



Asian Research Association

INTERNATIONAL RESEARCH JOURNAL OF MULTIDISCIPLINARY TECHNOVATION



Comparative Evaluation of Tree-Based Algorithms for Postpartum Mental Health Risk: a Methodological Benchmark

T. Om Bhavani Shankar ^a, R. Prathiba ^a, V. Keerthi Vasan ^a, T. Sivakumar ^{a,*}, K. Kanagaraj ^a, K. Sudhakar ^a

^a School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu – 632 014, India

* Corresponding Author Email: sivakumar.t@vit.ac.in

DOI: <https://doi.org/10.54392/irjmt26518>

Received: 22-12-2025; Revised: 08-09-2026; Accepted: 20-09-2026; Published: 30-09-2026



Abstract: Gradient-boosting classifiers now underpin most screening tools proposed for postpartum mental health, but published comparisons rarely separate what the modelling pipeline contributes from what the clinical construct itself contributes. Here we benchmark XGBoost, CatBoost, and LightGBM inside a single leakage-controlled pipeline, asking whether performance gaps trace back to identifiable structural properties rather than inconsistent data handling. A public postpartum survey dataset of 1,503 records was split 80:20 into development and test sets, with encoding and imputation fitted on the development set alone, and each classifier tuned via stratified five-fold cross-validation over a shared search space, isolating architecture as the sole variable. LightGBM's histogram-based, leaf-wise growth produced the strongest held-out discrimination (accuracy 88.4%, weighted F1 88.4%, balanced accuracy 87.8%, macro F1 88.0%), ahead of XGBoost (86.0%) and CatBoost (82.4%), with its advantage concentrated in the Moderate and High categories (81 of 92 High-category records correctly classified, versus 80 and 76). The selected classifier was integrated into a four-layer architecture, preprocessing, predictive inference, dialogue management, and deterministic recommendation, keeping the model's output separate from user-facing text and independently auditable. We report the full search space and the structural mechanisms plausibly explaining the ranking, and state plainly that training time, memory, and inference latency were not measured, naming this as required future work under TRIPOD+AI. The results demonstrate a reproducible method for comparing boosting architectures on a fixed screening task, they do not establish diagnostic validity, calibrated suicide-risk estimation, or clinical readiness, which require external validation beyond this internal benchmark.

Keywords: Gradient Boosting, Xgboost, Lightgbm, Catboost, Hyperparameter Optimization, Postpartum Mental Health, Algorithmic Architecture, Computational Performance, Safety-Aware Artificial Intelligence, Conversational Agent

1. Introduction

Gradient-boosted decision trees remain the dominant family of models for structured, tabular health data, with XGBoost, LightGBM, and CatBoost together accounting for most applied clinical-prediction studies on tabular cohorts over the past three years [1]. All three share an additive-ensemble inductive bias fit by gradient descent, yet diverge in how they grow trees, encode categorical predictors, and regularize against overfitting [2]. XGBoost grows trees level by level using an exact or histogram-approximated greedy search with second-order loss information [3] LightGBM grows trees leaf-wise using gradient-based one-side sampling and exclusive feature bundling [4] CatBoost uses ordered boosting and permutation-driven target statistics to remove the leakage that arises when the same data compute both a categorical encoding and the gradient

step [5]. These structural differences let three algorithms trained on identical rows, under the same loss function and search space, rank differently on a held-out partition [6]. Tuning strategy and stopping rule alone can shift that ranking by as much as algorithm choice itself [7], a finding a related large-scale study confirms [8]. In one health-services task, the choice among nine hyperparameter-optimization methods alone changed discrimination and feature-importance rankings enough to alter which predictors a clinician would see [9]. A benchmark that omits search space, encoding, and leakage control cannot be reproduced or support a claim about which architecture suits a given regime [10].

Postpartum mental-health screening adds a construct-validity problem. A scoping review of AI approaches to postpartum depression documents a fast-growing literature alongside persistent shortfalls in

external validation, calibration, and standardized reporting [11]. A review of low- and middle-income countries found machine learning the default approach even though few studies specify software environment, hyperparameters, or cost [12]. A 2025 review of AI-oriented PPD risk models reached a similar verdict: heterogeneity in outcome definition and validation design precludes any pooled statement about which algorithm class is superior [13]. An EHR study of over twelve thousand deliveries found even a carefully engineered random-forest model achieved only moderate discrimination once diagnosis codes and documentation gaps were reconciled into a single label, and a trimester-resolved cohort of 2,865 women found consistent discrimination only once the outcome was tied to a validated instrument, the EPDS, rather than a coarser label [14]. A 2025 meta-analysis confirmed accuracy varies by outcome definition far more than by algorithm family, why internal discrimination must be kept separate from clinical validity [15]. These reviews motivate the design adopted here: this study holds data, split, preprocessing, and search space fixed and varies only the boosting architecture, so any ranking reads as a statement about the algorithms rather than the study.

This study works with a three-category, self-reported item on suicide-attempt status from a public postpartum-survey dataset of 1,503 records, mapped onto interface labels Low, Moderate, and High as a communication convenience, not a clinical-severity claim; Section 3.3 details the mapping. A 2025 meta-analysis of 122 studies and over half a million women put global perinatal co-morbid anxiety and depression prevalence at roughly 9 percent, varying more by assessment timing and country than by any single risk factor [16]. US national-registry analyses find suicide accounts for a disproportionate share of pregnancy-associated deaths, linked to intimate-partner problems and substance use [17], and a Swedish cohort followed for eighteen years found a perinatal depression diagnosis carried a more than sevenfold hazard of subsequent suicidal behaviour [18]. A 2025 cohort study confirmed psychiatric history and prior ideation as dominant predictors, noting a single self-reported item omits recency, intent, and planning [19] while a 2026 explainable-ML study found a substantial share of cases unexplained by predictor set alone, matching the residual misclassification reported here for the High category [20]. These findings justify treating the source label as a self-reported response, and any derived score as internal discrimination rather than a validated suicide-risk estimate.

A second strand concerns communicating a boosting model's output once embedded in a conversational interface: a 2025 review of mental-health chatbots found only a minority of large-language-model systems had undergone clinical-efficacy testing, versus rule-based systems, more limited but substantially more auditable [21]. An evaluation of five language models

handling simulated crisis inputs found meaningful rates of unsafe responses to subtle signals, attributing safety to alignment engineering rather than model scale [22]; a rule-based youth agent's case series found users increasingly expect generative-level fluency, creating tension between auditability and naturalistic quality [23]. Accordingly, the boosting classifier produces only a probability vector and operational category, while every word of user-facing text is drawn from a fixed, clinician-reviewable rule table.

This study connects the strongest held-out discrimination among the three algorithms to a documented structural property rather than an unexplained result, through a leakage-controlled protocol isolating algorithmic from data-handling contributions, characterizes the selected classifier's class-resolved error structure, treating High as the primary quantity of clinical interest given the asymmetric costs of False Negatives and Positives, and specifies a four-layer architecture keeping predictive output auditable and separable from user-facing language, naming the training-time, memory, and latency evidence the source study never captured and how a future package should report it under TRIPOD+AI [24]. It does not propose a new algorithm or claim universal superiority; the ranking is conditional on this dataset, and offers instead a mechanism-linked account of why three architectures diverge, and a template for keeping predictive output separate from the language a screening tool shows a user.

2. Related Work

This section summarizes prior work most directly relevant to the present benchmark. Table 1 organizes representative studies by author, objective, methodology, dataset, and reported outcomes, alongside the specific limitation each leaves open; the subsections that follow synthesize this literature in more analytical depth.

Taken together, these studies span four distinct respects in which prior work bears on the design adopted here: outcome and predictor heterogeneity in PPD prediction, structural and computational properties of gradient-boosting architectures, hyperparameter-search and reproducibility practice, and the clinical stakes and predictors associated with perinatal suicide. No single study among them addresses all of these respects within one controlled, multi-algorithm boosting comparison.

