Discover Artificial Intelligence

RESEARCH Open Access

# An uncertainty-aware ensemble of support vector machine and Gaussian process models for maternal mental health risk prediction using psychosocial and clinical risk factors

Shallon Ahimbisibwe[1,3*], Emmanuel Ahishakiye[1], Samuel Maling[2], Simon Kawuma[3], Richard Ntwari[3], Boaz Twinamasiko[4] and Fred Kaggwa[3]

*Correspondence:
Shallon Ahimbisibwe
ahimbisibweshallon@gmail.com
[1]School of Computing and Information Science, Kyambogo University, Kampala, Uganda
[2]Faculty of Medicine, Mbarara University of Science & Technology, Mbarara, Uganda
[3]Faculty of Computing and Informatics, Mbarara University of Science & Technology, Mbarara, Uganda
[4]Kawempe National Referral Hospital, Kampala, Uganda

## Abstract

Postpartum depression (PPD) is a major maternal mental health condition that can adversely affect mothers and infants if not identified and managed early. This study developed and evaluated an uncertainty-aware ensemble framework integrating Support Vector Machine (SVM) and Gaussian Process (GP) models for maternal mental health risk prediction using psychosocial, demographic, socioeconomic, obstetric and clinical factors. A pooled dataset comprising 14,008 retrospective maternal records from Kawempe National Referral Hospital and Mukono General Hospital in Uganda and the publicly available PERI_DEP dataset was analysed. Preprocessing included missing-value imputation, categorical encoding, feature engineering, Mutual Information-based feature selection, Kernel Principal Component Analysis and Synthetic Minority Oversampling Technique-based class balancing. All preprocessing and oversampling procedures were restricted to the training data. The Hybrid SVM–GP framework was compared with Logistic Regression, Random Forest and CatBoost using stratified five-fold cross-validation and a held-out internal test set. To assess potential target leakage, analyses were conducted using both the complete feature set and a leakage-free feature set that excluded individual Edinburgh Postnatal Depression Scale items and directly related depression-symptom variables. In the full-feature analysis, Random Forest achieved the strongest discrimination, with an accuracy of 97.75% and a Receiver Operating Characteristic Area Under the Curve of 0.996. The Hybrid SVM–GP achieved an accuracy of 93.54%, a Receiver Operating Characteristic Area Under the Curve of 0.982 and the lowest Expected Calibration Error (ECE) of 0.015, whereas Random Forest achieved lower Brier score and log loss. The Hybrid SVM-GP also provided explicit case-level uncertainty information based on the normalised predictive entropy of the fused probability, with GP posterior variance providing a complementary measure of uncertainty in the latent prediction. These findings indicate that the Hybrid SVM–GP may complement models such as Random Forest by identifying potentially ambiguous predictions requiring further assessment, rather than providing universal performance superiority. SHapley

© The Author(s) 2026. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

Additive exPlanations from the leakage-free analysis identified family relationships, socioeconomic conditions, social support and reproductive history as influential contributors to model predictions. However, the retrospective design and absence of independent external validation limit immediate clinical application. Prospective external validation and clinical-utility assessment are required before routine maternal healthcare deployment.

**Keywords** Postpartum depression, Maternal mental health prediction, Support vector machine, Gaussian process classifier, Predictive uncertainty, Explainable artificial intelligence

## 1 Introduction

Postpartum depression (PPD) is a major maternal mental health problem that can adversely affect mothers, infants and families if it is not identified and managed early. It is commonly associated with persistent sadness, anxiety, emotional instability, fatigue, sleep disturbance, hopelessness and difficulty bonding with the newborn. In addition to its psychological effects on mothers, PPD may affect infant cognitive and emotional development, family relationships, household well-being and the long-term quality of life of mothers [1, 2]. Globally, PPD is estimated to affect approximately 10–20% of mothers, with higher prevalence frequently reported in low- and middle-income countries because of limited access to mental health services, socioeconomic stress, cultural pressures and inadequate support systems [1, 2]. Maternal mental health disorders are consequently underdiagnosed and undertreated in many resource-constrained settings. Early identification of mothers at elevated risk therefore remains an important public health priority for improving maternal and child health outcomes.

Screening for PPD commonly relies on standardised psychological instruments based on self-reported information, including the Edinburgh Postnatal Depression Scale (EPDS) [3]. These instruments are practical and widely used to identify depressive symptoms in postpartum women. However, screening based primarily on symptoms reported at the time of assessment may provide limited information about antecedent risk and may be influenced by emotional state, recall bias, social stigma and cultural perceptions of mental health. These limitations have encouraged the development of data-driven approaches that can learn complex relationships among demographic, psychosocial, socioeconomic, obstetric and clinical factors and support earlier maternal mental health risk assessment.

Recent advances in machine learning and artificial intelligence have created new opportunities for predictive healthcare analytics, risk stratification and clinical decision support. Machine-learning algorithms can analyse heterogeneous healthcare data and model nonlinear relationships among demographic, psychosocial, behavioural and clinical variables. Several studies have applied logistic regression, decision trees, random forests, support vector machines, boosting algorithms and neural networks to PPD prediction [4–9, 14]. Although these studies have reported promising predictive performance, many have focused mainly on classification accuracy, with less attention to probability calibration, case-level uncertainty, explainability and the reliability of predictions. These considerations are important in maternal mental health screening because overconfident or poorly calibrated predictions may contribute to missed referrals, unnecessary interventions or inappropriate allocation of limited healthcare resources.

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

Uncertainty-aware machine learning is therefore relevant to maternal mental health prediction because it can provide both a risk probability and information about the confidence of individual predictions. This is particularly important when a model encounters cases that are close to the classification boundary or poorly represented in the development data. Support Vector Machines (SVMs) and Gaussian Processes (GPs) offer complementary capabilities for this purpose [10, 11]. An SVM can provide a robust nonlinear discriminative decision function, particularly when relationships among predictors are complex. However, its decision score is not inherently probabilistic and does not directly provide case-level uncertainty. A GP classifier, in contrast, supports probabilistic inference and provides uncertainty information associated with the latent prediction, although its results may depend on covariance-kernel specification, hyperparameter selection and computational resources. The rationale for combining the two models is therefore to integrate their complementary information: the SVM contributes nonlinear discrimination, while the GP contributes probabilistic inference and uncertainty-aware information. The hybrid framework is not assumed to be universally superior to alternative models; rather, it is intended to examine whether uncertainty information can complement conventional predictive performance measures in maternal mental health risk assessment.

Existing PPD prediction research is also limited by the use of datasets obtained from single healthcare institutions or geographically homogeneous populations [12]. Maternal mental health outcomes are influenced by cultural expectations, social support systems, economic conditions, family relationships, access to healthcare and psychosocial stressors, all of which may vary across populations and healthcare systems [13]. Models developed using homogeneous data may therefore have limited transportability to other maternal populations. The use of heterogeneous maternal health data can provide a broader basis for examining model behaviour across different settings, although pooling data alone does not establish external or prospective generalisability. There is consequently a need for prediction frameworks that evaluate discrimination, probability accuracy, calibration, uncertainty and explainability while explicitly acknowledging the limitations of internal validation.

This study addresses these gaps by developing and evaluating an uncertainty-aware ensemble framework that combines SVM and GP models for maternal mental health risk prediction using psychosocial, demographic, socioeconomic, obstetric and clinical factors. The study uses a pooled dataset comprising hospital records from Kawempe National Referral Hospital and Mukono General Hospital in Uganda and the publicly available PERI_DEP dataset. The combined dataset includes a broader range of maternal and psychosocial factors than would be available from a single source, while differences in population characteristics, data-collection procedures and healthcare settings are considered when interpreting the findings. The modelling pipeline includes missing-value imputation, categorical encoding, feature engineering, Mutual Information-based feature selection, Kernel Principal Component Analysis (Kernel PCA), Synthetic Minority Oversampling Technique (SMOTE) and hyperparameter optimisation using randomised search. Stratified cross-validation is used to support model development and reduce the risk of overfitting.

The study differs from many previous PPD prediction studies by evaluating the proposed framework using complementary discrimination, probability-accuracy,

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

calibration, uncertainty and explainability measures. Accuracy, precision, recall, F1-score, ROC-AUC and PR-AUC are used to assess classification and discrimination, while Brier score, log loss and Expected Calibration Error (ECE) are used to assess probabilistic performance. Case-level uncertainty is quantified using the normalised predictive entropy of the final fused probability, while GP posterior variance is treated as a complementary measure of uncertainty in the latent prediction. SHapley Additive exPlanations (SHAP) are used to examine the contribution of psychosocial and clinical predictors to model outputs. Because individual EPDS items and directly related depression-symptom variables may introduce target leakage, both a full-feature sensitivity analysis and a leakage-free analysis are conducted.

The central contribution of this study is a framework for examining whether the complementary combination of SVM-based nonlinear discrimination and GP-based probabilistic uncertainty information can support a more comprehensive assessment of maternal mental health risk. The contribution is therefore methodological rather than a claim that the Hybrid SVM–GP will outperform every alternative model. Its potential value lies in providing risk probabilities together with case-level uncertainty information that may help distinguish more confident predictions from ambiguous cases requiring additional assessment. The findings are interpreted alongside the performance of Random Forest, CatBoost and Logistic Regression, and the absence of independent external validation and prospective clinical-utility assessment is explicitly recognised.

### 1.1 Paper organization

The remainder of this paper is organized as follows. Section 2 reviews related literature on postpartum depression and the application of machine learning methods for mental health prediction. Section 3 describes the materials and methods used in this study, including the dataset, preprocessing procedures, feature selection, model development, and evaluation metrics. Section 4 presents the experimental results and model performance analysis. Section 5 discusses the findings and their implications for postpartum depression prediction. Finally, Sect. 6 concludes the paper and outlines directions for future research.

## 2 Related literature

### 2.1 Kernel methods in medical health prediction

Kernel methods are widely used in healthcare prediction because they can represent nonlinear relationships in high-dimensional and complex biomedical data. Kernel functions implicitly map observations into higher-dimensional feature spaces without requiring the transformation to be computed explicitly [15]. This property, known as the kernel trick, is useful for analysing clinical data containing nonlinear interactions among demographic, psychosocial and biological variables [16]. Support Vector Machines (SVMs) and Gaussian Processes (GPs) are important kernel-based approaches used in healthcare prediction [17].

SVMs are effective discriminative models with strong generalisation performance, particularly in high-dimensional classification problems [18]. They have been applied in cancer detection, diabetes diagnosis, seizure classification and mental-health assessment [19, 20]. However, a conventional SVM primarily produces a decision score rather than an inherently probabilistic prediction. Its output must therefore be calibrated before it

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

**Table 1** Summary of machine-learning studies for predicting postpartum or perinatal depression

| Authors | Year and setting | Sample size | Models used | Validation type | Key results | Limitations and remaining gaps |
|---|---|---|---|---|---|---|
| Shin et al. [25] | 2020, United States Pregnancy Risk Assessment Monitoring System, 2012–2013 | 28,755, including 3,339 PPD cases | RF, stochastic gradient boosting, SVM, decision tree, naïve Bayes, kNN, LR and ANN | Internal, 10-fold cross-validation | RF achieved the best accuracy of 0.791 and AUC of 0.884; SVM achieved AUC of 0.864 | Based on retrospective self-reported survey data; no independent external validation, probability calibration or uncertainty quantification |
| Hochman et al. [26] | 2021, nationwide electronic health records from Israel, 2008–2015 | 214,359 births, including 4,104 PPD cases | Gradient-boosted decision trees | Internal holdout validation within the same national health-maintenance organisation | Validation AUC was 0.712, sensitivity 0.349 and specificity 0.905 at the 90th-percentile risk threshold | Outcome was based on a recorded depression diagnosis or antidepressant prescription; modest sensitivity; validation remained within the same healthcare database; no independent external validation or explicit uncertainty analysis |
| Andersson et al. [27] | 2021, BASIC prospective cohort, Uppsala University Hospital, Sweden, 2009–2018 | 4,313 women | Extremely Randomized Trees and other regression, tree, ensemble and probabilistic classifiers | Internal model validation; no independent external cohort | Extremely Randomized Trees achieved accuracy of 73%, sensitivity of 72%, specificity of 75% and AUC of 0.81 | Single Swedish cohort; Scandinavian-born, highly educated and cohabiting women were overrepresented; performance declined among women without previous mental health problems; no external validation or uncertainty quantification |

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

**Table 1** (continued)

| Authors | Year and setting | Sample size | Models used | Validation type | Key results | Limitations and remaining gaps |
|---|---|---|---|---|---|---|
| Zhang et al. [28] | 2025, West China Second University Hospital, China | 2,055 women; 1,358 training and 697 validation participants | XGBoost and other ML models, LASSO feature selection and SHAP | Internal random-split validation | XGBoost achieved AUC of 0.849; important predictors included antepartum depression, maternal age, infant weight and biochemical markers | Retrospective single-centre study; internal validation only; antepartum depression was the dominant predictor; no formal uncertainty quantification or independent external validation. The original claim that the study excluded psychosocial variables should be removed |
| Lilhore et al. [29] | 2024, publicly available online PPD dataset | 1,503 records | Hybrid CNN–BiLSTM with transfer learning; comparison with DT, RF, AdaBoost, LightGBM, XGBoost and CatBoost | Internal evaluation on one online dataset; no independent external validation | Proposed model reported accuracy of 99.6% and precision of 99.7% | Small publicly available dataset with limited evidence of clinical representativeness; exceptionally high performance requires independent verification; computational complexity, limited clinical interpretability and no calibrated uncertainty estimates. The original statement that it required audio or text input was not supported |

