Research Square

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

Comparative machine learning for prediction of probable depression among people living with HIV in a rural Ghanaian antiretroviral therapy programme: development, internal validation and SHAP-based interpretation

Vincent Uwumboriyhie Gmayinaam
vugmayinaam@uhas.edu.gh
University of Health and Allied Sciences
David Naboare
University of Health and Allied Sciences
Georgina Oduro Yeboah
University of Health and Allied Sciences
Patrick Elorm Dzissem
University of Health and Allied Sciences
Edem Kagbetor Kodzo
University of Health and Allied Sciences
Alfred Kwesi Manyeh
University of Health and Allied Sciences
AbdulZeid Yen Anafo
University of Mines and Technology

Research Article

Keywords: machine learning, probable depression, people living with HIV, PHQ-9, SHAP, clinical prediction model

Posted Date: November 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11154327/v1

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

Additional Declarations: No competing interests reported.

Comparative machine learning for prediction of probable depression among people living with HIV in a rural Ghanaian antiretroviral therapy programme: development, internal validation and SHAP-based interpretation

Vincent Uwumboriyhie Gmayinaam[1,2,*], David Naboare[3], Georgina Oduro Yeboah[1], Patrick Elorm Dzissem[1], Edem Kagbetor Kodzo[1], Alfred Kwesi Manyeh[1], AbdulZeid Yen Anafo[4]

[1]Institute of Health Research, University of Health and Allied Sciences, Ho, Volta Region, Ghana
[2]Department of Epidemiology and Biostatistics, Fred Newton Binka School of Public Health, University of Health and Allied Sciences, Hohoe, Volta Region, Ghana
[3]Department of Population and Behavioural Sciences, Fred Newton Binka School of Public Health, University of Health and Allied Sciences, Hohoe, Volta Region, Ghana
[4]Department of Mathematical Sciences, University of Mines and Technology, Tarkwa, Ghana

*Correspondence: Vincent Uwumboriyhie Gmayinaam, Institute of Health Research, University of Health and Allied Sciences, PMB 31, Ho, Volta Region, Ghana. E-mail: (vugmayinaam@uhas.edu.gh)

Emails:

VUG: vugmayinaam@uhas.edu.gh

GOY: georginaoduroyeboah@gmail.com

PED: pedzissem@uhas.edu.gh

DN: davidnaboare2@gmail.com

EKK: eddyboltzmann@gmail.com

AKM: akmanyeh@uhas.edu.gh

AYA: ayanafo@umat.edu.gh

Abstract

Background: Depressive symptoms are common among people living with HIV (PLHIV) and are associated with poorer antiretroviral therapy (ART) adherence and viral suppression. In rural settings, routine symptom screening may be constrained by counsellor workload and limited specialist mental health capacity. This study developed and internally evaluated ten supervised machine-learning (ML) models for identifying PHQ-9-defined probable depression among PLHIV attending ART clinics in Kadjebi District, Ghana.

Methods: A facility-based cross-sectional dataset of 264 adults receiving ART was analysed. Probable depression was defined as a Patient Health Questionnaire-9 (PHQ-9) score of 10 or above. Fifty-two predictors were engineered from sociodemographic, HIV-clinical, tuberculosis-service, physical-health, psychosocial, stigma and health-system variables; the nine PHQ-9 items were excluded from predictors. Data were partitioned 80:20 with stratification. Imputation and, where required, standardisation were fitted using the training partition, and SMOTE was applied only to the training data. Ten algorithms were tuned using five-fold cross-validation within the training partition. Performance on the untouched held-out test set was treated as the primary split-sample internal validation, with 5-fold x 4-repeat cross-validation of the training partition used as a secondary stability analysis. Performance metrics included AUC, sensitivity, specificity, precision, F1 score, Cohen's kappa, Brier score and calibration slope. Interpretability used permutation importance and SHapley Additive exPlanations (SHAP); decision-curve analysis was exploratory.

Results: PHQ-9-defined probable depression was present in 100 of 264 participants (37.9%). Random Forest was selected by the training-set cross-validation results. On the held-out test set, it achieved an AUC of 0.870, accuracy of 77.4%, sensitivity of 80.0%, specificity of 75.8%, F1 score of 0.727, Cohen's kappa of 0.536 and Brier score of 0.137. Its secondary repeated-cross-validation AUC was 0.893 (SD 0.041). Differences in cross-validated AUC among the leading algorithms were small relative to fold-to-fold variability, and no formal pairwise superiority testing was performed. Permutation importance and SHAP consistently highlighted comfort disclosing psychological difficulty, belief system, counselling and tailored mental-health service availability, perceived healthcare access and quality, sex, and perceived HIV impact. Exploratory decision-curve analysis suggested potential net benefit of model-guided PHQ-9 administration; at a threshold of 0.40, 48.9% of attendees would receive the PHQ-9 while 94% of PHQ-9-positive cases would be identified. This calculation concerns PHQ-9 administrations only and does not quantify total workflow burden. At the PHQ-9 >=5 sensitivity threshold, Random Forest had a secondary cross-validated AUC of 0.911.

Conclusion: Random Forest showed promising discrimination for PHQ-9-defined probable depression in this small, single-district development study, but the findings remain proof-of-concept. Relational and service-related variables contributed strongly to prediction. Threshold analyses suggest that risk stratification could reduce the number of PHQ-9 administrations, but total counselling burden and safety were not established. External validation, development of a more parsimonious routine-workflow model, and prospective evaluation are required before clinical use; a prediction model should not replace assessment of suicide or self-harm risk.

Keywords: machine learning; probable depression; people living with HIV; PHQ-9; SHAP; clinical prediction model

Background

Depression is the most frequently occurring mental disorder among people living with HIV (PLHIV) and carries consequences that extend across the whole HIV care continuum. A meta-analysis of 16 studies and 80,103 participants found that people without depressive symptoms were significantly more likely to achieve viral suppression than those with them [1], and treatment of depression improves antiretroviral therapy (ART) adherence [2]. Detection of depression is therefore not a peripheral concern of HIV programmes but a determinant of their core outcomes.

The burden across sub-Saharan Africa is substantial. Bernard et al. (2017), pooling 66 studies, reported a PHQ-9-based prevalence of 14% (95% CI 12-16) among ART-treated populations [3], while a more recent synthesis of African studies reported a pooled prevalence of depressive symptoms of 33.32% (95% CI 30.00-36.65), with female sex, perceived stigma, comorbid chronic illness, low CD4 count and poor social support as consistent determinants [4]. In Ghana, published evidence is sparse; John et al. (2024) reported depressive symptoms in 20.4% of 181 PLHIV in the Volta Region [5] .

Task-shared mental-health interventions show that non-specialists can deliver effective care in resource-constrained settings [6,7]. A separate challenge is systematic case identification in high-volume ART clinics. Universal PHQ-9 administration provides consistent symptom ascertainment but can add consultation time where staffing is limited. Risk stratification may be useful only if predictor information is already available or can be collected with less burden than the screening instrument itself. We therefore evaluated ML as a proof-of-concept prioritisation approach rather than as a replacement for clinical assessment.