2.1 Predictive Modelling of Postpartum Mental Health

Research on postpartum mental-health prediction spans registries, EHR cohorts, and survey repositories; the data source bears directly on what a reported figure means (Table 1: Ma, Sibbald, Huang, Qi, Kim) [14].

Table 1. Overview of literature most directly relevant to the present postpartum mental-health benchmark.

Author(s), Year	Objective	Methodology	Dataset	Outcomes	Limitation
Yildiz & Kalayci, 2024 [1]	Benchmark GBDT vs. tabular deep learning	TabNet, TabTransformer comparison	Multiple medical-diagnosis datasets	GBDT retained highest average rank	Advantage regime-specific, may not generalize to larger data
Qi <i>et al.</i> , 2025 [2]	Predict PPD across a larger cohort	Seven feature-selection strategies × six ML algorithms	Multicentre Chinese cohort	Predictor/outcome choice explained more variance than algorithm	Algorithm's own contribution not isolated
Ma <i>et al.</i> , 2025 [4]	Build a PPD classifier from EHR	Outcome built from EPDS, PHQ-9, diagnosis codes	12,284 deliveries, tertiary-care EHR	Accuracy depended on outcome-label construction	Diagnosis-code reliance introduced selection bias
Huang <i>et al.</i> , 2025 [5]	Identify PPD psychosocial predictors	Explainable GB with LASSO/Boruta selection	1,065 women, single-centre China	Predictor availability drives cross-study variance	Single-centre design
Kim, 2025 [6]	Predict PPD from routine variables	Conventional classifiers	Korean single-institution cohort	Comparable results with a smaller feature set	Smaller feature set, more conventional classifiers
Gonçalves & de Carvalho, 2025 [9]	Catalogue GBDT interpretability methods	Methodological review	Not applicable	Shapley methods align with additive-tree structure	Review, not empirical validation
Xia <i>et al.</i> , 2025 [13]	Systematic review of AI PPD risk models	Synthesis of published studies	Published PPD literature	Heterogeneity precludes pooled algorithm judgment	Few studies share a reproducible search space
Sibbald <i>et al.</i> , 2025 [14]	Prenatal, trimester-resolved PPD prediction	Lasso model tied to a validated EPDS score	2,865 women, Dutch cohort	Stable discrimination once tied to EPDS	Later-pregnancy predictors added little
Xie <i>et al.</i> , 2025 [15]	Quantify heterogeneity across ML PPD studies	Meta-analysis	Pooled published studies	Discrimination varies more by outcome/predictors than algorithm	Comparability limited without fuller reporting
Zivin <i>et al.</i> , 2024 [17]	Characterize perinatal suicide circumstances	National-registry analysis	US National Violent Death Reporting System	IPV, depressed mood, substance use, bereavement common	Descriptive, not a predictive model
Yu <i>et al.</i> , 2024 [18]	Link PPD diagnosis to suicide-behaviour risk	Population-register cohort, 18-year follow-up	Swedish population register	More than sevenfold hazard in first year	Registry diagnosis, not self-report
Goldman-Mellor <i>et al.</i> , 2025 [19]	Incidence/risk factors for suicide attempt, pregnancy/postpartum	Clinical-cohort analysis	Clinical cohort	Psychiatric history, prior ideation dominant	Single item omits recency, intent, planning
Hua <i>et al.</i> , 2025 [21]	Trace chatbot evolution, rule-based to LLM	Systematic review, 160 studies	160 published	Rule-based more validated, LLM early-stage	Reflects a fast-moving field

Table 1. Overview of literature most directly relevant to the present postpartum mental-health benchmark.

Author(s), Year	Objective	Methodology	Dataset	Outcomes	Limitation
			chatbot studies		
Arnaiz-Rodriguez <i>et al.</i> , 2026 [22]	Evaluate LLM safety on crisis inputs	Five LLMs evaluated	Simulated crisis inputs	Unsafe responses to subtle signals	Simulated, not clinical inputs
Meaney <i>et al.</i> , 2025 [25]	Compare HPO strategies for a GB model	Nine strategies on one XGBoost model	Health-services prediction task	Optimization method alone changed rankings	Single-algorithm scope
Gupta <i>et al.</i> , 2024 [26]	Compare CatBoost, LightGBM for PPD	Direct algorithm comparison	Separate postpartum survey dataset	Similar qualitative ranking to present study	Not numerically comparable (differing pipelines)

Across these studies, predictor availability and outcome-label construction, not algorithm choice, explain the largest share of between-study variance. A 2024 study directly comparing CatBoost and LightGBM for early PPD detection on a separate survey dataset reported a similar qualitative ranking to the one obtained here, lending independent, cross-dataset support to the structural explanation developed, though differing predictor sets and pipelines make direct numerical comparison inappropriate [26].

Systematic syntheses published in 2025 quantify this heterogeneity directly (Xie, Xia, Table 1), converging on the same barrier: too few studies report a shared, reproducible search space to support any pooled judgment about which algorithm performs best. This motivates the design adopted here: because outcome definition and predictor set already introduce substantial between-study heterogeneity, a controlled, single-dataset comparison holding outcome, predictors, split, and preprocessing constant is the only design capable of isolating the algorithm's own contribution to performance.

Suicide-related outcomes in the perinatal period form a distinct, clinically consequential sub-literature. National-registry analysis found perinatal decedents significantly more likely to have experienced intimate-partner problems, depressed mood, substance use, and recent bereavement. A Swedish cohort followed for eighteen years found a perinatal depression diagnosis carried a more than sevenfold hazard of subsequent suicidal behaviour in the first year after diagnosis. A 2025 clinical-cohort analysis confirmed psychiatric history and prior ideation as dominant predictors, noting a single item omits recency, intent, and planning. A 2026 explainable-ML study found roughly forty percent of cases unexplained by predictor set alone, matching the unexplained High-category misclassification reported here, and a separate Swedish study linked perinatal

depression to increased all-cause mortality, extending the stakes beyond suicide alone [27]. Outside the perinatal setting, a social-media-based suicide-risk algorithm required explicit external validation before its internally strong discrimination could be treated as clinically meaningful [28]. Two 2025 systematic reviews of adolescent suicide-risk prediction converged: ensemble methods show the strongest internal discrimination though external validation remains rare [29], and effect sizes vary substantially by outcome type and validation status [30].

2.2 Gradient-Boosting Architectures for Structured Clinical Data

A benchmark against tabular deep-learning architectures (TabNet, TabTransformer) found gradient-boosted implementations retain the highest average performance rank at this data scale, traced to lower computational cost and more stable optimization (Table 1). Applied work validates each algorithm in its own domain: an interpretable LightGBM model reported comparable discrimination to XGBoost at reduced tuning cost for coronary heart disease [31], and Shapley-explained XGBoost remained competitive for pressure-injury prediction in diabetic ICU patients [32]. On the present dataset, whose predictors are categorical and low-cardinality, ordered target statistics offer CatBoost little structural advantage relative to the leakage risk they solve, one plausible explanation for its weaker held-out performance (Section 5.2).

2.3 Hyperparameter Optimization, Search-Space Design, and Reproducibility

A gradient-boosting comparison is only as informative as the hyperparameter search behind it, since an under-tuned algorithm can appear structurally inferior for reasons unrelated to its architecture (Table 1, Meaney) [25]. Complementary work on Bayesian-

optimized XGBoost for diabetes-onset prediction found Bayesian search outperformed grid and random search at comparable cost, arguing the search strategy itself, not only the space, is part of the comparison [33].

Handling of class imbalance and missing data is a second, frequently under-reported source of between-study variance. A 2024 comparison of resampling strategies across observational health datasets found resampling choice interacted with algorithm choice inconsistently across datasets, arguing against any single strategy as a universal default [34] the present study applied none, relying on stratified splitting and cross-validation instead, with limitations for the smallest class (Moderate, 67 records) discussed in Section 6. A 2025 evaluation of missingness-handling methods for EHR-based prediction models found simple, development-set-derived imputation performs comparably to more elaborate schemes at lower cost, supporting the approach adopted [35]. More broadly, a 2025 analysis of reproducibility barriers in machine-learning research identified incomplete reporting of code, data, and software environments as the largest obstacle to verification.

