

Original Research

Received 15 June 2026

Revised 24 July 2026

Accepted 08 Aug 2026

Natural Resources for Human Health 2026, Vol. 6 (15s): 830–847

KEYWORDS:

Postpartum depression; Machine learning; Artificial intelligence; Predictive model; Explainable AI; Risk prediction; Systematic review

Artificial Intelligence-Based Prediction of Postpartum Depression Using Structured Data: A Systematic Review

Ms. Suma N G¹, Dr. S Saravana Kumar²

¹Department of AI & Robotics School of AI & Advanced Computing Presidency University Bengaluru, Karnataka, India

Email: suma.2023cse0047@presidencyuniversity.in

ORCID: 0009-0007-6798-1649

²School of Information Science Presidency University Bengaluru, Karnataka, India

Email: saravanakumar.s@presidencyuniversity.in

ORCID: 0000-0001-5686-1000

ABSTRACT: Background: PPD is considered a substantial public health concern due to its negative consequences on both mother and child's well-being. Despite the potential for earlier recognition of PPD, early prediction of women who will develop this disease proves challenging due to the complexity of the PPD condition which includes demographics, clinical, obstetric and psychological variables. Machine learning (ML) is a promising tool for such risk prediction, but its clinical adoption is hindered by variability across data collection, pre-processing, model development, validation, and interpretability.

Objective: The aim of this systematic review was to evaluate systematically the ML-based prediction models for PPD based on structured data, with special focus on data set characteristics, predictor variables, pre-processing techniques, features selection and engineering, development, validation, performance and explainability.

Methods: Systematic literature review was performed following PRISMA 2020 guidelines in relevant electronic databases for studies published between 2015 and 2026. Studies included in this review were original articles written in English language which involve human participants applying ML-based prediction or classification models for PPD based on structured data. Study screening and selection were done independently by two authors resolving conflicts through discussion. Data extraction involved study characteristics, data source, predictors, pre-processing techniques, algorithms, validation strategy, performance measures and explainability.

Results: Nineteen studies met the inclusion criteria out of the initial set of 3,805 articles. Among the included studies, there was an increasing tendency towards the use of ensemble learning, deep learning, and hybrid models as well as conventional machine learning methods including Logistic Regression, Random Forest, Support Vector Machines, and XGBoost. Psychosocial variables – involving prenatal anxiety, depressive symptoms, stress, sleep disorders, social support, and previous history of mental health issues – appeared to be significant predictors in all studies. Nonetheless, considerable heterogeneity was present in the studies in relation to the adopted pre-processing steps, features selection techniques, outcome definition, and validation procedure. A formal four-domain PROBAST assessment found 18 of the 19 studies (94.7%) at overall high risk of bias, most often driven by inadequate sample size relative to the number of candidate predictors, validation on the training set or on internal cross-validation alone,

or use of non-validated proxy outcome measures; only one study, which combined a large multi-site sample with external validation and reported calibration, was judged low risk across all domains.

Conclusion: There is great potential for applying machine learning for predicting PPD risk, but the evidence base remains at high risk of bias overall, and the current literature is limited by considerable methodological variability, scarce external validation, and problems concerning generalizability and clinical interpretability of the models. Substantial improvements in study design, sample size justification, and validation practice are needed before these models are ready for clinical translation.

1. INTRODUCTION

Postpartum depression (PPD) is one of the most common mental health disorders among mothers, occurring in the first year after childbirth and having many implications for mothers, infants, and families[1]. Mothers suffering from PPD usually have constant mood swings, anxiety, fatigue, sleep problems, difficulties with bonding with their babies, and low quality of life[2]. In case of lack of diagnosis and treatment, postpartum depression may influence child's development psychologically, emotionally, and behaviourally, and additionally increase family burden in the long run[3]. Despite the existence of successful treatment methods, early diagnosis of postpartum depression is difficult because of overlapping of symptoms of postpartum depression with normal postpartum symptoms, lack of reporting due to stigma, and not being diagnosed in a clinic[4]. Thus, accurate detection of those who are at risk before or after delivery became the main goal.

In the past, the evaluation of risk for postpartum depression (PPD) has been based on clinical judgement, screenings, and statistics[5, 6]. The use of screening methods such as the Edinburgh Postnatal Depression Scale (EPDS) has aided in increasing the detection of PPD; nevertheless, it detects the occurrence of the disorder rather than being able to predict it[7, 8]. In recent times, due to an increase in healthcare data that is well organized, the development and implementation of machine learning (ML) methods in predicting PPD has become possible[9]. Unlike traditional statistics, ML methods are able to model the complex nonlinear interactions of risk factors including demographic, obstetric, clinical, behavioural, and psychosocial factors[10].

Recent research has taken into account the progress made in the field by studying a variety of machine learning approaches such as classic classifiers, ensembles, deep learning architectures, and hybrids[11]. At the same time, there is an increasing focus on methodological factors other than choice of the algorithm including data preprocessing, feature engineering, feature selection, validation methods, performance assessment, and explanation of the process. Together, all these factors form a machine learning pipeline and impact greatly the reliability, reproducibility, explainability, and clinical relevance of the prediction model[12]. In such a way, assessment of model performance alone gives just a superficial overview of the methodological and translational value of machine learning PPD prediction studies.

While the number of studies using machine learning for prediction of PPD has increased significantly recently, the available literature reviews have mainly focused on comparisons of prediction algorithms or performance assessment of the latter. Relatively attention has been paid to a critical analysis of the full machine learning pipeline including datasets characteristics, predictors' choice, data preprocessing, feature engineering, feature selection, validation methods, performance assessment, and explanation. Given that these factors greatly impact the robustness and clinical relevance of the model, a comprehensive review of the whole pipeline is needed.

This systematic review examines contemporary literature on the use of machine learning for predictive modelling of postpartum depression with structured data. While evaluating the performance of the predictive models, this review critically evaluates each step in the machine learning process, from characteristics of datasets used, selection of predictors, preprocessing of data, feature engineering, model building, validation, performance measurement, and explainability of the model (refer to Figure 1). Using such an approach, the paper aims to establish trends in current literature, define gaps in research and provide evidence-based guidelines for improving machine learning-based predictions of postpartum depression.



Figure 1: AI prediction pipeline

2. METHODOLOGY

2.1 Review Design

The objective of this systematic review is to provide an overview of ML-based prediction models for PPD using structured data. The design and presentation of findings of the review comply with PRISMA 2020 guidelines[13]. Due to the large degree of heterogeneity among the studies included in this review regarding data sources, predictor variables, data preprocessing techniques, feature engineering/feature selection, model building, evaluation, and performance measurement, a qualitative analysis was undertaken in place of a quantitative meta-analysis. No prospective registration of the review was undertaken.

