xue, Jun Cao, Jieyu Jin, Ting Wang, Jia Li, et al. Machine learning-based prediction of preeclampsia using first-trimester inflammatory markers and red blood cell indices. *BMC Pregnancy and Childbirth*, 25(1):1083, 2025.

[19] Amene Ranjbar, Farideh Montazeri, Sepideh Rezaei Ghamsari, Vahid Mehrnoush, Nasibeh Roozbeh, and Fatemeh Darsareh. Machine learning models for predicting preeclampsia: a systematic review. *BMC Pregnancy and Childbirth*, 24(1):6, 2024.

[20] Christopher WG Redman and Anne Cathrine Staff. Preeclampsia, biomarkers, syncytiotrophoblast stress, and placental capacity. *American journal of obstetrics and gynecology*, 213(4):S9–e1, 2015.

[21] Eric AP Steegers, Peter Von Dadelszen, Johannes J Duvekot, and Robert Pijnenborg. Preeclampsia. *The lancet*, 376(9741):631–644, 2010.

[22] Herdiantri Sufriyana, Wu Yu-Wei, and Emily Chia-Yu Su. Prediction of preeclampsia and intrauterine growth restriction: development of machine learning models on a prospective cohort. *JMIR medical informatics*, 8(5), 2020.

[23] Stefan Verlohren, Ignacio Herraiz, Olav Lapaire, Dietmar Schlembach, Manfred Moertl, Harald Zeisler, Pavel Calda, Wolfgang Holzgreve, Alberto Galindo, Theresa Engels, et al. The sflt-1/plgf ratio in different types of hypertensive pregnancy disorders and its prognostic potential in preeclamptic patients. *American journal of obstetrics and gynecology*, 206(1):58–e1, 2012.

[24] Harald Zeisler, Elisa Llurba, Frederic Chantraine, Manu Vatish, Anne Cathrine Staff, Maria Sennström, Matts Olovsson, Shaun P Brennecke, Holger Stepan, Deirdre Allegranza, et al. Predictive value of the sflt-1: Plgf ratio in women with suspected preeclampsia. *New England Journal of Medicine*, 374(1):13–22, 2016.