Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Evaluating the predictive power of machine learning in preeclampsia

Fatemeh Abdi
Iran University of Medical Sciences
Nasibeh Roozbeh
Hormozgan University of Medical Sciences
Fatemeh Darsareh
famadarsareh@yahoo.com
Hormozgan University of Medical Sciences
Vahid Mehrnoush
Hormozgan University of Medical Sciences
Ali Haghighat
Shiraz University
Anna Nami
Shiraz University
Farideh Montazeri
Hormozgan University of Medical Sciences
Mojdeh Banaei
Hormozgan University of Medical Sciences

Article

Keywords: preeclampsia, machine learning, artificia inteligince, random forest

Posted Date: April 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9292292/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Given that machine learning is one of the most effective approaches for identifying disease risk factors, this study aimed to explore the predictors of preeclampsia through machine learning.

Methods

This retrospective study evaluated data collected over a span of two years [2020–2022] that was retrieved from electronic health records from a tertiary care medical center in Bandar Abbas, Iran. It consists of 36 features, reflecting both demographic and medical characteristics of 8,888 patients who gave birth at our center during the study period. The original dataset contained a target variable that categorized women into two distinct groups: “preeclamptic” and “non-preeclamptic”. The input data were incorporated into 12 machine learning models. The area under the curve (AUC), accuracy, precision, Brier score, recall, F1 score, precision-recall (PR-AUC) were employed to evaluate the model's performance.

Results

The incidence of preeclampsia was 6.5%. Due to imbalanced data, the Synthetic Minority Over-sampling Technique (SMOTE) approach was utilized to run the models. Among all models Random Forest stands out as the top model across multiple measures with an accuracy of 0.993, AUC of 0.972, a Brier Score of 0.016, and a PR-AUC of 0.935, reflecting the least error and highlighting this model’s superior accuracy and dependability in predicting preeclampsia. Previous preeclampsi, gestational age, parity, maternal education, thyroid dysfunction, place of residency, diabetes, iron deficiency anemia, newborn sex, and maternal age showed to be the most significant features in predicting preeclampsia. Previous preeclampsia by far, was the most important predictor.

Conclusions

Utilizing the SMOTE approach to balance the data revealed that the Random Forest model stands out as the top model across multiple measures showing superior accuracy and dependability in predicting preeclampsia. Previous preeclampsia was the most important predictor.

Background

Preeclampsia is characterized by gestational hypertension occurring after 20 weeks of pregnancy, which significantly impacts maternal and neonatal mortality worldwide [1]. The incidence of preeclampsia varies between 1.8% and 16.7%. Worldwide, approximately 12% of mothers succumb to preeclampsia due to unidentified reasons, various risk factors, and numerous potential pathogenic phenotypes [2]. Preeclampsia is estimated to result in 46,000 to 76,000 maternal fatalities and 500,000 neonatal fatalities each year [3]. Demographic and clinical risk factors such as maternal age, ethnicity, previous pregnancies with preeclampsia, maternal metabolic syndrome, and various first-trimester biochemical and ultrasound indicators have been identified as contributors to the onset of preeclampsia. These factors have been included in numerous algorithms with varying performance

levels, especially for identifying late-onset preeclampsia [4]. Consequently, it is essential to create more efficient screening techniques. A substantial amount of knowledge has been gathered on possible risk markers and algorithms, and it is recommended that future research should focus on extensive prospective studies utilizing patient registries and innovative data fusion techniques to create effective, reliable, and clinically applicable screening algorithms for predicting preeclampsia.

Employing big data in this manner could also facilitate the formulation of algorithms capable of identifying pregnancies that would benefit from preventative care and be beneficial in creating and executing tailored preventive treatments [5]. In recent years, there has been a growing use of artificial intelligence in medicine and healthcare. The application of artificial intelligence in obstetrics and gynecology has garnered interest from the scientific community [4–6]. Lately, machine learning, a crucial part of artificial intelligence, has been extensively utilized in various medical areas, resulting in progress in disease diagnosis and prediction [6]. In supervised machine learning, a model is initially trained using a range of features linked to a defined outcome. The model is capable of predicting the outcome using new data. When examining a discrete outcome, the developed model is referred to as a classification algorithm [7]. Numerous machine learning techniques have been proposed to enhance the accuracy of data classification. In contrast to conventional parametric statistical methods, machine learning approaches do not require distributional assumptions regarding the dataset and are highly effective for large datasets. Nonetheless, each proposed machine learning approach possesses distinct traits for classification and outcome prediction, and its effectiveness can vary depending on conditions and datasets [8]. Given that machine learning is one of the most effective approaches for identifying disease risk factors, this study aimed to explore the predictors of preeclampsia through machine learning to lower maternal mortality rates.

Methods

This retrospective study adhered to the Declaration of Helsinki and was conducted following approval from the ethics committee. The dataset represents a span of two years [2020–2022] and comprises clinical data retrieved from electronic health records from a tertiary care medical center in Bandar Abbas, Iran. It consists of 36 features (nationality, age, place of residency, maternal education, attendance to birth class, medical insurance, body mass index, gestational age, parity, multiple pregnancy, smoking status, substance use, alcohol consumption, history of abortion, history of neonatal death, history of intrauterine fetal death, history of infertility, history of chronic hypertension, history of cardiovascular disease, history of iron deficiency anemia, history of hemoglobinopathy, history of hepatitis B, history of HIV, history of COVID-19, history of hypothyroidism, history of systemic lupus erythematosus or antiphospholipid syndrome, history of gestational diabetes, history of overt diabetes, history of urinary infections during pregnancy, history of flu during pregnancy, COVID-19 vaccination, history of preeclampsia in previous pregnancies, history of anticoagulant therapy during the current pregnancy, and newborn sex), reflecting both demographic and medical characteristics of 8,888 patients who gave birth in our center during the study period. As we obtained our data from the national electronic health records where all variables had to be completed mandatorily, there were no absent value in the data to address.

The original dataset contained a target variable that categorized women into two distinct groups: “preeclamptic” and “non-preeclamptic”. Preeclampsia was characterized as systolic blood pressure (SBP) ≥ 140 mm Hg and diastolic blood pressure (DBP) ≥ 90 mm Hg. Proteinuria was characterized by the presence of one or more of the following: random urine dipstick findings of at least 1 + on two separate instances, 24-hour proteinuria ≥ 300 mg, a urine protein/creatinine ratio of 30 mg/mmol, or any other newly appearing signs of organ dysfunction related to PE in the absence of proteinuria [9].