A 2024 review of empathic conversational agents similarly found technical performance detecting user affect varied widely, with classification accuracy rarely reported in reproducible detail [36]. A 2025 study of chatbot agents on suicidal-ideation detection found inconsistent escalation behaviour for ambiguous input, directly relevant to the Moderate category here [37]. Broader ethical analysis argues for regulatory transparency distinguishing rule-based logic from generative production [38], and two scoping reviews confirm generative architectures introduce evaluation

and safety burdens deterministic ones do not [39], reinforcing the rationale for keeping the classifier and its language functionally separate and auditable [40].

3. Materials and Methods

3.1 Study Design and Intended Use

This study is a secondary, algorithm-focused analysis of a publicly available, structured postpartum survey dataset, designed to separate what gradient-boosting architecture contributes to predictive performance from what data handling contributes. The workflow comprises four stages (Figure 1): development-only preprocessing, model development under a shared hyperparameter search space, held-out evaluation, and integration of the selected classifier into a four-layer chatbot architecture. Every non-algorithmic pipeline element, the train-test split, feature encoding, imputation rule, cross-validation folds, and search space, is held constant, so the resulting ranking reflects the three architectures rather than an uneven comparison.

The resulting prototype is not intended to diagnose postpartum depression, assess imminent suicide risk, replace a validated screening instrument, or function as a stand-alone crisis service. It is best regarded as an internal technical-validation study of three candidate architectures, together with an interface proof of concept for keeping predictive inference separate from user-facing safety language, no prospective recruitment, clinician adjudication, or clinical-workflow evaluation was performed.

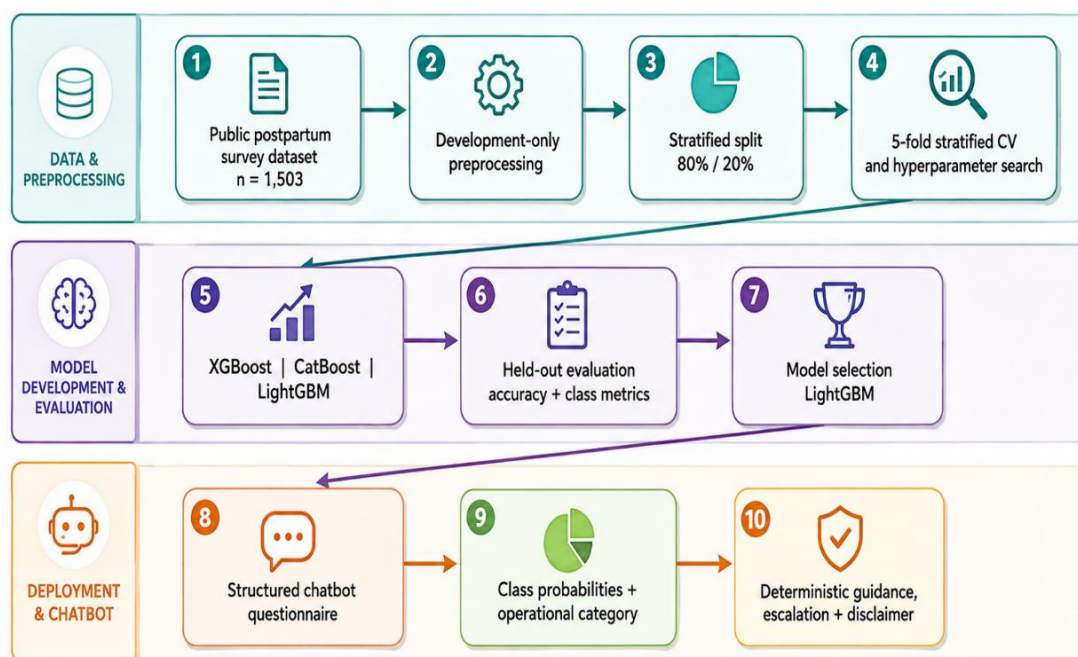


Figure 1. Development, evaluation, and deployment workflow.

Table 2. Predictor groups and outcome variable used in the predictive workflow.

Variable group	Variables	Role in model
Demographic	Age category	Predictor
Emotional symptoms	Sadness or tearfulness, anxiety, guilt	Predictors
Behavioural and cognitive symptoms	Sleep difficulty, concentration or decision difficulty, overeating or loss of appetite	Predictors
Relational symptoms	Irritability toward baby or partner, difficulty bonding with baby	Predictors
Outcome	Self-reported suicide-attempt status: No, Not interested to say, Yes	Three-class target, mapped to interface labels only for chatbot output

3.2 Data Source and Variables

The dataset was obtained from the publicly available Kaggle postpartum-depression repository and contains 1,503 records [41]. Predictors represent demographic, emotional, behavioural, cognitive, and relational characteristics (Table 2), a categorical self-reported item on suicide-attempt status served as the prediction target.

All nine predictors and the outcome are categorical with low cardinality (two to five levels each), a structural property directly relevant since it constrains which categorical-handling mechanism each candidate algorithm can meaningfully exercise.

The dataset contains missing values. In the source analysis, missingness in affected variables was reported to be approximately 2 to 7 percent, distributed unevenly across predictors rather than concentrated in one variable.

The target variable is a survey response rather than a clinically adjudicated diagnosis or prospective event, and this distinction defines the scope of every claim in this paper: the algorithm classifies records according to the source responses No, Not interested to say, and Yes, and provides neither a validated severity score nor a calibrated probability of imminent suicidal behaviour.

The 1,503 records were divided into development and held-out test sets using a stratified 80:20 split (1,202 development, 301 test), preserving the relative distribution of the three target classes [42]. All model-selection and tuning was confined to the development set; the held-out set was used once for final evaluation, preventing the optimistic bias that arises when test-set performance influences model-selection decisions.

Categorical responses were encoded using a consistent mapping; missing categorical values were imputed with the development-set modal value and missing numerical values with the development-set mean, then applied unchanged to test data and chatbot inputs, reducing information leakage. A 2025 evaluation of missingness-handling strategies for EHR-based prediction models found that at this size and structure,

simple development-set-derived imputation performs comparably to more expensive multiple-imputation procedures at a fraction of the cost, supporting this as a defensible default [43]. Because the candidate models are tree-based, feature scaling was not required; the essential requirement was consistency of encoding and imputation between training and inference, since drift between the two mappings would silently corrupt downstream predictions without producing an error.

No synthetic class-balancing was applied; stratified splitting and cross-validation preserved observed class proportions instead, since resampling strategies interact with algorithm choice in dataset-dependent ways, which would have added a poorly controlled source of variation to a comparison meant to isolate the algorithm's own contribution. Consequences for the smallest class, Moderate (67 held-out records), are discussed further below.

The three candidate classifiers share an additive-ensemble, gradient-boosted-tree inductive bias but differ in three structural respects directly relevant to interpreting the ranking: tree-growth strategy, treatment of categorical predictors, and the mechanism used to control overfitting during boosting.

XGBoost's level-wise growth tends to produce more balanced trees than leaf-wise growth at a given maximum depth, at the cost of evaluating more candidate splits per iteration. Both techniques reduce the data points and features examined per split without proportionally degrading split quality, the mechanism identified as the source of its characteristic speed and memory advantage on structured tabular data. CatBoost's ordered boosting and ordered target statistics prevent the target leakage that arises when the same rows both encode a categorical feature and compute the gradient step using that encoding, but this mechanism targets high-cardinality categorical features specifically. On this dataset, where every predictor is categorical but of very low cardinality (at most five levels), the leakage risk it solves is intrinsically small while its computational overhead is still incurred, one plausible, falsifiable explanation for CatBoost's comparatively weaker held-out performance.

Table 3. Structural and computational properties of the three candidate gradient-boosting architectures.

