British Journal of Hospital Medicine

Br. J. Hosp. Med. (Lond) **2026**; 87(8): 56968
https://doi.org/10.31083/BJHM56968

Article

Building a Machine Learning Prediction Model for Postpartum Depression Based on Ensemble Learning Strategies

Fengyuan Zhang[1,†], Zhengcheng Tu[2,†], Xiang Hong[3], Xiaoli Wu[4], Suli Wang[4,*,§], Hong Yu[1,*,§]

[1]Department of Obstetrics and Gynecology, Zhongda Hospital, School of Medicine, Southeast University, 210009 Nanjing, Jiangsu, China
[2]School of Translational Medicine, Faculty of Medicine, Nursing and Health Sciences, Monash University, Melbourne, VIC 3800, Australia
[3]Department of Epidemiology and Health Statistics, Key Laboratory of Environmental Medicine and Engineering of Ministry of Education, School of Public Health, Southeast University, 210009 Nanjing, Jiangsu, China
[4]Department of Obstetrics and Gynecology, Women's Hospital of Nanjing Medical University, Nanjing Women and Children's Healthcare Hospital, 210004 Nanjing, Jiangsu, China
*Correspondence: 798906297@qq.com (Suli Wang); yuhong650325@sina.com (Hong Yu)
[†]These authors contributed equally.
[§]These authors contributed equally.
Academic Editor: Ashok Roy
Submitted: 17 December 2025 Revised: 3 March 2026 Accepted: 18 March 2026 Published: 12 August 2026

Abstract

Aims/Background: Postpartum depression (PPD) is a common perinatal mental health disorder that leads to adverse maternal and infant outcomes. Existing screening strategies rely on self-reported symptom presence, potentially delaying early identification. This study aims to integrate biological, psychological, and social predictive factors using an ensemble learning (EL) strategy to develop a superior early prediction machine learning (ML) model for PPD. **Methods:** By integrating sociodemographic characteristics, clinical baseline data, and psychological test results of 1323 postpartum women, PPD status was classified using the Edinburgh Postnatal Depression Scale (EPDS). Eight selection methods were used, and eleven base ML models were optimized through EL strategies (Voting and Stacking) to identify the model with the optimal predictive performance. **Results:** Twenty predictor features were selected for the model construction. The Voting top 5+XGBoost (eXtreme Gradient Boosting, Weight) model demonstrated the best overall performance (area under the curve [AUC] = 0.835) and the highest specificity (0.906), indicating that it is more suitable for differential diagnosis following an initial positive screening result. The Voting (Synthetic Minority Over-sampling Technique [SMOTE]) model performed exceptionally well in sensitivity (0.758) and clinical net benefit at low-risk thresholds, indicating its suitability for early screening for PPD. **Conclusion:** The results demonstrate that ensemble learning models outperform individual ML prediction models. Voting top 5+XGBoost (Weight) and Voting (SMOTE) show respective advantages in predictive specificity and sensitivity. This suite of models can provide distinct data-driven prediction strategies tailored to different clinical application contexts.

Keywords: postpartum depression; machine learning; ensemble learning; prediction methods

1. Introduction

Perinatal mental health affects up to 20% of pregnant women and postpartum women, which constitutes a major public health concern [1]. Among the causes of suicide among pregnant women and new mothers, 15% are attributable to postpartum depression (PPD) [2], and most occur within the first year after childbirth [3]. Moreover, PPD can have long-term negative effects on infants' emotional and cognitive development, family relationships, and social functioning [4]. Comprehensive screening and proactive identification of maternal mental health conditions are among the most commonly used methods for preventing PPD. The Edinburgh Postnatal Depression Scale (EPDS) is currently the most widely used screening tool for PPD, with both sensitivity and specificity exceeding 80% [5]. The results primarily rely on self-reported symptoms. This approach may lead to delayed identification or underdetection due to stigma, cultural factors, or limited symptom awareness. Moreover, the current screening strategy is typically applied after symptom onset, limiting its value for early prevention [6]. Therefore, constructing effective predictive models based on pre-symptomatic predictive factors can enhance early screening for PPD.

Social support and sleep quality are critical factors influencing the onset and progression of PPD [7]. Research indicates that low levels of social support during the perinatal period, particularly insufficient emotional and social support from partners and families, significantly increase maternal loneliness and stress [8]. Concurrently, physiological changes and childcare demands often lead to sleep disturbances during this period. These disturbances can further disrupt neuroendocrine balance in emotional regulation [9], increasing the risk of developing PPD. In addition to these sociopsychological factors, a series of objective, measurable physiological indicators-including inflam-

mation markers, thyroid function [10], postpartum anemia [11] that can be collected during pregnancy and the early postpartum period have also been identified as important biological factors for PPD risk. Machine learning (ML) methods can help integrate multi-domain predictive factors to build predictive models.

ML refers to mathematical frameworks that use algorithms to automatically learn patterns and rules from data, thereby enabling predictions or decisions on unseen data [12,13]. Ensemble learning (EL) strategies represent a core methodology for improving model performance in ML. By combining the results of multiple base models, EL significantly improves overall accuracy, robustness, and generalization while reducing the risk of overfitting. In ML, ensemble methods are particularly well-suited for predicting disease outcomes [14,15]. To date, no studies on PPD have reported using EL strategies for model optimization.

This study aims to develop and validate interpretable ML models for the earlier identification of women at high risk of PPD by integrating biological, psychological, and social factors and applying EL strategies. Based on the above objectives, we propose the following hypotheses: (1) By integrating diverse feature selection methods to identify robust predictors and systematically combining a broad range of foundational ML models through EL strategies, the resulting framework will achieve superior and more reliable PPD prediction compared to traditional approaches. (2) This method is expected to generate distinct models with clinically complementary strengths, offering better solutions for early screening (high sensitivity) and confirmatory evaluation (high specificity).

2. Methods

2.1 Study Design and Participants

This study selected women who underwent a 42-day postpartum follow-up visit at Zhongda Hospital and Nanjing Women and Children's Healthcare Hospital from June to December 2024. The inclusion criteria were as follows: (1) participants aged 18 years or older, who received regular prenatal check-ups, delivery and postpartum examinations at the participating hospitals and who had complete clinical information; (2) at 42 days postpartum, depressive symptoms over the past week were assessed using the EPDS, sociodemographic information was collected a structured questionnaire, the social support rating scale (SSRS) was used to evaluate the level of social support over the past year, and the Pittsburgh Sleep Quality Index (PSQI) was applied to assess sleep quality during the first month postpartum; (3) no serious cardiovascular, neurological, pulmonary, or renal diseases; (4) no history of substance abuse; (5) provision of signed informed consent. The exclusion criteria were as follows: (1) twin pregnancies; (2) preconception diabetes; (3) pre-pregnancy hypertensive disorders; (4) hyperthyroidism; (5) those who have had a miscarriage or stillbirth; (6) diagnosis of schizophrenia and bipolar disorder; (7) family history of mental illness; (8) lack of clinical information; (9) missing questionnaire information. Following the application of these inclusion and exclusion criteria, 1323 women were enrolled in the study.

All scales were administered under the supervision of trained physicians. Retrospective data included clinical information such as maternal baseline characteristics, neonatal outcomes, liver and renal function, biochemical parameters, and thyroid function during pregnancy, and complete blood count results during the postpartum period. Participants were divided into PPD and control groups based on EPDS scores ($\geq$10 and <10, respectively).

2.2 Instruments

2.2.1 Edinburgh Postnatal Depression Scale (EPDS)

The Edinburgh Postnatal Depression Scale (EPDS), developed by Cox et al. in 1987 [16] and adapted into Chinese by Lee et al. in 1998 [17], contains 10 items assessing symptoms such as loss of interest/pleasure, anxiety, sleep disturbance, sadness, and self-harm thoughts. Each item is scored from 0 to 3, with higher scores indicating greater severity of depressive symptoms. In this study, a cutoff score of 10 was used to identify PPD.