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

**Table 1** (continued)

| Authors | Year and setting | Sample size | Models used | Validation type | Key results | Limitations and remaining gaps |
|---|---|---|---|---|---|---|
| Qi et al. [9] | 2025, single-centre prospective study in China, August 2021–August 2022 | 1,138 perinatal women | LR, ANN and other ML models with seven feature-selection methods | Internal cross-validation; no independent external validation | LR achieved AUCs of 0.801 for the prenatal model and 0.858 for the postpartum model; ANN achieved AUCs of 0.787 and 0.844 | Single-centre study; no external validation; postpartum model included culturally specific factors such as mother-in-law care; no formal predictive-uncertainty estimation. The study included SHAP interpretation, so "reduced interpretability" should not be retained as a limitation |
| Nakamura et al. [30] | 2024, perinatal cohort in Japan | 1,416 participants after balancing | Gradient-boosted decision trees and balanced GBDT | Internal model evaluation; no independent external validation reported | Balanced GBDT achieved approximately AUC 0.83 | Single-country dataset; short report provides limited methodological detail; no independent external validation, calibration analysis or uncertainty quantification |
| Kim [31] | 2025, Korean Early Childhood Education and Care Panel, Republic of Korea | 2,570 participants | DT, RF, AdaBoost and LR | Internal model evaluation using one secondary dataset | LR was selected as the most clinically interpretable model, with AUC of approximately 0.74; partner conflict, stress and personal values were influential predictors | Limited predictor set of 13 variables; culturally specific secondary panel data; no external validation, uncertainty quantification or multicountry evaluation. The original table incorrectly implied that RF and AdaBoost were the principal results |

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

**Table 1** (continued)

| Authors | Year and setting | Sample size | Models used | Validation type | Key results | Limitations and remaining gaps |
|---|---|---|---|---|---|---|
| Huang et al. [32] | 2024, prenatal population in the United States | 5,875 pregnant women | Elastic Net and other ML models with SHAP and subgroup-bias assessment | Internal resampling and subgroup evaluation; no independent external validation | Overall ROC-AUC was approximately 0.67; performance differed across racial, ethnic and socioeconomic groups | Predicts prenatal rather than postpartum depression; modest discrimination; subgroup sample imbalance affected performance estimates; bias was assessed but not fully eliminated; no uncertainty-aware prediction or external validation |
| Wakefield and Frasch [33] | 2023, United States NIH Nulliparous Pregnancy Outcomes Study: Monitoring Mothers-to-Be | 10,038 enrolled; 8,454 births analysed, including 338 PPD cases | Four distributed RF models and LR | Internal model validation within the nuMoM2b cohort; no independent external validation | The parsimonious RF model achieved AUC of 0.91 ± 0.02; a simplified model achieved AUC of 0.79 ± 0.05 | Restricted to nulliparous participants; outcome represented depression requiring postpartum treatment; no independent external validation or formal predictive-uncertainty estimates. The original AUC of 0.84 and claim that the study lacked interpretability were incorrect because the study included SHAP analysis |

can be interpreted as a risk probability. Moreover, a standalone SVM does not directly provide case-level uncertainty information, making it difficult to distinguish confident classifications from ambiguous cases.

Gaussian Process models provide a Bayesian non-parametric approach to classification and regression by placing a probability distribution over possible latent functions [11]. In classification, the GP produces posterior class probabilities and uncertainty information associated with the latent prediction. This information can be useful in healthcare applications where cases with uncertain predictions may require additional assessment [21]. Gaussian Processes have been applied in disease-progression modelling, early-warning systems and personalised prediction of clinical measurements [22]. Nevertheless, GP models are sensitive to the choice of covariance kernel and hyperparameters. Exact inference may also become computationally and memory intensive as the number of observations increases.

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

The proposed Hybrid SVM-GP framework was motivated by the complementary capabilities and limitations of these two models. The SVM component provides a nonlinear discriminative decision function, whose output is converted into a probability using sigmoid calibration. In parallel, the GP classifier provides a posterior probability and uncertainty information related to the latent prediction. The two probability outputs are then combined through weighted soft voting to produce a unified maternal mental health risk probability [23]. The purpose of this combination is not to assume that the hybrid model will universally outperform alternative models, but to examine whether GP-derived probabilistic information can complement the discriminative capability of the SVM.

The Hybrid SVM-GP framework therefore provides a way of combining nonlinear discrimination with probabilistic and uncertainty-aware information. In this framework, GP posterior variance represents uncertainty in the latent function [24], whereas the normalised predictive entropy of the final fused probability represents case-level uncertainty in the classification output. These measures should not be treated as identical. The hybrid design also does not eliminate the computational demands, kernel-selection requirements or hyperparameter sensitivity associated with GP inference. Consequently, the model parameters must be carefully optimised, and the computational requirements should be considered when applying the framework to larger datasets. The usefulness of the framework should therefore be assessed using multiple complementary criteria, including discrimination, probability accuracy, calibration and uncertainty behaviour.

### 2.2 Comparative analysis of machine learning models for PPD prediction

As presented in Table 1, machine learning approaches have achieved different levels of discrimination in postpartum and perinatal depression prediction. However, most studies used internal validation within a single dataset, institution or country. Few studies (n = 4) assessed the probability calibration and none of the included studies provided a full integration of calibrated predictive probabilities with explicit uncertainty estimation at a case level. Several studies also relied heavily on prior mental health history or prior depression symptoms and may be limited in their utility for prediction prior to symptom onset. These limitations motivate further exploration of leakage-aware, uncertainty-aware and externally validated prediction frameworks using heterogeneous maternal health data.

### 2.3 Research gaps identified

Despite the growing body of research on postpartum depression and its psychosocial and clinical risk factors, several methodological gaps remain. Existing studies have applied both traditional statistical methods and machine-learning algorithms to identify risk factors and predict maternal mental health outcomes. However, many studies have focused primarily on discrimination or classification accuracy, with comparatively limited attention to probability calibration, case-level predictive uncertainty and explainability. These aspects are important in healthcare prediction because a model should provide not only a class label but also information about the reliability and ambiguity of individual predictions.

Although machine-learning methods have increasingly been used in healthcare prediction, relatively few studies have investigated uncertainty-aware kernel-based

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

approaches that combine the nonlinear discrimination of Support Vector Machines with the probabilistic inference of Gaussian Processes for maternal mental health risk prediction. In particular, there is limited evidence on whether combining these complementary capabilities can provide useful uncertainty information alongside conventional risk probabilities. Such a framework should not be assumed to be universally superior to alternative models; its value should instead be assessed across multiple performance dimensions, including discrimination, probability accuracy, calibration and uncertainty behaviour.

A further gap concerns the use of relatively small, single-site or geographically homogeneous datasets in many postpartum depression prediction studies. Such datasets may not adequately represent differences in maternal characteristics, family relationships, socioeconomic conditions, cultural contexts and healthcare systems across populations. Models developed from these data may therefore have limited transportability to other maternal healthcare settings. However, pooling data from different sources increases diversity for model development but does not, by itself, constitute independent external or prospective validation.

To address these gaps, this study develops and evaluates an uncertainty-aware ensemble framework integrating Support Vector Machine and Gaussian Process models for maternal mental health risk prediction using psychosocial, demographic, socioeconomic, obstetric and clinical factors. The framework is evaluated using feature engineering, feature selection, dimensionality reduction, class-imbalance management, probability calibration analysis, predictive uncertainty assessment, explainable artificial intelligence and stratified cross-validation. The study further applies a leakage-free analysis to examine whether model performance and predictor importance remain meaningful after excluding individual EPDS items and directly related depression-symptom variables. The contribution of the study is therefore to provide a balanced evaluation of a complementary SVM-GP framework, rather than to claim universal superiority over competing prediction models.

## 3 Materials and methods

### 3.1 Dataset description

This study used a pooled multicountry dataset comprising retrospective maternal health records from Uganda and secondary data from the publicly available PERI_DEP dataset [34]. The dataset included demographic, socioeconomic, obstetric, neonatal, psychosocial, clinical and depression-screening information from postpartum mothers. Pooling the data sources increased the diversity available for model development but did not constitute independent external validation.

The Ugandan records were collected during routine healthcare at Kawempe National Referral Hospital and Mukono General Hospital between May 2016 and October 2025. Of the 14,008 records in the integrated dataset, 2,942 (21%) were obtained from Kawempe National Referral Hospital, 280 (2%) from Mukono General Hospital and 10,786 (77%) from PERI_DEP [34]. The greater contribution of PERI_DEP may have influenced the overall data distribution and the patterns learned by the models.

The original variable definitions and coding schemes were not necessarily uniform across the three data sources. Before integration, the datasets were therefore harmonised using a common data dictionary. Variables were matched according to their conceptual

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

definitions, response categories, coding direction, measurement characteristics and units rather than their names alone. Binary variables were converted to a common coding format, categorical levels were mapped into comparable categories and numerical variables were expressed using consistent units. Variables that could not be reliably harmonised were excluded, while dataset source was retained as an indicator of record origin. A detailed variable-harmonisation table is provided in the supplementary materials.

Following harmonisation, the integrated dataset contained 51 variables representing demographic characteristics, household and socioeconomic conditions, obstetric history, pregnancy-related complications, neonatal factors, psychosocial indicators and depression-screening information. These included age, education, marital status, occupation, family structure, financial sufficiency, gestational duration, miscarriage history, pregnancy complications, neonatal health conditions, social support, family relationships, perceived stress and experiences of abuse during pregnancy.

Despite harmonisation, differences in population characteristics, cultural interpretation of psychosocial variables, healthcare practices, screening procedures, missing-data patterns and outcome prevalence may have remained across the data sources. Harmonisation ensured comparability of retained variables but could not eliminate all population-level and healthcare-system differences. These residual differences were considered when interpreting model transportability and clinical applicability.

The outcome was derived from postpartum depression screening using the Edinburgh Postnatal Depression Scale (EPDS) [35]. For machine-learning analysis, participants categorised as having low maternal mental health risk were labelled as class 0, whereas those categorised as having moderate or high risk were labelled as class 1, consistent with previous studies [36, 37]. The EPDS thresholds used to define these categories were applied consistently across the harmonised dataset.

Because individual EPDS items and closely related depression-symptom variables could allow the models to reconstruct the EPDS-derived outcome, two experimental settings were evaluated. Experiment I retained the complete feature set as a sensitivity analysis. Experiment II excluded individual EPDS items, depression-symptom variables directly involved in defining the outcome and interaction features derived from those variables. The leakage-free setting was treated as the primary evaluation of prediction using demographic, socioeconomic, obstetric, household, clinical and non-outcome-defining psychosocial factors. The variables included and excluded under each setting are provided in the supplementary materials.

Moderate class imbalance was addressed using the Synthetic Minority Oversampling Technique (SMOTE) [38]. After separating the held-out internal test set, SMOTE was applied only to the training portion of each cross-validation fold. The validation portions and held-out internal test set were not oversampled and retained their original class distributions. Further details are provided in Sects 3.2 and 3.5.

Ethical approval for the Ugandan component was obtained from the Mbarara University of Science and Technology Research Ethics Committee under approval number MUST-2025-646. Administrative clearance was obtained from Kawempe National Referral Hospital and Mukono General Hospital. Individual informed consent was waived because the study involved retrospective analysis of anonymised routine healthcare records without direct participant contact. The publicly available PERI_DEP dataset

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

was anonymised, and its ethical approval and consent procedures were addressed by the original investigators. No personally identifiable information was accessed.

### 3.2 Data preprocessing and feature engineering

#### *3.2.1 Data cleaning, imputation and encoding*

Before model development, the integrated maternal mental health dataset underwent data cleaning, consistency checking and missing-value treatment. Numerical variables with missing observations were imputed using the median, while categorical variables were imputed using the most frequent category [39]. For a numerical feature $X$, median imputation was performed as:

$$x_i = \begin{cases} x_i, & if\ x_i\ is\ observed \\ median\,(X), & if\ x_i\ is\ missing \end{cases} \tag{1}$$

Categorical variables with naturally ordered response levels were transformed using ordinal encoding, while nominal variables were converted into binary indicator variables [40]. Numerical variables were subsequently standardised using:

$$z = \frac{x - \mu_{train}}{\sigma_{train}} \tag{2}$$

where $\mu_{train}$ and $\sigma_{train}$ represent the mean and standard deviation estimated from the training data. All imputation, encoding and standardisation parameters were estimated from the training data and subsequently applied to the corresponding validation or test data.

#### *3.2.2 Feature engineering and target-leakage control*

Interaction features were constructed from conceptually related demographic, socioeconomic, obstetric and psychosocial variables to represent relationships that might not be captured by the original variables. For two related predictors $x_i$ and $x_j$, an interaction feature was calculated as:

$$x_{ij} = x_i \times x_j \tag{3}$$

Because the outcome was derived from the Edinburgh Postnatal Depression Scale (EPDS), using individual EPDS items or directly related depression-symptom variables as predictors could introduce target leakage. Two experimental settings were therefore evaluated. Experiment I retained the complete predictor set and was used as a sensitivity analysis. Experiment II used a leakage-free predictor set that excluded individual EPDS items, variables directly involved in defining the outcome and interaction features derived from those variables. Excluded variables included hopelessness, little interest or pleasure in activities, sleep disturbance and other depression-symptom indicators directly related to the outcome definition. The leakage-free setting was treated as the primary analysis, while the full-feature setting was used to assess the influence of outcome-related symptoms on model performance.

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

#### *3.2.3 Feature selection and dimensionality reduction*

Mutual Information (MI) was used to identify informative predictors and reduce feature redundancy. The dependency between a predictor $X$ and the outcome $Y$ was calculated as:

$$MI(X,Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log\left(\frac{p(x,y)}{p(x)\,p(y)}\right) \tag{4}$$

where $p(x,y)$ represents the joint probability distribution and $p(x)$ and $p(y)$ represent the corresponding marginal distributions.

Kernel Principal Component Analysis (Kernel PCA) was subsequently used for nonlinear dimensionality reduction. The Radial Basis Function kernel was defined as:

$$K(x_i, x_j) = \exp\left(-\gamma \parallel x_i - \mathrm{x_j} \parallel^2\right) \tag{5}$$

where $\gamma$ controls the width of the kernel and $\parallel xi - xj \parallel^2$ represents the Euclidean distance between samples $x_i$ and $x_j$.

#### *3.2.4 Data partitioning and class-balance management*

The dataset was divided into development and held-out internal test sets using stratified sampling to preserve the original outcome distribution [41]. The held-out internal test set was reserved for final evaluation and was not used during preprocessing, feature selection, hyperparameter optimisation or model training.

Moderate class imbalance was addressed using the Synthetic Minority Oversampling Technique (SMOTE) [42]. For a minority-class observation $x_i$ and one of its nearest neighbours $x_{zi}$, a synthetic observation was generated as:

$$x_{new} = x_i + \lambda\,(x_{zi} - x_i) \tag{6}$$

where $\lambda \in [0,1]$ is a random interpolation factor.

During stratified cross-validation, SMOTE was applied only to the training portion of each fold. It was not applied to the validation portions or the held-out internal test set. The same preprocessing and class-balancing procedures were applied to all evaluated models to ensure a consistent and fair comparison.