Property	XGBoost	LightGBM	CatBoost
Tree-growth strategy	Level-wise (depth-wise)	Leaf-wise (best-first), depth-constrained	Level-wise, symmetric (oblivious) trees
Split-finding	Exact greedy or histogram-based approximation	Histogram-based with gradient-based one-side sampling	Histogram-based with ordered boosting
Categorical-feature handling	Requires pre-encoding (e.g., one-hot or target encoding)	Native support, less leakage-protected	Native support via ordered target statistics
Primary overfitting control	L1/L2 leaf-weight regularization, second-order gradients	Leaf-wise growth with max-depth cap, GOSS, EFB	Ordered boosting to prevent target leakage / prediction shift
Approximate per-tree split-finding cost	$O(n \cdot d \cdot \log n)$ (exact) or $O(n \cdot d)$ (histogram), n = instances, d = features	$O(n \cdot d)$ with n effectively reduced by GOSS sampling	$O(n \cdot d)$ per permutation, multiplied by number of supporting models
Structural fit to this dataset	Neutral: low-cardinality categorical predictors do not require its exact-search strength	Favourable: symmetric, low-cardinality feature space suits histogram + leaf-wise growth without overfitting risk	Unfavourable: leakage-protection overhead is paid without a correspondingly large leakage risk to offset it

Table 3 summarizes these properties alongside their approximate computational-complexity characterization, drawing on the reviews cited above. The complexity expressions describe known, published behaviour and are explicitly not derived from wall-clock measurements on the present dataset.

3.3 Hyperparameter Search Space and Optimization Procedure

Hyperparameter selection was performed on the 1,202-record development set using stratified five-fold cross-validation, holding the optimization procedure constant across algorithms so any performance difference reflects architecture rather than an uneven search. The shared search space (Table 4) contained learning rates of 0.01, 0.05, and 0.10, maximum tree depths of 4, 6, and 8, and regularization values of 0.1, 0.5, and 1.0, applied identically across all three libraries' native parameters, with a fixed random seed of 42 and early stopping at 50 iterations.

Performance was evaluated on the 301-record held-out test set. Accuracy represents the proportion of all records classified correctly. For each class c , precision is $TP_c / (TP_c + FP_c)$, recall is $TP_c / (TP_c + FN_c)$, and F1-score is the harmonic mean of precision and recall. Weighted precision, recall, and F1-score were calculated by weighting each class-specific value by its support. These metrics were reported in the source analysis.

To strengthen interpretation without introducing new data, macro precision, macro recall, and macro F1-

score were recalculated directly from the reported confusion matrices; macro recall equals balanced accuracy here, the unweighted mean of class-specific recalls, and these secondary metrics depend only on reported cell counts, not record-level prediction vectors [44]. Paired statistical comparisons, resampling-based confidence intervals, ROC-AUC, calibration, and decision-curve analysis are not reported because the required prediction-level data were unavailable. Figure 2 shows the safety aware separation of predictive inference and recommendation logic.

The chatbot was implemented in Python and presents the assessment items sequentially, validating and encoding each response before passing it through the fitted preprocessing pipeline to the selected LightGBM classifier; Algorithm 1 details the full inference procedure.

The operational logic is summarized in Algorithm 1:

Algorithm 1. Inference and recommendation procedure used by the prototype chatbot.

Input: structured user responses

1. Validate completeness and permitted response values.
2. Apply the encoding and imputation rules estimated from the development set.
3. Generate the three-class probability vector using the selected LightGBM model.
4. Select the class with the maximum predicted probability.
5. Map the source response to the interface state: No -> Low, Not interested to say -> Moderate, Yes -> High.

6. Retrieve the predefined guidance and escalation message assigned to that operational category.
 7. Display the operational category, model scores, guidance, and non-diagnostic disclaimer.
- Output: operational category and deterministic guidance.

4. Results

4.1 Dataset Split and Class Composition

The full dataset held 1,503 records. An 80:20 stratified split produced 1,202 development and 301 held-out test records, each confusion matrix contained exactly 301 observations, confirming all three models were evaluated on the same test set. The held-out set contained 142 Low records (47.2%), 67 Moderate (22.3%), and 92 High (30.6%), describing test-set composition only, not population prevalence.

4.2 Overall Model Performance

LightGBM reached the highest held-out accuracy (88.4%) and weighted F1-score (88.4%), ahead of XGBoost (86.0%, 86.0%) and CatBoost (82.4%, 82.3%) as listed in Table 5. These are descriptive results from a single held-out split; a paired

significance test was not performed because record-level prediction outputs were unavailable.

Metrics recalculated from the confusion matrices produced the same ranking: balanced accuracy of 87.8% and macro F1 of 88.0% for LightGBM, versus 85.2%/85.5% for XGBoost and 80.5%/81.4% for CatBoost, as listed in Table 6. Figure 3 shows Held-out test-set performance of XGBoost, CatBoost, and LightGBM across four aggregate metrics. This agreement between weighted and macro-level measures indicates LightGBM's advantage is distributed across all three categories, not concentrated in the largest class. Section 5.3 interprets this ranking against the structural properties in Table 3.

4.3 Confusion-Matrix Analysis and Error Structure

For LightGBM, 128 of 142 Low records, 57 of 67 Moderate records, and 81 of 92 High records were classified correctly. XGBoost correctly classified 125, 54, and 80 records in the corresponding classes, whereas CatBoost correctly classified 124, 48, and 76. LightGBM therefore had the largest diagonal count for each class in the reported test set (Figure 4).

Table 4. Shared hyperparameter search space applied identically to XGBoost, LightGBM, and CatBoost.

Hyperparameter	Search values	Applied consistently via
Learning rate (η)	0.01, 0.05, 0.10	learning_rate (all three libraries)
Maximum tree depth	4, 6, 8	max_depth (XGBoost, CatBoost), max_depth constraint on leaf-wise growth (LightGBM)
Regularization strength	0.1, 0.5, 1.0	reg_lambda (XGBoost), l2_leaf_reg (CatBoost), lambda_l2 (LightGBM)
Cross-validation design	Stratified 5-fold, development set only	Identical fold assignment across all three algorithms
Random seed	42 (fixed)	All stochastic elements (fold assignment, subsampling, permutations)
Early stopping	Patience of 50 boosting iterations	Monitored on held-out cross-validation fold