2.2.2 Social Support Rating Scale (SSRS)

SSRS was developed by Shuiyuan Xiao based on the sociocultural context of China [18]. The scale comprises 10 items across three dimensions: objective support, subjective support, and the utilization of social support. Higher total scores indicate greater levels of perceived social support.

2.2.3 Pittsburgh Sleep Quality Index (PSQI)

PSQI is a commonly used tool for assessing sleep quality [19], which consists of 19 items across seven components. Higher scores indicate poorer sleep quality; a total score of $\leq$7 reflects good sleep quality, whereas >7 indicates sleep disturbance [20].

2.2.4 Sociodemographic Characteristics and Clinical Baseline Data

This study collected sociodemographic data using a self-designed questionnaire administered in Chinese. The questionnaire included the following domains: age, marital status, educational level, occupation, postpartum working condition, smoking and drinking habits, primary caregiver, exposure to stressful events during pregnancy, exercise situation during pregnancy, evaluation of marriage and living environment, help-seeking behavior and willingness to disclose personal concerns and so on.

2.3 Feature Selection

This study performed collinearity and correlation analysis on all variables. It used 8 methods to screen for

predictive factors, avoiding bias from a single method and enhancing the robustness of the screening results.

(1) Stepwise regression (SR): It is a feature selection method that iteratively selects variables to build a parsimonious model based on predefined statistical criteria. It includes three main approaches: forward selection (FS), backward selection (BS), and bidirectional elimination (BE).

(2) Least absolute shrinkage and selection operator (LASSO): It uses an L1-regularized regression method, which compresses the coefficients of unimportant variables to zero through a penalty term, automatically selects key features to reduce overfitting, and enhances model interpretability.

(3) Random forest (RF): It efficiently identifies key features by integrating multiple decision trees, taking into account nonlinear relationships and interactions between features, and is suitable for high-dimensional data and resistant to overfitting. Mean decrease in accuracy (MDA) and mean decrease in Gini (MDG) coefficient were used to quantify feature importance and to guide feature selection.

(4) eXtreme Gradient Boosting (XGBoost): It is a gradient boosting-based EL method that can be used to identify important features through built-in feature importance metrics (Gain, Weight, and Cover) and regularization mechanisms. It can capture nonlinear relationships and feature interactions, making it suitable for high-dimensional data, and delivering strong predictive performance.

(5) Support vector machine-recursive feature elimination (SVM-RFE): It is a feature selection method that combines support vector machines with a recursive elimination strategy to iteratively remove less important features based on their contribution to the decision function, thereby identifying the most informative features.

Considering that feature factors play a crucial role in most feature selection methods (effectively filtering out method-specific noise, enhancing feature robustness and reproducibility, and implementing the concept of EL strategy during feature selection) [21]. Balancing selection rigor with information completeness (avoiding the loss of potentially important predictors due to overly stringent criteria or the introduction of redundancy and noise from overly lenient criteria) [22], we ultimately included predictive factors that occurred at least 6 times.

2.4 Model Development and Evaluation

The entire dataset was randomly split into independent training and testing sets at an 8:2 ratio. Then, the training set was subjected to stratified 10-fold cross-validation to maintain balanced data distribution and reduce bias in model evaluation. Simultaneously, 2 imbalance mitigation techniques-Synthetic Minority Over-sampling Technique (SMOTE) and sample weighting (Weight)-were employed to address the class imbalance between the PPD and control groups in the training set. Eleven basic ML methods, including categorical boosting (CatBoost), deep learning (DL), extremely randomized trees, gradient boosting machine, light gradient boosting machine, generalized linear model (GLM), K-Nearest Neighbors, naive bayes, RF, support vector machine (SVM), and XGBoost, were selected to train the models on the training set.

In the training set, 2 EL strategies-Voting and Stacking-were employed to integrate multiple base learners. We first grouped the samples using balancing strategies (SMOTE and Weight) and selected the top 5-performing models from each group. These yielded 2 independent Voting models (Voting-SMOTE and Voting-Weight) to evaluate the effectiveness of collective Voting. Next, we selected the optimal model from each base algorithm based on the area under the curve (AUC), and then constructed Voting top 3 and Voting top 5 by incorporating the top 3 and top 5 models, respectively. The 11 filtered base models and 2 auxiliary soft Voting EL models (Voting top 3 and Voting top 5) constitute the first-level learners. From these candidates, up to 6 models with the highest cross-validated AUC were selected, and all possible combinations of 2 to 5 models were then constructed, yielding a maximum of 56 stacking candidates. Out-of-fold (OOF) predictions were generated using 10-fold stratified cross-validation to form meta-features, which were subsequently integrated using a RidgeClassifierCV as the second-level meta-learner.

The discrimination metrics for all models include the AUC, precision, sensitivity, specificity, positive likelihood ratio (PLR), negative likelihood ratio (NLR), positive predictive value (PPV), negative predictive value (NPV), F1 score, and Brier score. The models' discriminative ability, predictive probability, clinical utility, and practical application value are evaluated using receiver operating characteristic (ROC) curves, calibration curves, decision curve analysis (DCA), and clinical impact curves. SHapley Additive exPlanations (SHAP) and a nomogram were used to quantify the explanatory model's power. The flowchart is shown in Fig. 1.

2.5 Data Processing and Statistical Analysis

To ensure data integrity, dual independent data entry was implemented. Predictors with >10% missing values or anomalies were excluded, while remaining missing data were imputed using missForest. Normality of continuous variables was assessed using the Shapiro-Wilk test. Variables were characterized as mean ± SD (standard deviation, normal distribution), median (Q1–Q3) (skewed distribution), or frequencies (categorical variables). Correlation analyses employed Pearson or Cramer's V for feature redundancy screening and multicollinearity assessment. Univariate screening (*t*-tests/Mann-Whitney U tests/χ^2 tests; $p < 0.05$) identified features for binary logistic regression (with Z-score normalization), ultimately generating a predictive nomogram. All analyses were conducted using SPSS 26.0 (IBM, Armonk, NY, USA), R 4.4.1 (R Foun-