2.2 Search Strategy

A thorough search of relevant literature was performed through four electronic databases, namely PubMed, Scopus, Web of Science, and IEEE Xplore. The literature published between January 2015 and 2026 was considered relevant for inclusion. In the search process, controlled vocabulary, where available, was used together with the free text associated with the postpartum depression, prediction, and machine learning. The main search concepts included: ("postpartum depression" OR "postnatal depression" OR PPD), (prediction OR predictive model OR risk prediction OR classification), and ("machine learning" OR

"artificial intelligence" OR "deep learning" OR "data mining" OR "classification model"). The search strategy was adjusted according to the indexing features and syntax of each database while retaining the main search concepts.

2.3 Eligibility Criteria

Study inclusion criteria were met by studies that: (i) were original research articles; (ii) were written in English; (iii) studied humans; (iv) focused on postpartum depression as a predictor outcome; (v) developed, validated, or evaluated prediction/classification models using machine learning algorithms; and (vi) used structured data, such as demographic, clinical, obstetric, psychosocial, behavioural, questionnaire, or electronic medical records. Exclusion criteria included: (i) review articles, editorials, conference abstracts, commentaries, and case reports; (ii) qualitative studies; (iii) those using unstructured text data like social media posts and free-text clinical narratives; and (iv) non-English language studies.

2.4 Study Selection

All retrieved articles were entered into the Zotero reference management software[14]. Title and abstract screening and full-text assessment were done by two separate reviewers. Any inconsistencies found were discussed and a consensus was reached. A PRISMA 2020 flow chart was used to map the entire process of article selection.

2.5 Data Extraction

To ensure that there was standardization of the way the methodological information was extracted from the selected articles, a structured framework was used to guide this process. The data were comprised of information on authorship, article characteristics, dataset characteristics, sample sizes, tools used for assessment of outcomes, predictive variables used, data cleaning techniques, feature engineering techniques, feature selection techniques, machine learning models, validation techniques, evaluation metrics, XAI techniques, and the main results in the study. Data extraction was done independently by two researchers, and any disagreements were sorted out through discussions.

2.6 Quality Assessment

Methodological quality was appraised using the Prediction model Risk Of Bias ASsessment Tool (PROBAST), which was developed specifically for systematic reviews of prediction model studies. All four PROBAST domains (Participants, Predictors, Outcome, and Analysis) were judged for each of the 19 included studies based on detailed extraction of recruitment method, blinding and timing of predictor and outcome ascertainment, sample size relative to the number of candidate predictors (events per variable, EPV), and handling of overfitting; full results are reported in Section 3.8.1 and Table 3. This review was not prospectively registered (e.g., with PROSPERO), which is acknowledged as a limitation in Section 4.8. Given the significant heterogeneity among study designs, data sets, predictive tasks, and reporting practices, a structured methodological analysis was considered more appropriate than pooled effect-size estimation.

2.7 Data Synthesis

As a result of substantial differences in research design, data sets, predictors, machine learning algorithms, and performance measures, quantitative meta-analysis was considered infeasible. Thus, findings were qualitatively summarized through analysis of each step of the machine learning process in order to identify similar patterns and assess the strengths and weaknesses of different studies.

3. RESULTS

A total of 19 studies met the criteria for inclusion and therefore formed the basis for this systematic review. The synthesis of the studies aimed at offering an overall view of the pipeline of the machine learning process in the prediction of postpartum depression using structured data. This synthesis includes information on the studies and datasets used, outcomes, predictor variables, data pre-processing, feature engineering, feature selection, prediction models, validation, performance evaluation, and explainability.

3.1 Study Selection

The literature search yielded 3,805 papers from the selected electronic databases. After exclusion of 900 duplicate papers, 2,905 unique papers were left to be screened for titles and abstracts. Out of these papers, 2,853 papers were excluded due to non-

compliance with the eligibility criteria, while 52 full-text papers were examined based on eligibility. After applying the full-text eligibility criteria, 33 papers were found ineligible as they were either based on unstructured data, review papers, conference papers, or written in other languages besides English. Finally, 19 papers that met the inclusion criteria were found and were included in the qualitative synthesis.

3.2 Features of Studies

All 19 studies included in this systematic review were carried out from 2020 till 2026, which reflects increased interest in the use of machine learning methods for predicting postpartum depression. All studies covered various geographic locations such as Asia, Europe, North America, and the Middle East, where the majority of studies were conducted in China, the USA, Sweden, and India. In most studies, an observational study design was used, including prospective and retrospective cohort studies, cross-sectional studies, and quantitative surveys, while some of the most recent studies were aimed at developing new hybrid and deep learning architectures. The inclusion of various geographic locations and study designs demonstrates wide application of machine learning methods under various healthcare systems. Features of the included studies, including country, data source, study design, sample size, and methodology of outcome assessment, are provided in Table 1.

3.3 Data Set Features

In the studies reviewed, several structured data sources were employed, such as questionnaire-based data sets, data sets derived from electronic health records, clinical registries, and publically available data sets. The questionnaire was the most commonly used data source, while some studies relied on data sets collected from electronic health records of hospitals or from clinical records along with the use of structured surveys.

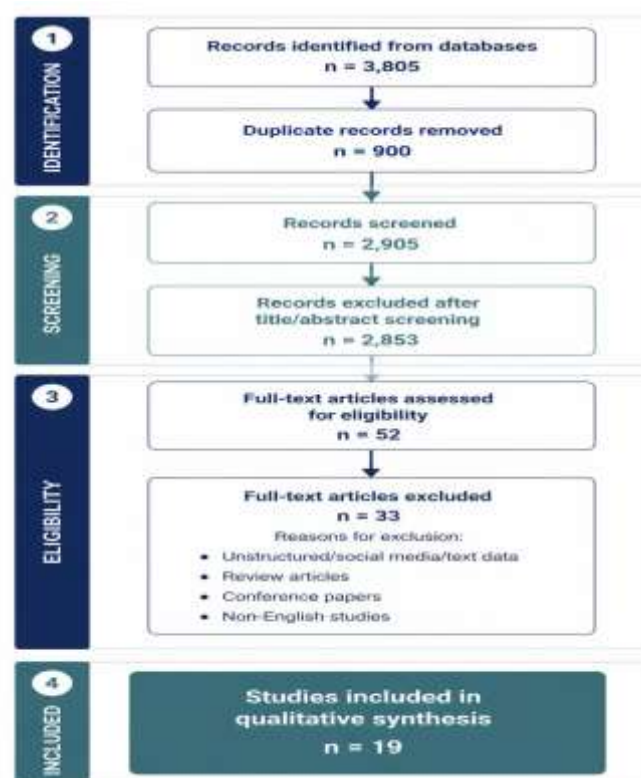


Figure 2. PRISMA 2020 Flow Diagram

The sample sizes varied widely, being anywhere between 64 and 29,168 participants. Even though many studies used large scale population registries or health records databases, other used smaller cohorts to develop their prediction model. It reflects the diversity of structured data sources that can be found for machine learning-based PPD prediction.