The input data were incorporated into 12 machine learning models. The models outlined below include a diverse array of simple, ensemble, and enhanced algorithms, enabling a comparison of their effectiveness. These frameworks included:

1. Logistic Regression [10], a linear model used for binary classification that yields probabilistic outcomes. It was selected for its straightforwardness and natural interpretability.
2. K-Nearest Neighbors (KNN) [11], a basic instance-based method that categorizes a data point according to the predominant class within its nearest neighbors. It is simple to execute and serves as an excellent standard for evaluating additional sophisticated models.
3. Decision Tree [12], which employs a tree-shaped framework of decisions and outcomes to categorize the information. It is easy to grasp and analyze, rendering it appropriate for preliminary data examination.
4. Random Forest [13], a collective approach that constructs several decision trees and merges their results to enhance precision and prevent overfitting. It is capable of managing extensive datasets with high dimensionality and provides a solid equilibrium between bias and variance.
5. Radial Kernel Support Vector Machine (RBF SVM) [14], a non-linear model utilizing the kernel trick to address intricate data distributions. It efficiently handles high-dimensional data and needs little memory, making it ideal for complicated and varied obstetric datasets.
6. Gradient Boosting [15], is an ensemble method that constructs models in succession, with each new model addressing the mistakes of the prior one. It serves as a robust instrument for enhancing precision.
7. XGBoost (Extreme Gradient Boosting) [16], a sophisticated gradient boosting algorithm recognized for its effectiveness and superior performance on structured datasets. It effectively manages missing data and provides comprehensive hyperparameter optimization.
8. AdaBoost (Adaptive Boosting) [17], an ensemble technique that merges various weak classifiers into a strong classifier by concentrating on the hardest instances to categorize.
9. Naive Bayes, a statistical classifier derived from Bayes theorem, presupposing independence among the input variables. It is particularly effective for small datasets containing categorical characteristics
10. CatBoost [18], a gradient boosted model designed specifically for categorical data. It provides excellent performance in complex predictions by using special techniques to handle nonlinear data.
11. Neural Network [19], a deep learning model that is suitable for more complex and nonlinear data. This algorithm was used to optimize the weights and achieve the best predictions. Neural network is very suitable for complex problems and nonlinear data.
12. Linear Support Vector Machine (Linear SVM) [20] is a powerful model for classification problems, especially suitable for linear and simple data. This model performs well for simple, linear data and is usually the best choice for linear problems.

We performed 10-fold cross-validation while creating the models. The training dataset and internal validation dataset were randomly divided into 10 groups through stratified random sampling. The variable for stratification in this randomization was the desired outcome. The models underwent training by combining eight groups and were validated internally using the remaining group each time. This procedure was carried out 10 times until every group had served as the validation set. We utilized 10-fold cross-validation, beginning with feature selection. Nevertheless, to effectively determine the optimal parameter tuning for each algorithm, we employed test split validation with an 80:20 ratio for both the training and validation data. In the final comparison of all algorithms, we used 10-fold cross-validation as well. The performance of each model was assessed and compared using the test data set. The area beneath the receiver operating characteristic curve (AUC) was employed to evaluate the model's

discrimination capability. Calibration was assessed through the slope, intercept, and Brier score of the calibration graph. Ultimately, we also presented the accuracy, precision, recall, F1 score, precision-recall (PR-AUC) and 95% confidence interval for these 12 algorithms.

The dataset features multiple columns with absent values. We created a method to address these missing values according to their quantity. If a specific feature column contains over 40% of its values missing, we eliminate that column to ensure data integrity. COVID-19 vaccination was eliminated from the analysis due to 63% missing data. As all the variables are categorical, we utilized mode imputation, substituting missing values with the most frequently occurring category. This method guarantees that our dataset is both comprehensive and maintains the overall quality and consistency of the information. The mode imputation method was used for history of anticoagulant therapy during the current pregnancy with missing data of 27%. All statistical analyses were performed using Python software (version 3.7.0).

Results

The main goal of this project was to develop machine learning-based modeling to predict the incidence of preeclampsia. One of the main challenges of the present data was the severe imbalance of the classes, with about 93.5% of the samples belonging to the non-preeclampsia class and only about 6.5% belonging to the preeclampsia class. This imbalance can lead to severe bias in the evaluation criteria and create a false impression of good model performance. For this reason, the models were evaluated in three different scenarios.

1. Raw data (without balancing)
2. Undersampling
3. SMOTE (Synthetic Minority Over-sampling Technique) balancing

Performance of models on raw data

In the first stage, different machine learning models were trained on raw, unbalanced data, in which the majority class (non-preeclampsia) accounted for about 93% and the minority class (preeclampsia) for only about 7% of the total samples. The results showed that almost all models – including Logistic Regression, Random Forest, XGBoost and AdaBoost – achieved very high accuracy of around 99%. The Precision value was also reported to be equal to 1 in many models. At first glance, such performance seems very desirable; however, a closer examination of the metrics more sensitive to data imbalance, particularly Sensitivity (Recall), PR-AUC, and most importantly, the Confusion Matrix, reveals a different picture.

In this case, the number of True Negatives (TN) is very high, which naturally results from the dominance of the majority class. In contrast, although the number of False Positives (FP) is very low, there are still False Negatives (FN). From a clinical perspective, the presence of FN is of great importance because it means that women who are truly at risk of preeclampsia are not being identified. In such a situation, the model may appear statistically “excellent” but not clinically reliable. One of the important points that became apparent at this stage is the severe limitation of Accuracy and Precision as the main evaluation criteria in unbalanced data. In this project, it was observed that even models such as Naive Bayes with very low accuracy had high AUC and still produced FN, compared to models with accuracy close to 99%. This shows that high accuracy does not necessarily mean good clinical performance.

High Precision (even equal to 1) is also not necessarily indicative of optimal performance, as this metric focuses only on positive predictions and is not sensitive to missing cases (FN). In unbalanced data, a model that rarely makes positive predictions can have very high Precision but still fail to identify true patients. Overall, the results show that using unbalanced raw data leads to apparently very favorable performance of the models in terms of Accuracy, Precision, and Specificity, but this performance mainly reflects the dominance of the majority class. (Table 1)

Table 1
Comparison of evaluation metrics of all models based on raw data.