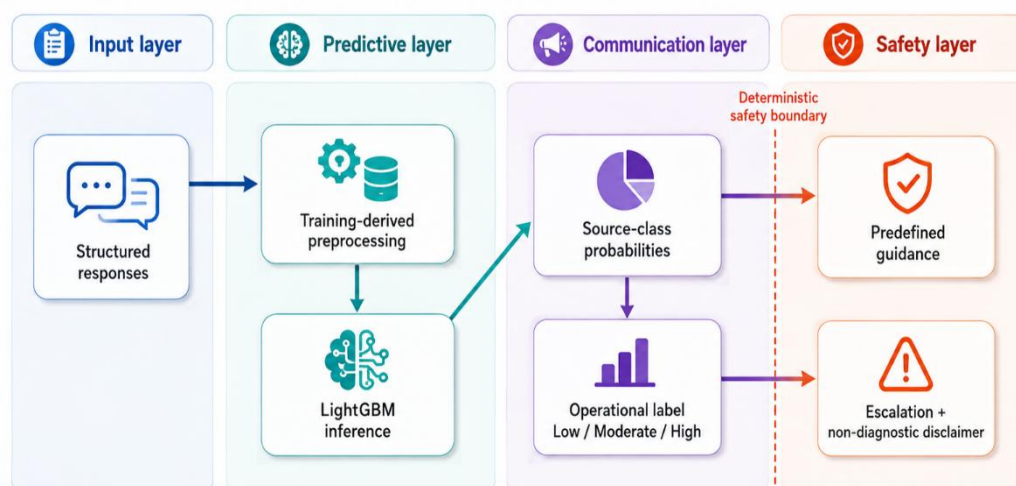
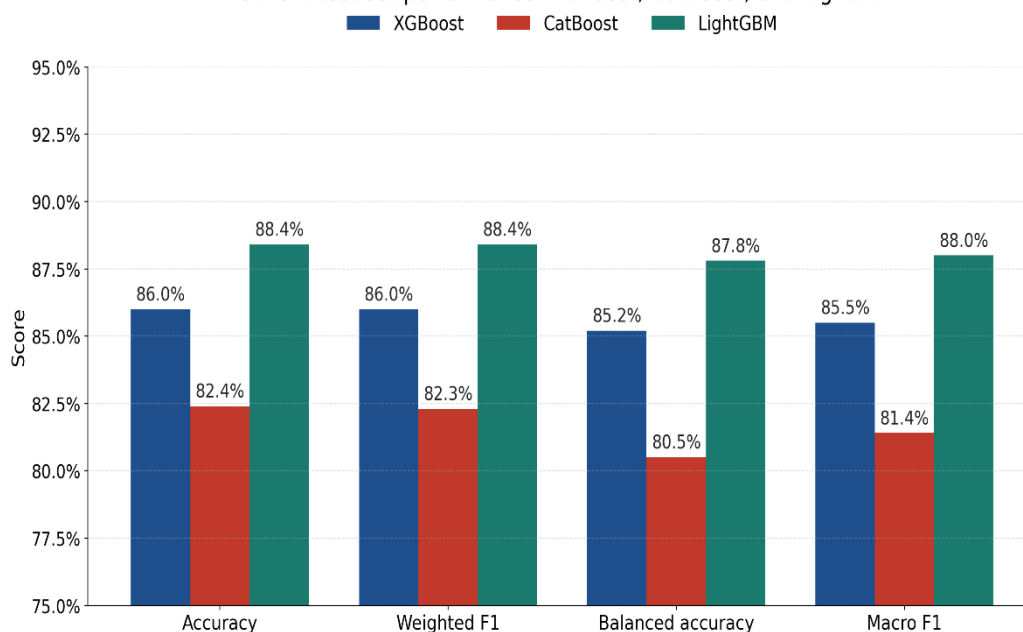
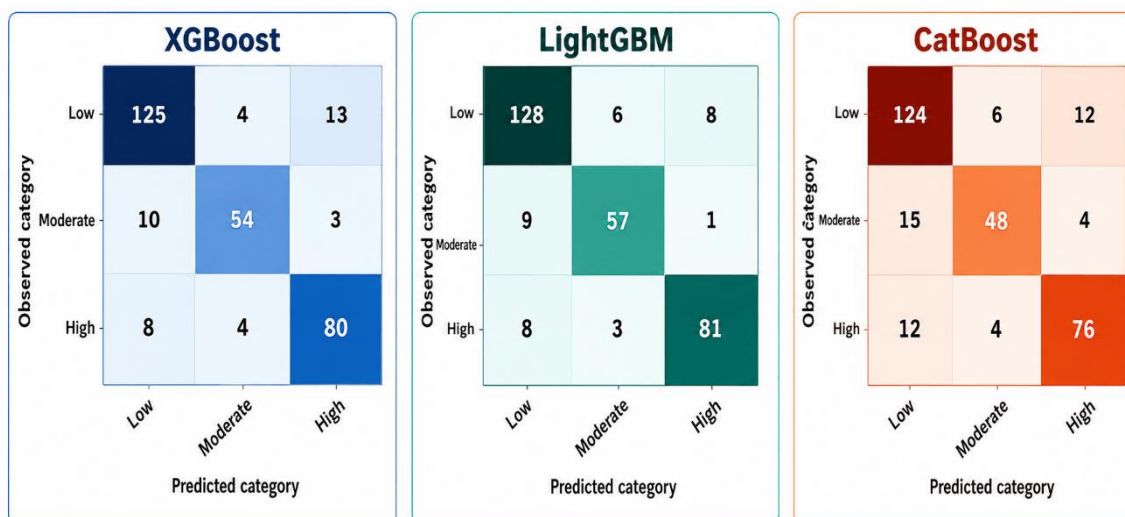


Figure 2. Safety-aware separation of predictive inference and recommendation logic.

Table 5. Held-out test-set performance.

Model	Accuracy	Weighted precision	Weighted recall	Weighted F1	Balanced accuracy	Macro F1
XGBoost	86.0%	86.1%	86.0%	86.0%	85.2%	85.5%
CatBoost	82.4%	82.4%	82.4%	82.3%	80.5%	81.4%
LightGBM	88.4%	88.4%	88.4%	88.4%	87.8%	88.0%

Held-out test-set performance: XGBoost, CatBoost, and LightGBM

**Figure 3.** Held-out test-set performance of XGBoost, CatBoost, and LightGBM across four aggregate metrics.**Figure 4.** Confusion matrices for XGBoost, LightGBM, and CatBoost on the 301-record held-out test set.**Table 6.** Class-specific LightGBM performance on the held-out test set. Operational labels are interface categories and are not clinical severity levels.

Operational category	Support	Precision	Recall	F1-score
Low	142	88.3%	90.1%	89.2%
Moderate	67	86.4%	85.1%	85.7%
High	92	90.0%	88.0%	89.0%

Table 7. Safety-oriented descriptive error profile for observed High-category records, derived directly from the reported confusion matrices.

Model	Observed High	Predicted High	High→Moderate	High→Low	Observed High not classified as High
XGBoost	92	80	4	8	12 (13.0%)
CatBoost	92	76	4	12	16 (17.4%)
LightGBM	92	81	3	8	11 (12.0%)

Table 7 lists the Safety oriented descriptive error profile. Error direction matters more than total error count for an interface that shapes escalation language. LightGBM and XGBoost each classified 8 observed High records as Low, versus 12 for CatBoost, LightGBM classified 3 High records as Moderate, versus 4 for both other models. In total, 11 of 92 observed High records were not predicted as High by LightGBM, versus 12 for XGBoost and 16 for CatBoost.

The Moderate class requires careful interpretation, since it corresponds to non-disclosure or ambiguity rather than a validated intermediate severity level. LightGBM produced 9 Moderate-to-Low errors, versus 10 for XGBoost and 15 for CatBoost.

High-category precision was 90.0% and recall 88.0% (81 of 92 records correctly identified), Low and Moderate recalls were 90.1% and 85.1%. Moderate had the lowest recall and F1-score, but this should not be read as weaker detection of “moderate PPD”: the source class is a non-disclosure response, not a validated symptom-severity level.

4.4 Reliability and Scope of Statistical Inference

A previously reported bootstrap confidence interval for High-category recall could not be independently verified because the number of resamples, the resampling unit, and the interval-construction method were not documented. Therefore, no confidence interval is reported for this metric. Statistical superiority is likewise not claimed for any of the evaluated algorithms, as a valid paired comparison would require record-level prediction outputs, which were not available.

These results are internal descriptive estimates from one stratified split, cross-validation supported model selection within development data but does not substitute for external validation. This elaborates on what further evidence clinical translation would require.

5. Discussion

This study compared three gradient-boosting classifiers under a common, leakage-controlled pipeline and integrated the best-performing model with a deterministic conversational layer. LightGBM achieved the highest held-out accuracy, weighted F1-score,

balanced accuracy, and macro F1-score; class-specific performance was broadly consistent across categories, though Moderate had the lowest recall and F1-score. More importantly, 11 of 92 observed High records were not classified as High, including 8 assigned to Low, errors that preclude any claim the system is safe for clinical deployment. In the prototype, the classifier's job ends at producing a probability vector and operational category; a rule engine determines what the user actually sees. Keeping these functions apart keeps the recommendation layer auditable and guarantees identical operational outputs always trigger the same predefined guidance, a meaningful engineering safeguard even though it cannot alone establish clinical safety.

Three structural properties, considered jointly, offer the most parsimonious account available. First, leaf-wise growth allocates additional splitting capacity to whichever region of the feature space the ensemble currently fits worst, which on a three-class problem with unevenly sized and unevenly separable classes would be expected to help more than level-wise growth, which expands all leaves uniformly regardless of local error. Second, gradient-based one-side sampling concentrates each iteration's information on the highest-gradient instances, which should help most exactly on the costly High-Low and High-Moderate boundaries examined. Third, CatBoost's ordered-boosting mechanism solves a target-leakage problem this dataset poses only weakly, given every predictor's low cardinality (at most five levels), so its computational overhead is paid without a correspondingly large benefit, a pattern consistent with, though not proof of, its comparatively lower rank in Table 5.