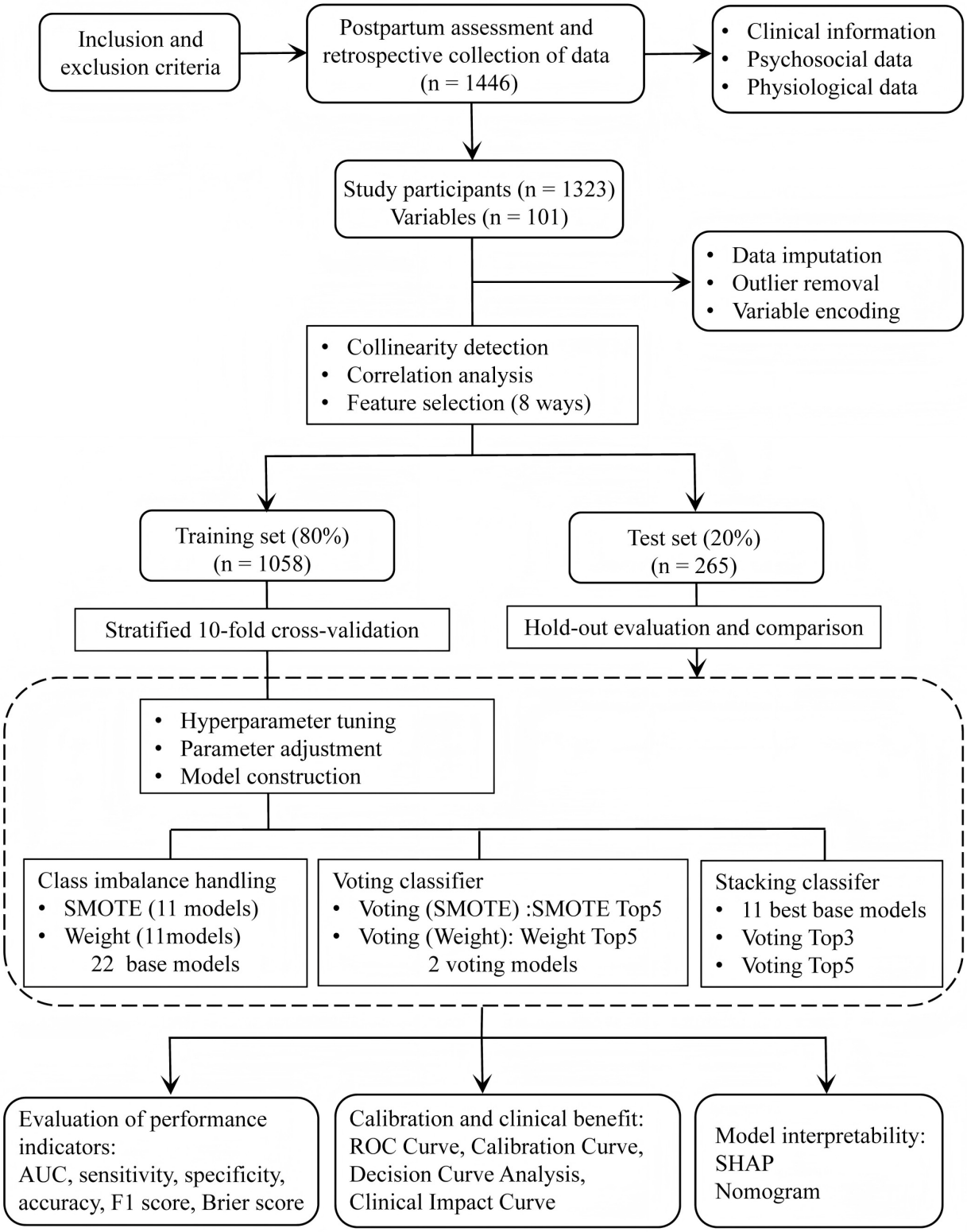

Fig. 1. Research flowchart. Abbreviations: SMOTE, Synthetic Minority Over-sampling Technique; AUC, area under the curve; ROC, receiver operating characteristic; SHAP, SHapley Additive exPlanations.

dation for Statistical Computing, Vienna, Austria), and Python 3.11.13 (Python Software Foundation, Wilmington, DE, USA).

2.5.1 Missing Data Imputation

After initial data cleaning, variables with more than 10% missing values were excluded. For the remaining variables, missing data were imputed using the missFor-

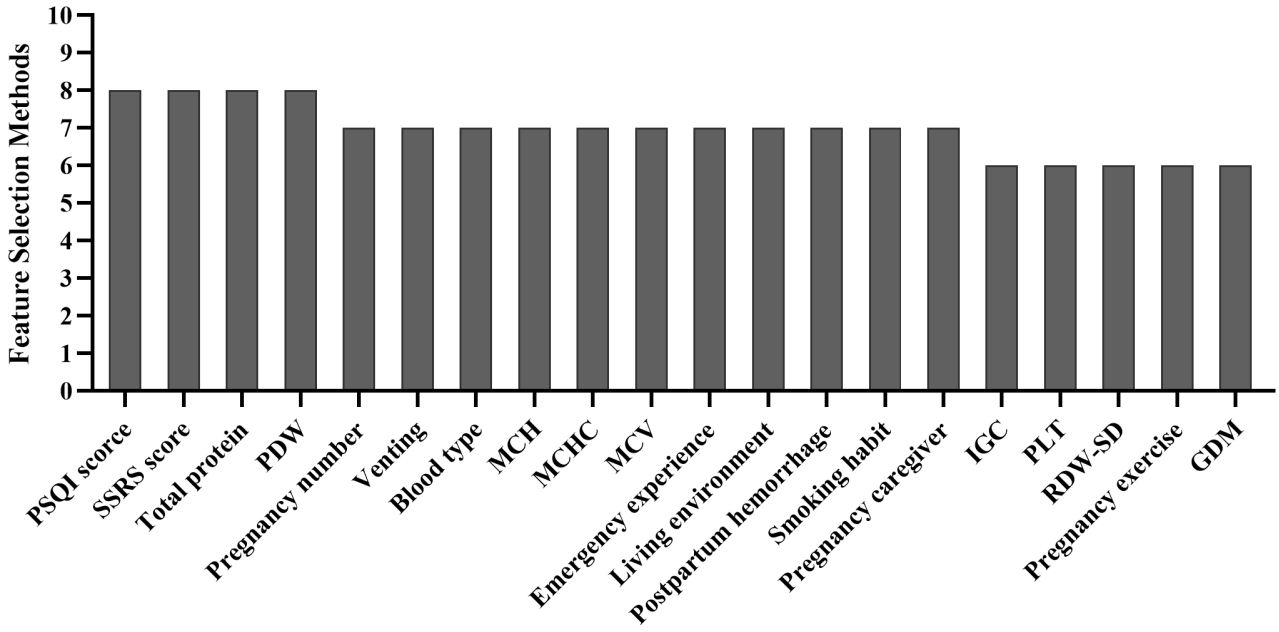

Fig. 2. The 20 factors ultimately incorporated into the machine learning model. Abbreviations: PSQI, Pittsburgh Sleep Quality Index; SSRS, social support rating scale; PDW, platelet distribution width; MCH, mean corpuscular hemoglobin; MCHC, mean corpuscular hemoglobin concentration; MCV, mean corpuscular volume; IGC, immature granulocyte count; PLT, platelet count; RDW-SD, red cell distribution width-standard deviation; GDM, gestational diabetes mellitus.

est algorithm, a nonparametric random forest-based method suitable for datasets with mixed continuous and categorical variables. To minimize the risk of information leakage, data were first partitioned into training and test sets, and the imputation procedure was conducted based on the training data. The same imputation scheme was then applied consistently across both the training and test sets. This approach does not rely on distributional assumptions and is robust to nonlinear relationships and interactions among predictors.

2.5.2 Hyperparameter Tuning

Hyperparameter optimization was performed for each of the 11 base ML models within the training set using a randomized search strategy. For each algorithm, a predefined hyperparameter search space was specified according to its methodological characteristics. Candidate configurations were evaluated using cross-validated performance, with AUC as the primary optimization criterion. The configuration achieving the highest cross-validated AUC was selected as the final tuned model.

2.5.3 Cross-Validation Strategy and Leakage Control

Model training and hyperparameter tuning were conducted using stratified 10-fold cross-validation, ensuring that the class distribution was preserved in each fold and reducing evaluation bias associated with class imbalance. All preprocessing steps were incorporated within the cross-validation framework and applied separately within each training fold to prevent information leakage. These steps included feature scaling for models sensitive to variable magnitude, appropriate encoding of categorical variables, and handling of class imbalance.

3. Results

3.1 Baseline Characteristics and Correlation Analysis

Among the 1323 participants, 311 (23.5%) women scored above the threshold (≥10) on the EPDS, indicating PPD. Women under the age of 35 accounted for 85.2%. The median age at marriage of the entire group was 27 years (26–29), 64.3% were primiparous, and 77.3% had complications during pregnancy. All study variables and comparative analyses were detailed in **Supplementary Tables 1–3**. The correlation analysis of all variables was presented in **Supplementary Fig. 1**. The top 20 collinear factors, ranked by variance inflation factor (VIF), were presented in **Supplementary Fig. 2**.