Model	Accuracy	AUC	Precision	Sensitivity	Specificity	F1	PR_AUC	Brier
AdaBoost	0.993 (0.991–0.994)	0.967 (0.937–0.972)	1.000 (1.000–1.000)	0.891 (0.870–0.913)	1.000 (1.000–1.000)	0.943 (0.923–0.955)	0.928 (0.890–0.940)	0.071 (0.066–0.076)
CatBoost	0.992 (0.990–0.994)	0.959 (0.937–0.972)	1.000 (0.976–1.000)	0.882 (0.870–0.913)	1.000 (0.998–1.000)	0.937 (0.920–0.955)	0.923 (0.881–0.935)	0.008 (0.006–0.010)
Decision Tree	0.993 (0.991–0.994)	0.948 (0.936–0.966)	1.000 (1.000–1.000)	0.891 (0.870–0.913)	1.000 (1.000–1.000)	0.943 (0.923–0.955)	0.927 (0.887–0.941)	0.007 (0.006–0.010)
KNN (5)	0.989 (0.985–0.992)	0.940 (0.916–0.952)	1.000 (1.000–1.000)	0.848 (0.773–0.886)	1.000 (1.000–1.000)	0.913 (0.872–0.930)	0.913 (0.874–0.925)	0.011 (0.009–0.016)
KNN (7)	0.989 (0.984–0.991)	0.940 (0.923–0.954)	1.000 (1.000–1.000)	0.837 (0.756–0.870)	1.000 (1.000–1.000)	0.907 (0.861–0.927)	0.911 (0.872–0.926)	0.011 (0.009–0.015)
LightGBM	0.990 (0.990–0.993)	0.950 (0.927–0.959)	0.976 (0.953–0.994)	0.882 (0.870–0.913)	0.998 (0.997–1.000)	0.921 (0.918–0.946)	0.908 (0.879–0.930)	0.009 (0.007–0.010)
Linear SVM	0.993 (0.991–0.994)	0.944 (0.932–0.957)	1.000 (1.000–1.000)	0.891 (0.870–0.913)	1.000 (1.000–1.000)	0.943 (0.923–0.955)	0.901 (0.884–0.921)	0.007 (0.006–0.009)
Logistic Regression	0.993 (0.991–0.994)	0.964 (0.947–0.970)	1.000 (1.000–1.000)	0.891 (0.870–0.913)	1.000 (1.000–1.000)	0.943 (0.923–0.955)	0.926 (0.892–0.944)	0.007 (0.006–0.009)
Naive Bayes	0.085 (0.078–0.092)	0.947 (0.933–0.965)	0.066 (0.066–0.067)	1.000 (0.984–1.000)	0.022 (0.015–0.030)	0.124 (0.123–0.125)	0.774 (0.746–0.808)	0.913 (0.905–0.919)
Neural Net	0.992 (0.990–0.994)	0.945 (0.928–0.959)	1.000 (0.982–1.000)	0.882 (0.853–0.908)	1.000 (0.999–1.000)	0.936 (0.919–0.952)	0.918 (0.883–0.936)	0.008 (0.006–0.010)
RBF SVM	0.991 (0.986–0.993)	0.941 (0.935–0.954)	1.000 (1.000–1.000)	0.859 (0.805–0.891)	1.000 (1.000–1.000)	0.924 (0.884–0.943)	0.905 (0.870–0.930)	0.009 (0.007–0.012)
Random Forest	0.993 (0.991–0.994)	0.966 (0.942–0.987)	1.000 (1.000–1.000)	0.891 (0.870–0.913)	1.000 (1.000–1.000)	0.943 (0.923–0.955)	0.920 (0.890–0.942)	0.009 (0.008–0.011)
XGBoost	0.993 (0.991–0.994)	0.959 (0.930–0.976)	1.000 (1.000–1.000)	0.891 (0.870–0.913)	1.000 (1.000–1.000)	0.943 (0.923–0.955)	0.923 (0.884–0.942)	0.007 (0.006–0.010)

Data are presented with Confidence Intervals (CI).

Performance of models on under-sampled data

In order to deal with the problem of data imbalance, the under-sampling method was used. In this approach, the number of samples in the majority class (Non-Preeclampsia) was reduced to make the distribution of classes more balanced. The aim was to increase the ability of the models to correctly identify the minority class (Preeclampsia) and reduce the important clinical error False Negative (FN).

The results showed that after undersampling, unlike the raw data, the accuracy of the models decreased slightly (from about 99% to 96–98% in most models). This decrease in accuracy is not only not considered negative, but also indicates that the models are moving out of the "majority class bias" mode. In unbalanced data, high precision is often a misleading measure, while in more balanced data, reduced precision is more meaningful.

One of the most important achievements of under-sampling is the stability and improvement of Sensitivity (Recall) in most models. In almost all models, Sensitivity remained in the range of 0.88 to 0.91. This is very important from a clinical perspective, as it means a reduction in missed cases of preeclampsia. Overall, XGBoost, Random Forest, and Logistic Regression had more balanced performance in terms of Sensitivity, Precision, and AUC. Naive Bayes showed significant improvement over raw data, but still remained weaker in terms of Precision. KNN and Neural Network had good sensitivity but were associated with increased FP. Finally, linear and tree models showed more stable and interpretable behavior after undersampling.

Overall, the use of under-sampling allowed the models to avoid over-dependence on the majority class; significantly improve the ability to detect preeclampsia (the minority class), and exhibit more clinically appropriate behavior, even at the cost of a slight decrease in accuracy. Consequently, if the main goal is clinical screening and FN reduction, models trained with under-sampled data are more reliable and applicable than models trained on raw data, although they may require adjustment of decision thresholds or probability calibration. (Table 2)

Table 2
Comparison of evaluation metrics of all models based on under-sampling approach.