This is a prediction problem, and machine learning (ML) has been applied to analogous public health triage questions in the region with encouraging results. Alemayehu (2024) compared eight algorithms to predict measles vaccination dropout using Ethiopian Demographic and Health Survey data, with XGBoost achieving 73.9% accuracy and an AUC of 0.813 [8]. Demsash et al. (2023) applied eight algorithms to childhood vaccination in Ethiopia, with the PART rule learner reaching 95.53% accuracy [9]. In the mental health domain, Vu et al. (2025) compared six algorithms for depressive disorder using NHANES data, with XGBoost best at an AUC of 0.69, and SHAP identifying poverty-income ratio, sex and hypertension as leading contributors [10]. Comparable work in an African HIV treatment population has not, to our knowledge, been published.

Two methodological features distinguish credible prediction studies from the larger number that are not reproducible. The first is that discrimination is reported alongside calibration: a model whose ranking is good but whose probabilities are systematically wrong cannot support a risk threshold, and calibration has been described as the Achilles heel of predictive analytics [11] . The second is that clinical utility is quantified, typically by decision-curve analysis, which expresses whether acting on the model beats the default strategies of treating everyone or no one [12]. Systematic review has also established that ML confers no automatic performance advantage over logistic regression in clinical prediction [13], making a like-for-like comparison across algorithm families a necessary rather than an optional analysis.

This study therefore had four objectives: (i) to develop and internally evaluate ten supervised ML algorithms for identifying PHQ-9-defined probable depression among PLHIV attending rural Ghanaian ART clinics using variables obtainable without laboratory investigation; (ii) to compare discrimination, classification performance, agreement and calibration using an untouched held-out

test set, supplemented by repeated cross-validation as a stability analysis; (iii) to identify features contributing to predictions using permutation importance and SHapley Additive exPlanations; and (iv) to explore decision-curve net benefit and hypothetical PHQ-9 administration at candidate risk thresholds without assuming a reduction in total clinic workload.

Methods and materials

Study design and setting

This study used data from a facility-based cross-sectional survey conducted in February 2024 in the Kadjebi District, Oti Region, Ghana. The district covers approximately 949 km² with a population of 73,959 [14] and is predominantly agrarian, with subsistence farming and small-scale trading as the main livelihoods. Two facilities provide HIV treatment services in the district: Mary Theresa Hospital, Papase, and the Kadjebi Health Centre. No psychiatrist is posted to the district; mental health care is delivered by community and mental health nurses within general services. Reporting follows the TRIPOD+AI statement [15].

Study population and data source

The source population comprised all adults aged 18 years and over living with HIV and receiving ART at the two district facilities. Participants who were acutely unwell and required immediate clinical attention were excluded, both because participation would have delayed care and because acute illness confounds the somatic items of the PHQ-9. The analytical dataset comprised 264 participants and 78 variables.

Sampling size and sampling procedure

District Health Management Information System records indicated that 461 adults living with HIV were receiving care at the two participating facilities. The minimum sample for the parent cross-sectional survey was estimated using Yamane's finite-population formula, $n = N/[1 + N(e^2)]$, with $N = 461$ and a 5% precision level, giving 215 participants. To allow for an anticipated 20% non-response, the recruitment target was adjusted to 269 participants (215/0.80). A total of 264 participants completed the study, representing 98.1% of the adjusted recruitment target.

Updated lists of adults receiving ART at the two facilities formed the sampling frame. A computerised random-number procedure in Microsoft Excel was used to select 269 potential participants. Eligible selected participants were approached at the participating facilities, and 264 completed the study procedures. Clinical information was extracted using KoboCollect, and the structured questionnaire was administered face-to-face with support from facility staff.

Study variables

Target variable

The target variable was PHQ-9-defined probable depression, coded 1 for a PHQ-9 score >=10 and 0 for a score <10. The PHQ-9 comprises nine items scored from 0 (not at all) to 3 (nearly every day) over the preceding two weeks, giving a total score from 0 to 27 [16]. A score >=10 was used to identify probable depression, consistent with the conventional threshold for clinically significant depressive symptoms [16]. A sensitivity analysis repeated the modelling pipeline using a threshold

of >=5, representing at least mild depressive symptoms. Internal consistency of the nine items in this sample was good (Cronbach alpha = 0.839).

Features

Fifty-two encoded features were derived from seven domains: sociodemographic (18), HIV clinical (4), tuberculosis services (4), physical health and behaviour (7), health-system and services (8), cultural and stigma (6), and psychological non-PHQ items (5). WHO clinical stage and CD4 count were represented as documentation/knowledge indicators because 76.1% and 92.0% of participants, respectively, did not know these values. The nine PHQ-9 items, PHQ-9 total and severity category were excluded from the predictor set to prevent direct target leakage. Nominal variables were one-hot encoded with the first level dropped; binary and ordered variables were retained as integers. The complete encoded feature dictionary is presented in Table 1.

Modelling environment

All analyses were conducted in Python 3.11. Data manipulation used pandas and NumPy; model development, tuning and evaluation used scikit-learn [17]; gradient-boosting implementations used xgboost, lightgbm and catboost; class-imbalance handling used imbalanced-learn; interpretability used shap; and figures were produced with Matplotlib. The analysis code is publicly available in the GitHub repository specified in the Data and code availability section.

Data pre-processing

Pre-processing followed five steps. First, the dataset was inspected for exact duplicate records; three were identified and retained because identical questionnaire response patterns can occur, and sensitivity checks indicated that their removal did not materially alter results. Second, features with more than 50% missingness were flagged for removal; no encoded feature exceeded this threshold. Third, remaining missing values were imputed using the median estimated from the 80% training partition, and the fitted imputer was applied unchanged to the held-out test partition. Fourth, continuous and ordinal features were standardised using parameters estimated from the training partition for scale-sensitive algorithms; tree-based learners used unstandardised values. Fifth, the dataset was split into training (80%, n=211, 80 events) and held-out test (20%, n=53, 20 events) partitions with outcome stratification and a fixed random seed. Imputation and scaling were fitted before the internal training-set cross-validation rather than re-estimated within each cross-validation fold; consequently, cross-validation results are treated as model-selection/stability estimates rather than as the primary unbiased performance estimate. The held-out test partition was not used during pre-processing, tuning or model selection (Figure 1).