3.2 Selection of Predictor Variables

Among the 8 feature factor screening methods, 5 methods (stepwise regression-forward selection [SR-FS], stepwise regression-backwards selection [SR-BS], stepwise regression-bidirectional elimination [SR-BE], LASSO, SVM-RFE) selected 21, 30, 21, 16, and 15 predictive variables, respectively. In comparison, 3 methods (random forest-mean decrease accuracy [RF-MDA], random forest-mean decrease in Gini [RF-MDG], XGBoost) included all factors and ranked them (Table 1, **Supplementary Figs. 3–5**). The 20 predictive variables in six or more

Table 1. Predictor variables of PPD using different selection methods.

Method	Predictor variables	Number of identified predictors
SR-FS	Blood type, pregnancy caregiver, emergency experience, IGC, living environment, MCH, MCHC, MCV, partner employment, PDW, PLT, postpartum hemorrhage, GDM, pregnancy exercise, pregnancy number, PSQI score, RDW-SD, smoking habit, SSRS score, total protein, venting	21
SR-BS	Blood type, pregnancy caregiver, creatinine, emergency experience, eosinophil percentage, IGC, living environment, lymphocyte percentage, MCH, MCHC, MCV, miscarriage history, monocyte percentage, neutrophil percentage, neutrophil/lymphocyte ratio, partner employment, PDW, PLT, postpartum hemorrhage, assisted reproduction, GDM, preeclampsia, pregnancy number, PSQI score, RDW-SD, smoking habit, SSRS score, total protein, uric acid, venting	30
SR-BE	Blood type, pregnancy caregiver, emergency experience, IGC, living environment, MCH, MCHC, MCV, partner employment, PDW, PLT, postpartum hemorrhage, GDM, pregnancy exercise, pregnancy number, PSQI score, RDW-SD, smoking habit, SSRS score, total protein, venting	21
LASSO	Blood type, living environment, pregnancy caregiver, emergency experience, help seeking, marital evaluation, MCV, PDW, postpartum hemorrhage, pregnancy exercise, pregnancy number, PSQI score, smoking habit, SSRS score, total protein, venting	16
RF-MDA[#] (top 20)	PSQI score, living environment, SSRS score, emergency experience, help seeking, marital evaluation, venting, pregnancy caregiver, mental history, neutrophil percentage, lymphocyte percentage, WBC, anti TPO, neutrophil count, neutrophil/lymphocyte ratio, religion, MCH, PLCR, anticipated caregiver, MPV	101
RF-MDG[#] (top 20)	PSQI score, SSRS score, living environment, emergency experience, homocysteine, anti TPO, PLT/lymphocyte ratio, creatinine, total cholesterol, ALT, urea, uric acid, TSH, G-GT, ALP, triglycerides, total protein, marital evaluation, direct bilirubin, FT4	101
XGBoost[#] (top 20)	PSQI score, living environment, SSRS score, emergency experience, anti-TPO, HDL, TSH, urea, AST, ALP, glucose, lymphocyte count, LDH, PLT/lymphocyte ratio, MCV, RDW-SD, total bile acids, uric acid, G-GT, ALT	101
SVM-RFE	basophil_count, creatinine, eosinophil percentage, lymphocyte percentage, MCH, MCHC, monocyte percentage, neutrophil percentage, PDW, PSQI score, SSRS score, total protein, triglycerides, urea, uric acid	15

Abbreviations: PPD, postpartum depression; SR-FS, stepwise regression-forward selection; SR-BS, stepwise regression-backward selection; SR-BE, stepwise regression-bidirectional elimination; LASSO, least absolute shrinkage and selection operator; RF-MDA, random forest-mean decrease accuracy; RF-MDG, random forest-mean decrease in Gini; XGBoost, eXtreme Gradient Boosting; SVM-RFE, support vector machine-recursive feature elimination; IGC, immature granulocyte count; MCH, mean corpuscular hemoglobin; MCHC, mean corpuscular hemoglobin concentration; MPV, mean platelet volume; MCV, mean corpuscular volume; PLT, platelet count; PLCR, platelet large cell ratio; GDM, gestational diabetes mellitus; PSQI, Pittsburgh Sleep Quality Index; SSRS, social support rating scale; RDW-SD, red cell distribution width-standard deviation; PDW, platelet distribution width; ALT, alanine aminotransferase; AST, aspartate aminotransferase; ALP, alkaline phosphatase; TPO, thyroid peroxidase; FT4, free thyroxine 4; TSH, thyroid-Stimulating hormone; HDL, high-density lipoprotein; G-GT, gamma-glutamyl transferase; LDH, lactate dehydrogenase; WBC, white blood cell count; [#], RF (MDG and MDA) and XGBoost only rank features without filtering them. The table displays only the top 20 indicators. When taking the intersection of the subsequent 8 models, these three models separately include 101 indicators.

methods were selected as the final predictive variables, as shown in Fig. 2. The immature granulocyte count (IGC) was selected by six feature-engineering methods, indicating that it is a strong predictor. Although its adjusted generalized variance inflation factor (GVIF) was high, indicating strong collinearity with other variables, no other variables with similarly high GVIFs were included in the final model. Therefore, the risk of "group collinearity" was considered extremely low, and therefore, IGC was retained.

3.3 Model Development and Comparison

The performance of the top 5 integrated PPD prediction models in the test set is shown in Table 2 and Fig. 3. While showing comparable AUCs (0.833–0.835), these models exhibit a notable trade-off between sensitivity and specificity: the Voting top 5+XGBoost (Weight) achieved the highest AUC (0.835), whereas Voting (SMOTE) achieved the highest sensitivity (0.758) and

Table 2. Performance metrics of the top 5 models in the test set.

Model	AUC	Accuracy	SEN	SPE	PPV	NPV	PLR	NLR	F1	Brier
Voting top 5+XGBoost (Weight)	0.835	0.808	0.484	0.906	0.612	0.852	5.170	0.569	0.541	0.162
Voting (SMOTE)	0.834	0.785	0.758	0.793	0.528	0.915	3.664	0.305	0.623	0.153
Voting top 3+CatBoost (Weight)+RF (Weight)	0.833	0.808	0.468	0.911	0.617	0.849	5.275	0.584	0.532	0.162
Voting top 5+GLM (Weight)	0.833	0.800	0.484	0.897	0.588	0.850	4.677	0.576	0.531	0.162
Voting top 3+CAT (Weight)	0.833	0.808	0.468	0.911	0.617	0.849	5.275	0.584	0.532	0.161

Abbreviations: RF, random forest; XGBoost, eXtreme Gradient Boosting; CatBoost, categorical boosting; GLM, generalized linear model; SMOTE, Synthetic Minority Over-sampling Technique; AUC, area under the curve; SEN, sensitivity; SPE, specificity; PPV, positive predictive value; NPV, negative predictive value; PLR, positive likelihood ratio; NLR, negative likelihood ratio. Weight, Sample Weight; Voting, Voting Ensemble Model.

F1-score (0.623). SMOTE strategy significantly improved sensitivity but reduced the specificity of models, while Weight strategy maintained high specificity (0.897–0.911) at lower sensitivity (0.468–0.484) of models. All models demonstrated good calibration (Brier scores: 0.153–0.162), with Voting (SMOTE) showing superior calibration and the highest net benefit in DCA at low thresholds (0.05–0.30), indicating enhanced clinical utility for early screening. This model's high sensitivity enables efficient identification of true positives, potentially reducing misdiagnosis rates in practice. Additional performance metrics and clinical impact curves are presented in **Supplementary Tables 4–6** and **Supplementary Figs. 6,7**.

3.4 Model Interpretability

3.4.1 SHAP Analysis

The SHAP values were used to elucidate the impact of each variable on PPD prediction. As shown in Fig. 4A, the importance of the four factors (PSQI score, SSRS score, encountering emergencies during pregnancy, and an average evaluation of the living environment) is significantly higher than that of the other factors. The strength and direction of each predictive factor are shown in Fig. 4B. In the positive SHAP value region, the concentration of red samples is high, indicating that the variable is a predictive factor. Conversely, in the negative SHAP value region, the concentration of red samples is high, indicating that the variable is a protective factor. Fig. 4C,D show the SHAP values at the individual level.