Model	Accuracy	AUC	Precision	Sensitivity	Specificity	F1	PR_AUC	Brier
AdaBoost	0.993 (0.991–0.994)	0.961 (0.938–0.972)	1.000 (1.000–1.000)	0.891 (0.870–0.913)	1.000 (1.000–1.000)	0.943 (0.923–0.955)	0.928 (0.888–0.939)	0.158 (0.156–0.160)
CatBoost	0.973 (0.960–0.981)	0.952 (0.934–0.963)	0.743 (0.642–0.862)	0.891 (0.870–0.913)	0.979 (0.965–0.990)	0.817 (0.740–0.855)	0.918 (0.880–0.935)	0.028 (0.024–0.033)
Decision Tree	0.992 (0.985–0.994)	0.959 (0.926–0.967)	1.000 (0.961–1.000)	0.891 (0.870–0.913)	1.000 (0.997–1.000)	0.931 (0.870–0.955)	0.923 (0.890–0.934)	0.024 (0.022–0.026)
KNN (5)	0.963 (0.957–0.974)	0.949 (0.927–0.976)	0.690 (0.629–0.744)	0.880 (0.832–0.908)	0.974 (0.964–0.979)	0.742 (0.723–0.816)	0.904 (0.834–0.920)	0.036 (0.031–0.041)
KNN (7)	0.971 (0.961–0.979)	0.952 (0.928–0.973)	0.776 (0.651–0.798)	0.882 (0.853–0.913)	0.983 (0.967–0.985)	0.781 (0.747–0.845)	0.905 (0.850–0.922)	0.033 (0.031–0.039)
LightGBM	0.964 (0.957–0.971)	0.943 (0.926–0.967)	0.675 (0.638–0.709)	0.891 (0.870–0.913)	0.970 (0.967–0.974)	0.763 (0.729–0.808)	0.910 (0.882–0.936)	0.036 (0.029–0.042)
Linear SVM	0.988 (0.983–0.994)	0.938 (0.917–0.955)	1.000 (0.853–1.000)	0.891 (0.870–0.913)	1.000 (0.990–1.000)	0.910 (0.865–0.952)	0.915 (0.885–0.922)	0.020 (0.016–0.026)
Logistic Regression	0.987 (0.982–0.991)	0.956 (0.953–0.963)	0.950 (0.838–0.954)	0.891 (0.870–0.913)	0.997 (0.987–0.997)	0.896 (0.854–0.930)	0.919 (0.885–0.940)	0.021 (0.019–0.025)
Naive Bayes	0.974 (0.963–0.979)	0.953 (0.939–0.969)	0.752 (0.673–0.795)	0.882 (0.870–0.908)	0.980 (0.970–0.984)	0.816 (0.760–0.845)	0.830 (0.793–0.874)	0.026 (0.021–0.037)
Neural Net	0.958 (0.945–0.963)	0.939 (0.923–0.953)	0.639 (0.544–0.656)	0.891 (0.870–0.908)	0.964 (0.946–0.968)	0.719 (0.681–0.763)	0.852 (0.778–0.925)	0.038 (0.033–0.050)
RBF SVM	0.962 (0.953–0.967)	0.949 (0.930–0.967)	0.663 (0.596–0.696)	0.891 (0.870–0.908)	0.970 (0.958–0.973)	0.746 (0.706–0.780)	0.906 (0.862–0.918)	0.035 (0.031–0.042)
Random Forest	0.992 (0.990–0.994)	0.961 (0.944–0.983)	1.000 (1.000–1.000)	0.891 (0.870–0.913)	1.000 (1.000–1.000)	0.937 (0.918–0.952)	0.928 (0.891–0.941)	0.050 (0.047–0.053)
XGBoost	0.977 (0.972–0.981)	0.954 (0.927–0.972)	0.790 (0.745–0.834)	0.891 (0.870–0.913)	0.983 (0.979–0.989)	0.833 (0.801–0.864)	0.917 (0.881–0.941)	0.028 (0.024–0.033)

Data are presented with Confidence Intervals (CI).

Performance of models using SMOTE

Due to the imbalance of the data (low prevalence of preeclampsia), the SMOTE method was used in this study to increase the minority class samples. The goal of using SMOTE was to improve the models' ability to identify true cases of preeclampsia and reduce the dangerous False Negative clinical error, without eliminating majority class information.

After applying SMOTE, the performance pattern of the models showed a significant change compared to the raw data. Sensitivity (Recall) remained at a high level in most models (around 0.88–0.89), indicating the appropriate power of the models in identifying patients with preeclampsia. Compared to the raw data, Precision decreased in some models, indicating an increase in False Positives; however, this decrease is considered more clinically acceptable than the increase in False Negatives. Accuracy remained high (around 98–99%), but was no longer the dominant criterion for judging the performance of the models. The F1-score improved or remained at a reasonable level in most models, indicating a better balance between Precision and Sensitivity. PR-AUC increased in many models; this is particularly important because PR-AUC is a more accurate measure of imbalanced data than ROC-AUC. Brier Score decreased in most models, indicating improved calibration of the models' output probabilities after SMOTE.

From a clinical perspective, using SMOTE allowed the models to correctly identify more patients as having preeclampsia. The probability of missing high-risk patients (FN) was reduced. In return, the number of unnecessary referrals (FP) increased slightly. In the context of screening and preventive care, this trade-off is perfectly acceptable, as the clinical cost of a false negative is much greater than a false positive.

Overall, the results showed that using SMOTE significantly improved the behavior of models in the face of data imbalance. This method increased the power of models in identifying the minority class without removing real data, and resulted in models with high sensitivity, better PR-AUC, and more appropriate probability calibration. Among the models examined, Random Forest, XGBoost, and AdaBoost showed the best overall performance after applying SMOTE and can be suggested as suitable options for developing clinical decision support systems in predicting preeclampsia.

Comparing different models considering the three scenarios described above showed that the random forest model applying the SMOTE approach had the best performance in predicting preeclampsia. (Table 3)

Table 3
Comparison of evaluation metrics of all models based on raw data.

Model	Accuracy	AUC	Precision	Sensitivity	Specificity	F1	PR_AUC	Brier
AdaBoost	0.993 (0.990–0.994)	0.970 (0.928–0.980)	1.000 (0.982–1.000)	0.891 (0.870–0.913)	1.000 (0.999–1.000)	0.943 (0.918–0.955)	0.929 (0.886–0.942)	0.161 (0.161–0.163)
CatBoost	0.993 (0.991–0.994)	0.955 (0.924–0.962)	1.000 (1.000–1.000)	0.891 (0.870–0.913)	1.000 (1.000–1.000)	0.943 (0.923–0.955)	0.920 (0.886–0.936)	0.007 (0.006–0.010)
Decision Tree	0.992 (0.989–0.994)	0.954 (0.933–0.967)	1.000 (0.954–1.000)	0.891 (0.870–0.913)	1.000 (0.997–1.000)	0.932 (0.913–0.955)	0.927 (0.890–0.939)	0.013 (0.010–0.014)
KNN (5)	0.977 (0.974–0.981)	0.936 (0.920–0.952)	0.817 (0.754–0.847)	0.880 (0.834–0.908)	0.986 (0.979–0.989)	0.835 (0.801–0.860)	0.890 (0.862–0.916)	0.018 (0.016–0.022)
KNN (7)	0.980 (0.969–0.980)	0.934 (0.927–0.951)	0.797 (0.724–0.830)	0.880 (0.840–0.908)	0.983 (0.979–0.988)	0.846 (0.785–0.856)	0.897 (0.864–0.912)	0.019 (0.019–0.024)
LightGBM	0.991 (0.989–0.994)	0.956 (0.930–0.958)	0.976 (0.960–1.000)	0.882 (0.870–0.913)	0.998 (0.997–1.000)	0.926 (0.913–0.955)	0.907 (0.883–0.929)	0.009 (0.006–0.010)
Linear SVM	0.712 (0.446–0.827)	0.861 (0.831–0.899)	0.168 (0.099–0.245)	0.817 (0.724–0.870)	0.703 (0.414–0.831)	0.279 (0.177–0.372)	0.713 (0.549–0.787)	0.185 (0.150–0.271)
Logistic Regression	0.992 (0.988–0.994)	0.962 (0.937–0.970)	1.000 (0.940–1.000)	0.891 (0.870–0.913)	1.000 (0.996–1.000)	0.930 (0.904–0.952)	0.926 (0.898–0.943)	0.018 (0.016–0.020)
Naive Bayes	0.093 (0.091–0.100)	0.953 (0.938–0.970)	0.067 (0.066–0.067)	1.000 (0.979–1.000)	0.030 (0.029–0.038)	0.125 (0.124–0.126)	0.795 (0.770–0.845)	0.907 (0.900–0.909)
Neural Net	0.982 (0.981–0.986)	0.911 (0.900–0.929)	0.871 (0.853–0.916)	0.839 (0.826–0.880)	0.991 (0.990–0.995)	0.863 (0.849–0.879)	0.879 (0.847–0.911)	0.018 (0.016–0.019)
RBF SVM	0.979 (0.976–0.983)	0.921 (0.906–0.939)	0.823 (0.783–0.884)	0.859 (0.804–0.908)	0.987 (0.982–0.992)	0.850 (0.813–0.865)	0.883 (0.838–0.914)	0.026 (0.023–0.030)
Random Forest	0.993 (0.991–0.994)	0.972 (0.941–0.983)	1.000 (1.000–1.000)	0.891 (0.870–0.913)	1.000 (1.000–1.000)	0.943 (0.923–0.955)	0.935 (0.892–0.946)	0.016 (0.014–0.018)
XGBoost	0.993 (0.991–0.994)	0.961 (0.924–0.971)	1.000 (1.000–1.000)	0.891 (0.870–0.913)	1.000 (1.000–1.000)	0.943 (0.923–0.955)	0.926 (0.888–0.938)	0.007 (0.006–0.010)