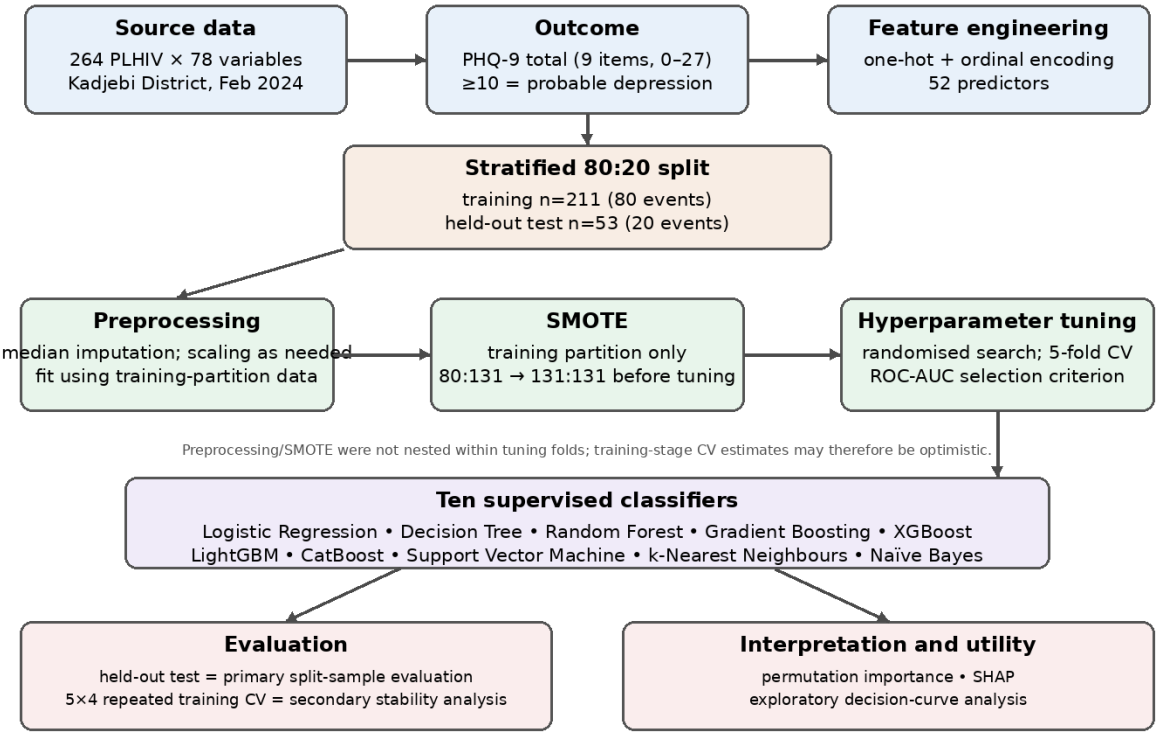


Figure 1: Analytical workflow for the development, internal validation and interpretation of the machine learning models. PLHIV=people living with HIV; SMOTE=Synthetic Minority Over-sampling Technique; CV=cross-validation; SHAP=SHapley Additive exPlanations.

Correlation assessment and exploratory feature reduction

Questionnaire-derived predictors can contain near-redundant information that affects coefficient stability and feature-attribution methods. We therefore examined pairwise dependence and performed an exploratory correlation-clustering sensitivity analysis to assess whether a smaller feature set produced materially different results. This analysis was not used to define the primary model.

The Spearman rank-correlation matrix was computed across all 52 features (Figure S1A). Correlation was converted to distance as d(i,j) = 1 - |rho(i,j)|, and Ward's minimum-variance linkage was used to form feature clusters, cut at a distance of 0.7 (Figure S1B). One representative per cluster was retained using the largest absolute Wald z statistic from a univariable logistic regression against the outcome, yielding 33 features. Because the representative within each cluster was selected using outcome information from the full dataset rather than within resampling folds, all reduced-feature results are explicitly exploratory and may be optimistic.

Variance inflation factors (VIF) were computed on the retained set as $VIF_j = 1 / (1 - R^2_j)$, where R^2_j is the coefficient of determination from regressing feature j on all the remaining features. Values above 5 are conventionally taken to indicate problematic collinearity and above 10 severe collinearity.

The modelling pipeline was repeated on the reduced 33-feature set as an exploratory sensitivity analysis. These reduced-set results were used to assess robustness and parsimony only; they were

not used as the primary internal-validation estimate or as evidence of superiority over the full-feature model.

Handling class imbalance

At the PHQ-9 >=10 threshold, the outcome distribution was 100 PHQ-9-positive to 164 PHQ-9-negative participants, with a training-partition ratio of 80:131. SMOTE [18] generated synthetic minority-class observations in the training partition, balancing it to 131:131. The held-out test partition retained its natural class distribution. In the implemented pipeline, SMOTE was applied to the training partition before the five-fold hyperparameter-search folds rather than separately within each fold. This can make training-set cross-validation estimates optimistic, so those estimates are used for model selection and stability description only. The untouched held-out test set provides the primary split-sample performance assessment. Because oversampling can also distort predicted probabilities [19], calibration and Brier score were reported.

Machine learning algorithms and their operationalisation

Ten supervised classification algorithms spanning linear/probabilistic, instance-based, kernel-based and tree-ensemble approaches were implemented. Throughout, x denotes the feature vector for a participant, y in {0,1} the observed PHQ-9-defined probable-depression status, p-hat the predicted probability, n the number of training observations and d the number of features.

Logistic Regression

Logistic regression models the log-odds of depression as a linear function of the predictors:

$$\text{logit}(p) = \ln\left[\, p / (1 - p) \,\right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

Coefficients were estimated by minimising the L2-penalised negative log-likelihood, which shrinks coefficients towards zero and controls variance in a high-dimensional, small-sample setting:

$$\text{minimise} \quad -\Sigma\left[\, y \ln p + (1 - y) \ln (1 - p) \,\right] + (1 / 2C)\, \Sigma\, \beta^2$$

where C is the inverse regularisation strength, tuned over the grid {0.01, 0.03, 0.1, 0.3, 1, 3, 10}. Features were z-standardised beforehand so that the penalty applied equally across predictors. This model served as the interpretable statistical benchmark against which the machine learning algorithms were judged.

Decision Tree

A classification tree recursively partitions the feature space by selecting, at each node, the feature and cut-point that maximise the reduction in node impurity. Gini impurity at node t is:

$$G(t) = 1 - \Sigma\, p_k(t)^2$$

where $p_k(t)$ is the proportion of class k at node t. The chosen split maximises $\Delta G = G(\text{parent}) - [\, (n_l/n) \,.\, G(\text{left}) + (n_r/n) \,.\, G(\text{right}) \,]$. Growth was constrained by maximum depth and minimum leaf size to limit overfitting. The tree is fully transparent but has high variance and is included principally as the single-learner baseline for the ensembles that follow.

Random Forest

Random Forest reduces the variance of a single tree by averaging over an ensemble of decorrelated trees. B bootstrap samples are drawn from the training data; a tree T^b is grown on each, with a random subset of $\sqrt{d}$ features considered at every split. The ensemble prediction is the average of the per-tree class probabilities [20]:

$$p(x) = (1 / B) \,.\, \Sigma\, T^b(x\,;\, \Theta^b), \quad b = 1 \ldots B$$

where Θ^b denotes the bootstrap sample and feature subset drawn for tree b. Averaging over decorrelated trees reduces variance without inflating bias, which is why the method performs well when d is large relative to n. The number of trees, maximum depth, minimum leaf size and the feature-subset rule were tuned.

Gradient Boosting

Gradient boosting builds the model additively, each new tree fitted to the negative gradient (the pseudo-residuals) of the loss with respect to the current prediction:

$$F_m(x) = F_{m-1}(x) + \nu \,.\, h_m(x)$$

where ν is the learning rate (shrinkage) and each h_m is a regression tree fitted to the negative gradient of the loss evaluated at the current model, that is to the pseudo-residuals $-\partial L(y, F)/\partial F$. The loss is the log-loss $L(y, F) = -[\, y \ln \sigma(F) + (1 - y) \ln(1 - \sigma(F)) \,]$, with σ the logistic function. Stochastic subsampling of rows at each iteration adds further regularisation. Learning rate, number of stages, depth and subsample fraction were tuned.