Table 1. Characteristics of Included Studies and Postpartum Depression Outcome Assessment

Study	Country	Data Source	Study Design	Sample Size	Outcome Assessment
[15]Andersson et al. (2021)	Sweden	Questionnaires and medical records (BASIC study)	Population-based prospective cohort study.	4,277	Depressive symptoms at 6 weeks postpartum (EPDS $\geq$ 12).
[16]Clapp et al. (2025)	USA	Electronic Health Records (EHR)	Retrospective cohort.	29,168	Composite outcome: mood disorder code, antidepressant prescription, or EPDS $\geq$ 13
[17]Fazraningtyas et al. (2025)	Indonesia	Questionnaires (primary data)	Quantitative study with resampling and feature selection.	317	PPD occurrences (at risk vs. not at risk).
[18]Gopalakrishnan et al. (2022)	India	Questionnaires	Online survey followed by ML validation.	132	PPD symptoms at 6 weeks postpartum (EPDS $\geq$ 12).
[19]Gulmohamed et al. (2026)	Saudi Arabia	Structured questionnaires	Hybrid deep learning framework development.	64	PPD status (Depressed / Non-Depressed).
[20]Huang et al. (2023)	USA	Electronic Medical Records (EMR)	Retrospective analysis with bias evaluation.	2,414	Prenatal depressive symptom severity.
[21]Huang et al. (2025)	China	Clinical records and questionnaires	Retrospective analysis.	1,065	PPD risk (EPDS score $>$ 9).
[22]Kafi et al. (2026)	Switzerland	Questionnaire (Public dataset)	Hybrid framework integrating XAI with ablation study.	402	PPD label (clinical threshold EPDS $>$ 13).
[23]Kim (2025)	Republic of Korea	Registry (K-ECEC-P)	Quantitative, cross-sectional design.	2,570	Depression status (1 = PPD, 0 = no PPD).
[24]Lin and Zhou (2025)	Kaggle (Hospital dataset)	Questionnaire (Kaggle public dataset)	Simulation experiment and comparative analysis.	1,503	Anxiousness (used as marker for PPD class).
[25]Nasim et al. (2024)	Kaggle (Medical hospital data)	Questionnaire (Kaggle dataset)	Meta-learning model development (MDKR).	1,503	"Feeling Anxious" predictive marker.
[26]Pillai and Chinnasamy (2025)	India	Questionnaires and e-forms	Hybrid data mining	233	Presence of PPD.

			classification study.		
[27]Qi et al. (2025)	China	Clinical registries and follow-up data	Prospective cohort study.	1,138	PPD at 6 weeks (EPDS cutoff > 11).
[28]Shen et al. (2023)	Sweden	Surveys and medical journals (BASIC study)	Deep reinforcement learning framework development.	4,313	PPD identification status.
[29]Shin et al. (2020)	USA	Registry (PRAMS)	Retrospective cohort study.	28,755	PPD based on PHQ-2 responses.
[30]Shivaprasad et al. (2024)	Kaggle (Public dataset)	Questionnaire (Kaggle repository)	Interdisciplinary diagnostic framework development.	1,503	"Suicide_attempt" predictor attribute.
[31]Tang et al. (2024)	Sweden	Surveys and medical records (BASIC study)	Reinforcement learning with advanced pre-training.	4,313	PPD identification.
[32]Zafar et al. (2025)	Pakistan	Questionnaires (PERI_DEP)	Novel dataset curation and deep learning modeling.	14,008	Perinatal depression status.
[33]Zhang et al. (2020)	China	Trimester-specific questionnaires	Prospective cohort study.	508	EPDS score within 42 days after delivery.

3.4 Outcome Variables and Predictors

3.4.1 Outcome Variables

The majority of studies used postpartum depression screening scales to define postpartum depression, and the most common was the Edinburgh Postnatal Depression Scale (EPDS)[7]. Cut-off values for EPDS varied among the studies due to different criteria used to define the condition[34]. The other studies used Patient Health Questionnaire (PHQ-2 or PHQ-9), composite outcomes which combined diagnostic and medication codes, or symptom categories[35]. While postpartum depression was the main outcome variable in most of the studies, some studies used postpartum depressive symptoms or other indicators based on anxiety.

3.4.2 Predictors

Wide variability was found in terms of the predictors used in the studies included in this review. Such predictors were usually classified into demographic, obstetric and medical, psychosocial, behavioural, and mental health variables. Commonly studied predictors included maternal age, education, body mass index, parity, mode of delivery, mental health history, prenatal anxiety and depression, sleep quality, social support, stress, and relationship variables. Some studies used laboratory data, medication information, psychometric measures, or individual symptoms obtained through questionnaires. Generally, great variability was seen in terms of types and numbers of predictors for model development.

3.5 Data Preprocessing

The data preprocessing methods among the selected studies ranged significantly when it came to dealing with missing values, outliers, categorical variables, feature scaling, class imbalance, and data splitting. Some of these studies provided extensive data preprocessing procedures, while some provided only few details about the methodologies applied.

The missing data were most commonly handled by excluding cases or variables that exceeded certain missing values criteria, as well as through imputation using MICE, KNN, missForest, and simple imputation (mean, median, mode). Outliers were not often discussed in these studies and were mainly determined through data exploration, IQR method, or Z-score. The categorical variables were coded using label encoding, one-hot encoding, or numerical coding, while continuous variables were scaled using Min–Max or Z-score scaling.

A number of researchers took class imbalance into consideration where SMOTE was the most used algorithm. The other techniques included random oversampling, bootstrapping, ADASYN, random under sampling, cost-sensitive learning, and reward-based methods. Splitting of data into training and testing sets at various proportions accompanied by k-fold cross validation were common practices in the development of models. Other preprocessing techniques involved duplicate elimination, feature selection, checking for bias, and cleaning up datasets prior to modelling.

3.6 Feature Engineering and Feature Selection

Feature engineering and feature selection gained popularity among researchers for achieving better model results and for data reduction purposes. Standard feature engineering approaches included categorical encoding, normalization/standardization, handling missing values, artificial oversampling such as SMOTE, and creating new clinical variables by combining other existing ones. In some more sophisticated cases, feature extraction was done using neural networks, attention models, or data augmentation.