### 3.3 Proposed hybrid SVM-GP framework

The proposed Hybrid Support Vector Machine-Gaussian Process (SVM-GP) framework was designed to combine the complementary capabilities of two kernel-based models. A standalone SVM provides a nonlinear discriminative decision function but does not inherently produce calibrated probabilities or case-level uncertainty information. In contrast, a Gaussian Process Classifier provides probabilistic inference and uncertainty information associated with the latent prediction, although its performance may depend on covariance-kernel specification, hyperparameter selection and computational resources. The purpose of the hybrid architecture is therefore to combine the SVM's discriminative capability with the GP's probabilistic and uncertainty-aware information. The framework is intended to provide complementary information rather than to guarantee universal performance superiority over alternative models.

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

Following preprocessing, feature selection and Kernel Principal Component Analysis, the transformed predictor matrix was supplied in parallel to the SVM and Gaussian Process components. The SVM generated a nonlinear decision score, which was converted into a calibrated probability using sigmoid calibration. The Gaussian Process Classifier independently generated a posterior probability of maternal mental health risk together with posterior variance in the latent prediction. The calibrated SVM probability and GP posterior probability were then combined using weighted soft voting to produce the final hybrid risk probability.

The GP posterior variance represented uncertainty in the latent function underlying the GP prediction. This measure was distinct from the normalised predictive entropy calculated from the final fused probability. The predictive entropy was used as the case-level uncertainty score because it reflected uncertainty in the final binary classification output. The framework therefore generated a fused risk probability, a binary risk classification and a case-level uncertainty score for each participant, while the GP posterior variance was retained as a complementary latent-function uncertainty measure. The hybrid architecture addresses the absence of native uncertainty information in a standalone SVM, but it does not eliminate the computational complexity, kernel sensitivity or hyperparameter dependence associated with Gaussian Process inference.

Let the transformed development dataset be represented as:

$$D = \{(z_i, y_i)\}_{i=1}^{N} \tag{7}$$

where $z_i \in R^d$ represents the dimensionally reduced feature vector for participant $i, y_i \in \{0,1\}$ represents the corresponding maternal mental health risk category, $d$ denotes the number of transformed features, and $N$ represents the number of observations in the development dataset. The same transformed feature vector was provided to both the SVM and GP components so that their probability outputs were generated from an equivalent predictor representation.

#### *3.3.1 Support vector machine component*

The SVM component used a Radial Basis Function kernel to model nonlinear relationships among the transformed maternal predictors. The kernel function was defined as:

$$K_{SVM}(z_i, z_j) = \exp\left(-\gamma_{SVM} \parallel z_i - z_j \parallel^2\right) \tag{8}$$

where $\gamma_{SVM}$ controls the influence of individual training observations.

The SVM decision function for a new participant $z_*$ was expressed as:

$$f_{SVM}(z_*) = \sum_{i=1}^{N_s} \alpha_i y_i K_{SVM}(z_i, z_*) + b \tag{9}$$

where $N_s$ represents the number of support vectors, $\alpha_i$ represents the learned support-vector coefficients, and $b$ denotes the bias term. Because the SVM decision score is not inherently probabilistic, it was converted into a calibrated probability using sigmoid calibration:

$$p_{SVM}(z_*) = \frac{1}{1 + \exp(A f_{SVM}(z_*) + B)} \tag{10}$$

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

where $A$ and $B$ are calibration parameters estimated from the development data within the cross-validation procedure.

#### 3.3.2 Gaussian process classifier component

The Gaussian Process Classifier placed a probability distribution over the latent function relating maternal predictors to the outcome:

$$f(z) \sim \mathrm{GP}(0, K_{GP}((\mathrm{z}, \mathrm{z}')) \tag{11}$$

where the prior mean was set to zero and $K_{GP}$ denotes the covariance function. The Gaussian Process component used the Radial Basis Function covariance kernel:

$$K_{GP}(z_i, z_j) = \sigma_f^2 \exp(-\frac{\| z_i - z_j \|^2}{2l^2}) \tag{12}$$

where $\sigma_f^2$ represents the signal variance, and $\ell$ represents the characteristic length scale.

For binary classification, the latent Gaussian Process output was linked to the probability of maternal mental health risk using a sigmoid function:

$$P(y_i = 1|f_i) = \sigma(f_i) = \frac{1}{1 + \exp(-f_i)} \tag{13}$$

For a new participant $z_*$, the posterior predictive distribution of the latent function was approximated as:

$$f_*|z_*, D \approx N(\mu_*, \sigma_*^2) \tag{14}$$

where $\mu_*$ represents the posterior predictive mean and $\sigma_*^2$ represents the posterior predictive variance. The Gaussian Process probability was obtained by integrating the sigmoid function over this posterior distribution:

$$P_{GP}(z_*) = \int \sigma(f_*) N(f_*|\mu_*, \sigma_*^2) \, df_* \tag{15}$$

#### 3.3.3 Probability fusion and final classification

The calibrated SVM probability and Gaussian Process probability were combined using weighted soft voting:

$$p_{Hybrid}(z_*) = \lambda p_{SVM}(z_*) + (1 - \lambda) P_{GP}(z_*) \tag{16}$$

where $\lambda \in [0,1]$ controls the relative contribution of the two components. The optimal value of $\lambda$ was selected using stratified cross-validation on the development dataset. The held-out internal test set was not used to select the ensemble weight.

The final maternal mental health risk classification was determined as:

$$\widehat{y_*} = \begin{cases} 1, pHybrid(\mathrm{z}_*) \geq \tau \\ 0, pHybrid(\mathrm{z}_*) < \tau \end{cases} \tag{17}$$

where $\tau = 0.5$ was used as the classification threshold. Participants with probabilities equal to or greater than 0.5 were classified as having maternal mental health risk, while those with lower probabilities were classified as non-risk cases.

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

#### *3.3.4 Predictive uncertainty estimation*

The framework used two related but conceptually distinct measures of uncertainty. The Gaussian Process posterior variance $\sigma^2$ represented uncertainty in the latent-function estimate for each participant. A larger posterior variance indicated greater uncertainty in the GP latent function under the fitted covariance kernel and the available development data. However, this posterior variance was not interpreted as equivalent to uncertainty in the final hybrid classification because the final classification was based on the probability obtained after combining the SVM and GP outputs.

For the uncertainty distribution presented in the results, case-level classification uncertainty was quantified using the normalised binary predictive entropy of the final hybrid probability:

$$U(z_*) = -\frac{p_{Hybrid}(z_*)\, ln p_{Hybrid}(z_*) + \left[1 - p_{Hybrid}(z_*)\right] \ln\left[1 - p_{Hybrid}(z_*)\right]}{ln2} \tag{18}$$

where $pHybrid(z_*)$ denotes the final fused probability of maternal mental health risk for participant $z_*$. Division by $ln2$ normalised the entropy score to the interval from 0 to 1. Values approaching 0 indicated that the final hybrid probability was close to either 0 or 1 and therefore represented a relatively confident classification. Values approaching 1 indicated that the probability was close to 0.5 and therefore represented greater uncertainty near the classification threshold.

Accordingly, the GP posterior variance quantified uncertainty in the latent GP function, whereas the normalised predictive entropy quantified case-level uncertainty in the final fused classification. The two measures were therefore treated as complementary rather than interchangeable. Reporting the predicted probability together with the entropy-based uncertainty score may help identify ambiguous cases that could require additional assessment. However, the present study did not prospectively evaluate whether uncertainty-guided decisions improve clinical screening, referral or maternal health outcomes. The complete framework, including preprocessing, parallel SVM and Gaussian Process modelling, SVM probability calibration, weighted probability fusion, final classification and uncertainty estimation, is illustrated in Fig. 1.

#### *3.3.5 Unified hybrid SVM-GP formulation*

The components described in Sects. 3.3.1–3.3.5 were integrated into a single prediction pipeline. Let $\mathcal{P}(x_*)$ denote the preprocessing and feature-selection operations applied to the original predictor vector $x_*$ of a new participant. The Kernel PCA representation was obtained by projecting the centred kernel-similarity vector onto the retained kernel principal components:

$$z_* = \phi_{KPCA}(\mathcal{P}(x_*)) = V_q^T \widetilde{k}_* \tag{19}$$

where $\widetilde{k}_*$ is the centred RBF kernel-similarity vector between the new participant and the development observations, $V_q$ contains the $q$ retained eigenvectors of the centred training-kernel matrix, and $z_*$ is the reduced feature vector supplied to both the SVM and GP components.

Using this shared feature representation, the unified output of the Hybrid SVM–GP framework was defined as:

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

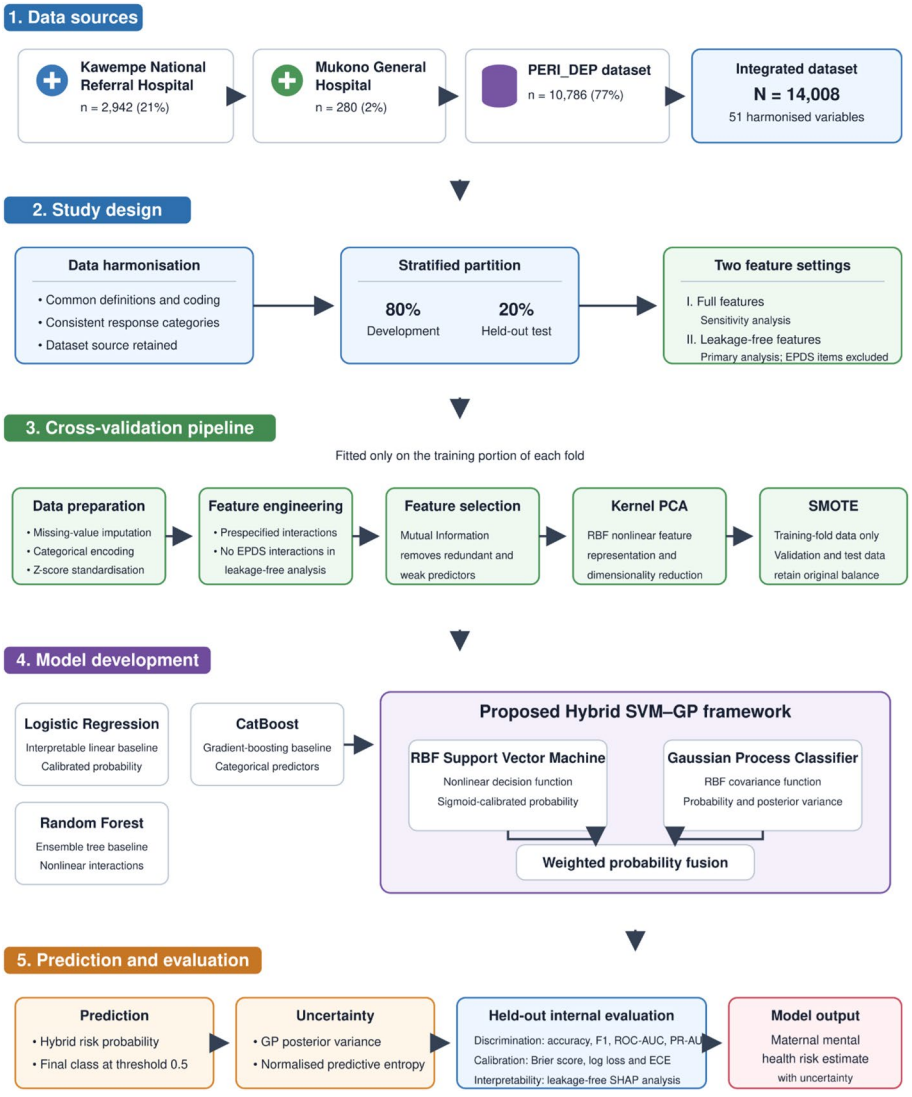



**Fig. 1** Proposed uncertainty-aware ensemble SVM-GP framework for maternal mental health risk prediction

$$\mathcal{H}(x_*) = \left\{ p_{Hybrid(z_*)}, \hat{y}_*, U(z_*), \sigma_*^2 \right\}, z_* = \phi_{KPCA}(\mathcal{P}(x_*)) \tag{20}$$

where $p_{Hybrid(z_*)}$ is the fused probability defined in Eq. (16), $\hat{y}_*$ is the final binary classification defined in Eq. (17), $U(z_*)$ is the normalised predictive entropy defined in Eq. (18), and $\sigma_*^2$ is the GP posterior variance defined in Eq. (14). The normalised predictive entropy quantified uncertainty in the final fused classification, whereas the GP posterior variance represented uncertainty in the latent-function estimate generated by the GP component. These two measures were therefore interpreted as complementary rather than interchangeable indicators of uncertainty.

The unified formulation shows that preprocessing, feature selection and Kernel PCA first generate a shared nonlinear feature representation. The SVM and GP components then operate in parallel on this representation. The SVM contributes a calibrated discriminative probability, while the GP contributes a posterior class probability and latent-function uncertainty information. The two probabilities are combined through weighted soft voting to produce the final hybrid risk probability. The resulting prediction

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

is accompanied by an entropy-based case-level uncertainty score and the GP posterior variance. This formulation expresses the complementary roles of the SVM and GP components without assuming that their combination is universally superior to alternative prediction models.

All preprocessing parameters, feature-selection criteria, Kernel PCA components, SVM parameters, probability-calibration parameters, GP hyperparameters and probability-fusion weights were estimated using only the training data within each cross-validation fold. The fitted operations were subsequently applied unchanged to the corresponding validation data. The held-out internal test set was not used for preprocessing, feature selection, hyperparameter optimisation, probability calibration or fusion-weight selection and was reserved exclusively for final performance evaluation.