XGBoost

XGBoost optimises a regularised objective in which model complexity is penalised explicitly [21]:

$$L = \Sigma\, l(y, \hat{y}) + \Sigma\, \Omega(f), \quad \Omega(f) = \gamma T + \tfrac{1}{2}\lambda \|w\|^2$$

where T is the number of leaves, w the vector of leaf weights, γ the complexity cost per leaf and λ the L2 penalty on leaf weights. The objective is expanded to second order, giving the optimal leaf weight $w_j^* = -G_j / (H_j + \lambda)$, where G_j and H_j are the sums of first- and second-order gradients in leaf j. This second-order optimisation, combined with column subsampling and shrinkage, is what distinguishes it from classical gradient boosting.

LightGBM

LightGBM optimises the same gradient-boosting objective but differs in two implementation choices [22]. Continuous features are bucketed into histograms so that split finding is O(#bins) rather than O(n), and trees are grown leaf-wise rather than level-wise, at each step splitting the leaf with the largest expected variance gain:

$$V(j, s) = (1 / n) \,.\, [\, (\Sigma_l\, g)^2 / n_l \;+\; (\Sigma_r\, g)^2 / n_r \,]$$

where g_i are the gradients. Leaf-wise growth yields lower loss for a fixed number of leaves but overfits more readily on small samples, which is why the number of leaves and minimum child sample size were tuned conservatively here.

CatBoost

CatBoost addresses the prediction shift that arises when the same observations are used both to estimate gradients and to fit the tree that consumes them. It employs ordered boosting, in which

the model used to compute the residual for observation i is trained only on observations preceding i in a random permutation, and ordered target statistics for categorical encoding [23]:

$$\hat{x} = (\Sigma\, ^{I}[\, x_j = x_i\,] \,.\, y_j \;+\; a.P\,) \;/\; (\Sigma\, ^{I}[\, x_j = x_i\,] \;+\; a\,), \quad \text{over j with } \sigma(j) < \sigma(i)$$

where σ is the permutation, P a prior and a a smoothing weight. It also grows symmetric (oblivious) trees, in which the same split is applied across an entire level, acting as a strong regulariser.

Support Vector Machine

The support vector machine constructs the maximum-margin separating hyperplane in a feature space induced by a kernel, solving the soft-margin problem [24]:

$$\text{minimise } \tfrac{1}{2}\|w\|^2 + C \Sigma\, \xi \quad \text{subject to} \quad y\, (\, w \,.\, \varphi(x) + b\,) \geq 1 - \xi, \quad \xi \geq 0$$

A radial basis function kernel was used, which computes inner products in an infinite-dimensional space without explicit mapping:

$$K(x_i, x_j) = \exp\, (\, -\gamma\, \|x_i - x_j\|^2\,)$$

C and γ were tuned. Because the decision function is a signed distance rather than a probability, probabilities were obtained by Platt scaling; as reported in the Results, this transformation left the model badly calibrated at the default classification threshold.

k-Nearest Neighbours

k-Nearest Neighbours is a non-parametric, instance-based learner that estimates the class probability from the labels of the k closest training observations:

$$p(y = 1 \mid x) = (\Sigma\, w_i \,.\, ^{I}[\, y_i = 1\,]) \,/\, (\Sigma\, w_i\,), \quad i \in N_k(x), \quad w_i = 1 \,/\, d(x, x_i)$$

$$d(x, x_i) = [\, \Sigma\, |\, x_j - x_{ij}\, |^{\,p}\,] \;\text{ raised to the power } 1/p$$

where $N_k(x)$ is the set of k nearest neighbours under that Minkowski distance of order p. Distance weighting was applied so that closer neighbours contribute more. Features were standardised, since the method is scale-sensitive by construction. The number of neighbours, weighting scheme and distance order p were tuned.

Gaussian Naïve Bayes

Naïve Bayes applies Bayes' theorem under the simplifying assumption that features are conditionally independent given the class:

$$P(y \mid x) \propto P(y) \,.\, \prod P(x_j \mid y), \quad \text{with each } P(x_j \mid y) \text{ Gaussian of mean } \mu \text{ and variance } \sigma^2$$

with class-conditional means μ and variances σ^2 estimated from the training data, and a variance-smoothing term added for numerical stability. The conditional-independence assumption is strongly violated by the correlated questionnaire items in this dataset, which is the expected explanation for its poor calibration in the Results.

SHapley Additive exPlanations

For interpretation, the SHAP value of feature j for a given prediction is its average marginal contribution across all orderings of the feature set [25]:

$$\varphi_j = \Sigma \; [\, |S|! \,.\, (|F| - |S| - 1)! \,/\, |F|! \,] \,.\, [\, f(S \cup j) - f(S) \,], \quad \text{over all } S \subseteq F \text{ excluding } j$$

where F is the full feature set and f(S) the model output using only the features in S. SHAP values are additive, so that $f(x) = \varphi_0 + \Sigma_j \varphi_j$, with φ_0 the expected model output. Exact values were computed with TreeExplainer for the tree ensembles and LinearExplainer for the penalised logistic regression.

Hyperparameter tuning

Hyperparameters were optimised by randomised search over algorithm-specific grids, with candidate configurations evaluated by stratified five-fold cross-validation in the training data and ROC-AUC as the selection criterion. The selected configuration for each algorithm is reported in Table S1 in supplementary file S1. All model selection occurred within the 80% training partition; the held-out test set was reserved for the final split-sample evaluation. Because preprocessing and SMOTE were not nested within each tuning fold in the implemented pipeline, tuning-fold AUCs are not interpreted as unbiased estimates of out-of-sample performance.

Model performance evaluation

Performance was reported on two complementary footings. The primary internal-validation assessment was the single evaluation on the untouched held-out test set (n=53, 20 PHQ-9-positive events) after all model selection had been completed in the training partition. A secondary repeated stratified cross-validation analysis (five folds x four repeats) on the unresampled training partition was used to describe model-ranking stability. Because preprocessing parameters and tuned hyperparameters were fixed before this repeated cross-validation, these cross-validation estimates are descriptive stability measures rather than fully nested validation estimates.

Metrics were derived from the confusion matrix. Discrimination was quantified by the area under the receiver-operating-characteristic curve (AUC). Chance-corrected agreement was quantified by Cohen's kappa; probabilistic accuracy by the Brier score; and calibration by regressing the observed outcome on the predicted linear predictor in the held-out test partition. A calibration slope of 1.00 indicates ideal calibration, values below 1 generally indicate overly extreme predictions/overfitting, and values above 1 indicate predictions that are not sufficiently extreme. Given the small test set, calibration slopes were interpreted cautiously.