Feature selection methods varied from traditional statistical approaches like Forward Selection, Backward Elimination, LASSO regression, and correlation analysis to more advanced machine learning approaches including Mean Decrease in Impurity, Random Forest importance, Relief, Boruta, Recursive Feature Elimination, and Association Rule Mining. Some studies combined feature engineering and feature selection into one preprocessing step while others were vague about this aspect of feature optimization.

3.7 Prediction Models

A broad range of machine learning models was considered to predict postpartum depression. The popular machine learning models that were used as baseline models include Logistic Regression, Decision Trees, Support Vector Machines, Naïve Bayes, K-Nearest Neighbors, and Artificial Neural Networks. On the other hand, the ensemble machine learning models such as Random Forest, XGBoost, Gradient Boosting, AdaBoost, CatBoost, and LightGBM were evaluated most often, and these models exhibited better predictive accuracy. The artificial intelligence models evaluated and best algorithms considered are shown in Table 2.

Table 2. Artificial Intelligence Models, Performance, Validation, and Explainability in Postpartum Depression Prediction

Study	AI Models Evaluated	Best Model Performance	Best Reported Performance Values	Validation Strategy	Explainability
[15]Andersson et al. (2021)	Ridge and LASSO Regression, GBM, DRF, XRT, Naïve Bayes, Stacked Ensembles.	Extreme Randomized Forest (XRT)	Accuracy: 73%; AUC: 81%; SENS: 72%; SPEC: 75%.	K-Fold Cross-Validation	Mean Decrease in Impurity (MDI) / Gini Importance.
[16]Clapp et al. (2025)	Logistic Regression (primary), XGBoost, Elastic Net.	Logistic Regression (with prenatal EPDS)	AUROC: 0.750; Brier: 0.073; PPV: 24.4%; NPV: 94.7%.	Not reported	LR coefficients and standardized mean differences.

[17]Fazraningtyas et al. (2025)	C4.5 Decision Tree, GBT, XGBoost.	XGBoost (with resampling)	Accuracy: 87.57%; Precision: 87.46%; Recall: 88.42%; AUC: 86.79%.	10-fold CV (stratified)	Feature importance from Forward Selection/Backward Elimination.
[18]Gopalakrishnan et al. (2022)	Naive Bayes, DRF, XRT, Ridge and LASSO Regression, Stacked Ensembles, GBM.	Random Forest	Accuracy: 71%; AUC: 80%; SENS: 69%; SPEC: 74%.	Not reported	Mean Decrease in Impurity (MDI) / Gini Index.
[19]Gulmohamed et al. (2026)	Hybrid CNN-BiLSTM-Attention (proposed), SVM, RF, LR, ANN, DNN.	Hybrid CNN-BiLSTM-Attention model	Accuracy: 96.8%; Sensitivity: 96.1%; Specificity: 97.4%; AUC: 0.98.	10-fold CV,	SHAP, LIME, and Attention mechanism weights.
[20]Huang et al. (2023)	Random Forest, Elastic Net, XGBoost.	Elastic Net	ROCAUC: 0.67 (Total), 0.85 (NHW); PRAUC: 0.50 (Total), 0.70 (NHW).	5-fold CV	Shapley Additive Explanations (SHAP).
[21]Huang et al. (2025)	SVM, XGBoost, CatBoost, Naive Bayes, RF, Logistic Regression, LightGBM, AdaBoost.	XGBoost	Accuracy: 0.95; AUC: 0.955; Precision: 0.945; Specificity: 0.985.	10-fold CV	SHAP (bar charts, beeswarm plots, force plots).
[22]Kafi et al. (2026)	LR, SVM, KNN, DT, SGD, RF, Extra Trees, GBM, XGB, AdaBoost, MLP, Stacking Ensembles, DNN, TabNet.	Tuned Random Forest (on 33 SHAP features)	Accuracy: 94.1%; Precision: 94.5%; Recall: 93.9%; F1: 94.0%; AUC: 0.978.	5-fold CV	SHAP, LIME, ARM, and Voting Ensemble consensus.
[23]Kim (2025)	Decision Tree, Random Forest, AdaBoost,	Logistic Regression	AUC-ROC: 0.74; Accuracy: 77.04%; Recall:	Not reported	Linear combination of independent variables (LR).

	Logistic Regression.		97.44%; F1: 86.56%.		
[24]Lin and Zhou (2025)	Ensemble NN (FCNN + DNN), SVM, KNN, DT, RF, NBC, LR, LDA, AdaBoost, CNN, LSTM.	Ensemble Neural Network (FCNN + DNN)	Accuracy: 0.933; Precision: 0.958; Recall: 0.939; F1: 0.948; MCC: 0.855.	Accuracy on training set	Connection patterns among neurons in FCNN.
[25]Nasim et al. (2024)	Meta DT-KN-RF (MDKR), DT, KNC, RF, SGD, Linear Regression, NB, BR, SVM, MLP, LoR, GBC.	Meta DT-KN-RF (MDKR)	Accuracy: 99%; Precision: 99%; Recall: 99%; F1: 99%.	10-fold CV,	Clear binary rules (DT component); contributory factor analysis.
[26]Pillai and Chinnasamy (2025)	SVM, ANN, Hybrid Classifier (Voting: SVM + ANN).	Hybrid Classifier	Accuracy: 0.86; Precision: 0.86; Recall: 0.85; Specificity: 0.8679; F1: 0.86.	Not reported	Correlation Heatmap analysis.
[27]Qi et al. (2025)	LR, DT, RF, XGBoost, SVM, ANN.	Logistic Regression (LR) and Artificial Neural Network (ANN)	AUC: 0.858; Sensitivity: 0.768; Specificity: 0.827; F1: 0.715; Brier: 0.137.	5-fold CV	SHAP (for ANN), Feature importance ranking, Nomograms.
[28]Shen et al. (2023)	Deep RL (ANN Agent + DE), Naïve Bayes, KNN, SVM, RF, Logistic Regression, Decision Tree.	Proposed Deep RL + DE model	Accuracy: 0.8907; Recall: 0.9053; Precision: 0.8847; F-measure: 0.8844; G-means: 0.9066.	Not reported	Differential Evolution (DE) weight analysis; expert qualitative review.
[29]Shin et al. (2020)	RF, GBM, SVM, RPART, NB, kNN, LR, NNET, AdaBoost.	Random Forest (RF)	AUC: 0.884; Sensitivity: 0.732; Specificity: 0.865; Accuracy: 0.791; F1: 0.776.	10-fold CV	varImp (variable importance) ranking.
[30]Shivaprasad et al. (2024)	RF, LR, CatBoost, LGBM, XGB, AdaBoost, DT,	KNN and Customized Stack Model	Accuracy: 0.97; Recall: 1.00; Precision: 0.94; F1: 0.97; MCC: 0.9515.	5-fold CV	ELI5, SHAP, LIME, and Anchor.