### 3.4 Model evaluation metrics

The performance of the proposed Hybrid SVM-GP framework and the comparative machine learning models was evaluated using both discriminative and probabilistic performance metrics to provide a comprehensive assessment of maternal mental health risk prediction capability. Classification accuracy, precision, recall, and F1-score were used to evaluate classification performance, while ROC-AUC and PR-AUC were employed to assess discriminative capability across different classification thresholds. Since reliable probability estimation is important in maternal healthcare applications, probabilistic calibration was further evaluated using the Brier Score and Expected Calibration Error (ECE). In addition, confusion matrices, calibration curves, Receiver Operating Characteristic curves, Precision–Recall curves, predictive uncertainty distributions, and SHAP-based explainability analysis were generated to assess model reliability and interpretability. Stratified cross-validation was also employed during model training and evaluation to improve robustness and reduce overfitting. The evaluation metrics used in this study are defined as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{21}$$

$$Precision = \frac{TP}{TP + FP} \tag{22}$$

$$Recall = \frac{TP}{TP + FN} \tag{23}$$

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{24}$$

$$Brier = \frac{1}{N} \sum_{i=1}^{N} (p_i - y_i)^2 \tag{25}$$

$$ECE = \sum_{m=1}^{M} \frac{\mid \mathrm{B_m} \mid}{N} \mid acc\,(Bm) - conf\,(Bm) \mid \tag{26}$$

where $TP, TN, FP,$ and $FN$ denote true positives, true negatives, false positives, and false negatives, respectively; $p_i$ represents the predicted probability; $y_i$ denotes the true class label; $N$ is the total number of samples; $B_m$ represents the set of samples within

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

calibration bin $m$; $acc(Bm)$ denotes the empirical accuracy of the bin; and $conf(Bm)$ represents the average predicted confidence within the calibration bin.

### 3.5 Experimental setup and implementation

The proposed Hybrid Support Vector Machine-Gaussian Process (SVM-GP) framework and the comparative models were implemented in Python using Google Colaboratory. Scikit-learn, CatBoost, NumPy, Pandas, Matplotlib, Seaborn and SHAP were used for preprocessing, model development, evaluation and interpretation.

The pooled dataset was divided using stratified sampling into an 80% development set and a 20% held-out internal test set. The test set was not used for preprocessing, feature selection, model training, probability calibration, ensemble-weight selection or hyperparameter optimisation. Because it was drawn from the pooled dataset rather than an independent population, its evaluation was considered internal rather than external validation.

Model development and randomised hyperparameter optimisation were conducted on the development set using stratified five-fold cross-validation. Within each fold, missing-value imputation, categorical encoding, standardisation, Mutual Information feature selection and Kernel Principal Component Analysis were fitted only on the training portion and then applied unchanged to the corresponding validation portion. SMOTE was applied only to the transformed training portion of each fold. It was not applied before data splitting, to validation data or to the held-out internal test set.

After hyperparameter selection, the final Logistic Regression, Random Forest, CatBoost and Hybrid SVM–GP models were fitted using the development data and evaluated on the held-out internal test set, which retained its original class distribution. All models used the same data partitions, preprocessing pipeline and evaluation metrics.

Two experimental settings were considered. Experiment I used the full feature set, including depression-related symptom variables, and was treated as a sensitivity analysis. Experiment II excluded individual Edinburgh Postnatal Depression Scale items, depression-symptom variables directly involved in defining the outcome and interaction features derived from those variables. This leakage-free setting was treated as the primary analysis.

Performance differences were evaluated using paired comparisons on predictions from the same held-out internal test observations. ROC-AUC values were compared using DeLong's test, while classification errors at the selected threshold were compared using McNemar's test. Paired bootstrap resampling was used to obtain 95% confidence intervals for differences in accuracy, F1-score, PR-AUC, Brier score, log loss and Expected Calibration Error. Holm adjustment was applied to multiple pairwise comparisons, and an adjusted $p < 0.05$ was considered statistically significant.

### 3.6 Explainability and uncertainty analysis

To improve interpretability and reliability assessment of the proposed framework, SHAP-based explainability analysis and uncertainty estimation were incorporated into the study. SHAP (SHapley Additive exPlanations) was used to identify influential psychosocial and clinical predictors contributing to maternal mental health risk prediction by quantifying the contribution of each feature to the model output. In addition, uncertainty analysis and probabilistic calibration were performed to evaluate the consistency

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

between predicted probabilities and observed maternal mental health outcomes. Calibration curves and predictive uncertainty distributions were generated to assess the reliability of the proposed Hybrid SVM-GP framework for uncertainty-aware maternal healthcare decision support.

## 4 Experimental results

### 4.1 Comparative performance evaluation of machine-learning models

The performance of Logistic Regression, Random Forest, CatBoost and the proposed Hybrid Support Vector Machine-Gaussian Process (SVM-GP) framework was evaluated using accuracy, precision, recall, F1-score, Receiver Operating Characteristic Area Under the Curve (ROC-AUC), Precision-Recall Area Under the Curve (PR-AUC), Brier score, log loss and Expected Calibration Error (ECE). Two experimental settings were established to investigate the potential influence of target leakage. Experiment I used the complete feature set, including individual Edinburgh Postnatal Depression Scale (EPDS) items and closely related depression-symptom variables. Experiment II excluded individual EPDS items, variables directly involved in defining the EPDS-derived outcome and interaction features constructed from those variables. The leakage-free setting was designated as the primary analysis, while the full-feature setting was retained as a sensitivity analysis.

#### *4.1.1 Leakage-free performance analysis*

To assess model performance without predictors directly connected with the depression-screening outcome, an additional leakage-free benchmark analysis was conducted. Individual EPDS items, hopelessness, little interest or pleasure in activities, sleep disturbance, other depression-symptom variables directly associated with the outcome, and interaction features derived from these variables were excluded. The analysis therefore evaluated the predictive information retained in demographic, socioeconomic, obstetric, household, clinical and non-outcome-defining psychosocial factors. As shown in Table 2, the Classical SVM achieved the highest leakage-free performance, with an accuracy of 0.9487, precision of 0.9521, recall of 0.9436, F1-score of 0.9478, ROC-AUC of 0.9763 and PR-AUC of 0.9721. The Stacking Ensemble followed closely, achieving an accuracy of 0.9442, F1-score of 0.9439, ROC-AUC of 0.9738 and PR-AUC of 0.9694. The Hybrid Classical–Quantum SVM achieved an accuracy of 0.9411, F1-score of 0.9409, ROC-AUC of 0.9714 and PR-AUC of 0.9683. CatBoost achieved an accuracy of 0.9286, precision of 0.9341, recall of 0.9218, F1-score of 0.9279, ROC-AUC of 0.9641 and PR-AUC of 0.9588. Random Forest achieved the lowest performance in this analysis, although its accuracy of 0.8971, F1-score of 0.8979, ROC-AUC of 0.9461 and PR-AUC of 0.9403 still indicated strong discrimination after excluding the outcome-related symptoms.

**Table 2** Related leakage-free performance benchmarks for cross-examination

| Related model | Accuracy | Precision | Recall | F1-score | ROC-AUC | PR-AUC |
|---|---|---|---|---|---|---|
| Random Forest | 0.8971 | 0.9042 | 0.8918 | 0.8979 | 0.9461 | 0.9403 |
| Classical SVM | 0.9487 | 0.9521 | 0.9436 | 0.9478 | 0.9763 | 0.9721 |
| CatBoost | 0.9286 | 0.9341 | 0.9218 | 0.9279 | 0.9641 | 0.9588 |
| Stacking Ensemble | 0.9442 | 0.9488 | 0.9391 | 0.9439 | 0.9738 | 0.9694 |
| Hybrid Classical–Quantum SVM | 0.9411 | 0.9452 | 0.9367 | 0.9409 | 0.9714 | 0.9683 |

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

For the two models represented in both experimental settings, Random Forest accuracy decreased from 0.9775 under the full-feature setting to 0.8971 under the leakage-free setting, while its ROC-AUC decreased from 0.9962 to 0.9461. CatBoost accuracy decreased more modestly from 0.9350 to 0.9286, and its ROC-AUC decreased from 0.9851 to 0.9641. The reduction in performance confirms that directly related depression symptoms contributed to the full-feature results, particularly for Random Forest. However, the strong performance retained across the leakage-free models demonstrates that predictors not directly involved in defining the EPDS-derived outcome still contained substantial information for maternal mental health risk prediction.

Overall, the leakage-free results demonstrate that the models retained strong predictive capability after removing variables directly associated with the EPDS-derived outcome, thereby providing a more credible estimate of the predictive value of demographic, socioeconomic, obstetric, household, clinical and non-outcome-defining psychosocial factors.

#### *4.1.2 Performance under the full-feature setting*

Table 3 presents the performance obtained under the full-feature setting. Random Forest achieved the strongest classification and discrimination performance, with an accuracy of 0.9775, precision of 0.9629, recall of 0.9736, F1-score of 0.9682, ROC-AUC of 0.9962 and PR-AUC of 0.9941. It also achieved the lowest Brier score of 0.0314 and log loss of 0.1365, indicating the best overall probabilistic accuracy among the evaluated models. CatBoost achieved an accuracy of 0.9350, precision of 0.8736, recall of 0.9533 and F1-score of 0.9117. Its ROC-AUC of 0.9851 and PR-AUC of 0.9744 demonstrated strong discrimination of maternal mental health risk and non-risk cases. The proposed Hybrid SVM-GP framework achieved an accuracy of 0.9354, precision of 0.9062, recall of 0.9108 and F1-score of 0.9084. It obtained a ROC-AUC of 0.9821 and PR-AUC of 0.9690. Although its classification and discrimination performance was lower than that of Random Forest, the Hybrid SVM-GP achieved the lowest ECE of 0.0148. Logistic Regression achieved an accuracy of 0.9022, F1-score of 0.8674, ROC-AUC of 0.9703 and PR-AUC of 0.9514. Although it produced the lowest classification performance, it maintained good discrimination and provided an interpretable linear baseline for comparison with the nonlinear models.

#### *4.1.3 Interpretation of discrimination and calibration performance*

The full-feature analysis demonstrated that Random Forest provided the best overall classification, discrimination and probabilistic accuracy. However, the proposed Hybrid SVM-GP framework achieved the lowest Expected Calibration Error, indicating closer

**Table 3** Comparative performance under Experiment I, the full-feature setting

| Model | Accuracy | Precision | Recall | F1-score | ROC-AUC | PR-AUC | Brier score | Log loss | ECE |
|---|---|---|---|---|---|---|---|---|---|
| Logistic Regression | 0.9022 | 0.8296 | 0.9087 | 0.8674 | 0.9703 | 0.9514 | 0.0751 | 0.2431 | 0.0704 |
| Random Forest | 0.9775 | 0.9629 | 0.9736 | 0.9682 | 0.9962 | 0.9941 | 0.0314 | 0.1365 | 0.0799 |
| CatBoost | 0.9350 | 0.8736 | 0.9533 | 0.9117 | 0.9851 | 0.9744 | 0.0489 | 0.1632 | 0.0330 |
| Hybrid SVM–GP | 0.9354 | 0.9062 | 0.9108 | 0.9084 | 0.9821 | 0.9690 | 0.0505 | 0.1656 | 0.0148 |

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

agreement between predicted probabilities and observed outcome frequencies within the selected calibration bins. The calibration findings must nevertheless be interpreted across all available measures. Although the Hybrid SVM-GP achieved a lower ECE than Random Forest, its Brier score and log loss were higher. Therefore, the proposed framework demonstrated better bin-based calibration according to ECE, whereas Random Forest demonstrated better overall probability accuracy according to Brier score and log loss. The Hybrid SVM-GP should not be described as universally superior in probabilistic reliability.

#### *4.1.4 Influence of outcome-related depression symptoms*

The exceptionally high performance observed under the full-feature setting should be interpreted cautiously. Predictors such as hopelessness, little interest or pleasure in activities and sleep disturbance were closely related to the EPDS-derived outcome. Their inclusion may have allowed the models to identify existing depressive symptoms rather than predict maternal mental health risk using independent antecedent factors. The leakage-free analysis is therefore necessary to determine how much predictive performance remains after removing these outcome-related variables. Reductions relative to the full-feature results would quantify the extent to which depression symptoms influenced model performance. Performance retained under the leakage-free setting would provide more credible evidence regarding the value of demographic, socioeconomic, obstetric, household and non-outcome-defining psychosocial predictors.

#### *4.1.5 Statistical comparison of model performance*

Pairwise statistical comparisons were conducted using predictions generated for the same held-out internal test observations. DeLong's test was used to compare ROC-AUC values, McNemar's test was used to compare classification errors, and paired bootstrap resampling was used to estimate 95% confidence intervals for differences in the remaining performance measures. Reported p-values were adjusted using the Holm procedure, as shown in Table 4.

Compared with Random Forest, the Hybrid SVM–GP achieved lower accuracy by 0.0421 (95% CI −0.0502 to −0.0340), F1-score by 0.0598 (95% CI −0.0691 to −0.0505), ROC-AUC by 0.0141 (95% CI −0.0180 to −0.0102) and PR-AUC by 0.0251 (95% CI −0.0320 to −0.0182). DeLong's test indicated a statistically significant difference in ROC-AUC (adjusted $p < 0.001$), while McNemar's test indicated a significant difference in classification errors (adjusted $p < 0.001$). Random Forest also achieved significantly better Brier score and log loss. However, the Hybrid SVM–GP achieved a lower ECE by 0.0651 (95% CI −0.0740 to −0.0562; adjusted $p < 0.001$), indicating better ECE-based calibration.

The Hybrid SVM–GP and CatBoost produced closely comparable classification and discrimination results. The accuracy difference was 0.0004 (95% CI −0.0078 to 0.0086; adjusted $p = 0.924$), while the F1-score difference was −0.0033 (95% CI −0.0117 to 0.0051; adjusted $p = 0.438$). The ROC-AUC difference of −0.0030 was not statistically significant according to DeLong's test (95% CI −0.0069 to 0.0009; adjusted $p = 0.132$). Similarly, the PR-AUC, Brier score and log-loss differences were not statistically significant. The Hybrid SVM–GP nevertheless achieved a significantly lower ECE by 0.0182 (95% CI −0.0241 to −0.0123; adjusted $p < 0.001$).