Definitions of the confusion-matrix elements and derived performance metrics

1. True positive (TP): a participant meeting the PHQ-9 >=10 threshold correctly classified as positive.
2. True negative (TN): a participant below the PHQ-9 >=10 threshold correctly classified as negative.
3. False positive (FP): a participant below the PHQ-9 >=10 threshold incorrectly classified as positive.
4. False negative (FN): a participant meeting the PHQ-9 >=10 threshold incorrectly classified as negative.
5. Accuracy = (TP + TN) / (TP + TN + FP + FN)
6. Sensitivity (recall, true-positive rate) = TP / (TP + FN)
7. Specificity (true-negative rate) = TN / (TN + FP)
8. Precision (positive predictive value) = TP / (TP + FP)

9. F1 score = $2 \times (\text{Precision} \times \text{Sensitivity}) / (\text{Precision} + \text{Sensitivity})$
10. Cohen's kappa = $(p_o - p_e) / (1 - p_e)$, where p_o is the observed and p_e the expected agreement
11. Brier score = $(1/n) \Sigma (p_i - y_i)^2$, where p_i is the predicted probability and y_i the observed outcome

Model interpretability

Two complementary interpretability methods were applied to the Random Forest model selected from the training-stage comparison. Permutation importance measured the decrease in held-out test-set AUC when a feature was randomly shuffled and was repeated 50 times. SHapley Additive exPlanations (SHAP) [25] attributed each individual prediction to its contributing features. SHAP values were computed with TreeExplainer and summarised using mean absolute SHAP values and beeswarm plots.

Exploratory decision-curve analysis

Decision-curve analysis was performed on out-of-fold predicted probabilities over the full dataset to explore model net benefit across threshold probabilities from 0.05 to 0.75 relative to PHQ-9 administration for all attendees or for none [12]. The proportion of attendees who would receive the PHQ-9 and the proportion of PHQ-9-positive cases detected were tabulated across candidate thresholds. These calculations quantify potential PHQ-9 administrations only; they do not include the time required to collect model predictors and therefore should not be interpreted as a demonstrated reduction in total clinic workload. The analysis also did not establish the safety of withholding PHQ-9 item 9 or other urgent mental-health assessment from people classified below a risk threshold.

Ethical considerations

Ethical approval was granted by the University of Health and Allied Sciences Research and Ethics Committee, approval number UHAS-REC A.2 [6] 23-24. Administrative clearance was obtained from the Kadjebi District Health Directorate and the management of both participating facilities. All participants provided written informed consent; those unable to write indicated consent by thumbprint in the presence of an impartial witness. Data were de-identified before analysis. Participants screening positive for moderate or worse depressive symptoms received immediate psychosocial support from a mental health officer, who was a member of the study team.

Results

Characteristics of the study participants

The analytical dataset comprised 264 adults receiving ART. Two hundred (75.8%) were women and 64 (24.2%) men; median age was 39 years (IQR 31.5-49), with men older than women (median 44 vs 38 years). Educational attainment was low: 116 participants (43.9%) had no formal education, 104 (39.4%) primary or junior high school, 30 (11.4%) senior high school and only 14 (5.3%) tertiary education. One hundred and ninety-four (73.5%) were self-employed as traders or farmers and 62 (23.5%) were unemployed. One hundred and sixty-five (62.5%) were unmarried. Christianity (135, 51.1%) and Islam (107, 40.5%) predominated. Clinical documentation was sparse: 201 participants (76.1%) did not know their WHO clinical stage and 243 (92.0%) did not

know their CD4 count, which is why these variables entered the model as documentation indicators rather than clinical values. Full characteristics are given in Table 2.

Prevalence and severity of depressive symptoms

Median PHQ-9 score was 7 (IQR 4-12; range 0-22). PHQ-9-defined probable depression (score >=10) was present in 100 of 264 participants (37.9%, 95% CI 32.2-43.9); at the lower >=5 threshold used in the sensitivity analysis, 181 (68.6%) had at least mild depressive symptoms. Women had a higher proportion meeting the >=10 threshold than men (44.0% vs 18.8%; chi-square=12.09, p=0.0005). By severity band, 83 participants (31.4%) had none or minimal symptoms, 81 (30.7%) mild, 69 (26.1%) moderate, 29 (11.0%) moderately severe and 2 (0.8%) severe. The score distribution and severity bands are shown in Figure 2.

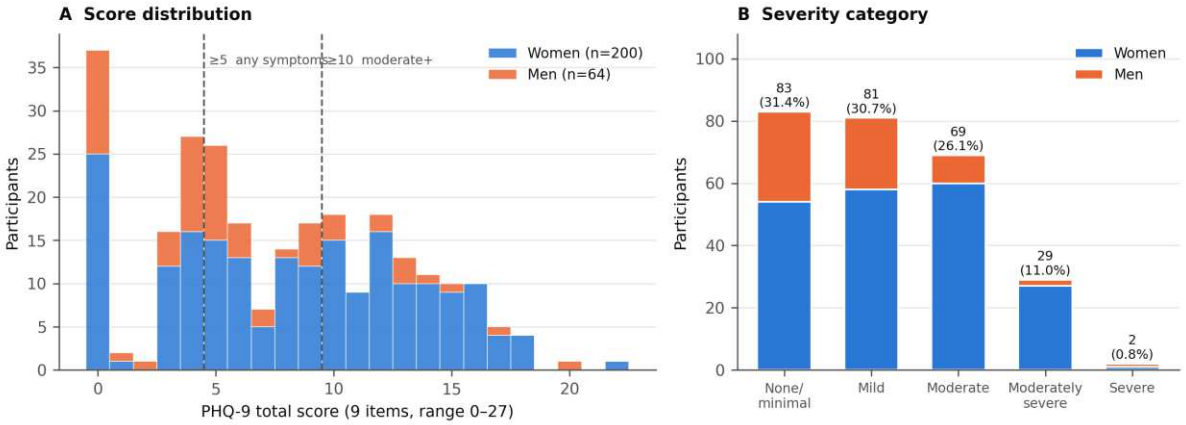


Figure 2: Distribution of PHQ-9 total scores (panel A) and severity categories (panel B) among 264 people living with HIV, Kadjebi District, Ghana, February 2024. Dashed vertical lines in panel A mark the two classification thresholds evaluated.

Class distribution before and after resampling

The training partition contained 131 PHQ-9-negative and 80 PHQ-9-positive observations, a positive-class proportion of 37.9%. SMOTE generated 51 synthetic positive-class observations, balancing the training partition to 131:131 (Figure 3). The held-out test partition was not resampled and retained 33 PHQ-9-negative and 20 PHQ-9-positive participants.

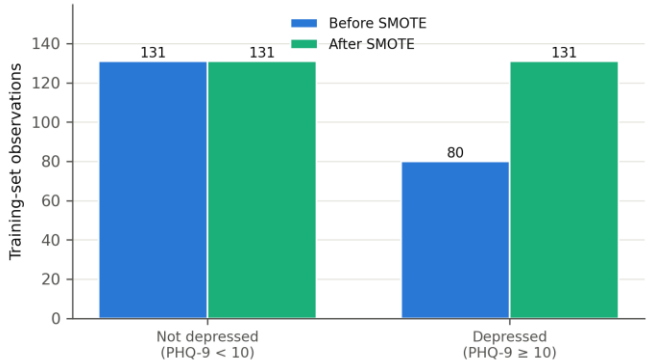


Figure 3: Training-partition class distribution before and after application of the Synthetic Minority Over-sampling Technique (SMOTE). The held-out test partition was not resampled.