	KNN, Stack Ensemble.				
[31]Tang et al. (2024)	Proposed ML-ABC + PPO (ANN agent), NB, KNN, SVM, RF, LR, DT.	Proposed ML-ABC + PPO model	Accuracy: 0.91; F-measure: 0.88; G-means: 0.90; Recall: 0.93; AUC: 0.89.	Not reported	ML-ABC for weight refinement and pattern recognition.
[32]Zafar et al. (2025)	Hybrid RNN-LSTM, RNN, LSTM, CNN-BiLSTM, LSTM-CNN, RF, DT, GBT.	Hybrid RNN-LSTM (with SMOTE)	Accuracy: 96%; Precision: 96%; Recall: 98%; F1: 97%; AUC: 99%; Brier: 0.03.	Not reported	Architectural focus on capture short and long-term sequential dependency.
[33]Zhang et al. (2020)	SVM, Random Forest.	SVM (with FFS-RF selection)	Sensitivity: 0.69; ROC-AUC: 0.78; SPEC: 0.83; PPV: 0.68; NPV: 0.84.	5-fold CV	FFS-RF importance score ranking.

In recent times, many studies have been conducted using deep learning structures such as CNN, LSTM, RNN, and MLP to learn about the complicated nonlinear relationship between structured clinical datasets. The use of hybrid structures, which include combinations of machine learning, deep learning, or reinforcement learning techniques, has also become popular, where optimization algorithms, attention algorithms, or meta-learning are included in the methods used. Most studies conduct experiments using several algorithms in order to find out the best-suited one for a particular dataset since there is no universal prediction algorithm.

3.8 Validation Approaches Used

Most studies applied internal validation with the help of separate training and testing datasets supplemented by cross-validation, where the most popular methods included 5-fold and 10-fold cross-validation. Splitting of the dataset occurred at various ratios, such as 80:20, 70:30, 75:25, and 60:40 splits. Moreover, some researchers compared various options for the split ratio before implementing the optimal one.

Optimization of hyperparameters often took place through grid search, Bayesian optimization, or iterative optimization. Different approaches were used for the minimization of overfitting, such as feature selection, normalization, regularization, dropout layers, batch normalization, dimensionality reduction, oversampling, and ensemble learning. Yet, external validation was rare, with only a few studies applying the test of models on an independent data set.

3.8.1 PROBAST-Informed Risk of Bias Assessment

A four-domain PROBAST assessment (Participants, Predictors, Outcome, Analysis) was completed for all 19 included studies using detailed data extraction on recruitment method, blinding, outcome ascertainment, sample size relative to candidate predictors (events per variable, EPV), and handling of overfitting (Table 3). Eighteen of the 19 studies (94.7%) were judged at overall high risk of bias, and only Clapp et al. [16] was judged low risk across all four domains, reflecting its large multi-site electronic health record sample (development n=15,018; validation n=14,150), an outcome definition specified a priori, blinded and temporally separated predictor and outcome ascertainment, external validation across three independent hospitals, and reported calibration (Brier score 0.073). This proportion of high-risk studies is consistent with PROBAST-based reviews of clinical prediction models in other fields, where the large majority of published models are typically judged at high risk of bias, most often driven by the Analysis domain.

Table 3. PROBAST-Informed Risk of Bias Assessment by Domain

Study	Participants	Predictors	Outcome	Analysis	Overall
[15] Andersson (2021)	Low	Low	Low	High – EPV borderline (~10–20); no calibration; internal CV only	High
[16] Clapp (2025)	Low	Low	Low	Low – EPV>50; external validation across 3 hospitals; calibration reported (Brier 0.073)	Low
[17] Fazraningtyas (2025)	High – purposive sampling	High – unblinded, concurrent	High – unblinded, concurrent, variable timing	High – EPV=2.77	High
[18] Gopalakrishnan (2022)	High – >65% respondent loss	High – partial blinding, concurrent	High – multistage overlap with predictors	High – EPV<2.5	High
[19] Gulmohamed (2026)	High – small N=64, unclear sampling	High – unblinded, concurrent	High – inconsistent timing	High – EPV<1.0	High
[20] Huang (2023)	High – outcome-based exclusion of ambiguous PHQ-9 scores (5–8)	Low – blinded, before outcome	Low – validated PHQ-9, blinded	Low – regularization, training-split selection	High
[21] Huang (2025)	Low	High – unblinded, concurrent with EPDS	High – unblinded, concurrent	High – EPV ~5–10	High
[22] Kafi (2026)	Low	High – unblinded, cross-sectional	High – unblinded, concurrent	High – EPV=2.28–3.24	High
[23] Kim (2025)	Low – systematic stratified sampling, 89% response	Low – blinded, before outcome	Low – blinded, consistent	Low – EPV=150.1	Low
[24] Lin and Zhou (2025)	High – Kaggle, unverified provenance	High – unblinded, concurrent	High – “feeling anxious” proxy, not a validated PPD scale	Unclear – see note on Table 2 discrepancy	High
[25] Nasim (2024)	High – Kaggle, unverified provenance	High – unblinded, concurrent	High – “feeling anxious” proxy, not a validated PPD scale	High – feature pre-selection before split	High
[26] Pillai and Chinnasamy (2025)	High – convenience sample, small N	High – unblinded, concurrent	High – no validated instrument, inconsistent timing	High – feature selection on full dataset pre-split	High

[27] Qi (2025)	Low – prospective cohort, appropriate exclusions	Low – blinded for prenatal (PN-PPD) model	Low – validated EPDS, blinded	High – feature selection across full dataset before split	High
[28] Shen (2023)	Low	Low	Low	High – EPV borderline; no calibration; internal only	High
[29] Shin (2020)	High – >60% dropped via complete-case deletion	High – unblinded, recall bias	High – unblinded recall, variable window	Low – EPV=26.5, split-then-select	High
[30] Shivaprasad (2024)	High – Kaggle, unverified provenance	High – unblinded, concurrent	High – “suicide attempt” item as PPD proxy	High – univariate selection on full dataset pre-split	High
[31] Tang (2024)	Low	Low	Low	High – EPV borderline; only proxy calibration (loss curves)	High
[32] Zafar (2025)	High – convenience enrollment	High – unblinded, cross-sectional	High – antenatal/postnatal cases combined into one target	Low – EPV>100, split-then-model, calibration reported	High
[33] Zhang (2020)	High – 54.9% attrition (demographics checked, not fully reassuring)	Low – blinded, before outcome	Low – validated EPDS, blinded	High – EPV=4.4–15.7, selection before split	High