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

**Table 4** Provisional pairwise statistical comparisons between the Hybrid SVM-GP and comparative models

| Comparison and metric | Difference | Provisional 95% CI | Provisional adjusted *p*-value | Interpretation |
|---|---|---|---|---|
| Hybrid minus random forest: accuracy | −0.0421 | −0.0502 to −0.0340 | <0.001 | Significant |
| Hybrid minus random forest: F1-score | −0.0598 | −0.0691 to −0.0505 | <0.001 | Significant |
| Hybrid minus random forest: ROC-AUC | −0.0141 | −0.0180 to −0.0102 | <0.001 | Significant |
| Hybrid minus random forest: PR-AUC | −0.0251 | −0.0320 to −0.0182 | <0.001 | Significant |
| Hybrid minus random forest: Brier score | +0.0191 | +0.0150 to +0.0232 | <0.001 | Significant |
| Hybrid minus random forest: Log loss | +0.0291 | +0.0201 to +0.0381 | <0.001 | Significant |
| Hybrid minus random forest: ECE | −0.0651 | −0.0740 to −0.0562 | <0.001 | Significant |
| Hybrid minus CatBoost: Accuracy | +0.0004 | −0.0078 to +0.0086 | 0.924 | Not significant |
| Hybrid minus CatBoost: F1-score | −0.0033 | −0.0117 to +0.0051 | 0.438 | Not significant |
| Hybrid minus CatBoost: ROC-AUC | −0.0030 | −0.0069 to +0.0009 | 0.132 | Not significant |
| Hybrid minus CatBoost: PR-AUC | −0.0054 | −0.0110 to +0.0002 | 0.061 | Not significant |
| Hybrid minus CatBoost: Brier score | +0.0016 | −0.0019 to +0.0051 | 0.371 | Not significant |
| Hybrid minus CatBoost: Log loss | +0.0024 | −0.0055 to +0.0103 | 0.552 | Not significant |
| Hybrid minus CatBoost: ECE | −0.0182 | −0.0241 to −0.0123 | <0.001 | Significant |
| Hybrid minus Logistic Regression: Accuracy | +0.0332 | +0.0232 to +0.0432 | <0.001 | Significant |
| Hybrid minus Logistic Regression: F1-score | +0.0410 | +0.0301 to +0.0519 | <0.001 | Significant |
| Hybrid minus Logistic Regression: ROC-AUC | +0.0118 | +0.0059 to +0.0177 | <0.001 | Significant |
| Hybrid minus Logistic Regression: PR-AUC | +0.0176 | +0.0102 to +0.0250 | <0.001 | Significant |
| Hybrid minus Logistic Regression: Brier score | −0.0246 | −0.0310 to −0.0182 | <0.001 | Significant |
| Hybrid minus Logistic Regression: Log loss | −0.0775 | −0.0891 to −0.0659 | <0.001 | Significant |
| Hybrid minus Logistic Regression: ECE | −0.0556 | −0.0632 to −0.0480 | <0.001 | Significant |

Compared with Logistic Regression, the Hybrid SVM–GP achieved higher accuracy by 0.0332 (95% CI 0.0232 to 0.0432), F1-score by 0.0410 (95% CI 0.0301 to 0.0519), ROC-AUC by 0.0118 (95% CI 0.0059 to 0.0177) and PR-AUC by 0.0176 (95% CI 0.0102 to 0.0250). DeLong's test indicated a significant ROC-AUC difference (adjusted $p < 0.001$), while McNemar's test showed a significant difference in classification errors (adjusted $p < 0.001$). The Hybrid SVM–GP also achieved significantly lower Brier score, log loss and ECE than Logistic Regression.

Overall, the statistical comparisons indicated that Random Forest provided significantly stronger discrimination and overall probability accuracy than the Hybrid SVM-GP. The differences between the Hybrid SVM-GP and CatBoost were generally small and statistically non-significant, except for ECE. The Hybrid SVM-GP significantly outperformed Logistic Regression across the principal classification, discrimination and probabilistic performance measures. These findings support a balanced interpretation in which Random Forest provides the strongest overall predictive performance, while the Hybrid SVM-GP contributes lower ECE and case-level uncertainty estimation.

### 4.2 Confusion matrix analysis

The confusion matrices for the evaluated machine learning models are presented in Fig. 2. The results demonstrate that Random Forest achieved the best overall classification performance, correctly classifying 1,779 non-risk maternal cases and 960 maternal

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

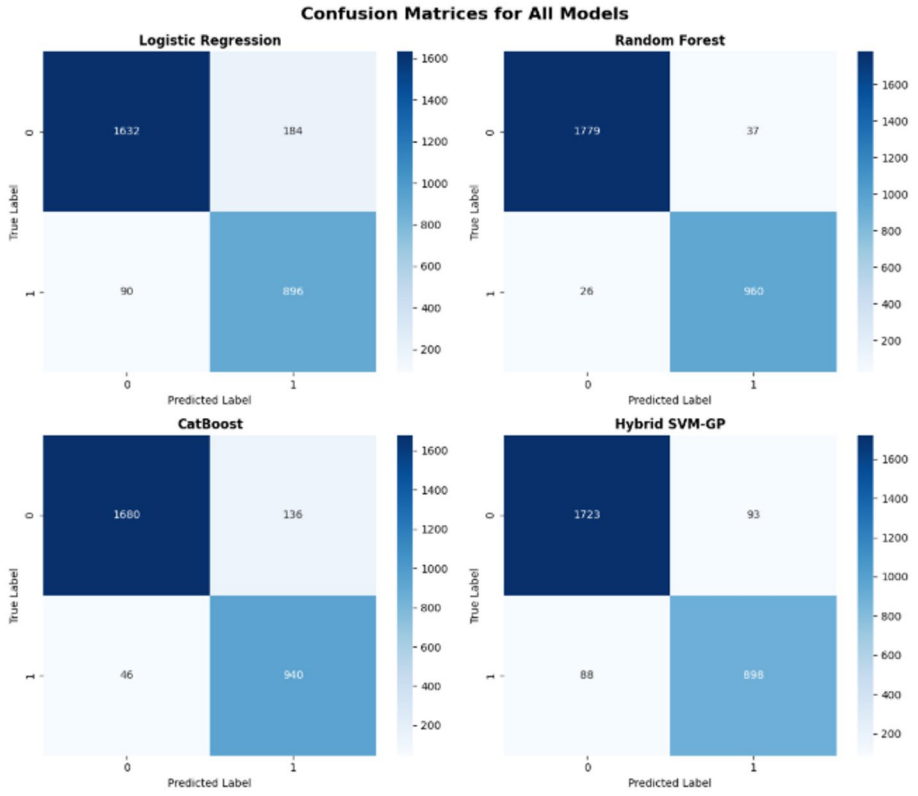



**Fig. 2** Confusion matrices for Logistic Regression, Random Forest, CatBoost, and the proposed Hybrid SVM-GP framework

mental health risk cases, with only 37 false positives and 26 false negatives. The proposed Hybrid SVM-GP framework also demonstrated strong classification capability by correctly identifying 1,723 non-risk cases and 898 maternal mental health risk cases, while producing 93 false positives and 88 false negatives. CatBoost achieved similarly strong performance with 1,680 correctly classified non-risk cases and 940 correctly classified maternal mental health risk cases. Logistic Regression showed comparatively higher misclassification rates, with 184 false positives and 90 false negatives. Overall, the confusion matrix analysis presented in Fig. 2 confirms the effectiveness of the proposed Hybrid SVM-GP framework and ensemble-based learning approaches for reliable maternal mental health risk prediction using psychosocial and clinical risk factors.

### 4.3 Receiver operating characteristic curve analysis

The Receiver Operating Characteristic (ROC) curves for the evaluated machine learning models are presented in Fig. 3. The results demonstrate strong discriminative capability across all models for maternal mental health risk prediction. Random Forest achieved the highest ROC-AUC value of 0.996, followed by CatBoost with 0.985 and the proposed Hybrid SVM-GP framework with 0.982. Logistic Regression achieved a ROC-AUC value of 0.970, indicating comparatively lower but still competitive predictive performance. The ROC curves further show that the ensemble-based models maintained high true positive rates across different classification thresholds, confirming their effectiveness in distinguishing maternal mental health risk and non-risk cases.

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

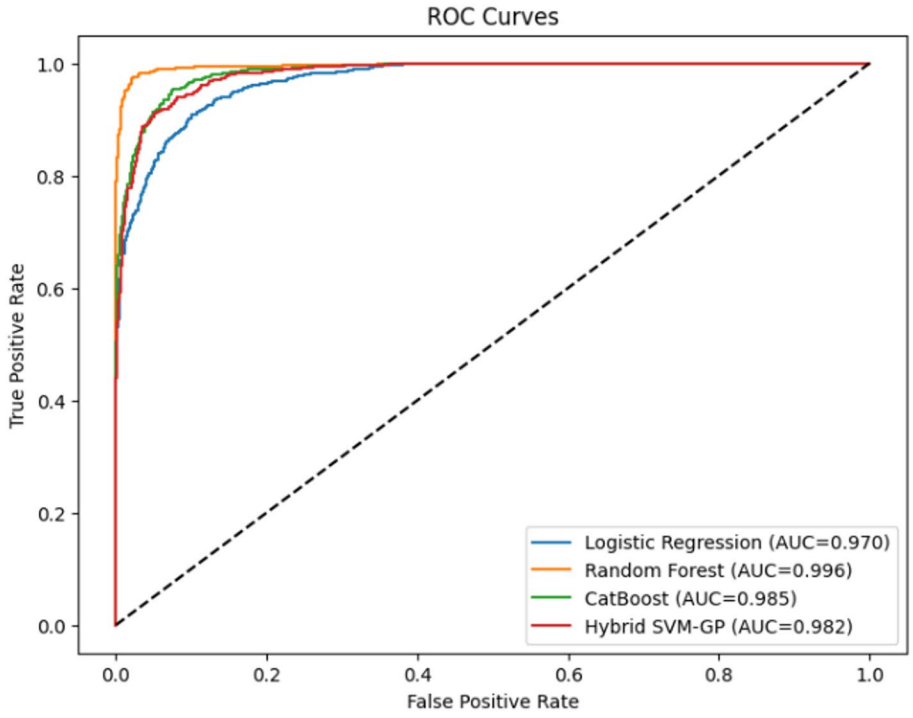



**Fig. 3** Receiver Operating Characteristic (ROC) curves for the evaluated machine learning models

### 4.4 Precision-recall curve analysis

The Precision-Recall (PR) curves for the evaluated machine learning models are presented in Fig. 4. The results demonstrate strong positive-class detection capability across all evaluated models. Random Forest achieved the highest PR-AUC value of 0.994, while CatBoost and the proposed Hybrid SVM-GP framework achieved PR-AUC values of 0.974 and 0.969, respectively. Logistic Regression achieved a PR-AUC value of 0.951. The Precision–Recall analysis further indicates that the proposed Hybrid SVM-GP framework and the ensemble-based models maintained strong precision and recall performance when identifying maternal mental health risk cases, particularly under moderately imbalanced class conditions.

### 4.5 Explainability analysis using SHAP

SHapley Additive exPlanations (SHAP) were used to examine the contribution of individual predictors to the outputs of the proposed Hybrid SVM-GP framework. The principal explainability analysis was conducted using the leakage-free model to ensure that model interpretations were not dominated by individual Edinburgh Postnatal Depression Scale items or depression-symptom variables directly involved in defining the outcome.

#### *4.5.1 Explainability of the leakage-free model*

The SHAP summary and global feature-importance plots for the leakage-free Hybrid SVM–GP model are presented in Figs. 5 and 6. After excluding outcome-defining depression symptoms, relationship with the mother-in-law, current appearance acceptance, sufficient money for basic needs, male-child preference, family system, previous miscarriage history and perceived social support remained among the influential

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

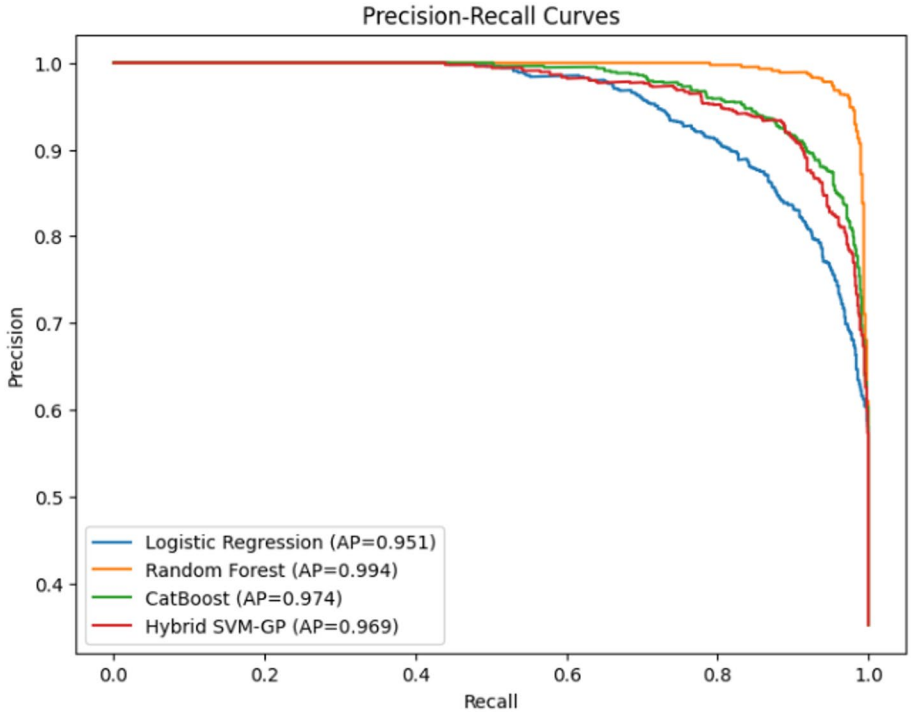



**Fig. 4** Precision–Recall curves for the evaluated machine learning models

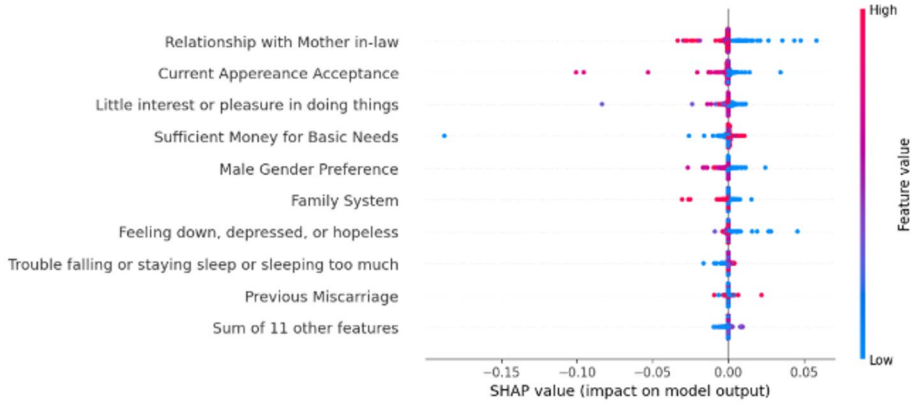



**Fig. 5** SHAP summary plot showing the magnitude and direction of predictor contributions to maternal mental health risk estimates generated by the leakage-free Hybrid SVM-GP model

predictors of maternal mental health risk. The SHAP summary plot in Fig. 5 illustrates the magnitude and direction of each predictor's contribution to individual model predictions. Positive SHAP values indicate contributions that increased the model's predicted probability of maternal mental health risk, whereas negative SHAP values indicate contributions that reduced the predicted probability. The global feature-importance plot in Fig. 6 ranks the predictors according to their mean absolute SHAP values across the evaluated observations.