Data are presented with Confidence Intervals (CI).

Analysis of the importance of features in the Random Forest model after applying SMOTE

Figure 1 shows the ten most important features that the Random Forest model (after balancing the data with SMOTE) used to predict preeclampsia. The importance of each feature indicates its contribution to reducing the model's decision-making uncertainty; in other words, the higher the importance of a feature, the stronger its role in distinguishing people with and without preeclampsia. Previous preeclampsi, gestational age, parity, maternal education, thyroid dysfunction, place of residency, diabetes, iron deficiency anemia, newborn sex, and maternal age showed to be the most significant features in prediction preeclampsia. Previous preeclamsia by far, it is the most important predictor in the model.

The overall interpretation of the SHAP (SHapley Additive exPlanationsdiagram) shows how each feature affects the output of the Random Forest model (after applying SMOTE) in predicting preeclampsia. In this graph, the horizontal axis represents SHAP values, with positive values indicating an increase in the probability of predicting preeclampsia and negative values indicating a decrease in this probability. The color of the points also represents the value of the feature, with red indicating higher values and blue indicating lower values of the features. The order of displaying variables from top to bottom is determined based on the average absolute value of SHAP values and actually reflects the true importance of each feature in the model's decision-making process. (Fig. 2)

Discussion

In medical decision-making (patient risk evaluation, diagnosis), it is crucial that the choices are made efficiently and reliably. Machine learning models have been developed to evaluate large collections of clinical and physiological information to identify women at significant risk of developing preeclampsia before the onset of symptoms [4, 9]. These models aim to detect patterns that might be too subtle for humans to notice, thereby enhancing early diagnosis and allowing for prompt intervention. In this study, we applied different machine learning models to predict preeclampsia. Our model incorporated predictors primarily derived from medical history, which was the most commonly utilized predictor in models for predicting preeclampsia [9].

According to the results obtained from comparing different machine learning models on unbalanced and balanced data, it was found that using the SMOTE method provides more stable and clinically reliable performance than undersampling. SMOTE strengthened the minority class without removing real data and was able to strike a good balance between reducing false negative errors and controlling false positives, while undersampling, by removing part of the majority class data, caused information loss and greater fluctuation in model performance [21].

Based on SMOTE approach, among the various machine learning models, Random Forest, XGBoost, and AdaBoost have been identified as reliable and efficient methods for predicting preeclampsia, providing high classification accuracy along with a clear representation of the medical data. These models can assist the obstetrician in addressing the inquiries: “Will this pregnant patient experience preeclampsia?”

Random forest has gained popularity as a machine learning technique in clinical research due to its simplicity and interpretability, high accuracy, rapid training time, and strong capabilities in avoiding overfitting. Its broad applicability is also driven by its versatility, as it accommodates both classification and regression tasks [22, 23]. XGBoost model has also been introduced as an accurate and relaiable model in predicting preeclampsia [24]. A recent systematic review revelaed that Elastic net, stochastic gradient boosting, XGBoost, and Random forest were among the best models to predict preeclampsia [9].

Factors that have historically been associated with the onset of preeclampsia include a prior occurrence of preeclampsia, existing chronic kidney disease, hypertension, diabetes, autoimmune conditions like systemic lupus

erythematosus and antiphospholipid syndrome, advanced maternal age (over 40 years), and a body mass index exceeding 35 kg/m2, all of which have been connected to a heightened risk of developing preeclampsia [25–27]. In our study previous preeclampsi, gestational age, parity, maternal education, thyroid dysfunction, place of residency, diabetes, iron deficiency anemia, newborn sex, and maternal age showed to be the most significant features in prediction preeclampsia. Previous preeclamsia by far, it is the most important predictor in the model. Its high significance (predominantly over other features) indicates that a history of preeclampsia is the strongest risk factor for recurrence of the disease in subsequent pregnancies. This finding is fully consistent with clinical evidence and epidemiological studies and strengthens the clinical validity of the model [27, 28].

Gestational age is the second most important feature of the model. This indicates that the timing of symptoms and stages of pregnancy progression play a key role in identifying preeclampsia. The importance of this variable may reflect the association of preeclampsia with placental developmental disorders and hemodynamic changes during pregnancy [27, 29–31].

The inclusion of parity among the important characteristics suggests that previous pregnancy history and delivery experience may influence the risk of preeclampsia. This result is consistent with clinical findings that preeclampsia is more common in first pregnancies [27, 32].

The importance of maternal education, although less than direct clinical factors, suggests that socioeconomic factors can indirectly affect pregnancy health and access to prenatal care [33].

The presence of thyroid disorders among the contributing features indicates the role of underlying hormonal diseases in increasing the risk of preeclampsia, a factor that has also been reported in many clinical studies [34, 35].

The importance of the place of residency in predicting preeclampsia is relatively lower than other factors, but its presence in the list of top ten characteristics suggests that environmental differences, access to health services, or social conditions can have a limited but significant effect [27, 36].