Exploratory correlation clustering and feature reduction

Spearman correlations among the 52 features were generally modest (mean $|\rho| = 0.122$) but included a small number of strongly redundant pairs: two pairs at $|\rho| \geq 0.9$, four at ≥ 0.8 and six at ≥ 0.7, with a maximum of 0.922. The strongest redundancies were structurally explicable. The two occupation indicators were near-perfectly inversely related ($\rho = -0.922$), being complementary categories of the same variable. A tightly correlated block linked the tuberculosis-service and physical-health items: being on tuberculosis preventive treatment correlated with the chills-and-fever item at $\rho = 0.902$ and with comorbid illness at 0.842, chills-and-fever with comorbid illness at 0.804 and having been TB-screened with having been asked for a TB test at 0.792. This block reflects the fact, established in a companion analysis of the same cohort, that tuberculosis screening acts as the gate to every downstream tuberculosis service, so that the whole block moves together.

Ward's linkage on the distance $1 - |\rho|$, cut at 0.7, produced 33 clusters from the 52 features (Figure S1). Retaining one representative per cluster gave a reduced set of 33 features. Variance inflation factors on this set were uniformly acceptable: the maximum VIF was 2.42 (tuberculosis preventive treatment), no feature exceeded 5, and the mean VIF was below 1.5, confirming that the residual collinearity is not of a magnitude that would destabilise coefficient-based estimation (Table S2).

In the exploratory reduced-feature analysis, discrimination was similar to the full-feature models (Table S3). Random Forest and CatBoost each had a secondary cross-validated AUC of 0.887, compared with 0.893 for Random Forest on the full set Table S3. Some threshold-based metrics and calibration estimates also improved in the reduced set. However, because the representative feature within each correlation cluster was selected using outcome information from the full dataset, these reduced-set estimates may be optimistic and are not treated as validated evidence of improved performance. The full 52-feature analysis therefore remains the primary analysis.

Comparative performance of the machine learning algorithms

Across the secondary repeated training-set cross-validation analysis, AUC values ranged from 0.825 to 0.893. Random Forest had the highest mean cross-validated AUC (0.893, SD 0.041), followed by k-Nearest Neighbours (0.890, SD 0.045) and Support Vector Machine (0.888, SD 0.039). Decision Tree had the lowest mean AUC (0.825, SD 0.054). These values describe training-stage ranking stability; primary performance estimates are from the held-out test set. Full metrics are presented in Table 3 and Figure 4.

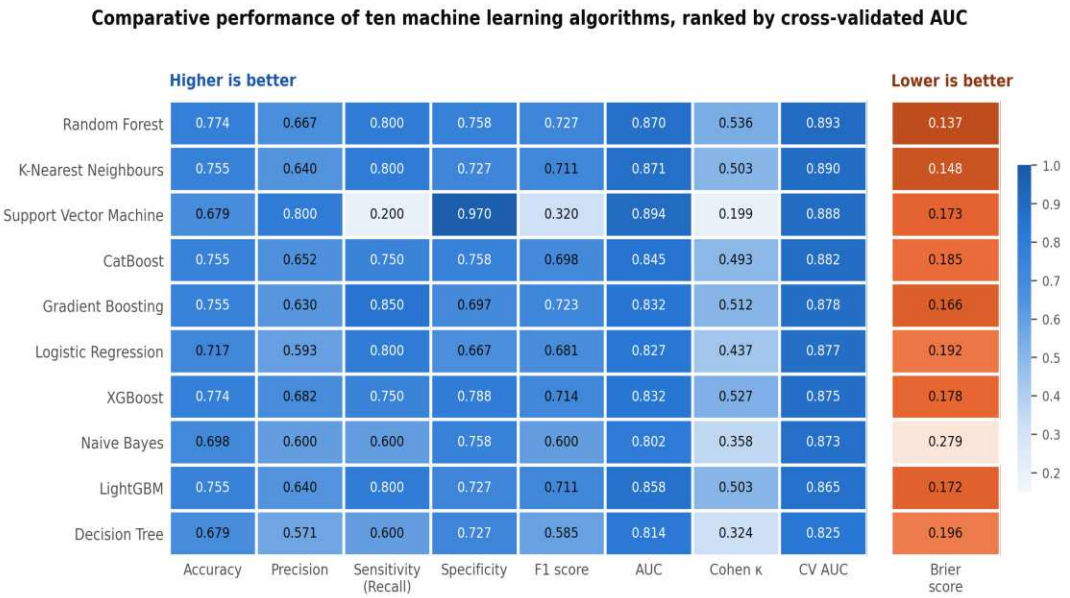


Figure 4: Held-out test-set performance of the ten algorithms across classification, discrimination, agreement and calibration-related metrics, ordered by the secondary cross-validated AUC. Darker cells indicate better performance except for Brier score, for which lower values are better.

On the held-out test set, Random Forest achieved an accuracy of 77.4%, sensitivity 80.0%, specificity 75.8%, precision 66.7%, F1 score 0.727, AUC 0.870, Cohen's kappa 0.536 and Brier score 0.137. Its kappa falls in the moderate agreement band of Landis and Koch (1977) and was the highest of any algorithm. XGBoost matched its accuracy (77.4%) with higher specificity (78.8%) but lower sensitivity (75.0%).The Random Forest model at candidate risk thresholds can be found in table 4 below

The Support Vector Machine illustrates why discrimination alone is insufficient for model selection. It had the highest held-out test-set AUC (0.894) but low sensitivity at the default 0.5 threshold (20.0%, 4 of 20 PHQ-9-positive participants), Cohen's kappa of 0.199 and a calibration slope of 2.945. Thus, although it ranked participants well, its default probability threshold produced poor case detection and would require recalibration or threshold adjustment before practical use.

Differences in secondary cross-validated AUC among the leading algorithms were small: the top seven ranged from 0.875 to 0.893, a span of 0.018 that was smaller than the fold-to-fold standard deviation for each of those models (0.036-0.046). No formal paired comparison of AUCs was performed, so these findings should be interpreted as showing similar performance rather than statistical equivalence. Regularised Logistic Regression (0.877) performed within the same range as several ensemble methods, consistent with evidence that ML methods do not automatically outperform regression in clinical prediction [13].

Receiver-operating-characteristic analysis

ROC curves for all ten algorithms on the held-out test set are shown in Figure 5A, with secondary repeated-cross-validation AUC and fold-to-fold variability in Figure 5B. The overlapping variability across the leading models illustrates that their ranking is not stable enough to support strong claims of algorithmic superiority.