3.4.2 Nomogram Analysis

The nomogram (**Supplementary Fig. 8**) incorporates 11 predictive factors (**Supplementary Table 7**), which quantified PPD risk through weighted scoring: smoking (19 points), pregnancy inactivity (9 points), suboptimal living environment (32 points), pregnancy emergencies (28 points), lack of prenatal support (43 points), passive communication (20 points), and blood type B (10 points). Higher PSQI scores, platelet count (PLT), and mean corpuscular volume (MCV) values, along with lower SSRS scores, were associated with a proportional increase in the top-axis points and corresponding PPD risk.

3.4.3 Sample Size Assessment

Supplementary Fig. 9 shows that the AUC exhibited a stage-dependent growth pattern as the sample size increased. It rose steadily at first and then plateaued at approximately 400 cases, suggesting that this threshold achieves optimal predictive performance with no further improvement from additional sample expansion.

4. Discussion

This study integrates biological, psychological, and social factors to apply an EL strategy to screen and validate reliable ML models for the early prediction of PPD. Compared to previous studies [23,24,25], this research employs 8 feature selection methods to enhance the reliability of feature selection and reduce the bias that may be introduced by a single method [26]. Furthermore, by applying an EL strategy, the ML model's robustness, stability, and ability to capture complex data patterns of the ML model were fundamentally strengthened. This systematically enhanced the potential for discovering optimal predictive combinations while ensuring the generalizability and methodological rigor of the findings. Ultimately, we selected 2 ML models suitable for different clinical scenarios.

PSQI scores, SSRS scores, platelet distribution width (PDW), and total protein serve as the core predictors identified by consensus across 8 feature selection methods, demonstrating their extreme robustness as PPD predictors. Research indicates that sleep disorders lead to an abnormal elevation in cortisol levels through sustained activation of the hypothalamic-pituitary-adrenal (HPA) axis, thereby promoting the release of inflammatory cells [27,28]. Increased cortisol reactivity and delayed recovery may represent the biological basis of depression, contributing to worsening depressive symptoms over time [29]. SSRS captures perceived social support, which buffers postpartum stress and reduces the incidence of PPD by regulating psychological resilience and facilitating access to practical resources [30]. More mechanistically novel are the hematological markers. Platelets have increasingly been found to play a crucial role in inflammation, and research indicates that they serve as a bridge linking inflammatory responses to mental disorders [31]. PDW serves as a key indicator re-

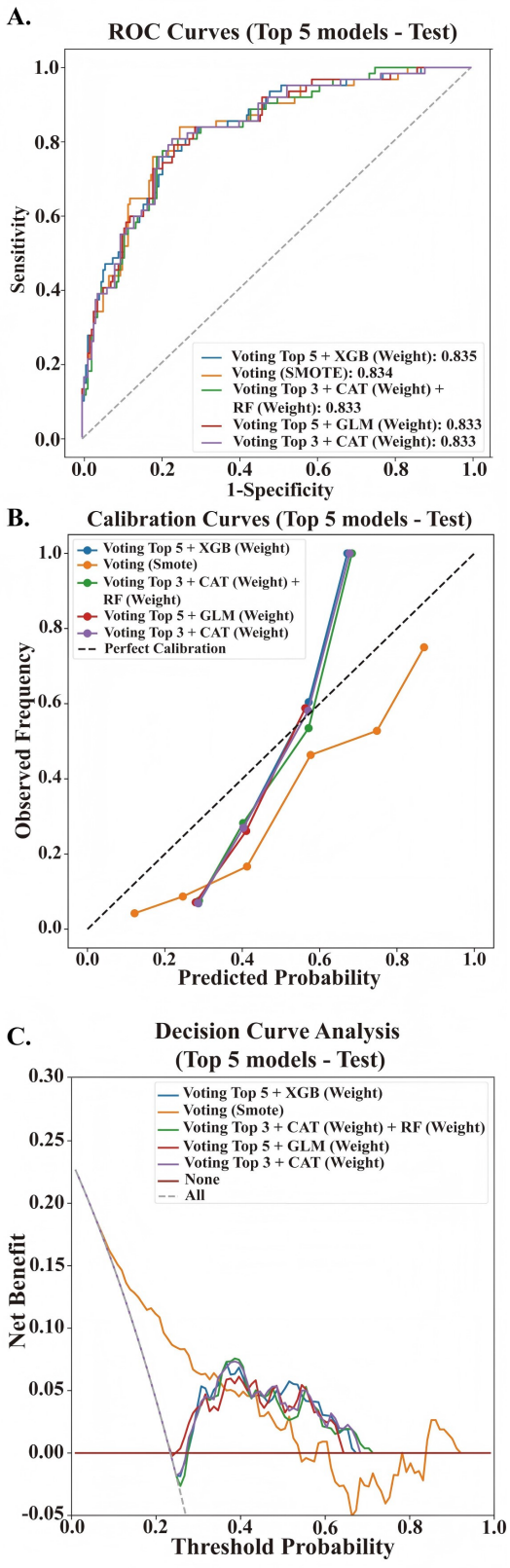

Fig. 3. Performance of the top 5 models on the test set. (A) ROC curves of the top 5 models. (B) Calibration curves of the top 5 models. (C) Decision curve analysis of the top 5 models.

flecting platelet activation and heterogeneity. Its alterations reflect subclinical systemic inflammation and endothelial dysfunction [32,33,34], both of which are associated with neuroinflammation and emotional regulation [35,36]. Simultaneously, total protein levels, while nonspecific, could signal underlying nutritional status, chronic low-grade inflammation, or stress-related physiological dysregulation [37]. Reproducible predictors of PPD across different screening strategies suggest that PPD involves complex pathophysiological processes. Interactions among social support, sleep disturbances, inflammation, and other factors may increase susceptibility to PPD.

In this study, we addressed the sample imbalance issue in the training set using 2 mechanisms: SMOTE and Weight. SMOTE alleviates class imbalance by synthesizing minority-class samples, effectively expanding the model's decision boundaries for high-risk cases [38]. This strategy tends to enhance sensitivity and reduce false-negative rates, making it particularly suitable for early screening or risk-stratification tasks. It mitigates intervention delays caused by missed potential PPD cases. In contrast, sample weighting does not alter the data distribution; instead, it introduces a penalty for misclassifying minority samples during model optimization. This approach encourages the model to make more conservative positive predictions, thereby improving specificity and reducing false positives. This makes it more suitable for confirmatory assessment or differential diagnosis following a preliminary positive screening, particularly in resource-constrained clinical settings [39]. Among final models, the Voting top 5+XGBoost (Weight) model exhibited optimal overall discriminative performance on the test set (AUC = 0.835) and high specificity (0.906).

In contrast, the Voting (SMOTE) model excelled in sensitivity (0.758) and clinical net benefit at low-risk thresholds. High sensitivity indicates that the model can maximize the identification of potential patients, making it suitable for early screening and reducing the risk of missed diagnoses. High specificity indicates that the model prioritizes diagnostic accuracy, making it suitable for confirmatory assessments and helping minimize unnecessary driven by false positives in resource-constrained settings. Our results indicate that SMOTE-based and weight-based ensemble models are not mutually exclusive but possess complementary strengths. This dual strategy enables flexible model selection across clinical scenarios, achieving a balanced trade-off between sensitivity and specificity in a targeted and clinically meaningful manner.