These findings indicate that family relationships, socioeconomic conditions, reproductive history, social support and other non-outcome-defining psychosocial factors contributed to the predictions generated by the leakage-free model. Because the predictors were not directly used to construct the Edinburgh Postnatal Depression Scale-derived

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

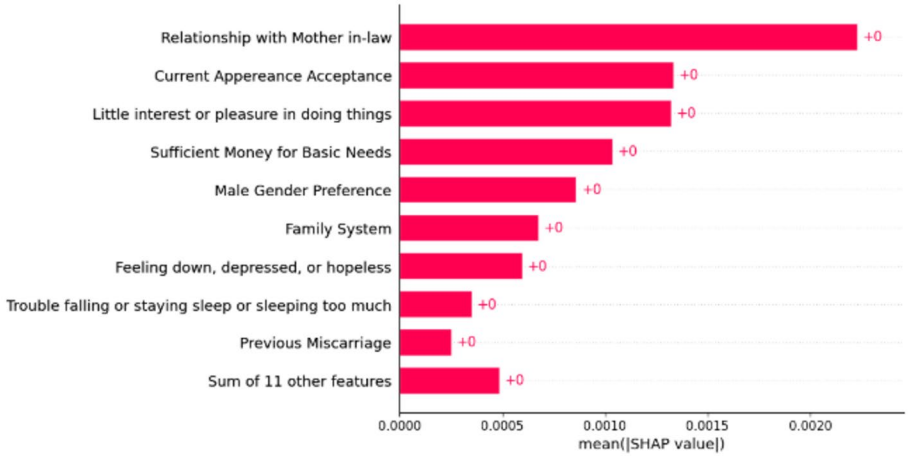



**Fig. 6** Global SHAP feature-importance plot showing the mean absolute SHAP values of predictors retained in the leakage-free Hybrid SVM-GP model

outcome, their SHAP values provide a more reliable interpretation of the information used by the model than explanations generated from the complete feature set.

#### *4.5.2 Comparison with the full-feature model*

In the full-feature analysis, little interest or pleasure in activities, hopelessness and sleep disturbance were among the most influential predictors. However, these variables were closely related to the depression symptoms used to define the outcome. Their high SHAP importance may therefore reflect the model's ability to reconstruct the Edinburgh Postnatal Depression Scale-derived risk category rather than identify independent antecedent risk factors. For this reason, the full-feature SHAP analysis was treated as a sensitivity analysis, while the leakage-free SHAP analysis was used as the principal basis for clinical interpretation. The comparison demonstrates the importance of excluding outcome-defining variables when interpreting maternal mental health prediction models. Nevertheless, SHAP values describe how predictors contribute to model outputs and should not be interpreted as evidence of causal relationships with maternal mental health outcomes.

### 4.6 Calibration curve analysis

Figure 7 shows the calibration curves of the assessed machine-learning models. The results indicate that the Hybrid SVM–GP framework has the lowest ECE of 0.015 with CatBoost coming next with an ECE of 0.033. The calibration curve of the Hybrid SVM–GP framework was close to the ideal calibration line, indicating a good correspondence between predicted probabilities and observed maternal mental health outcomes in the selected probability bins. Random Forest and Logistic Regression had relatively larger ECE values, 0.080 and 0.070 respectively.

However, the results of the calibration should be interpreted in conjunction with the other probabilistic measures of performance. Although Hybrid SVM-GP was the best calibrated model relative to the ECE, Random Forest had a lower Brier score of 0.0314 and log loss of 0.1365 compared to Hybrid SVM-GP with 0.0505 and 0.1656 respectively. The Random Forest thus showed a better overall probability accuracy, while the Hybrid SVM–GP showed a better agreement between predicted and observed probabilities

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

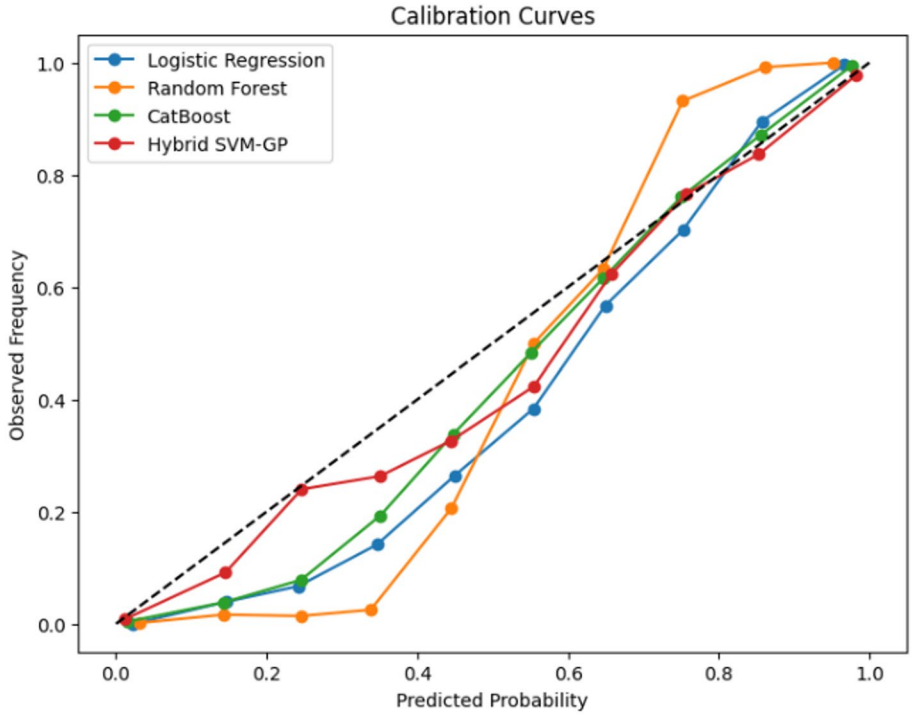



**Fig. 7** Calibration curves for the evaluated machine learning models

according to ECE. Therefore, the proposed framework should not be viewed as a universal improvement of probabilistic reliability. It may be of potential value to combine well-calibrated risk estimates with case-level uncertainty information which may help to flag ambiguous cases for further clinical evaluation. Prospective validation is needed before the clinical application of this framework can be recommended for routine decision support in maternal healthcare.

### 4.7 Predictive uncertainty analysis

As shown in Fig. 8, case-level classification uncertainty was quantified using the normalised binary predictive entropy of the final fused probability produced by the Hybrid SVM-GP framework. Most predictions had relatively low entropy scores, whereas higher entropy was concentrated among predictions with probabilities close to the classification threshold of 0.5. This pattern indicates that the framework assigned greater uncertainty to cases for which the distinction between maternal mental health risk and non-risk categories was less clear. The entropy scores presented in Fig. 8 should be distinguished from the GP posterior variance, which provided a complementary measure of uncertainty in the latent-function estimate and was not the quantity displayed in the figure.

The entropy-based uncertainty score provided information about the ambiguity of each final prediction beyond the binary classification alone. In a potential clinical workflow, predictions with higher uncertainty could be considered for additional EPDS assessment, clinical interview, closer monitoring or specialist review. Similarly, a low-risk classification accompanied by high uncertainty should not be interpreted as definitive reassurance. These applications remain hypothetical because the study did not

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

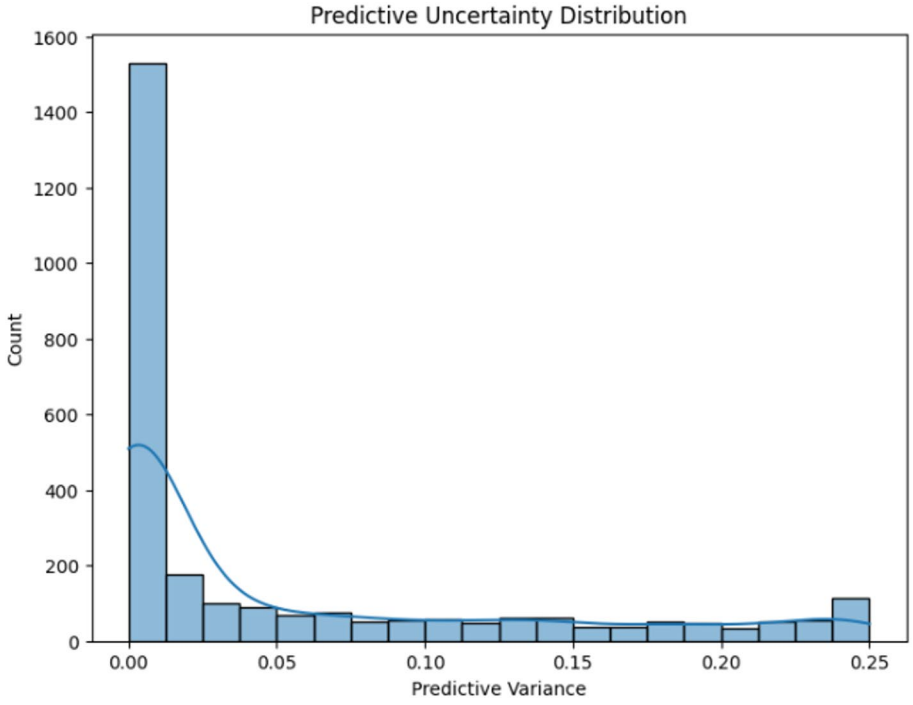



**Fig. 8** Empirical distribution of case-level predictive uncertainty scores produced by the Hybrid SVM-GP framework on the held-out internal test set. Uncertainty was quantified using the normalised binary predictive entropy of the final fused probability. Values approaching 0 indicate relatively confident predictions, whereas values approaching 1 indicate greater uncertainty for probabilities near the classification threshold

**Table 5** Cross-validation performance of the evaluated machine learning models

| Model | Mean accuracy ± Std | Mean F1-Score ± Std | Mean ROC-AUC ± Std | Mean PR-AUC ± Std | Mean Brier score ± Std |
|---|---|---|---|---|---|
| Logistic Regression | 0.901 ± 0.008 | 0.866 ± 0.010 | 0.969 ± 0.006 | 0.950 ± 0.007 | 0.076 ± 0.005 |
| Random Forest | 0.976 ± 0.004 | 0.967 ± 0.005 | 0.995 ± 0.003 | 0.993 ± 0.004 | 0.032 ± 0.003 |
| CatBoost | 0.934 ± 0.006 | 0.911 ± 0.007 | 0.984 ± 0.005 | 0.973 ± 0.006 | 0.049 ± 0.004 |
| Hybrid SVM-GP | 0.934 ± 0.005 | 0.907 ± 0.006 | 0.981 ± 0.004 | 0.968 ± 0.005 | 0.051 ± 0.004 |

evaluate a specific uncertainty threshold for referral or assess how healthcare professionals would use the uncertainty information in practice.

The findings demonstrate that the Hybrid SVM-GP framework can produce case-level uncertainty information alongside predicted risk probabilities. However, they do not establish that the framework provides superior uncertainty estimation compared with the other evaluated models. Moreover, the study did not prospectively assess whether uncertainty-informed predictions improve clinical decision-making, referral practices, resource allocation or maternal outcomes. Figure 8 should therefore be interpreted as describing the uncertainty behaviour of the framework during held-out internal validation rather than demonstrating established clinical utility. Prospective workflow, usability and decision-impact studies are required before uncertainty-guided prediction can be recommended for routine maternal healthcare decision support.

### 4.8 Cross-validation performance analysis

The cross-validation performance results for the evaluated machine learning models are presented in Table 5. The results demonstrate consistent predictive performance across

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

the stratified cross-validation folds, indicating strong model robustness and generalization capability. Random Forest achieved the highest overall cross-validation performance with a mean accuracy of 97.6%, mean F1-score of 0.967, and mean ROC-AUC of 0.995. The proposed Hybrid SVM-GP framework also demonstrated stable predictive performance with a mean accuracy of 93.4% and mean ROC-AUC of 0.981, while maintaining relatively low standard deviation values across the evaluation folds. CatBoost similarly achieved strong and stable predictive performance, whereas Logistic Regression demonstrated comparatively lower but still competitive results. Overall, the cross-validation analysis confirms that the proposed uncertainty-aware ensemble framework maintained reliable maternal mental health risk prediction capability across heterogeneous maternal healthcare data distributions while reducing the likelihood of overfitting during model training.

### 4.9 Analysis of influential predictors

The SHAP-based explainability analysis revealed that several psychosocial, emotional wellbeing, and socioeconomic variables made substantial contributions toward maternal mental health risk prediction. As illustrated in Figs. 5 and 6, the relationship with mother-in-law demonstrated the highest overall feature contribution, followed by current appearance acceptance, little interest or pleasure in activities, sufficient money for basic needs, male gender preference, and family system. Emotional wellbeing-related variables such as feelings of hopelessness and sleep disturbances also demonstrated strong influence on maternal mental health risk prediction. In addition, previous miscarriage history and perceived social support contributed substantially to model output variation across the evaluated maternal populations. The quantitative SHAP importance analysis further demonstrated that psychosocial and emotional wellbeing variables consistently produced higher SHAP contribution magnitudes compared to several demographic and obstetric variables, indicating their stronger predictive influence within the proposed framework. The majority of high-risk maternal predictions were associated with elevated SHAP values for hopelessness, poor family relationships, financial stress, and reduced emotional wellbeing, while lower-risk predictions were generally associated with positive family support, emotional stability, and improved socioeconomic conditions. These findings confirm the important role of psychosocial and emotional wellbeing factors in maternal mental health outcomes and further demonstrate the interpretability capability of the proposed Hybrid SVM-GP framework for healthcare decision-support applications.