Diabetes, as one of the important metabolic diseases, also played an acceptable role in this model. This result is fully consistent with previous correlation analyses and statistical tests [37, 38].

The importance of iron deficiency anemia indicates that nutritional status and anemia can play a role as a concomitant factor in the occurrence of preeclampsia, although the severity of its effect is less than that of the main factors [39, 40].

According to our findings, the low significance of fetal sex indicates that the baby's sex has a limited role in predicting preeclampsia and is mostly considered as a secondary factor. However some studies revealed that women who gave birth to a male newborn had higher odds of developing severe preeclampsia [41, 42].

Although maternal age is less important than other characteristics in the Random Forest model, it is considered an important factor in many studies. This suggests that the effect of maternal age is likely exerted through interactions with other factors (such as medical history or pregnancy conditions) [43–45].

One of the strengths of this research is the use of multiple variables with a large sample size. Another important point is the use of an accurate maternal pregnancy and childbirth information system that minimized the number of missing data. On the other hand, the imbalance of the data (small number of cases versus controls) is one of

the limitations of the research, which was addressed to ensure the validity of the analysis results through three scenarios described in the results section. The biggest limitation of this study is the external validation that has not been evaluated due to the time-consuming process and high cost. These limitations should be considered in further studies.

Conclusion

Machine learning algorithms such as Random Forest, XGBoost, and AdaBoost outperformed other prediction methods, offering higher accuracy and the ability to analyze complex, nonlinear data. However, several challenges limit the clinical applicability of these models, including the imbalanced data and the lack of external validation. Using SMOTE approach to balance the data showed that Random Forest with high accuracy and reliability can be considered as a predictive models in preeclamsia. Previous preeclampsi, gestational age, parity, maternal education, thyroid dysfunction, place of residency, diabetes, iron deficiency anemia, newborn sex, and maternal age showed to be the most significant features in prediction preeclampsia. Previous preeclampsia by far, was the most important predictor.

Abbreviations

XGBoost
Extreme gradient boost
KNN
K-nearest neighbors
SVM
Super vector machine
AUC
Area beneath the receiver operating characteristic curve
PA-AUC
Precision-recall of the area beneath the receiver operating characteristic curve
TN
True Negatives
FN
False Negative
TP
True Positive
FP
False positive
SMOTE
Synthetic Minority Over-sampling Technique

Declarations

Ethics approval and consent to participate

This research adhered to the Declaration of Helsinki and was carried out following approval from the ethics committee. The study was approved by the Ethics and Research Committee of Iran University of Medical Sciences

(IR.IUMS.REC.1404.395). The Ethics and Research Committee of Iran University of Medical Sciences waived the requirement for informed consent for participation because the study was retrospective in nature. Statistical analysis was conducted with patient confidentiality, adhering to ethics committee regulations.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Aknowlegment

All the authors aknowledge Iran University of Medical Sciences for their support.

Funding

Iran University of Medical Sciences.

Author Contribution

F.A. and F.D. wrote the proposal. F.M. contributed significantly to data collection. The findings were analyzed and interpreted by A.H. and A.N. F.D. was responsible for the manuscript's writing and editing. V.M. and N.R. and M.B. assessed the manuscript's scientific content critically. The final manuscript for submission was read and approved by all authors.

Acknowledgment

All of the authors acknowledged Iran University of Medical Sciences.

Data Availability

The datasets generated and analyzed during the current study are available from the corresponding author upon reasonable request.

References

1. Rahnemaei, F. A., Fashami, M. A., Abdi, F. & Abbasi, M. Factors effective in the prevention of Preeclampsia:A systematic review. *Taiwan. J. Obstet. Gynecol.* **59** (2), 173–182. 10.1016/j.tjog.2020.01.002 (2020).
2. Abdi, F., Aghaie, Z., Rahnemaie, F. S. & Alimoradi, Z. A Systematic Review of First Trimester Biochemical and Molecular Predictive Tests for Preeclampsia. *Curr. Hypertens. Rev.* **14** (1), 21–28. 10.2174/1573402114666180416160425 (2018).

3. Chiang, Y. T., Seow, K. M. & Chen, K. H. The Pathophysiological, Genetic, and Hormonal Changes in Preeclampsia: A Systematic Review of the Molecular Mechanisms. *Int. J. Mol. Sci.* **25** (8), 4532. 10.3390/ijms25084532 (2024).
4. Malakooti, N., Mehrnoush, V., Abdi, F., Farashah, M. S. V. & Darsareh, F. Development of a machine learning model to identify the predictors of the neonatal intensive care unit admission. *Sci. Rep.* **15** (1), 20914. 10.1038/s41598-025-06651-0 (2025).
5. Hedley, P. L., Hagen, C. M., Wilstrup, C. & Christiansen, M. The use of artificial intelligence and machine learning methods in early pregnancy pre-eclampsia screening: A systematic review protocol. *PLoS One*. **18** (4), e0272465. 10.1371/journal.pone.0272465 (2023).
6. Rezaei, N. et al. Machine Learning Approach to Predict Emergency Cesarean Sections Among Nulliparous Women. *Cureus* **17** (8), e90501. 10.7759/cureus.90501 (2025).
7. Safarzadeh, S., Ardabili, N. S., Farashah, M. V., Roozbeh, N. & Darsareh, F. Predicting mother and newborn skin-to-skin contact using a machine learning approach. *BMC Pregnancy Childbirth*. **25** (1), 182. 10.1186/s12884-025-07313-9 (2025).
8. Banaei, M. et al. Utilizing machine learning to predict the risk factors of episiotomy in parturient women. *AJOG Glob Rep.* **5** (1), 100420. 10.1016/j.xagr.2024.100420 (2024).
9. Ranjbar, A., Ghamsari, S., Boujarzade, B., Mehrnoush, V. & Darsareh, F. Predicting risk of postpartum hemorrhage using machine learning approach: A systematic review. *GOCM* **3** (3), 170–174. doi.org/10.1016/j.gocm.2023.07.002 (2023).
10. Sperandei, S. Understanding logistic regression analysis. Biochem Med (Zagreb). ;24(1):12 – 8. doi: 10.11613/BM.2014.003. PMID: 24627710 (2014).
11. Zhang, Z. Introduction to machine learning: k-nearest neighbors. *Ann. Transl Med.* **4** (11), 218. 10.21037/atm.2016.03.37 (2016).
12. Song, Y. Y. & Lu, Y. Decision tree methods: applications for classification and prediction. *Shanghai Arch. Psychiatry*. **27** (2), 130–135. 10.11919/j.issn.1002-0829.215044 (2015).
13. Rigatti, S. J. & Random Forest *J. Insur Med.* ;**47**(1):31–39. doi: 10.17849/insm-47-01-31-39.1. (2017).
14. Montesinos López, O. A., Montesinos López, A. & Crossa, J. Multivariate Statistical Machine Learning Methods for Genomic Prediction [Internet]. Cham (CH): Springer; Chapter 9, Support Vector Machines and Support Vector Regression. 2022 Jan 14. Available from: https://www.ncbi.nlm.nih.gov/books/NBK583961/ (2022). 10.1007/978-3-030-89010-0_9
15. Boldini, D., Grisoni, F., Kuhn, D., Friedrich, L. & Sieber, S. A. Practical guidelines for the use of gradient boosting for molecular property prediction. *J. Cheminform*. **15** (1), 73. 10.1186/s13321-023-00743-7 (2023).
16. Wiens, M., Verone-Boyle, A., Henscheid, N., Podichetty, J. T. & Burton, J. A Tutorial and Use Case Example of the eXtreme Gradient Boosting (XGBoost) Artificial Intelligence Algorithm for Drug Development Applications. *Clin. Transl Sci.* **18** (3), e70172. 10.1111/cts.70172 (2025).
17. Freund, Y. & Schapire, R. E. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *J. Comput. Syst. Sci.* **55** (1), 119–139 (1997).
18. Hancock, J. T. & Khoshgoftaar, T. M. CatBoost for big data: an interdisciplinary review. *J. Big Data*. **7** (1), 94. 10.1186/s40537-020-00369-8 (2020).
19. Kufel, J. et al. What Is Machine Learning, Artificial Neural Networks and Deep Learning?-Examples of Practical Applications in Medicine. *Diagnostics (Basel)*. **13** (15), 2582. 10.3390/diagnostics13152582 (2023).