At the same time, we found that EL models generally outperform single-base ML models, but not all stacking approaches-combining multiple independent models-yield additional performance gains. This demonstrates that the effectiveness of integration is highly dependent on the diversity and complementarity of base models. If the combined models are highly correlated, exhibit similar error patterns, or individual models perform too poorly, the

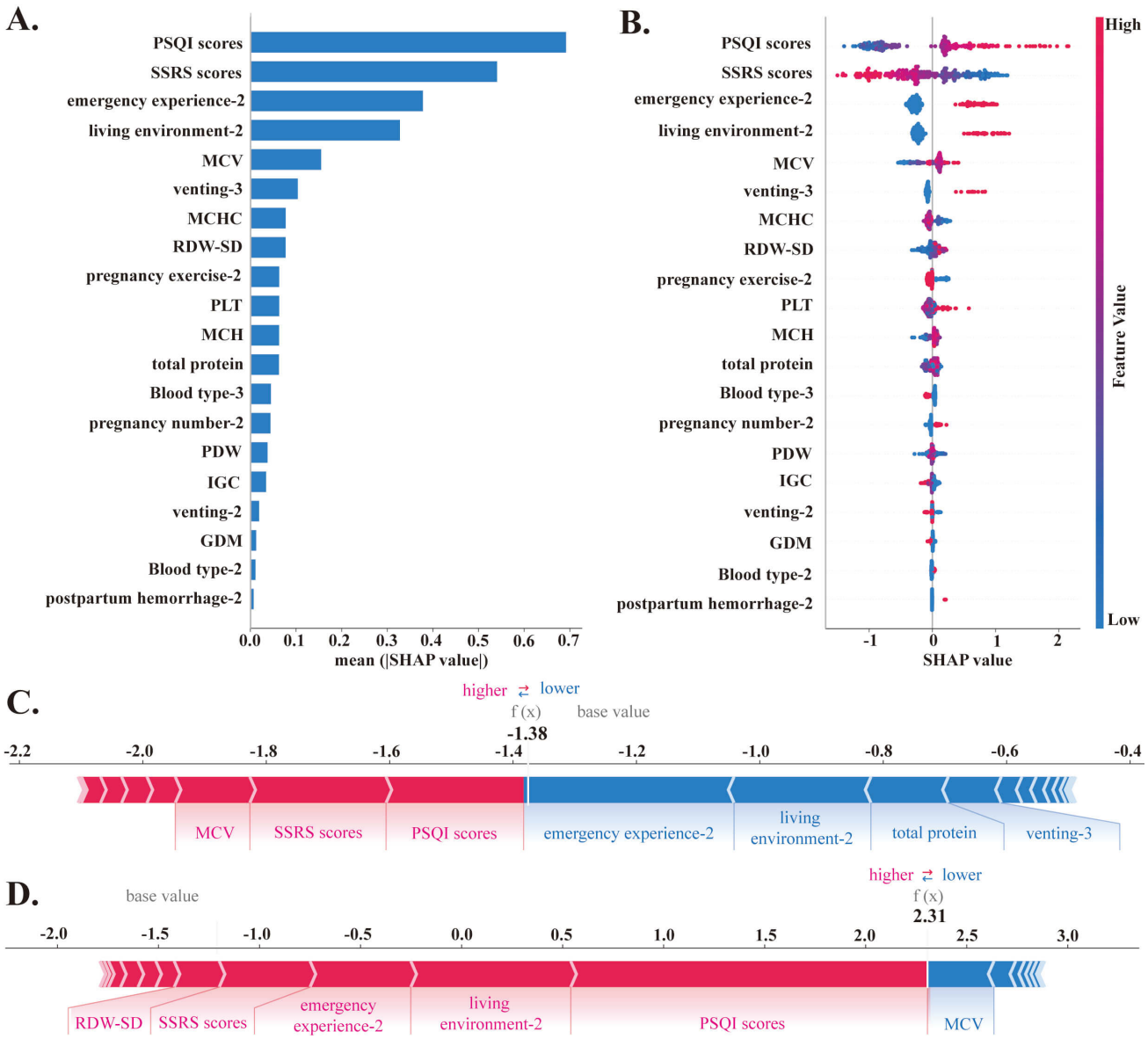

Fig. 4. Analyzing the interpretability of important variables. (A) Distribution of each feature factor based on the mean SHAP value. (B) Summary of SHAP values for each feature factor. (C,D) Two SHAP force plots: The color indicates the contribution of each feature factor. Blue indicates that the feature factor has a negative impact on the prediction, while red indicates that the feature factor has a positive impact on the prediction. The length of the color bar represents the strength of the contribution. Abbreviations: PSQI, Pittsburgh Sleep Quality Index; SSRS, social support rating scale; PDW, platelet distribution width; MCH, mean corpuscular hemoglobin; MCHC, mean corpuscular hemoglobin concentration; MCV, mean corpuscular volume; IGC, immature granulocyte count; PLT, platelet count; RDW-SD, red cell distribution width-standard deviation; GDM, gestational diabetes mellitus. emergency experience-2, suffering emergency events during pregnancy; living environment-2, the evaluation of the living environment is average; venting-2, confide only in those who are extremely close; venting-3, only confide when asked; pregnancy exercise-2, regular exercise during pregnancy; Blood type-2, B; Blood type-3, O; pregnancy number-2, the second pregnancy; postpartum hemorrhage-2, experiencing postpartum hemorrhage.

stacked architecture may not only fail to improve decision quality but also amplify shared errors or introduce unnecessary complexity [40,41,42]. This study incorporates 11 heterogeneous base algorithms to proactively create low-correlation conditions among prediction errors, thereby meeting the prerequisites for effective EL.

Using SHAP analysis, we identified PSQI scores and SSRS scores as the two most significant factors in this model. Social support and sleep disorders are key areas currently amenable to intervention. The key hematological indicators identified by this model, such as MCV, mean corpuscular hemoglobin concentration (MCHC), red cell distribution width-standard deviation (RDW-SD), PLT, mean

corpuscular hemoglobin (MCH), and IGC, provide multidimensional insights into the biological basis of PPD. Among these, changes in MCV, MCH, and MCHC indicate subclinical nutritional and metabolic abnormalities, particularly in iron and vitamin B metabolism. Deficiencies in folate or vitamin B12 impair neurotransmitter synthesis and are associated with perinatal mental health [43,44,45]. The changes in RDW-SD, PLT, and IGC indicate chronic low-grade inflammation [46]. Physiological stress, tissue repair processes, and psychological pressure during the perinatal period collectively create a pro-inflammatory environment, leading to sustained activation of the innate immune system and linking to central emotional regulation [47,48]. These alterations in peripheral blood parameters may serve as measurable intermediate biomarkers linking physiological stress during pregnancy, nutritional depletion, and dysfunction of postpartum emotional-regulation neural circuits. This suggests the potential for nutritional interventions or anti-inflammatory strategies to prevent PPD. The nomogram analysis revealed that blood types B and A were associated with a higher PPD risk than AB or O, largely consistent with previous domestic research [49].

Limitations

There are certain limitations to note. Firstly, although the sample size (n = 1323) is sufficient to ensure stable predictive performance, the study population was entirely drawn from 2 hospitals in Nanjing, China, with 97.7% Han Chinese, indicating a degree of homogeneity. The validity and applicability of the model and conclusions require further verification when applied to populations in other regions or with different ethnic or cultural backgrounds. Secondly, PPD classification was based on EPDS scores rather than clinical diagnostic interviews. Despite being administered under professional supervision, the self-report nature of the EPDS may introduce reporting and social desirability biases, leading to potential misclassification. Future studies should incorporate standardized psychiatric diagnoses or longitudinal assessments to further validate the proposed models. Finally, a key limitation of this study is that the developed ML model has been validated only on internal datasets and has not yet been tested on an independent external cohort. Consequently, the model's demonstrated superior performance may exhibit a degree of optimistic bias, and its ability to generalize to other populations, different healthcare institutions, or broader geographic regions remains uncertain. A primary step for future research is to collect multicenter, diverse prospective cohort data to rigorously validate our model externally, thereby confirming its clinical applicability, universality, and robustness.