Several recurring patterns emerged. First, three studies ([24] Lin and Zhou, [25] Nasim et al., and [30] Shivaprasad et al.) all drew on the same public Kaggle survey dataset (attributed to Mosaraf, 2023) and used a single questionnaire item (“feeling anxious” in two cases, “suicide attempt” in one) as a proxy label for postpartum depression rather than a validated screening instrument or clinical diagnosis; this construct mismatch was judged high risk on the Outcome domain in all three cases regardless of the strong discrimination metrics reported. Second, several studies conducted feature selection or exploratory analysis (e.g., correlation heatmaps, univariate tests, wrapper methods) on the entire dataset prior to splitting into training and test partitions ([26], [27], [30], [33]), which can inflate apparent performance through data leakage and was judged high risk on the Analysis domain even where sample size was otherwise adequate. Third, Huang et al. [20] excluded participants with intermediate PHQ-9 scores (5–8) specifically to sharpen class separation between low-risk and at-risk groups; because this exclusion was applied based on the outcome value itself, it was judged high risk on the Participants domain despite otherwise strong methodology (blinded, temporally separated predictors; regularized, split-based model development).

3.9 Performance Evaluation

The performance of the models was measured through a large number of different evaluation metrics. Amongst those, discrimination measures were the most prevalent, such as accuracy, Area Under the Receiver Operating Characteristic Curve (AUC/ROC-AUC), precision, recall, sensitivity, specificity, F1-score, and Matthews Correlation Coefficient. A number of papers also reported Positive Predictive Value, Negative Predictive Value, Precision–Recall AUC, Geometric Mean, or runtime metrics.

However, only a small number of studies used measures for evaluating calibration, such as the Brier score, calibration plot, or Decision Curve Analysis for the clinical usefulness assessment. Since postpartum depression datasets are usually imbalanced, several works performed imbalance-aware performance evaluation by measuring F1-score, Precision–Recall curves, or using oversampling methods like SMOTE and ADASYN during the model training process. Different datasets and algorithms achieved various levels of predictive performance; however, newer ensemble and hybrid models generally showed the best performance results.

3.10 Explainability

Explainability was gradually introduced to machine learning-based postpartum depression prediction models, but the methods differed across studies. Classic models, including Logistic Regression and Decision Trees, offered inbuilt explainability via model parameters or decision trees. Ensemble and deep learning models, on the other hand, used mostly post-hoc explainability tools.

SHAP became one of the most common explainability tools, followed by LIME, feature importance, variable importance, and correlation analysis[36]. Some recent studies used a combination of several explainability approaches in order to achieve both global and local explanation of model predictions. Predictors found to be constantly significant across all studied papers were prenatal anxiety and depression, mental history, sleep, stress, social support, relationships, personality, and maternal demographics.

3.11 General Methodological Trends

As a whole, the reviewed studies show a gradual improvement in machine learning methodologies used for predicting postpartum depression. Although initial studies utilized standard machine learning models along with simplistic methodologies for data preprocessing and validation, recent studies have utilized sophisticated ensemble and deep learning models, thorough feature selection methods, strong validation methods, and explainable artificial intelligence techniques. However, even with such improvements in methodological procedures, there is still considerable heterogeneity within the machine learning process, especially within data preprocessing, feature engineering, feature selection, and validation methods.

4. DISCUSSION

4.1 Evolution of Machine Learning Methods Used for PPD Prediction

The reviewed papers showcase a continuous development in machine learning methods used for predicting PPD by shifting from statistical methods and classical machine learning algorithms to ensemble learning, deep learning, and hybrid models[11]. The most common baseline machine learning methods applied in the selected literature include logistic regression, decision trees, support vector machines, and naïve bayes, with random forest, xgboost, deep neural networks, and reinforcement learning models becoming more popular in recent research to identify interactions between predictors[37–39].

Even though this evolution proves methodological progress in the field, an increased complexity of the models does not necessarily mean that they become more clinically useful. Some research reports high accuracy of predictions based on complex methods designed on the basis of limited datasets without external validation of the models developed. Thus, future advances in this area should focus more on methodological soundness and clinical application.

4.2 Dataset Characteristics and Selection of Predictors

The majority of studies reviewed have used structured questionnaire-based datasets, with relatively few studies using electronic health records (EHRs), clinical registries, or multi-source data. Structured questionnaires provide a practical and effective approach for capturing psychosocial, behavioural, and self-report measures that are crucial in estimating the risk of developing PPD. On the other hand, the use of self-reports poses a risk of recall bias, while EHRs will give objective clinical information without psychosocial details. A combination of both sources will provide a better estimate of PPD risk profiles.

There was significant heterogeneity in the predictor selection process across the different studies. The number, type, and time at which variables are selected for inclusion in the model building process differs significantly. Even though factors such as demographics, obstetric, clinical, and psychosocial are used, there is no standardized set of predictors. This being said, it is

evident that psychosocial factors such as prenatal anxiety, depression, stress, sleep problems, social support, and pre-existing mental health issues emerged as predictors of PPD risk.

4.3 Data Preprocessing and Feature Engineering

Data preprocessing is an important step in machine learning models' development, although there was found to be quite a lot of variation in both data processing approach and its documentation in the reviewed literature. Some papers implemented elaborate preprocessing pipelines including handling of missing data, feature scaling, encoding, and class imbalance, while others were not descriptive enough about their data preprocessing procedure. This might impact reproducibility and make comparisons between different models impossible.

As compared to preprocessing, feature engineering was addressed quite sparsely in the reviewed literature. In most cases, the researchers used basic data transformations, whereas only a few used more sophisticated methods to generate features using domain knowledge, deep learning features extraction, or optimization. Given that predictors used in the analysis have a significant impact on performance of models, future research needs to consider the issues of preprocessing and feature engineering.

4.4 Feature Selection and Modelling

In the reviewed articles, the authors used various feature selection methodologies ranging from filter-based, wrapper-based, embedded and model-driven. Nonetheless, there was no common framework for conducting feature selection processes in the reviewed studies, which is indicative of lack of established practices of finding clinically important predictors in predicting PPD. This inconsistency of the feature selection methodology could have influenced different model performances.

On the topic of modelling, there have been more recent studies that investigated ensemble learning, deep learning and hybrid modelling among other machine learning techniques. While the use of such sophisticated models has the capability of discovering nonlinear associations between variables in structured health datasets, there are several considerations when choosing which technique to apply, such as size of the dataset and its computability.

4.5 Validation and Clinical Generalizability

Although various methods of internal validation have been used widely in many studies such as train-test splitting and cross-validation, the application of external validation was documented in relatively few studies. While internal validation offers important insights into the model performance within the study cohort, it is insufficient in proving the ability of the model to preserve its performance in different clinical settings. Thus, the lack of sufficient external validation represents one of the significant challenges that restrict the clinical applicability of ML-based PPD prediction models.