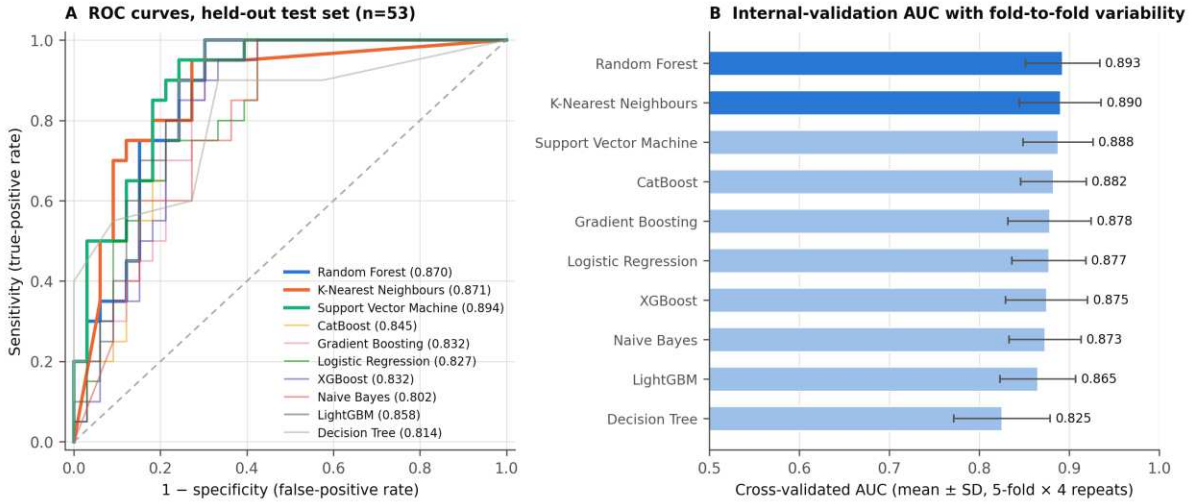


Figure 5: (A) Receiver-operating-characteristic curves for the ten algorithms on the held-out test set (n=53). (B) Secondary repeated-cross-validation AUC (mean +/- standard deviation over five folds x four repeats) in the training partition.

Confusion matrices

Confusion matrices at the default 0.5 classification threshold are shown in Figure 6. Random Forest correctly classified 16 of 20 PHQ-9-positive and 25 of 33 PHQ-9-negative participants. Gradient Boosting achieved the highest sensitivity (85.0%, 17 of 20) at the cost of ten false positives. The Support Vector Machine produced four true positives and sixteen false negatives.

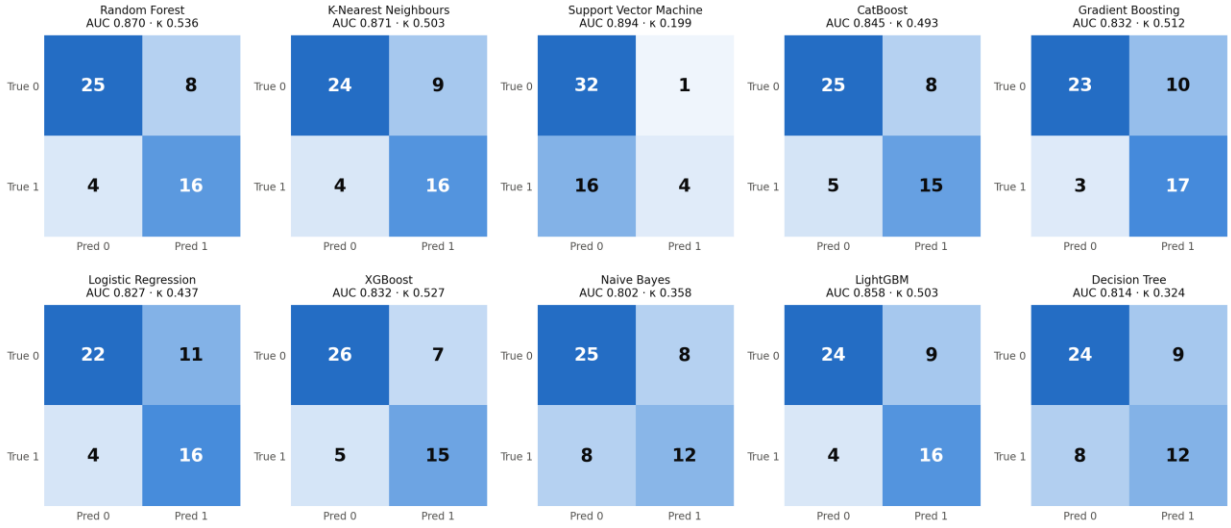


Figure 6: Confusion matrices for all ten algorithms on the held-out test set at the default classification threshold of 0.5, ordered by cross-validated AUC. Rows are true class, columns predicted class.

Calibration

Held-out test-set calibration slopes varied substantially across algorithms (Figure 7). Random Forest (1.381) and Gradient Boosting (1.565) had slopes above 1, whereas Logistic Regression (0.640), XGBoost (0.511), LightGBM (0.479), CatBoost (0.409), Naive Bayes (0.090), k-Nearest Neighbours (0.162) and Decision Tree (0.177) had slopes below 1. Brier scores ranged from 0.137

for Random Forest to 0.279 for Naive Bayes. These results show that discrimination and probability calibration can diverge; however, the test set contained only 53 participants, so exact calibration slopes are imprecise and require confirmation in larger external datasets. See Table S4 fort the Out-of-fold calibration statistics for the five best-performing algorithms.

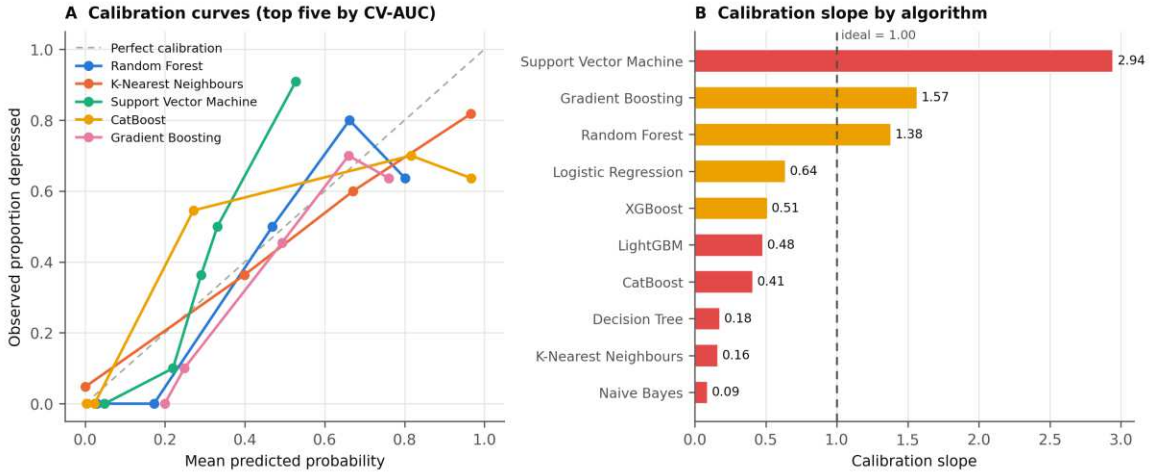


Figure 7: (A) Calibration curves for the five highest-ranked algorithms; the dashed diagonal represents perfect calibration. (B) Calibration slope for all ten algorithms; a slope of 1.00 is ideal, values below 1 indicate predictions that are too extreme and values above 1 indicate predictions that are not extreme enough.

Feature importance

Permutation importance for the Random Forest model, computed over 50 permutation repeats on the held-out test set, is shown in Figure 8. The six most important features were comfort in disclosing psychological difficulty (mean AUC decrease 0.0167), belief system (0.0143), availability of tailored mental health services (0.0102), availability of counselling (0.0089), perceived healthcare access (0.0081) and senior high school education (0.0073). Perceived cultural influence on health beliefs, medication side-effects, overall HIV impact, comorbid illness and prior tuberculosis screening followed (Table 5).