This account is a plausible, literature-grounded reading, not a controlled ablation; settling it conclusively would mean refitting LightGBM with leaf-wise growth disabled, or CatBoost with ordered boosting swapped for plain boosting, on this exact dataset. Recent work applying interpretable LightGBM models to cardiovascular data reports a comparable pattern, leaf-wise growth outperforming level-wise alternatives on datasets with few low-cardinality predictors, which fits but does not substitute for the ablation identified as necessary future work in Section 6.

The LightGBM accuracy of 88.4% should be interpreted only within the present dataset. A 2025 EHR study on a substantially larger cohort (12,284 deliveries)

with a clinically constructed outcome reported materially different discrimination, illustrating how strongly outcome construction and cohort composition shape performance independent of algorithm choice. A trimester-resolved Dutch cohort similarly found discrimination depended on how early in pregnancy the outcome was measured, complicating direct comparison with the present self-reported outcome. A 2025 meta-analysis confirmed between-study variance is driven more by outcome definition and predictor availability than algorithm family, supporting the decision to restrict comparison to the internal ranking.

This distinction also explains why the present results should not be read as evidence that LightGBM is superior to a validated screening scale such as the EPDS: the model is evaluated against the survey label available in this dataset, not against clinical assessment or diagnostic interview. Strong internal discrimination is therefore evidence of reproducible classification, not diagnostic equivalence, a distinction reinforced by a 2025 psychometric evaluation showing even the EPDS requires factor-level scrutiny before its items can be treated as equivalent ordinal severity indicators.

External validation, calibration, and standardized reporting remain recurring weaknesses across postpartum AI research generally, and the present prototype should accordingly be read as a transparent internal benchmark and testable software architecture, not evidence of clinical readiness. Of the three interface categories, Moderate rests on the shakiest conceptual ground: the underlying response, Not interested to say, signals non-disclosure or uncertainty rather than a middling level of depression, suicidality, or distress. High deserves similarly careful handling: a Yes response marks self-reported suicide-attempt status, but the record says nothing about recency, intent, planning, or immediacy, exactly what a single categorical item cannot capture [45]. Model confidence, outcome prevalence, and clinical urgency are separate considerations: a class may carry a high predicted probability while poorly calibrated, while a lower-probability event may still require urgent action. Recent calibration-focused audit of a deployed malnutrition-prediction model found strong aggregate discrimination concealed material miscalibration across demographic subgroups, invisible from discrimination metrics alone, and a 2025 study of algorithmic fairness drift found subgroup-level performance gaps can widen over time even when aggregate performance holds stable [46].

Moving toward clinical translation would take evidence well beyond this internal evaluation: a clinically defined outcome with prospective relevance (ideally linked to a validated instrument such as the EPDS or clinician assessment), external validation across independent cohorts following the TRIPOD+AI guideline [47], calibration and prevalence-sensitive measures

alongside discrimination metrics, since the two are known to diverge, and evaluation across clinically relevant subgroups repeated periodically, since subgroup-level fairness can drift over time without any change to the underlying model.

Since Shapley-value explanations align naturally with the additive-tree structure all three algorithms share, a clinically reviewed successor should report per-prediction and global feature attributions rather than treat the classifier as closed-box. A recent meta-analysis across 62 clinical decision-support studies found Shapley-value explanations consistently improved clinician trust and comprehension relative to unexplained outputs, a directly transferable argument for adding this layer before any clinician-facing pilot [48].

The interface and recommendation rules would require review by maternal mental-health professionals and evaluation with representative users, covering how users interpret the operational labels, accessibility, and reactions to escalation messages, an evaluation burden recent chatbot reviews find is met inconsistently even for far more mature systems. Data-governance procedures, informed consent, encryption, retention, audit trails, and locally appropriate emergency-response pathways would also need specifying before deployment, in step with regulatory analysis calling for standardized data-security disclosures for AI-based mental-health tools. None of this is optional; each is a precondition for safe clinical use.

6. Conclusion

This study developed a rule-based chatbot prototype built on a leakage-controlled, mechanism-linked comparison of XGBoost, CatBoost, and LightGBM on a publicly available postpartum survey dataset, predicting a three-category self-reported response on suicide-attempt status. LightGBM produced the highest accuracy (88.4%) and weighted F1-score (88.4%), consistent with its leaf-wise growth strategy and the low-cardinality categorical predictor set, and inconsistent with a large structural advantage for CatBoost's leakage protection on data posing little leakage risk. Despite this, 11 of 92 High-category records were classified into other categories, showing aggregate performance alone cannot establish clinical safety. The system was designed to keep predictive output and user-facing communication separate: the classifier generates class probabilities and selects an operational category, and the chatbot applies fixed rules for guidance and escalation, keeping Low, Moderate, and High as operational descriptors rather than clinically validated risk stages. These findings establish the technical feasibility of a shared, leakage-controlled boosting comparison feeding a deterministic recommendation layer, with a mechanism-linked account of why one architecture outperformed the others here. They do not establish clinical validity, calibrated risk estimation, or

deployment readiness, which would take clinically defined outcomes, independent and multicentre validation, prospective evaluation, subgroup-specific assessment, the computational-benchmarking study named in Table 4, clinician-reviewed escalation procedures, and formal usability, safety, and data-governance testing. Until then, the system should remain a research prototype, not a tool for diagnosis, emergency risk assessment, or treatment decisions.