Multicentre model development and independent external validation of these models with diverse populations need to be considered in future research efforts. Besides, there is a need for standardized validation and reporting frameworks to assess the robustness of the model and compare studies; adherence to established prediction-model reporting standards such as TRIPOD, and its AI-specific extension TRIPOD-AI, would improve comparability across studies and make external validation attempts more feasible for future researchers.

4.6 Explainability and Clinical Translatability

The growing prevalence of explainable artificial intelligence (XAI) methods is indicative of the understanding of the fact that the sole presence of predictive performance is not sufficient for clinical application. Some recent works used methods like SHAP and LIME for predictor selection and increased explainability of the models, while classical approaches like Logistic Regression inherently have interpretable model coefficients. Such techniques can help clinicians understand the results of machine learning algorithms and judge whether the discovered predictors are clinically reasonable.

Explainability is not always present in the literature, and there are cases when more complicated models provide little to no explanation of their logic. Moreover, explainability in itself is not enough for clinically valid algorithms since explanations have to be stable, clinically valid, and properly validated. Future research should focus on explainability as a mandatory step of machine learning algorithm development rather than just post hoc work. A good combination of predictive performance, interpretability, proper validation, and clinical applicability will be crucial for translation.

4.7 Research Gaps and Future Directions

Despite the high accuracy of predictive models reported in all of the above-mentioned papers, some of the issues in methodology still prevent their application in practice. First, many models are based on retrospective and single-centre data with low representativeness of demographic and geographical variables, limiting their generalizability in clinical practice. Second, the common use of self-reported measures such as EPDS and PHQ-9 as the outcomes can be problematic due to misclassification bias and calls for further studies that would include clinically diagnosed cases and long-term longitudinal data.

Another problem is the increasing complexity of the models built from relatively small datasets, which makes the risk of overfitting even higher and decreases the robustness of the models. For future research, multicentre collaboration and working with bigger and more representative structured databases along with standardized data preparation and model validation process as well as external validation are crucial.

Though explainable AI became one of the most popular topics in recent research, its application varies greatly between different studies. In future, explainability should be considered during the whole process of the models' development while accounting for fairness, transparency, algorithmic bias, and usability by clinicians.

4.8 Strengths and Weaknesses of This Systematic Review

There are several strengths of this systematic review. Firstly, it gives a thorough overview of the ML pipeline for the prediction of PPD from structured data, covering the properties of the dataset, preprocessing, feature engineering, feature selection, modelling, validation, performance evaluation, and explainability. In contrast to systematic reviews that concentrate solely on comparison of algorithms, this review evaluates methodological aspects of the full prediction pipeline enabling analysis of current trends, limitations, and possibilities for improvement in model building.

However, there are several weaknesses to be mentioned. First, only English-language peer-reviewed articles were taken into account. Second, this review was not prospectively registered (e.g., with PROSPERO). Third, although a full four-domain PROBAST assessment was completed (Section 3.8.1, Table 3), the domain-level judgments rely on the detail available in the published studies and reviewers' extraction rather than independent re-analysis of primary data, so some judgments (particularly for studies with sparse methodological reporting) carry residual uncertainty. Fourth, adherence of the primary studies to prediction-model reporting standards such as TRIPOD/TRIPOD-AI was not formally audited, although reporting gaps observed during data extraction (e.g., the Table 2/PROBAST-extraction discrepancies noted for two studies) suggest that adherence is likely inconsistent across the included literature. Finally, this review is limited to structured data and does not take into account ML models based on unstructured or multimodal datasets.

5. CONCLUSION

This systematic review reviewed the current evidence base for machine learning prediction models for postpartum depression based on structured data and evaluated the whole process of machine learning modelling, including data set properties, predictors, data preprocessing, features, development, validation, performance, and explainability. While there have been significant improvements in terms of prediction accuracy due to the use of ensemble, deep learning, and hybrid methods, there are still many differences in methodologies used in various studies, especially in relation to data preparation, feature selection, validation, and reporting.

These results imply that future progress in the field of postpartum depression prediction needs more than just algorithm developments. Standardization, reproducibility, interpretability, and clinical relevance of machine learning processes are essential factors for future success. The use of multicentre data sets, prospective studies, independent validation, and explainable AI techniques should be considered in order to create valid prediction models for early detection of PPD.

REFERENCES

- [1] Wang Z, Liu J, Shuai H, et al. Mapping global prevalence of depression among postpartum women. *Transl Psychiatry* 2021; 11: 543.
- [2] Khamidullina Z, Marat A, Muratbekova S, et al. Postpartum Depression Epidemiology, Risk Factors, Diagnosis, and Management: An Appraisal of the Current Knowledge and Future Perspectives. *JCM* 2025; 14: 2418.