5. Conclusion

This study integrates biological, psychological, and sociodemographic predictors to develop a high-performance EL models for PPD prediction. Through SHAP and nomogram analyses, the model's clinical interpretability is demonstrated. Adaptable to diverse clinical settings and integrable with electronic health record systems, this model provides an effective tool for early warning and intervention in PPD.

Key Points

•Undetected PPD carries severe risks for both mother and child, which elevates early detection to a critical public health priority.

• We propose a novel early prediction framework for PPD that synergistically combines ML method with biopsychosocial determinants, and refines model performance through EL strategies, thereby addressing the critical lag inherent in conventional symptom-based screening.

• Our framework supports tailored model selection for diverse clinical scenarios, bridging ML innovation with practical healthcare needs.

Availability of Data and Materials

The data that support the findings of this study are available from the corresponding authors upon reasonable request.

Author Contributions

FYZ was responsible for the study design, image processing and article writing. ZCT was responsible for study design and model building. XH was responsible for data curation and model interpretation. XLW was responsible for data collection, participants' management and postpartum guidance. SLW was responsible for research design and revision. HY was responsible for research design, revision and guidance. All authors contributed to the important editorial changes in the manuscript. All authors read and approved the final manuscript. All authors have participated sufficiently in the work and agreed to be accountable for all aspects of the work.

Ethics Approval and Consent to Participate

This study was performed in accordance with the "Declaration of Helsinki" and approved by the Ethics Committee of Zhongda Hospital (2023ZDSYLL438-p01) and the Ethics Committee of Nanjing Women and Children's Healthcare Hospital (2024KY-080-01). Informed consent was obtained from all participants prior to data collection.

Acknowledgment

We sincerely thank all participants for their valuable contribution to this study.

Funding

This study was supported by the Natural Science Foundation of Jiangsu Province (Project No. BK20231423) and the Jiangsu Provincial Preventive Medicine Research Project (Project No. Ym2023004).

Conflicts of Interest

Given his role as a Biostatistics Editor, Xiang Hong had no involvement in the peer review of this article and has no access to information regarding its peer review. Full responsibility for the editorial process for this article was delegated to Ashok Roy. Other authors declare no conflicts of interest.

Supplementary Material

Supplementary material associated with this article can be found, in the online version, at https://doi.org/10.31083/BJHM56968.

References

[1] Gimbel LA, Weingarten SJ, Smid MC, Hoffman MC. Maternal mental health as a major contributor to maternal mortality. Seminars in Perinatology. 2024; 48: 151943. https://doi.org/10.1016/j.semperi.2024.151943

[2] Trost SL, Beauregard JL, Smoots AN, Ko JY, Haight SC, Moore Simas TA, et al. Preventing Pregnancy-Related Mental Health Deaths: Insights From 14 US Maternal Mortality Review Committees, 2008-17. Health Affairs. 2021; 40: 1551–1559. https://doi.org/10.1377/hlthaff.2021.00615

[3] Appleby L. Suicide during pregnancy and in the first postnatal year. BMJ (Clinical Research Ed.). 1991; 302: 137–140. https://doi.org/10.1136/bmj.302.6769.137

[4] Saharoy R, Potdukhe A, Wanjari M, Taksande AB. Postpartum Depression and Maternal Care: Exploring the Complex Effects on Mothers and Infants. Cureus. 2023; 15: e41381. https://doi.org/10.7759/cureus.41381

[5] Park SH, Kim JI. Predictive validity of the Edinburgh postnatal depression scale and other tools for screening depression in pregnant and postpartum women: a systematic review and meta-analysis. Archives of Gynecology and Obstetrics. 2023; 307: 1331–1345. https://doi.org/10.1007/s00404-022-06525-0

[6] Matthey S, Agostini F. Using the Edinburgh Postnatal Depression Scale for women and men-some cautionary thoughts. Archives of Women's Mental Health. 2017; 20: 345–354. https://doi.org/10.1007/s00737-016-0710-9

[7] Zhao XH, Zhang ZH. Risk factors for postpartum depression: An evidence-based systematic review of systematic reviews and meta-analyses. Asian Journal of Psychiatry. 2020; 53: 102353. https://doi.org/10.1016/j.ajp.2020.102353

[8] Wong J, Fisher J. The role of traditional confinement practices in determining postpartum depression in women in Chinese cultures: a systematic review of the English language evidence. Journal of Affective Disorders. 2009; 116: 161–169. https://doi.org/10.1016/j.jad.2008.11.002

[9] Crowley SK, O'Buckley TK, Schiller CE, Stuebe A, Morrow AL, Girdler SS. Blunted neuroactive steroid and HPA axis responses to stress are associated with reduced sleep quality and negative affect in pregnancy: a pilot study. Psychopharmacology. 2016; 233: 1299–1310. https://doi.org/10.1007/s00213-016-4217-x

[10] Groer MW, Vaughan JH. Positive thyroid peroxidase antibody titer is associated with dysphoric moods during pregnancy and postpartum. Journal of Obstetric, Gynecologic, and Neonatal Nursing. 2013; 42: E26–E32. https://doi.org/10.1111/j.1552-6909.2012.01425.x

[11] Azami M, Badfar G, Khalighi Z, Qasemi P, Shohani M, Soleymani A, et al. The association between anemia and postpartum depression: A systematic review and meta-analysis. Caspian Journal of Internal Medicine. 2019; 10: 115–124. https://doi.org/10.22088/cjim.10.2.115

[12] Handelman GS, Kok HK, Chandra RV, Razavi AH, Lee MJ, Asadi H. eDoctor: machine learning and the future of medicine. Journal of Internal Medicine. 2018; 284: 603–619. https://doi.org/10.1111/joim.12822

[13] Shibue K. Artificial Intelligence and Machine Learning in Clinical Medicine. The New England Journal of Medicine. 2023; 388: 2398. https://doi.org/10.1056/NEJMc2305287

[14] Mahajan P, Uddin S, Hajati F, Moni MA. Ensemble Learning for Disease Prediction: A Review. Healthcare. 2023; 11: 1808. https://doi.org/10.3390/healthcare11121808

[15] Zhang J, Zhu H, Chen Y, Yang C, Cheng H, Li Y, et al. Ensemble machine learning approach for screening of coronary heart disease based on echocardiography and risk factors. BMC Medical Informatics and Decision Making. 2021; 21: 187. https://doi.org/10.1186/s12911-021-01535-5

[16] Cox JL, Holden JM, Sagovsky R. Detection of postnatal depression. Development of the 10-item Edinburgh Postnatal Depression Scale. The British Journal of Psychiatry. 1987; 150: 782–786. https://doi.org/10.1192/bjp.150.6.782

[17] Lee DT, Yip SK, Chiu HF, Leung TY, Chan KP, Chau IO, et al. Detecting postnatal depression in Chinese women. Validation of the Chinese version of the Edinburgh Postnatal Depression Scale. The British Journal of Psychiatry. 1998; 172: 433–437. https://doi.org/10.1192/bjp.172.5.433

[18] Gu Y, Hu J, Hu Y, Wang J. Social supports and mental health: a cross-sectional study on the correlation of self-consistency and congruence in China. BMC Health Services Research. 2016; 16: 207. https://doi.org/10.1186/s12913-016-1463-x

[19] Buysse DJ, Reynolds CF, 3rd, Monk TH, Berman SR, Kupfer DJ. The Pittsburgh Sleep Quality Index: a new instrument for psychiatric practice and research. Psychiatry Research. 1989; 28: 193–213. https://doi.org/10.1016/0165-1781(89)90047-4

[20] Wu YY, Zhang YC, Wu YF, Fu LY, Wu X, Huang LL, et al. Association between chronic disease and sleep quality and duration among 18-59-year-old adults in southern China. Southeast Asian Journal of Tropical Medicine and Public Health. 2019; 50: 912–924.