References

- [1] Y. Yildiz, A. Kalayci, Gradient Boosting Decision Trees on Medical Diagnosis over Tabular Data. IEEE International Conference on AI and Data Analytics (ICAD), IEEE, USA. <https://doi.org/10.1109/ICAD65464.2025.11114069>
- [2] R. Zhang, Y. Liu, Z. Zhang, R. Luo, B. Lv. Interpretable Machine Learning Model for Predicting Postpartum Depression: Retrospective Study. JMIR Medical Informatics, 13, (2025) e58649. <https://doi.org/10.2196/58649>
- [3] W. Qi, Y. Wang, Y. Wang, S. Huang, C. Li, H. Jin, J. Zuo, X. Cui, Z. Wei, Q. Guo, J. Hu. Prediction of Postpartum Depression in Women: Development and Validation of Multiple Machine Learning Models. Journal of Translational Medicine, 23, (2025) 291. <https://doi.org/10.1186/s12967-025-06289-6>
- [4] Z. Ma, M. Horvath, D.M. Stamilio, K. Sekyere, M.N. Gurcan. Building a Machine Learning Model to Predict Postpartum Depression from Electronic Health Records in a Tertiary Care Setting. Journal of Clinical Medicine, 14, (2025) 6644. <https://doi.org/10.3390/jcm14186644>
- [5] X. Huang, L. Zhang, C. Zhang, J. Li, C. Li. Postpartum Depression Risk Prediction Using Explainable Machine Learning Algorithms. Frontiers of Medicine, 12, (2025) 1565374. <http://dx.doi.org/10.3389/fmed.2025.1565374>
- [6] H. Kim. Predictive Analysis of Postpartum Depression Using Machine Learning. Healthcare, 13, (2025) 897. <https://doi.org/10.3390/healthcare13080897>
- [7] C. Sona Ajay, S. Juliet, A. Alex. Machine Learning and Survival Analysis Models for Postpartum Depression: A Comprehensive Risk Factor Analysis. In Proceedings of the 2024 10th International Conference on Advanced Computing and Communication Systems (ICACCS), India. <https://doi.org/10.1109/ICACCS60874.2024.10717207>
- [8] S. García-Méndez, F. de Arriba-Pérez. Detecting and Explaining Postpartum Depression in Real-Time with Generative Artificial Intelligence. Applied Artificial Intelligence, 39, (2025) 2515063. <https://doi.org/10.1080/08839514.2025.2515063>
- [9] V.S.F. Gonçalves, V.R. de Carvalho. A Review of Interpretability Methods for Gradient Boosting Decision Trees. Journal of the Brazilian Computer Society, 31, (2025) 639–653. <https://doi.org/10.5753/jbcs.2025.5324>
- [10] H. Semmelrock, T. Ross-Hellauer, S. Kopeinik, D. Theiler, A. Haberl, S. Thalmann, D. Kowald. Reproducibility in Machine-Learning-Based Research: Overview, Barriers, and Drivers. AI Magazine, 46(2), (2025) e70002. <https://doi.org/10.1002/aaai.70002>
- [11] M. Alkhateeb, A. Nayeem, A. Ahmed, M. Alsahli, J. Sheikh, A. Abd-Alrazaq. AI for Detecting and Predicting Postpartum Depression: Scoping Review. Journal of Medical Internet Research, 28, (2026) e77376. <https://doi.org/10.2196/77376>
- [12] U.S. Anaduaka, A.O. Oladosu, S. Katsande, C.S. Frempong, S. Awuku-Amador. Leveraging Artificial Intelligence in the Prediction, Diagnosis and Treatment of Depression and Anxiety among Perinatal Women in Low- and Middle-Income Countries: A Systematic Review. BMJ Mental Health, 28, (2025) e301445. <https://doi.org/10.1136/bmjment-2024-301445>
- [13] J. Xia, C. Chen, X. Lu, T. Zhang, T. Wang, Q. Wang, Q. Zhou. Artificial Intelligence-Oriented Predictive Model for the Risk of Postpartum Depression: A Systematic Review. Front. Public Health, 13, (2025) 1631705. <https://doi.org/10.3389/fpubh.2025.1631705>
- [14] L. Sibbald, M.I. van den Heuvel, M.R. Haas, C.J. van Lissa, H.J.A. van Bakel, J. Jongerling, L. P. Hulsbosch, L. Muskens, M. G. B. M. Boekhorst, I. Schwabe. Identifying Prenatal Risk Factors of Postpartum Depression with Machine Learning. Scientific Reports, 15, (2025) 34610. <https://doi.org/10.1038/s41598-025-18204-6>
- [15] Y. Xie, H. Zheng, W. Gan, C. Su, M. Shams, J. Yang. The Performance of Machine Learning Models in Predicting Postpartum Depression: A Meta-Analysis and Systematic Review. Journal of Reproductive and Infant Psychology, 43(5), (2025) 1093–1110. <https://doi.org/10.1080/02646838.2025.2517103>
- [16] L. Ou, Q. Shen, M. Xiao, W. Wang, T. He, B. Wang. Prevalence of Co-Morbid Anxiety and Depression in Pregnancy and Postpartum: A Systematic Review and Meta-Analysis. Psychological Medicine, 55, (2025) e93. <https://doi.org/10.1017/S0033291725000601>
- [17] K. Zivin, C. Zhong, A. Rodríguez-Putnam, E. Spring, Q. Cai, A. Miller, L. J. Johns, V. A. Kalesnikava, A. Courant, B. Mezuk. Suicide Mortality during the Perinatal Period. JAMA Netw Open, 7, (2024) e2418887. <https://doi.org/10.1001/jamanetworkopen.2024.18887>

- [18] H. Yu, Q. Shen, E. Bränn, Y. Yang, A.S. Oberg, U.A. Valdimarsdóttir, D. Lu. Perinatal Depression and Risk of Suicidal Behavior. *JAMA Netw. Open*, 7, (2024) e2350897. <https://doi.org/10.1001/jamanetworkopen.2023.50897>
- [19] S. Goldman-Mellor, M. Olfson, A. Gemmill, C. Margerison. Incidence and Risk Factors for Suicide Attempt during Pregnancy and the Postpartum Period. *The Journal of Clinical Psychiatry*, 86, (2025) 24m15633. <https://doi.org/10.4088/JCP.24m15633>
- [20] J. M. Martínez-Ramírez, R.A. Peinado-Molina, A. Hernández-Martínez, J.M. Martínez-Galiano. Predictive Models of Suicidal Ideation Risk in Perinatal-Stage Women Based on Sociodemographic and Clinical Data. *Frontiers in Artificial Intelligence Medicine and Public Health*, 9, (2026) 1774453. <https://doi.org/10.3389/frai.2026.1774453>
- [21] Y. Hua, S. Siddals, Z. Ma, I. Galatzer-Levy, W. Xia, C. Hau, H. Na, M. Flathers, J. Linardon, C. Ayubcha, J. Torous. Charting the Evolution of Artificial Intelligence Mental Health Chatbots from Rule-Based Systems to Large Language Models: A Systematic Review. *World Psychiatry*, 24(3), (2025) 383–394. <https://doi.org/10.1002/wps.21352>
- [22] A. Arnaiz-Rodriguez, M. Baidal, E. Derner, J. Layton Annable, M. Ball, M. Ince, E. Perez Vallejos, N. Oliver, Between Help and Harm: An Evaluation Study of Mental Health Crisis Handling by Large Language Models. *JMIR Mental Health*, 13(1), (2026) e88435. <https://doi.org/10.2196/88435>
- [23] A.-R. Wrightson-Hester, G. Anderson, J. Dunstan, P. M. McEvoy, C.J. Sutton, B. Myers, S.J. Egan, S.J. Tai, M. Johnston-Hollitt, W. Chen, T. Gedeon, J.C. Moullin, W. Mansell, A Rule-Based Conversational Agent for Mental Health and Well-Being in Young People: Formative Case Series during the Rise of Generative AI. *JMIR Formative Research*, 9, (2025) e69841. <https://doi.org/10.2196/69841>
- [24] G.S. Collins, K.G.M. Moons, P. Dhiman, R.D. Riley, A.L. Beam, B. van Calster, M. Ghassemi, X. Liu, J.B. Reitsma, M. van Smeden, A.L. Boulesteix, TRIPOD+AI Statement: Updated Guidance for Reporting Clinical Prediction Models That Use Regression or Machine Learning Methods. *BMJ*, 385, (2024) e078378. <https://doi.org/10.1136/bmj-2023-078378>
- [25] C. Meaney, X. Wang, J. Guan, T.A. Stukel, Comparison of Methods for Tuning Machine Learning Model Hyper-Parameters: with Application to Predicting High-Need High-Cost Health Care Users. *BMC Medical Research Methodology*, 25(1), (2025) 134. <https://doi.org/10.1186/s12874-025-02561-x>
- [26] V. Gupta, S. Tripathi, D. Singh, A. Bansal, (2024) Predictive Algorithms for Early Postpartum Depression Detection: CatBoost vs. LightGBM. In *Proceedings of the 2024 11th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO)*, IEEE, Noida, India. <https://doi.org/10.1109/ICRITO61523.2024.10522300>
- [27] N. Hagatulah, E. Bränn, A.S. Oberg, U.A. Valdimarsdóttir, D. Lu, Q. Shen, Perinatal Depression and Risk of Mortality: Nationwide, Register Based Study in Sweden. *BMJ*, 384, (2024) e075462. <https://doi.org/10.1136/bmj-2023-075462>
- [28] Z. Kaminsky, R. J. McQuaid, K. G. Hellemans, Z. R. Patterson, M. Saad, R. L. Gabrys, T. Kendzerska, A. Abizaid, R. Robillard, Machine Learning-Based Suicide Risk Prediction Model for Suicidal Trajectory on Social Media Following Suicidal Mentions: Independent Algorithm Validation. *Journal of Medical Internet Research*, 26, (2024) e49927. <https://doi.org/10.2196/49927>
- [29] L. Liu, Z. Li, Y. Hu, C. Li, S. He, S. Zhang, J. Gao, H. Zhu, G. Huang, Predictive Performance of Machine Learning for Suicide in Adolescents: Systematic Review and Meta-Analysis. *Journal of Medical Internet Research*, 27, (2025) e73052. <https://doi.org/10.2196/73052>
- [30] R. Li, Y. Yue, X. Gu, L. Xiong, M. Luo, L. Li, Risk Prediction Models for Adolescent Suicide: A Systematic Review and Meta-Analysis. *Psychiatry Research*, 347, (2025) 116405. <https://doi.org/10.1016/j.psychres.2025.116405>
- [31] L. Deng, K. Lu, H. Hu, An Interpretable LightGBM Model for Predicting Coronary Heart Disease: Enhancing Clinical Decision-Making with Machine Learning. *PLoS one*, 20(9), (2025) e0330377. <https://doi.org/10.1371/journal.pone.0330377>
- [32] J. Xu, T. Chen, X. Fang, L. Xia, X. Pan, Prediction Model of Pressure Injury Occurrence in Diabetic Patients during ICU Hospitalization: XGBoost Machine Learning Model Can Be Interpreted Based on SHAP. *Intensive and Critical Care Nursing*, 83, (2024) 103715. <https://doi.org/10.1016/j.iccn.2024.103715>
- [33] M.R. Khurshid, S. Manzoor, T. Sadiq, L. Hussain, M.S. Khan, A.K. Dutta, Unveiling Diabetes Onset: Optimized XGBoost with Bayesian Optimization for Enhanced Prediction. *PLoS ONE*, 20, (2025) e0310218. <https://doi.org/10.1371/journal.pone.0310218>
- [34] C. Yang, E.A. Fridgeirsson, J.A. Kors, J.M. Reps, P.R. Rijnbeek, Impact of Random Oversampling and Random Undersampling on the Performance of Prediction Models Developed Using Observational Health Data. *Journal of Big Data*, 11(1), (2024) 7. <https://doi.org/10.1186/s40537-023-00857-7>