20. Wu, J. & Yang, H. Linear regression-based efficient SVM learning for large-scale classification. *IEEE Trans. Neural Netw. Learn. Syst.* **26** (10), 2357–2369. 10.1109/TNNLS.2014.2382123 (2015).
21. Kivrak, M., Avci, U., Uzun, H. & Ardic, C. The Impact of the SMOTE Method on Machine Learning and Ensemble Learning Performance Results in Addressing Class Imbalance in Data Used for Predicting Total Testosterone Deficiency in Type 2 Diabetes Patients. *Diagnostics (Basel)*. **14** (23), 2634. 10.3390/diagnostics14232634 (2024).
22. Zhao, Q. et al. Early prediction of preeclampsia from clinical, multi-omics and laboratory data using random forest model. *BMC Pregnancy Childbirth*. **25** (1), 531. 10.1186/s12884-025-07582-4 (2025).
23. Wang, L. et al. Prediction Model of Adverse Pregnancy Outcome in Pre-Eclampsia Based on Logistic Regression and Random Forest Algorithm. *Altern. Ther. Health Med.* **30** (1), 142–147 (2024).
24. Kovacheva, V. P. et al. Preeclampsia Prediction Using Machine Learning and Polygenic Risk Scores From Clinical and Genetic Risk Factors in Early and Late Pregnancies. *Hypertension* **81** (2), 264–272. 10.1161/HYPERTENSIONAHA.123.21053 (2024).
25. Li, T. et al. Prediction model of preeclampsia using machine learning based methods: a population based cohort study in China. *Front. Endocrinol. (Lausanne)*. **15**, 1345573. 10.3389/fendo.2024.1345573 (2024).
26. Jhee, J. H. et al. Prediction model development of late-onset preeclampsia using machine learning-based methods. *PLoS One*. **14** (8), e0221202. 10.1371/journal.pone.0221202 (2019).
27. Mohamed Dkeen, N. O. et al. Artificial Intelligence Applications in Obstetric Risk Prediction: A Systematic Review of Machine Learning Models for Preeclampsia. *Cureus* **17** (5), e83961. 10.7759/cureus.83961 (2025).
28. Nie, Q. et al. Evaluation of outcomes and risk factors for recurrent preeclampsia in a subsequent pregnancy. *Arch. Gynecol. Obstet.* **310** (5), 2487–2495. 10.1007/s00404-024-07751-4 (2024).
29. Boadu, W. I. O. et al. Gestational Age-Specific Prevalence of Preeclampsia Among Pregnant Women in Ghana: A Hospital-Based Retrospective Observational Study. *Biomed. Res. Int.* **2025**, 4728838. 10.1155/bmri/4728838 (2025).
30. Haavaldsen, C., Thomas, O. & Eskild, A. Risk of pre-eclampsia according to gestational age across three decades: population-based study. *Ultrasound Obstet. Gynecol.* **7** 10.1002/uog.70160 (2026 Jan).
31. Ranjbar, A. et al. The Incidence and Outcomes of Late-Term Pregnancy. *Cureus* **15** (1), e33550. 10.7759/cureus.33550 (2023).
32. Long, P. A., Abell, D. A. & Beischer, N. A. Parity and pre-eclampsia. *Aust N Z. J. Obstet. Gynaecol.* **19** (4), 203–206. 10.1111/j.1479-828x.1979.tb01373.x (1979).
33. Johnson, S. M., Kelly, E. N., LaBrot, B. & Ryczak, K. Effect of Education on Adherence to Recommended Prenatal Practices among Indigenous Ngäbe-Buglé Communities of Panama. *Med. (Kaunas)*. **60** (7), 1055. 10.3390/medicina60071055 (2024).
34. Aryee, N. A. et al. Thyroid hormone profile in preeclampsia patients: a case control study. *Gynecol. Endocrinol.* **39** (1), 2186136. 10.1080/09513590.2023.2186136 (2023).
35. Hajifoghaha, M., Teshnizi, S. H., Forouhari, S. & Dabbaghmanesh, M. H. Association of thyroid function test abnormalities with preeclampsia: a systematic review and meta-analysis. *BMC Endocr. Disord*. **22** (1), 240. 10.1186/s12902-022-01154-9 (2022).
36. Mehrnoush, V. et al. Urban-rural differences in the pregnancy-related adverse outcome. *Gynecol. Obstetr Clin. Med.* **3** (1), 51–55. doi.org/10.1016/j.gocm.2022.12.001 (2023).