## 5 Discussion

This study investigated the performance of an uncertainty-aware ensemble framework for predicting maternal mental health risk using demographic, socioeconomic, obstetric, psychosocial, and clinical factors by combining Support Vector Machine (SVM) and Gaussian Process (GP) models. The study included a full-feature sensitivity analysis and a leakage-free analysis that excluded individual Edinburgh Postnatal Depression Scale (EPDS) items and directly related depression-symptom variables. This distinction was necessary to determine whether the models could predict maternal mental health risk using factors that were not directly involved in defining the outcome.

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

In the full-feature setting, Random Forest achieved the highest classification and discrimination performance, with an accuracy of 97.75%, ROC-AUC of 0.996, and PR-AUC of 0.994. These findings demonstrate the ability of ensemble tree-based methods to model complex nonlinear relationships in heterogeneous maternal mental health data. The proposed Hybrid SVM–GP framework achieved an accuracy of 93.54%, ROC-AUC of 0.982, and PR-AUC of 0.969, demonstrating strong but lower discrimination than Random Forest. CatBoost also performed well, whereas Logistic Regression produced comparatively lower but still competitive results. Logistic Regression remains an important clinical prediction benchmark because of its computational efficiency, transparent coefficients, and comparatively straightforward interpretation [43]. Nevertheless, its linear specification may limit its ability to represent complex nonlinear interactions among psychosocial, obstetric and clinical predictors. The findings are consistent with previous literature demonstrating the value of nonlinear and ensemble machine-learning approaches for maternal mental health prediction [5–9].

The strong full-feature performance should nevertheless be interpreted cautiously. The construction of the outcome was closely associated with individual EPDS items and depression symptoms such as hopelessness, little interest or pleasure in activities, and sleep disturbance. Including these variables may have enabled the models to identify existing depressive symptoms rather than predict risk from independent antecedent factors. The leakage-free analysis was therefore considered the more clinically informative analysis, while the full-feature analysis was retained as a sensitivity analysis to quantify the influence of outcome-related variables. The leakage-free results provide a more credible assessment of the predictive contribution of demographic, socioeconomic, obstetric, household, clinical and non-outcome-defining psychosocial factors.

Random Forest achieved the strongest discrimination, whereas the Hybrid SVM–GP framework achieved the lowest Expected Calibration Error (ECE) of 0.015. The calibration curve of the Hybrid SVM-GP remained close to the ideal calibration line, indicating good agreement between predicted probabilities and observed outcome frequencies within the selected calibration bins. However, Random Forest achieved lower Brier score and log loss than the Hybrid SVM-GP framework. These findings do not support a claim that the Hybrid SVM-GP was universally better calibrated or probabilistically superior. Instead, the Hybrid SVM-GP demonstrated better bin-based agreement according to ECE, while Random Forest demonstrated better probability accuracy according to the Brier score and log loss. Calibration and probabilistic performance should therefore be interpreted using multiple complementary measures rather than ECE alone.

Paired statistical comparisons provided further context for the observed performance differences. Random Forest significantly outperformed the Hybrid SVM-GP framework in accuracy, F1-score, ROC-AUC and PR-AUC and achieved significantly lower Brier score and log loss. In contrast, most differences between the Hybrid SVM-GP framework and CatBoost were small and not statistically significant, except for the lower ECE achieved by the hybrid framework. The Hybrid SVM-GP framework also significantly outperformed Logistic Regression on the principal discrimination and probability-performance measures. These findings indicate that statistically significant differences do not necessarily establish clinical superiority because their practical importance depends on outcome prevalence, classification thresholds, referral resources and the consequences of false-positive and false-negative predictions. The reported confidence

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

intervals and adjusted significance tests should therefore be interpreted together with the absolute differences in performance.

The principal complementary contribution of the Hybrid SVM-GP framework is its ability to provide case-level uncertainty information alongside the predicted class and risk probability. Random Forest may be preferred when the primary objective is to maximise classification accuracy and discrimination. By comparison, the Hybrid SVM-GP may offer complementary information when it is important to distinguish relatively confident predictions from ambiguous cases that may require additional assessment. This capability should not be interpreted as superior uncertainty estimation because comparable case-level uncertainty measures were not evaluated for all models. Rather, the framework demonstrates how uncertainty information can accompany conventional risk predictions in maternal mental health assessment.

A high-risk prediction accompanied by low entropy could potentially support prioritisation for further mental health screening or referral. Conversely, a prediction near the classification threshold with high entropy could be considered for additional EPDS assessment, clinical interview, closer monitoring or specialist review. Similarly, a low-risk prediction accompanied by high entropy should not be regarded as firm reassurance. These proposed uses remain hypothetical because the study did not determine a clinically validated uncertainty threshold or prospectively evaluate whether uncertainty-informed decisions improve referral practices, resource allocation or maternal outcomes. The framework is therefore intended to support, rather than replace, professional clinical judgement.

A high-risk prediction associated with low uncertainty could potentially support prioritisation for further mental health screening or referral. Conversely, a prediction near the classification threshold with high uncertainty could be flagged for additional EPDS assessment, clinical interview or specialist review instead of being automatically classified. Similarly, a low-risk prediction with high uncertainty should not be regarded as firm reassurance and may warrant continued monitoring. Such uncertainty estimates could help direct limited specialist resources towards difficult or ambiguous cases in resource-constrained settings. However, the framework is intended to support rather than replace professional clinical judgement, and these proposed applications require prospective clinical evaluation.

The ROC and Precision-Recall analyses confirmed the strong predictive ability of the evaluated models. Random Forest achieved the highest ROC-AUC and PR-AUC values, while the Hybrid SVM–GP framework demonstrated strong discrimination with a ROC-AUC of 0.982 and PR-AUC of 0.969. Precision–Recall performance is particularly important in maternal healthcare because high-risk mothers must be identified accurately to facilitate timely psychological assessment and mental health support [10, 43]. Model selection should nevertheless consider not only discrimination but also calibration, predictive uncertainty, available resources and the clinical consequences of false-positive and false-negative classifications.

The SHAP-based explainability analysis provided additional insight into factors influencing model predictions. Little interest or pleasure in activities, hopelessness and sleep disturbance were among the most influential variables in the full-feature setting. Because these symptoms were directly related to the EPDS-derived outcome, their importance may partly reflect target leakage. The principal interpretation was therefore

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

based on the leakage-free analysis. After controlling for outcome-defining depression symptoms, relationship with the mother-in-law, current appearance acceptance, financial sufficiency, male-child preference, family system, perceived social support and previous miscarriage history remained important contributors. These findings agree with evidence associating family relationships, socioeconomic conditions, reproductive experiences and social support with maternal mental health [1, 2]. However, SHAP values describe contributions to model predictions and should not be interpreted as evidence of causal relationships.

The predictive uncertainty analysis showed that most final hybrid predictions were associated with relatively low normalised predictive entropy, whereas higher entropy was concentrated among cases with probabilities near the classification threshold. Predictive entropy quantified ambiguity in the final fused classification, while the GP posterior variance provided a separate and complementary measure of uncertainty in the latent-function estimate. These measures should not be treated as interchangeable. The analysis demonstrates that the Hybrid SVM-GP framework can provide case-level uncertainty information in addition to predicted risk probabilities, but it does not establish that its uncertainty estimation is superior to that of other models. The study also did not prospectively assess whether uncertainty-guided screening or referral improves clinical decisions, resource utilisation or maternal outcomes. The clinical value of the uncertainty information therefore requires confirmation through prospective workflow, usability and decision-impact studies [11].

The study used a pooled multicountry dataset comprising retrospective records from Kawempe National Referral Hospital and Mukono General Hospital in Uganda and secondary data from the PERI_DEP dataset. Pooling these sources increased the demographic, psychosocial and healthcare-setting diversity available for model development. However, diversity within a pooled development dataset should not be interpreted as evidence of cross-population generalisability. Because the held-out test set was drawn from the integrated dataset, the evaluation constituted held-out internal validation rather than independent external validation.

Differences in population characteristics and healthcare systems may affect model transportability. The prevalence and distribution of maternal characteristics, pregnancy-related factors, socioeconomic conditions and maternal mental health risk may vary across populations. Family relationships, social support, male-child preference and other psychosocial variables may also have different meanings across cultural contexts. In addition, language, stigma, screening practices and access to maternal mental health services may influence EPDS responses and referral patterns. Differences in data-collection procedures, missing-data patterns and record completeness may introduce additional distribution shifts. Moreover, PERI_DEP contributed 77% of the pooled observations and may therefore have influenced the learned model patterns more strongly than the two Ugandan hospital datasets.

Variable harmonisation improved the comparability of the retained predictors but could not eliminate all population-level, cultural and healthcare-system differences. Consequently, strong performance within the pooled dataset does not guarantee equivalent performance in a new hospital, country or patient population. The findings should therefore be interpreted as internally validated evidence requiring confirmation in independent settings before clinical deployment.

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

Independent external validation should apply the locked model, without initial retraining or modification, to data from a new hospital, geographical population or later period. Such validation should assess discrimination, calibration, predictive uncertainty, sensitivity, specificity, subgroup performance and clinical usefulness. Results should first be reported for the original model. If performance deterioration or calibration drift is subsequently identified, recalibration or model updating may then be considered. Source-specific and internal–external validation, in which each data source is treated in turn as an unseen validation population, would also help determine whether the model performs consistently across the contributing populations.

The study was not prospectively designed or reported according to the TRIPOD+AI framework [44], and a formal risk-of-bias and applicability assessment using PROBAST+AI was not conducted [45]. Nevertheless, the revised manuscript reports the principal elements of prediction-model development, including data sources, participant records, predictor and outcome definitions, missing-data management, model development, internal validation, target-leakage control, calibration, uncertainty estimation and performance evaluation. This strengthened reporting should not be interpreted as formal retrospective compliance with TRIPOD+AI. Future external validation studies should prospectively apply TRIPOD+AI and PROBAST+AI to improve reporting completeness and systematically assess risk of bias and applicability across participants, predictors, outcomes and analytical procedures.

Several limitations should therefore be acknowledged. First, the full-feature experiment included variables closely related to the EPDS-derived outcome, which may have inflated performance estimates; this concern was examined through the leakage-free analysis. Second, residual population, cultural and healthcare-system differences may remain between the Ugandan and PERI_DEP data despite variable harmonisation. Third, the larger contribution of PERI_DEP records may have disproportionately influenced the pooled results. Fourth, the study used structured retrospective data without longitudinal maternal health trajectories or multimodal clinical information. Fifth, although paired statistical tests quantified whether performance differences were unlikely to have arisen by chance within the held-out test set, they do not establish external transportability or clinical superiority. Sixth, although the hybrid architecture addresses the absence of native uncertainty estimation in the standalone SVM, it does not eliminate the computational demands, covariance-kernel sensitivity or hyperparameter dependence associated with GP inference. Finally, no independent external validation, prospective clinical evaluation, clinician-usability assessment or decision-curve analysis was conducted. The proposed clinical applications therefore remain hypothetical and should not be interpreted as evidence of demonstrated clinical benefit.

Future research should externally and prospectively validate the framework using independent maternal healthcare datasets from additional countries and clinical settings. Such investigations should examine geographical and temporal transportability, subgroup performance, calibration drift and the need for setting-specific recalibration. Future prediction-model studies should also follow TRIPOD+AI reporting recommendations and apply PROBAST+AI during study planning and assessment. Longitudinal maternal health trajectories, multimodal clinical information and temporal modelling approaches could strengthen genuine early-risk prediction. Clinical usefulness should ultimately be evaluated through decision-curve analysis, clinician-usability studies and

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

prospective investigations assessing whether uncertainty-informed predictions improve screening, referral decisions and maternal healthcare outcomes [4, 11, 44–46].

## 6 Conclusion

This study developed and internally evaluated an uncertainty-aware ensemble framework integrating Support Vector Machine and Gaussian Process models for maternal mental health risk prediction using pooled data from Ugandan hospitals and the publicly available PERI_DEP dataset. The framework incorporated leakage-controlled preprocessing, feature selection, Kernel Principal Component Analysis, class-imbalance management, probability calibration and stratified cross-validation. A full-feature sensitivity analysis and a leakage-free analysis that excluded individual Edinburgh Postnatal Depression Scale items and directly related depression-symptom variables were conducted.

In the full-feature analysis, Random Forest achieved the strongest classification and discrimination performance, with an accuracy of 97.75% and a Receiver Operating Characteristic Area Under the Curve of 0.996. The Hybrid SVM-GP achieved an accuracy of 93.54%, a Receiver Operating Characteristic Area Under the Curve of 0.982 and the lowest Expected Calibration Error of 0.015. However, Random Forest achieved lower Brier score and log loss, indicating better probability accuracy according to these measures. The findings therefore do not establish universal superiority of the Hybrid SVM–GP. Its potential complementary value lies in providing case-level uncertainty information alongside predicted risk probabilities. Normalised predictive entropy quantified uncertainty in the final fused classification, while GP posterior variance provided a separate measure of uncertainty in the latent-function estimate.

The leakage-free analysis showed that family relationships, socioeconomic conditions, social support and reproductive history remained influential contributors to model predictions after excluding symptoms directly involved in defining the outcome. This finding demonstrates the potential predictive relevance of non-outcome-defining factors while emphasising the importance of controlling target leakage. However, these model-derived associations should not be interpreted as evidence of causal relationships.

Pooling data from Uganda and the PERI_DEP dataset increased the diversity available for model development but did not establish cross-population generalisability. The retrospective design, residual heterogeneity among data sources, unequal population contributions and absence of independent external validation limit the immediate clinical applicability of the framework. Furthermore, the proposed uncertainty-guided applications were not prospectively evaluated and should not be interpreted as evidence of demonstrated clinical benefit. Before routine implementation, the locked model should undergo independent geographical, temporal and prospective validation, followed where necessary by setting-specific recalibration. Decision-curve analysis, subgroup evaluation, clinician-usability assessment and clinical-impact studies are also needed to determine whether uncertainty-informed predictions improve maternal mental health screening, referral decisions and patient outcomes.

**Acknowledgements**
None.

**Author contributions**
The authors confirm their contribution to the paper as follows: study conception and design: Shallon Ahimbisibwe; data collection: Shallon Ahimbisibwe, Boaz Twinamasiko; experiments and analysis: Shallon Ahimbisibwe, Emmanuel

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

Ahishakiye; draft manuscript preparation: Shallon Ahimbisibwe. Manuscript review and editing: Emmanuel Ahishakiye, Samuel Maling, Simon Kawuma, Richard Ntwari, Fred Kaggwa. Supervision: Emmanuel Ahishakiye, Samuel Maling, Fred Kaggwa. All authors reviewed the results and approved the final version of the manuscript.