- [3] Alfayumi-Zeadna S, Masarwe S, Findling Y, et al. The impact of postpartum depression on infant development in the first year of life. *J Affect Disord* 2025; 388: 119558.
- [4] Ng'oma M, Waqas A, Seward N, et al. Identification, Prevention, and Treatment of Perinatal Depression in Low- and Middle-Income Countries: A Narrative Review. *Biological Psychiatry* 2026; 99: 769–778.
- [5] Yang X-W, Jiang X-L, Wu Y-L. Clinical investigation of postpartum depression risk factors and screening predictors. *World J Psychiatry* 2026; 16: 113101.
- [6] Heck JL. Screening for Postpartum Depression in American Indian/Alaska Native Women: A Comparison of Two Instruments. *Am Indian Alsk Native Ment Health Res* 2018; 25: 74–102.
- [7] Kamath K DrK, Safa S, Tantry SK. Evaluation Of Post Partum Depression Using Epds Scale. *Int J Res Publ Rev* 2024; 5: 4931–4936.
- [8] Meltzer-Brody S, Boschloo L, Jones I, et al. The EPDS-Lifetime: assessment of lifetime prevalence and risk factors for perinatal depression in a large cohort of depressed women. *Arch Womens Ment Health* 2013; 16: 465–473.
- [9] Belciug S, Iliescu D. Perinatal Depression and Artificial Intelligence. *Intelligent Systems Reference Library* 2023; 234: 239–244.
- [10] Fanos V, Dessì A, Deledda L, et al. Postpartum depression screening through artificial intelligence: preliminary data through the Talking About algorithm. *Journal of Pediatric and Neonatal Individualized Medicine* 2023; 12: 1–11.
- [11] Novick AM, Kwitowski M, Dempsey J, et al. Technology-Based Approaches for Supporting Perinatal Mental Health. *Current Psychiatry Reports* 2022; 24: 419–429.
- [12] Devi S, Dumbre D, Chavan R, et al. AI-Driven Prediction of Postpartum Depression: A Machine Learning Approach for Early Risk Assessment. Institute of Electrical and Electronics Engineers Inc. Epub ahead of print 2025. DOI: 10.1109/I2ITCON65200.2025.11210557.
- [13] Page MJ, Moher D, Bossuyt PM, et al. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. *BMJ* 2021; n160.
- [14] Behera M, Meher D. Zotero: An Overview of Open Source Citation Management Tool for Researchers.
- [15] Andersson S, Bathula DR, Iliadis SI, et al. Predicting women with depressive symptoms postpartum with machine learning methods. *Scientific Reports*; 11. Epub ahead of print 2021. DOI: 10.1038/s41598-021-86368-y.
- [16] Clapp MA, Castro VM, Verhaak P, et al. Stratifying Risk for Postpartum Depression at Time of Hospital Discharge. *Am J Psychiatry* 2025; 182: 551–559.
- [17] Fazraningtyas WA, Rahmatullah B, Naparin H, et al. A predictive model for postpartum depression: ensemble learning strategies in machine learning. *Indonesian Journal of Electrical Engineering and Computer Science* 2025; 37: 443–451.
- [18] Gopalakrishnan A, Venkataraman R, Gururajan R, et al. Predicting Women with Postpartum Depression Symptoms Using Machine Learning Techniques. *Mathematics*; 10. Epub ahead of print 2022. DOI: 10.3390/math10234570.
- [19] Gulmohamed RB, Sasikala N, Hetany WA, et al. AI Based Early Prediction of Postpartum Depression Using Multidimensional Maternal Health Data with Integrated Secure Data Storage and Privacy Protection Protocols. *Journal of Internet Services and Information Security* 2026; 16: 548–569.
- [20] Huang Y, Alvernaz S, Kim SJ, et al. Predicting prenatal depression and assessing model bias using machine learning models. *medRxiv: the preprint server for health sciences* 2023; 2023.07.17.23292587.
- [21] Huang X, Zhang L, Zhang C, et al. Postpartum depression risk prediction using explainable machine learning algorithms. *Frontiers in Medicine*; 12. Epub ahead of print 2025. DOI: 10.3389/fmed.2025.1565374.
- [22] Kafi GMAA, Suha SA, Ali G, et al. Transparent and tuned: XAI for postpartum depression screening with comprehensive feature engineering and real-time web-integration. *Network Modeling Analysis in Health Informatics and Bioinformatics*; 15. Epub ahead of print 2026. DOI: 10.1007/s13721-026-00783-7.
- [23] Kim H. Predictive Analysis of Postpartum Depression Using Machine Learning. *Healthcare (Switzerland)*; 13. Epub ahead of print 2025. DOI: 10.3390/healthcare13080897.

- [24] Lin Y, Zhou D. A method for predicting postpartum depression via an ensemble neural network model. *Frontiers in Public Health*; 13. Epub ahead of print 2025. DOI: 10.3389/fpubh.2025.1571522.
- [25] Nasim S, Sami Al-Shamayleh A, Thalji N, et al. Novel Meta Learning Approach for Detecting Postpartum Depression Disorder Using Questionnaire Data. *IEEE Access* 2024; 12: 101247–101259.
- [26] Pillai AP, Chinnasamy NV. Prediction Of Postpartum Depression With Dataset UsingHybrid Data Mining Classification Technique. *International Journal of Computational and Experimental Science and Engineering* 2025; 11: 475–483.
- [27] Qi W, Wang Y, Wang Y, et al. Prediction of postpartum depression in women: development and validation of multiple machine learning models. *Journal of Translational Medicine*; 23. Epub ahead of print 2025. DOI: 10.1186/s12967-025-06289-6.
- [28] Shen S, Qi S, Luo H. Automatic Model for Postpartum Depression Identification using Deep Reinforcement Learning and Differential Evolution Algorithm. *International Journal of Advanced Computer Science and Applications* 2023; 14: 154–166.
- [29] Shin D, Lee KJ, Adeluwa T, et al. Machine learning-based predictive modeling of postpartum depression. *Journal of Clinical Medicine* 2020; 9: 1–14.
- [30] Shivaprasad S, Chadaga K, Sampathila N, et al. Explainable machine learning methods to predict postpartum depression risk. *Systems Science and Control Engineering*; 12. Epub ahead of print 2024. DOI: 10.1080/21642583.2024.2427033.
- [31] Tang Y, Huang T, Yin X. Postpartum Depression Identification: Integrating Mutual Learning-based Artificial Bee Colony and Proximal Policy Optimization for Enhanced Diagnostic Precision. *International Journal of Advanced Computer Science and Applications* 2024; 15: 332–347.
- [32] Zafar A, Wasim M, Akram BA, et al. Prediction of perinatal depression among women in Pakistan using Hybrid RNN-LSTM model. *PeerJ Computer Science*; 11. Epub ahead of print 2025. DOI: 10.7717/peerj-cs.2673.
- [33] Zhang W, Liu H, Silenzio VMB, et al. Machine learning models for the prediction of postpartum depression: Application and comparison based on a cohort study. *JMIR Medical Informatics*; 8. Epub ahead of print 2020. DOI: 10.2196/15516.
- [34] Usuda K, Nishi D, Okazaki E, et al. Optimal cut-off score of the Edinburgh Postnatal Depression Scale for major depressive episode during pregnancy in Japan. *Psychiatry Clin Neurosci* 2017; 71: 836–842.
- [35] Ford J, Thomas F, Byng R, et al. Use of the Patient Health Questionnaire (PHQ-9) in Practice: Interactions between patients and physicians. *Qual Health Res* 2020; 30: 2146–2159.
- [36] Padmanabhan K, Lu M, Feng D, et al. Explainable machine learning in healthcare: methods, interpretation, and applications for clinical research. *J Am Med Inform Assoc* 2026; ocag077.
- [37] Sreeji S, Shirley CP. From Data to Diagnosis: A Review of Machine Learning Models for Postpartum Depression Prediction. Institute of Electrical and Electronics Engineers Inc. Epub ahead of print 2025. DOI: 10.1109/AIMLA63829.2025.11040647.
- [38] Prabhashwaree THKR, Wagarachchi NM. Predicting Mothers with Postpartum Depression using Machine Learning Approaches. Institute of Electrical and Electronics Engineers Inc., pp. 28–34.
- [39] Prabhashwaree THKR, Mihirini Wagarachchi N. Towards Machine Learning Approaches for Predicting Risk Level of Postpartum Depression. Institute of Electrical and Electronics Engineers Inc. Epub ahead of print 2022. DOI: 10.1109/SLAAI-ICAI56923.2022.10002477.