[21] He Z, Yu W. Stable feature selection for biomarker discovery. Computational Biology and Chemistry. 2010; 34: 215–225. https://doi.org/10.1016/j.compbiolchem.2010.07.002

[22] Bolón-Canedo V, Sánchez-Maroño N, Alonso-Betanzos A. A review of feature selection methods on synthetic data. Knowledge and Information Systems. 2013; 34: 483–519. https://doi.org/10.1007/s10115-012-0487-8

[23] Qi W, Wang Y, Wang Y, Huang S, Li C, Jin H, et al. Prediction of postpartum depression in women: development and validation of multiple machine learning models. Journal of Translational Medicine. 2025; 23: 291. https://doi.org/10.1186/s12967-025-06289-6

[24] Huang Y, Alvernaz S, Kim SJ, Maki P, Dai Y, Peñalver Bernabé B. Predicting Prenatal Depression and Assessing Model Bias Using Machine Learning Models. Biological Psychiatry Global Open Science. 2024; 4: 100376. https://doi.org/10.1016/j.bpsgos.2024.100376

[25] Matsuo S, Ushida T, Emoto R, Moriyama Y, Iitani Y, Nakamura N, et al. Machine learning prediction models for postpartum depression: A multicenter study in Japan. The Journal of Obstetrics and Gynaecology Research. 2022; 48: 1775–1785. https://doi.org/10.1111/jog.15266

[26] Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. Bioinformatics. 2007; 23: 2507–2517. https://doi.org/10.1093/bioinformatics/btm344

[27] Nguyen VT, Fields CJ, Ashley NT. Temporal dynamics of pro-inflammatory cytokines and serum corticosterone following

acute sleep fragmentation in male mice. PloS One. 2023; 18: e0288889. https://doi.org/10.1371/journal.pone.0288889
[28] Asarnow LD. Depression and sleep: what has the treatment research revealed and could the HPA axis be a potential mechanism? Current Opinion in Psychology. 2020; 34: 112–116. https://doi.org/10.1016/j.copsyc.2019.12.002
[29] LeMoult J, Chen MC, Foland-Ross LC, Burley HW, Gotlib IH. Concordance of mother-daughter diurnal cortisol production: Understanding the intergenerational transmission of risk for depression. Biological Psychology. 2015; 108: 98–104. https://doi.org/10.1016/j.biopsycho.2015.03.019
[30] Nie C, Dai Q, Zhao R, Dong Y, Chen Y, Ren H. The impact of resilience on psychological outcomes in women with threatened premature labor and spouses: a cross-sectional study in Southwest China. Health and Quality of Life Outcomes. 2017; 15: 26. https://doi.org/10.1186/s12955-017-0603-2
[31] Izzi B, Tirozzi A, Cerletti C, Donati MB, de Gaetano G, Hoylaerts MF, et al. Beyond Haemostasis and Thrombosis: Platelets in Depression and Its Co-Morbidities. International Journal of Molecular Sciences. 2020; 21: 8817. https://doi.org/10.3390/ijms21228817
[32] Karaduran K, Aydogdu A, Gelisin O, Gunpinar S. Predictive Value of Complete Blood Count Parameters for Alzheimer's Disease in Relation to Periodontal Status. Current Alzheimer Research. 2025; 22: 302–314. https://doi.org/10.2174/0115672050388220250511174043
[33] Zhang S, Qian H, Yang Q, Hu J, Gan C, Meng W. Relationship between the extent of dissection and platelet activation in acute aortic dissection. Journal of Cardiothoracic Surgery. 2015; 10: 162. https://doi.org/10.1186/s13019-015-0351-5
[34] Ghaffari S, Tajlil A, Aslanabadi N, Separham A, Sohrabi B, Saeidi G, et al. Clinical and laboratory predictors of coronary slow flow in coronary angiography. Perfusion. 2017; 32: 13–19. https://doi.org/10.1177/0267659116659918
[35] Pivato CA, Chandiramani R, Petrovic M, Nicolas J, Spirito A, Cao D, et al. Depression and ischemic heart disease. International Journal of Cardiology. 2022; 364: 9–15. https://doi.org/10.1016/j.ijcard.2022.05.056
[36] Celebi G, Gocmez SS, Ozer C, Duruksu G, Yazır Y, Utkan T. Propolis prevents vascular endothelial dysfunction by attenuating inflammation and oxidative damage in the chronic unpredictable stress model of depression in rats. The Journal of Pharmacy and Pharmacology. 2023; 75: 1418–1429. https://doi.org/10.1093/jpp/rgad071
[37] Soeters PB, Wolfe RR, Shenkin A. Hypoalbuminemia: Pathogenesis and Clinical Significance. JPEN. Journal of Parenteral and Enteral Nutrition. 2019; 43: 181–193. https://doi.org/10.1002/jpen.1451
[38] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research. 2002; 16: 321–357. https://doi.org/10.1613/jair.953
[39] Zadrozny B, Langford J, Abe N. Cost-sensitive learning by cost-proportionate example weighting. In 3rd IEEE International Conference on Data Mining (pp. 435–442). Melbourne, FL, USA. IEEE Computer Society. 2003. https://doi.org/10.1109/ICDM.2003.1250950
[40] Dietterich TG. Ensemble methods in machine learning. In Multiple Classifier Systems. Berlin, Heidelberg. Springer-Verlag. 2000.
[41] Ting KM, Witten IH. Issues in stacked generalization. Journal of Artificial Intelligence Research. 1999; 10: 271–289. https://doi.org/10.1613/jair.594
[42] Domingos P. A few useful things to know about machine learning. Communications of the ACM. 2012; 55: 78–87. https://doi.org/10.1145/2347736.2347755
[43] Ohsuga T, Egawa M, Takahashi Y, Ikeda Y, Tsuyuki K, Kanatani K, et al. Association between low MCV in early pregnancy and perinatal mental health in the Japan Environment and Children's Study and the possible effect of iron deficiency. Journal of Affective Disorders. 2024; 356: 34–40. https://doi.org/10.1016/j.jad.2024.04.013
[44] Islam N, Semmler A, Starling J, Voisey J. A Systematic Review of the Correlation Between Micronutrient Levels and Perinatal Depression. Nutrients. 2025; 17: 3479. https://doi.org/10.3390/nu17213479
[45] Bottiglieri T. Folate, vitamin B12, and neuropsychiatric disorders. Nutrition Reviews. 1996; 54: 382–390. https://doi.org/10.1111/j.1753-4887.1996.tb03851.x
[46] Wysokiński A, Szczepocka E. Platelet parameters (PLT, MPV, P-LCR) in patients with schizophrenia, unipolar depression and bipolar disorder. Psychiatry Research. 2016; 237: 238–245. https://doi.org/10.1016/j.psychres.2016.01.034
[47] Won E, Na KS, Kim YK. Associations between Melatonin, Neuroinflammation, and Brain Alterations in Depression. International Journal of Molecular Sciences. 2021; 23: 305. https://doi.org/10.3390/ijms23010305
[48] Lei AA, Phang VWX, Lee YZ, Kow ASF, Tham CL, Ho YC, et al. Chronic Stress-Associated Depressive Disorders: The Impact of HPA Axis Dysregulation and Neuroinflammation on the Hippocampus-A Mini Review. International Journal of Molecular Sciences. 2025; 26: 2940. https://doi.org/10.3390/ijms26072940
[49] Shi X, Ying Y, Yu Z, Xing M, Zhu J, Feng W, et al. Risk factors for postpartum depression in Chinese women: A cross-sectional study at 6 weeks postpartum. Journal of Psychosomatic Research. 2021; 140: 110295. https://doi.org/10.1016/j.jpsychores.2020.110295