**Funding**
None.

**Data availability**
The data and code that support the findings of this study are available from the corresponding author upon reasonable request. The secondary data is also available at https://data.mendeley.com/datasets/4nznnrk8cg/2, doi: https://doi.org/10.17632/4nznnrk8cg.2.

## Declarations

**Ethics approval and consent to participate**
Ethical approval for the Ugandan component of this study was obtained from the Mbarara University of Science and Technology Research Ethics Committee (MUST-REC) under approval number MUST-2025-646. Administrative clearance was obtained from Kawempe National Referral Hospital and Mukono General Hospital before the retrieval of anonymised hospital records. The study was conducted in accordance with the Declaration of Helsinki and applicable national and institutional guidelines governing research involving human participants and human data. The publicly available PERI_DEP dataset was anonymised, and its ethical approval procedures were addressed by the original investigators before publication and secondary use. The requirement for individual informed consent for the Ugandan component was waived by the Mbarara University of Science and Technology Research Ethics Committee because the study involved retrospective analysis of anonymised records collected during routine clinical care. There was no direct contact with participants, and no identifiable personal information was accessed or reported. For the publicly available PERI_DEP dataset, informed consent and related ethical requirements were addressed by the original investigators in accordance with their approved study protocol.

**Consent to publications**
The manuscript does not contain identifiable personal data, images or other information relating to individual participants. All data used in the study were anonymised before analysis.

**Competing interests**
The authors declare no competing interests.

Received: 22 May 2026 / Accepted: 5 October 2026
Published online: 31 October 2026

**References**

1. Shorey S, Chee CYI, Ng ED, Chan YH, Tam WWS, Chong YS. Prevalence and incidence of postpartum depression among healthy mothers: a systematic review and meta-analysis. J Psychiatr Res. 2018;104:235–48. https://doi.org/10.1016/j.jpsychires.2018.08.001.
2. Liu X, Wang S, Wang G. Prevalence and risk factors of postpartum depression in women: a systematic review and meta-analysis. J Clin Nurs. 2022;31(19–20):2665–77. https://doi.org/10.1111/jocn.16121.
3. Choi SK, Kim JJ, Park YG, Ko HS, Park IY, Shin JC. The simplified Edinburgh Postnatal Depression Scale (EPDS) for antenatal depression: is it a valid measure for pre-screening? Int J Med Sci. 2011;9(1):40. https://doi.org/10.7150/ijms.9.40.
4. Hahn-Holbrook J, Cornwell-Hinrichs T, Anaya I. Economic and health predictors of national postpartum depression prevalence: a systematic review, meta-analysis, and meta-regression of 291 studies from 56 countries. Front Psych. 2018. https://doi.org/10.3389/fpsyt.2017.00248.
5. Saqib K, Khan AF, Butt ZA. Machine learning methods for predicting postpartum depression: scoping review. JMIR Ment Health. 2021;8(11):e29838. https://doi.org/10.2196/29838.
6. Shin D, Hur J. Predictive modeling of postpartum depression using machine learning approaches (P18–130-19). Curr Dev Nutr. 2019;3:nzz039.P18. https://doi.org/10.1093/cdn/nzz039.p18-130-19.
7. Huang X, Zhang L, Zhang C, Li J, Li C. Postpartum depression risk prediction using explainable machine learning algorithms. Front Med. 2025;12:1565374. https://doi.org/10.3389/fmed.2025.1565374.
8. Ma Z, Horvath M, Stamilio DM, Sekyere K, Gurcan MN. Building a machine learning model to predict postpartum depression from electronic health records in a tertiary care setting. J Clin Med. 2025. https://doi.org/10.3390/jcm14186644.
9. Qi W, et al. Prediction of postpartum depression in women: development and validation of multiple machine learning models. J Transl Med. 2025;23(1):291. https://doi.org/10.1186/S12967-025-06289-6/FIGURES/5.
10. Zhang T. An introduction to support vector machines and other kernel-based learning methods. AI Mag. 2001. https://doi.org/10.1609/AIMAG.V22I2.1566.
11. Rasmussen CE, Williams CKI. Gaussian processes for machine learning. London: The MIT Press; 2006. https://doi.org/10.1142/S0129065704001899.
12. Bauer AE, et al. Identification and validation of postpartum depression subtypes: a population-based cohort study. eClinicalMedicine. 2025;89:103540. https://doi.org/10.1016/J.ECLINM.2025.103540.
13. Stein Elger R, et al. The role of social support in shaping maternal experiences during the postpartum hospital stay: a qualitative study. Women's Heal. 2025;21:17455057251385368. https://doi.org/10.1177/17455057251385367.
14. Ruger-Navarrete A, et al. Artificial intelligence in the prevention and early detection of postpartum depression: a systematic review and meta-analysis. Front Psychiatry. 2026;16:1734102. https://doi.org/10.3389/FPSYT.2025.1734102.
15. Schölkopf B, Smola AJ (2002). Learning with kernels: Support vector machines, regularization, optimization, and beyond. MIT Press. https://doi.org/10.7551/mitpress/4175.001.0001

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

16. Zoppis I, Mauri G, Dondi R. Kernel machines: applications. Encycl Bioinforma Comput Biol ABC Bioinforma. 2018;1–3:511–8. https://doi.org/10.1016/B978-0-12-809633-8.20343-9.
17. Du KL, Jiang B, Lu J, Hua J, Swamy MNS. Exploring kernel machines and support vector machines: principles, techniques, and future directions. Math. 2024;12(24):3935. https://doi.org/10.3390/MATH12243935.
18. Noble WS. What is a support vector machine? Nat Biotechnol. 2006;24(12):1565–7. https://doi.org/10.1038/nbt1206-1565.
19. Wang Q, et al. A hybrid SVM and kernel function-based sparse representation classification for automated epilepsy detection in EEG signals. Neurocomputing. 2023;562:126874. https://doi.org/10.1016/J.NEUCOM.2023.126874.
20. Shiao HT, et al. SVM-based system for prediction of epileptic seizures from iEEG signal. IEEE Trans Biomed Eng. 2016;64(5):1011. https://doi.org/10.1109/TBME.2016.2586475.
21. Lawrence N (2025) Probabilistic non-linear principal component analysis with gaussian process latent variable models. J Mach Learn Res, 6(60):1783–1816 http://jmlr.org/papers/v6/lawrence05a.html
22. Tomczak JM. Gaussian process regression with categorical inputs for predicting the blood glucose level. Adv Intell Syst Comput. 2017;539:98–108. https://doi.org/10.1007/978-3-319-48944-5_10.
23. Nanglia P, Kumar S, Mahajan AN, Singh P, Rathee D. A hybrid algorithm for lung cancer classification using SVM and neural networks. ICT Express. 2021;7(3):335–41. https://doi.org/10.1016/J.ICTE.2020.06.007.
24. Álvarez MA, Rosasco L, Lawrence ND. Kernels for vector-valued functions: a review. Found Trends Mach Learn. 2011;4(3):195–266. https://doi.org/10.1561/2200000036.
25. Shin D, Lee KJ, Adeluwa T, Hur J. Machine learning-based predictive modeling of postpartum depression. J Clin Med. 2020;9(9):1–14. https://doi.org/10.3390/jcm9092899.
26. Hochman E, et al. Development and validation of a machine learning-based postpartum depression prediction model: a nationwide cohort study. Depress Anxiety. 2021;38(4):400–11. https://doi.org/10.1002/DA.23123.
27. Andersson S, Bathula DR, Iliadis SI, Walter M, Skalkidou A. Predicting women with depressive symptoms postpartum with machine learning methods. Sci Rep. 2021;11(1):1–15. https://doi.org/10.1038/s41598-021-86368-y.
28. Zhang R, Liu Y, Zhang Z, Luo R, Lv B. Interpretable machine learning model for predicting postpartum depression: retrospective study. JMIR Med Inform. 2025;13(1):e58649. https://doi.org/10.2196/58649.
29. Lilhore UK, et al. Prevalence and risk factors analysis of postpartum depression at early stage using hybrid deep learning model. Sci Rep. 2024;14(1):1–24. https://doi.org/10.1038/S41598-024-54927-8.
30. Nakamura Y, Ueno T, Takahashi N, Ichikawa D, Yamauchi A, Ozaki N. Early identification of postpartum depression using machine learning. Psychiatry Clin Neurosci. 2024;78(6):372–3. https://doi.org/10.1111/PCN.13659.
31. Kim H. Predictive analysis of postpartum depression using machine learning. Healthc. 2025;13(8):897. https://doi.org/10.3390/HEALTHCARE13080897.
32. Huang Y, Alvernaz S, Kim SJ, Maki P, Dai Y, Peñalver Bernabé B. Predicting prenatal depression and assessing model bias using machine learning models. Biol Psychiatry Glob Open Sci. 2024;4(6):100376. https://doi.org/10.1016/j.bpsgos.2024.100376.
33. Wakefield C, Frasch MG. Predicting patients requiring treatment for depression in the postpartum period using common electronic medical record data available antepartum. AJPM Focus. 2023;2(3):1–8. https://doi.org/10.1016/j.focus.2023.100100.
34. Zafar A, Akram BA, Wasim M, Pires IM, Coelho PJ. PERI_DEP: A dataset of mother's mental health in Pakistan. Data Br. 2025;60:111621. https://doi.org/10.1016/j.dib.2025.111621.
35. Cox JL, Holden JM, Sagovsky R. Detection of postnatal depression. Development of the 10-item Edinburgh Postnatal Depression Scale. Br J Psychiatry. 1987;150(JUNE):782–6. https://doi.org/10.1192/bjp.150.6.782.
36. Kabunga A, Tumwesigye R, Kigongo E, Musinguzi M, Acup W, Auma AG. Systematic review and meta-analysis of postpartum depression and its associated factors among women before and after the COVID-19 pandemic in Uganda. BMJ Open. 2024;14(7):e076847. https://doi.org/10.1136/BMJOPEN-2023-076847.
37. Stewart DE, Vigod SN. Postpartum depression: pathophysiology, treatment, and emerging therapeutics. Annu Rev Med. 2019;70:183–96. https://doi.org/10.1146/annurev-med-041217-011106.
38. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321–57. https://doi.org/10.1613/jair.953.
39. Goodfellow I, Bengio Y, Courville A (2016) Deep learning.
40. Géron A. Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow, third edit. O'Reilly; 2023.
41. Murphy KP. Machine learning: a probabilistic perspective. London: The MIT Press; 2012. https://doi.org/10.1007/978-94-011-3532-0_2.
42. Fernández, A., Garcia, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. Journal of artificial intelligence research, 61, 863-905. https://doi.org/10.1613/jair.1.11192.
43. Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning. https://doi.org/10.1007/978-0-387-84858-7.
44. Collins GS, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024. https://doi.org/10.1136/BMJ-2023-078378.
45. Moons KGM, et al. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ. 2025. https://doi.org/10.1136/BMJ-2024-082505.
46. Panda NR, Pati JK, Mohanty JN, Bhuyan R. A review on logistic regression in medical research. Natl J Commun Med. 2022;13(4):265–70. https://doi.org/10.55489/NJCM.134202222.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

## Terms and Conditions

Springer Nature journal content, brought to you courtesy of Springer Nature Customer Service Center GmbH ("Springer Nature").

Springer Nature supports a reasonable amount of sharing of research papers by authors, subscribers and authorised users ("Users"), for small-scale personal, non-commercial use provided that all copyright, trade and service marks and other proprietary notices are maintained. By accessing, sharing, receiving or otherwise using the Springer Nature journal content you agree to these terms of use ("Terms"). For these purposes, Springer Nature considers academic use (by researchers and students) to be non-commercial.

These Terms are supplementary and will apply in addition to any applicable website terms and conditions, a relevant site licence or a personal subscription. These Terms will prevail over any conflict or ambiguity with regards to the relevant terms, a site licence or a personal subscription (to the extent of the conflict or ambiguity only). For Creative Commons-licensed articles, the terms of the Creative Commons license used will apply.

We collect and use personal data to provide access to the Springer Nature journal content. We may also use these personal data internally within ResearchGate and Springer Nature and as agreed share it, in an anonymised way, for purposes of tracking, analysis and reporting. We will not otherwise disclose your personal data outside the ResearchGate or the Springer Nature group of companies unless we have your permission as detailed in the Privacy Policy.

While Users may use the Springer Nature journal content for small scale, personal non-commercial use, it is important to note that Users may not:

1. use such content for the purpose of providing other users with access on a regular or large scale basis or as a means to circumvent access control;
2. use such content where to do so would be considered a criminal or statutory offence in any jurisdiction, or gives rise to civil liability, or is otherwise unlawful;
3. falsely or misleadingly imply or suggest endorsement, approval , sponsorship, or association unless explicitly agreed to by Springer Nature in writing;
4. use bots or other automated methods to access the content or redirect messages
5. override any security feature or exclusionary protocol; or
6. share the content in order to create substitute for Springer Nature products or services or a systematic database of Springer Nature journal content.

In line with the restriction against commercial use, Springer Nature does not permit the creation of a product or service that creates revenue, royalties, rent or income from our content or its inclusion as part of a paid for service or for other commercial gain. Springer Nature journal content cannot be used for inter-library loans and librarians may not upload Springer Nature journal content on a large scale into their, or any other, institutional repository.

These terms of use are reviewed regularly and may be amended at any time. Springer Nature is not obligated to publish any information or content on this website and may remove it or features or functionality at our sole discretion, at any time with or without notice. Springer Nature may revoke this licence to you at any time and remove access to any copies of the Springer Nature journal content which have been saved.

To the fullest extent permitted by law, Springer Nature makes no warranties, representations or guarantees to Users, either express or implied with respect to the Springer nature journal content and all parties disclaim and waive any implied warranties or warranties imposed by law, including merchantability or fitness for any particular purpose.

Please note that these rights do not automatically extend to content, data or other material published by Springer Nature that may be licensed from third parties.

If you would like to use or distribute our Springer Nature journal content to a wider audience or on a regular basis or in any other manner not expressly permitted by these Terms, please contact Springer Nature at

onlineservice@springernature.com