- [35] J. Digitale, D. Franzon, M.J. Pletcher, C.E. McCulloch, E.D. Gennatas, Methods for Addressing Missingness in Electronic Health Record Data for Clinical Prediction Models: Comparative Evaluation. *JMIR Medical Informatics*, 13(1), (2025) e79307. <https://doi.org/10.2196/79307>
- [36] R. Sanjeewa, R. Iyer, P. Apputhurai, N. Wickramasinghe, D. Meyer, Empathic Conversational Agent Platform Designs and Their Evaluation in the Context of Mental Health: Systematic Review. *JMIR Mental Health*, 11, (2024) e58974. <https://doi.org/10.2196/58974>
- [37] W. Pichowicz, M. Kotas, P. Piotrowski, Performance of Mental Health Chatbot Agents in Detecting and Managing Suicidal Ideation. *Scientific Reports*, 15(1), (2025) 31652. <https://doi.org/10.1038/s41598-025-17242-4>
- [38] A. Palmer, D. Schwan, Digital Mental Health Tools and AI Therapy Chatbots: A Balanced Approach to Regulation. *Hastings Center Report*, 55(3), (2025) 15–29. <https://doi.org/10.1002/hast.4979>
- [39] Y. Jin, J. Liu, P. Li, B. Wang, Y. Yan, H. Zhang, C. Ni, J. Wang, Y. Li, Y. Bu, Y. Wang, The Applications of Large Language Models in Mental Health: Scoping Review. *Journal of Medical Internet Research*, 27, (2025) e69284. <https://doi.org/10.2196/69284>
- [40] Y. Hua, H. Na, Z. Li, et al., A Scoping Review of Large Language Models for Generative Tasks in Mental Health Care. *npj digital medicine*, 8, (2025) 230. <https://doi.org/10.1038/s41746-025-01611-4>
- [41] Stolltho, Postpartum Depression Classification [Kaggle Notebook and Associated Input Data]. Kaggle, n.d. Available online: <https://www.kaggle.com/code/stolltho/postpartum-depression-classification/input>
- [42] E. Zain, Y. Watanabe, S. Takabayashi, L. Por, S. Fujita, S. Moriyama, A. Honma, N. Fukui, S. Boku, Psychometric Evaluation of the Japanese Edinburgh Postnatal Depression Scale for Screening Postpartum Anxiety. *Frontiers in Psychiatry*, 16, (2025) 1659497. <https://doi.org/10.3389/fpsyt.2025.1659497>
- [43] L. Liou, E. Scott, P. Parchure, Y. Ouyang, N. Egorova, R. Freeman, I. S. Hofer, G. N. Nadkarni, P. Timsina, A. Kia, M. A. Levin, Assessing Calibration and Bias of a Deployed Machine Learning Malnutrition Prediction Model within a Large Healthcare System. *npj Digit. Med.*, 7, (2024) 141. <https://doi.org/10.1038/s41746-024-01141-5>
- [44] I. Partheniadis, P. Talimtzis, A. Nikolakopoulou, A.-B. Haidich, Machine Learning-Based COVID-19 Prognostic Models Lag Behind in Reporting Quality: Findings from a TRIPOD/TRIPOD+AI Systematic Review. *Diagn. Progn. Res.*, 10, (2026) 3. <https://doi.org/10.1186/s41512-026-00218-x>
- [45] A. J. Millner, M. D. Lee, M. K. Nock, Single-Item Measurement of Suicidal Behaviors: Validity and Consequences of Misclassification. *PLoS ONE*, 10, (2015) e0141606. <https://doi.org/10.1371/journal.pone.0141606>
- [46] S.E. Davis, C. Dorn, D.J. Park, M.E. Matheny, Emerging Algorithmic Bias: Fairness Drift as the Next Dimension of Model Maintenance and Sustainability. *Journal of the American Medical Informatics Association*, 32, (2025) 845–854. <https://doi.org/10.1093/jamia/ocaf039>
- [47] K.G.M. Moons, J.A.A. Damen, T. Kaul, L. Hooft, C. Andaur Navarro, P. Dhiman, A.L. Beam, B. Van Calster, L.A. Celi, S. Denaxas, A.K. Denniston, M. Ghassemi, G. Heinze, A.P. Kengne, L. Maier-Hein, X. Liu, P. Logullo, M.D. McCradden, N. Liu, L. Oakden-Rayner, K. Singh, D.S. Ting, L. Wynants, B. Yang, J.B. Reitsma, R.D. Riley, G.S. Collins, M.V. Smeden, PROBAST+AI: An Updated Quality, Risk of Bias, and Applicability Assessment Tool for Prediction Models Using Regression or Artificial Intelligence Methods. *BMJ*, 388, (2025) e082505. <https://doi.org/10.1136/bmj-2024-082505>
- [48] Q. Abbas, W. Jeong, S.W. Lee, Explainable AI in Clinical Decision Support Systems: A Meta-Analysis of Methods, Applications, and Usability Challenges. *Healthcare*, 13(17), (2025) 2154. <https://doi.org/10.3390/healthcare13172154>

Acknowledgments

The authors thank the contributors who made the postpartum survey data publicly available. During the preparation of this manuscript, a generative AI-based writing assistant was used solely to support drafting, structural organization, language refinement, and proofreading, as detailed in the Declaration on the Use of Generative AI below, the assistant was not used to design the study, analyze data, generate results, or produce the scientific conclusions reported herein, which remain the sole responsibility of the named authors.

Authors Contribution Statement

T. Om Bhavani Shankar: Conceptualization, Methodology and Investigation. R. Prathiba: Data Curation, Software, and Validation. V. Keerthi Vasan: Formal Analysis and Visualization. T. Sivakumar: Supervision and Writing – Original Draft. K. Kanagaraj: Writing – Review & Editing. K. Sudhakar: Validation, Writing – Review & Editing. All the author read and approved the final version of the manuscript.

Competing Interests

The authors declare that there are no conflicts of interest regarding the publication of this manuscript.

Funding

The authors declare that no funds, grants or any other support were received during the preparation of this manuscript.

Data Availability

The dataset is publicly available through the Kaggle postpartum-depression repository cited in the manuscript. To support reproducibility, the analysis code should be archived alongside the exact downloaded dataset file, the retrieval date, and the dataset version used.

Declaration on the Use of Generative AI

During preparation of this manuscript, a generative AI-based language assistant was used to support (i) literature search and verification of externally published, peer-reviewed sources used to contextualize the study, and (ii) proofreading for clarity and consistency.

Has this article screened for similarity?

Yes

About the License

© The Author(s) 2026. The text of this article is open access and licensed under a Creative Commons Attribution 4.0 International License.