37. Weissgerber, T. L. & Mudd, L. M. Preeclampsia and diabetes. *Curr. Diab Rep.* **15** (3), 9. 10.1007/s11892-015-0579-4 (2015).
38. Hornová, M., Šimják, P. & Anderlová, K. Preeclampsia and diabetes mellitus. *Ceska Gynekol.* **88** (6), 467–471. 10.48095/cccg2023467 (2023). English.
39. Alahari, S. et al. JMJD6 Dysfunction Due to Iron Deficiency in Preeclampsia Disrupts Fibronectin Homeostasis Resulting in Diminished Trophoblast Migration. *Front. Cell. Dev. Biol.* **9**, 652607. 10.3389/fcell.2021.652607 (2021).
40. Safarzadeh, S., Banihashemi, F., Montazeri, F., Roozbeh, N. & Darsareh, F. Maternal and Neonatal Outcomes of Iron Deficiency Anemia: A Retrospective Cohort Study. *Cureus* **15** (12), e51365. 10.7759/cureus.51365 (2023).
41. Adam, G. K., Nasralla, K., Munder, O., AlHabardi, N. & Adam, I. Male newborn as a determinant of severe preeclampsia: a case-control study. *Sci. Rep.* **15** (1), 30054. 10.1038/s41598-025-16346-1 (2025).
42. Mirzakhani, H. & Weiss, S. T. Fetal sex and risk of preeclampsia: Dose maternal race matter? *J. Matern Fetal Neonatal Med.* **35** (17), 3379–3387. 10.1080/14767058.2020.1818221 (2022).
43. Gilboa, I. et al. The Association between Advanced Maternal Age and the Manifestations of Preeclampsia with Severe Features. *J. Clin. Med.* **12** (20), 6545. 10.3390/jcm12206545 (2023).
44. Ranjbar, A., Jahromi, M. S., Boujarzadeh, B., Roozbeh, N. & MehrnoushV, Darsareh, F. Pregnancy, childbirth and neonatal outcomes associated with adolescent pregnancy. *Gynecol. Obstetr Clin. Med.* **3**, 100–105. doi.org/10.1016/j.gocm.2023.02.002 (2023).
45. Shekari, M. et al. Advanced maternal age and adverse obstetrical and neonatal outcomes of singleton pregnancies. *Gynecol. Obstet. Clin. Med.* **2** (4), 175–180 (2022).

Figures

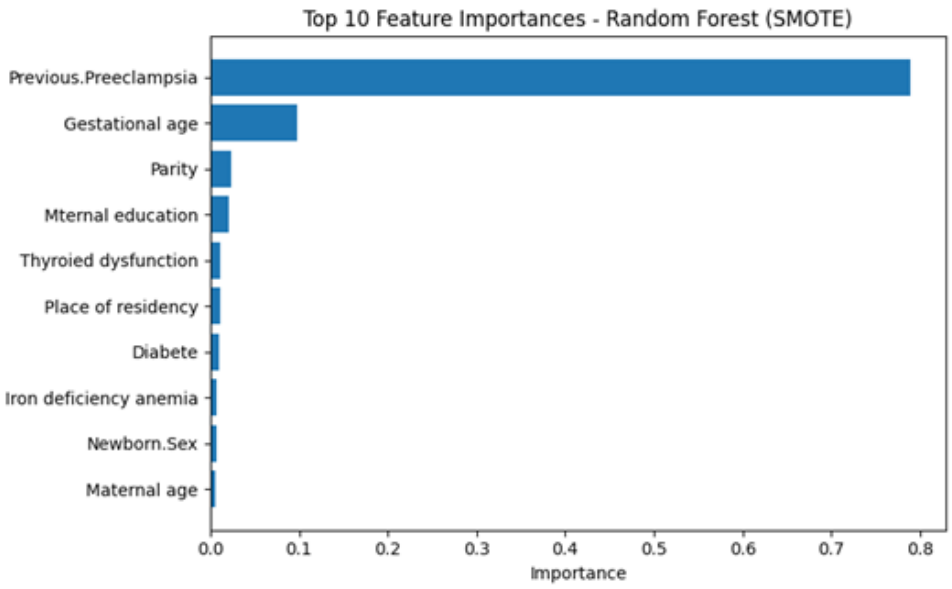


Figure 1

The most important features in predicting preeclampsia according to Random Forest model.

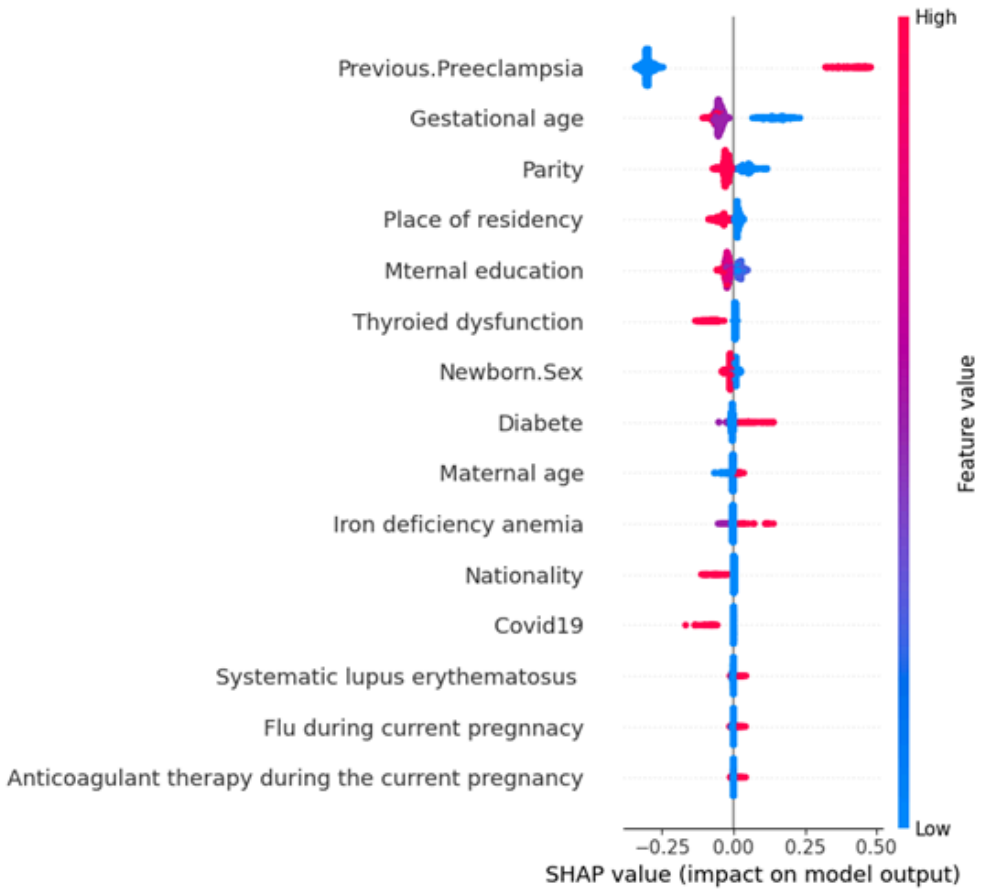


Figure 2

Visualization plots of SHAP values from an Random Forest model when predicting preeclampsia.