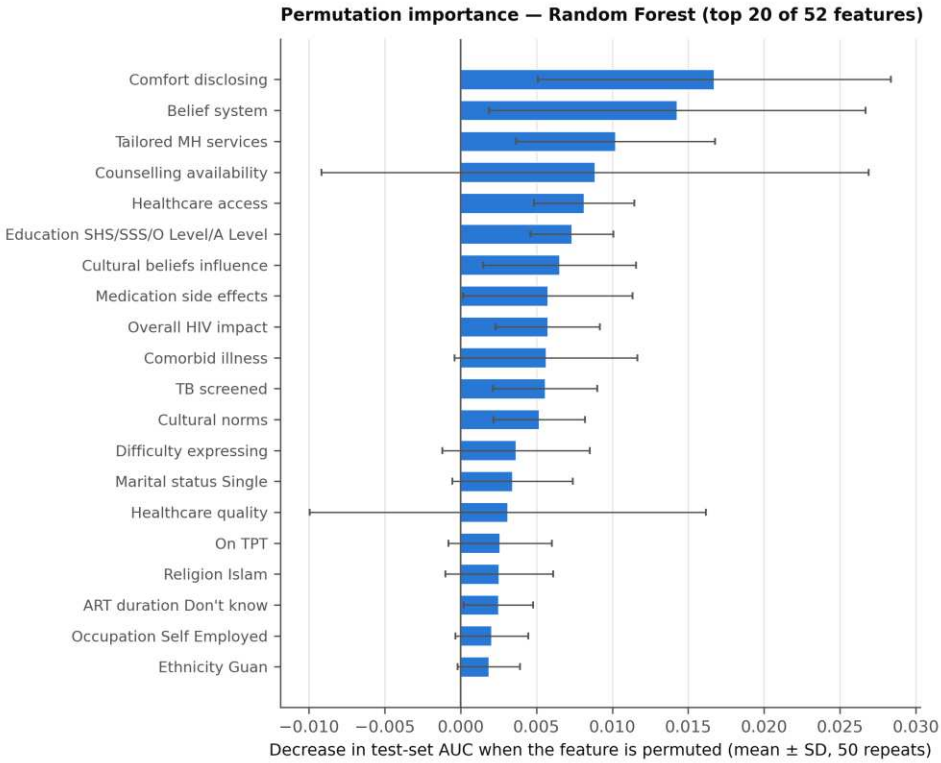


Figure 8: Permutation importance for the Random Forest model: mean decrease in held-out test-set AUC when each feature is randomly permuted over 50 repeats, with standard-deviation error bars. The top 20 of 52 features are shown.

SHAP interpretation

SHAP values computed with TreeExplainer are shown as a global importance ranking and as a beeswarm plot in Figure 9. The two interpretability methods converged closely, which is itself a reassuring signal: features identified by permuting the input and by decomposing the output agree on the same set.

The beeswarm plot adds direction and dispersion. Availability of counselling, perceived healthcare quality, belief system, comfort in disclosing psychological difficulty and availability of tailored mental health services showed the widest and most cleanly separated SHAP distributions, with high and low values of each falling on opposite sides of zero - the signature of a strong, consistently directional effect (Table 5). Female sex and the reported impact of HIV on daily life showed narrower but clearly asymmetric distributions. Age, cultural beliefs, tuberculosis preventive treatment and education contributed less and with greater overlap.

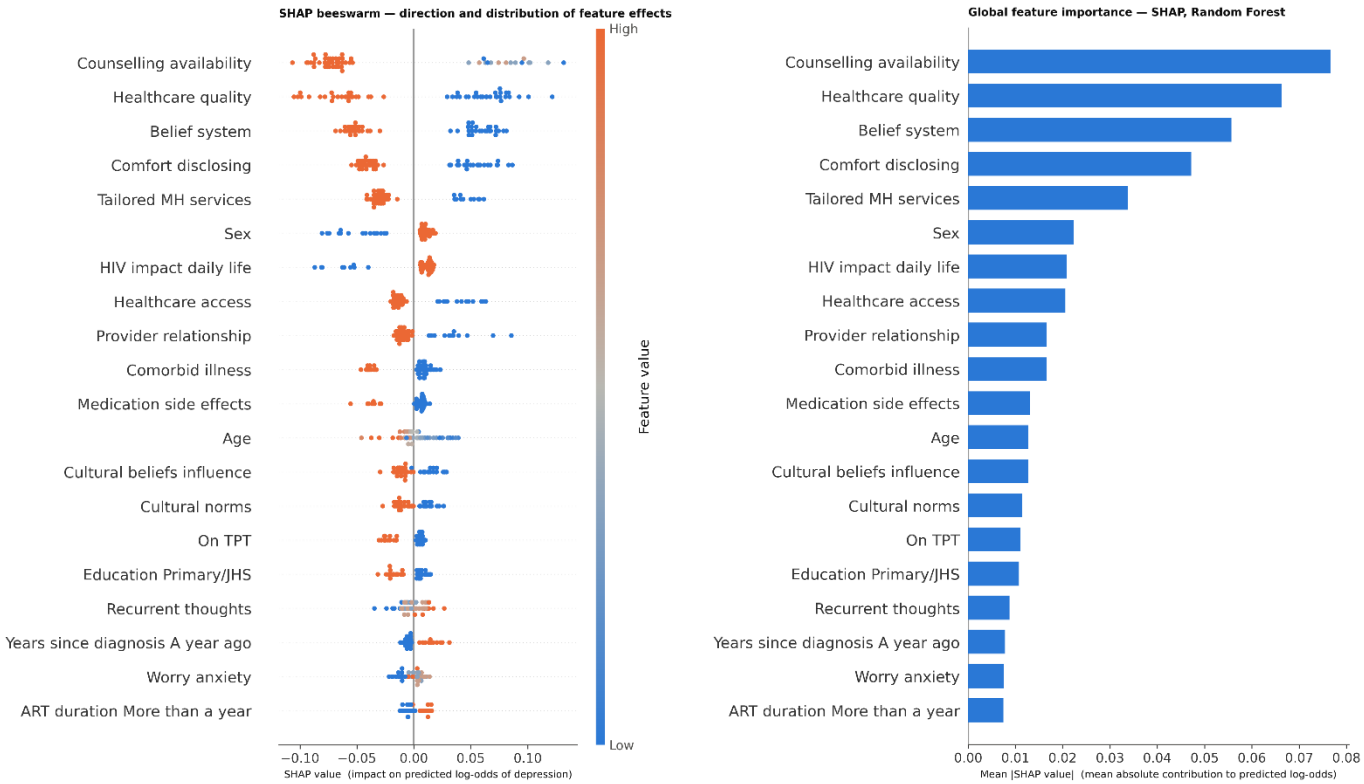


Figure 9: Global feature importance from SHAP analysis and SHAP beeswarm plot for the Random Forest model. Each point represents one participant; horizontal position gives the feature contribution to that participant's predicted model output, and colour represents the observed feature value.

Cross-model consensus on feature importance

Because a feature ranking produced by one algorithm may reflect that algorithm's inductive bias, SHAP attributions were also examined for five leading models spanning linear and tree-ensemble approaches: Random Forest, CatBoost, Gradient Boosting, penalised Logistic Regression and XGBoost. Beeswarm plots for all five are shown in Figure 10, and rank agreement is summarised in Figure 11. This cross-model analysis assesses internal consistency of feature attribution rather than external validity.

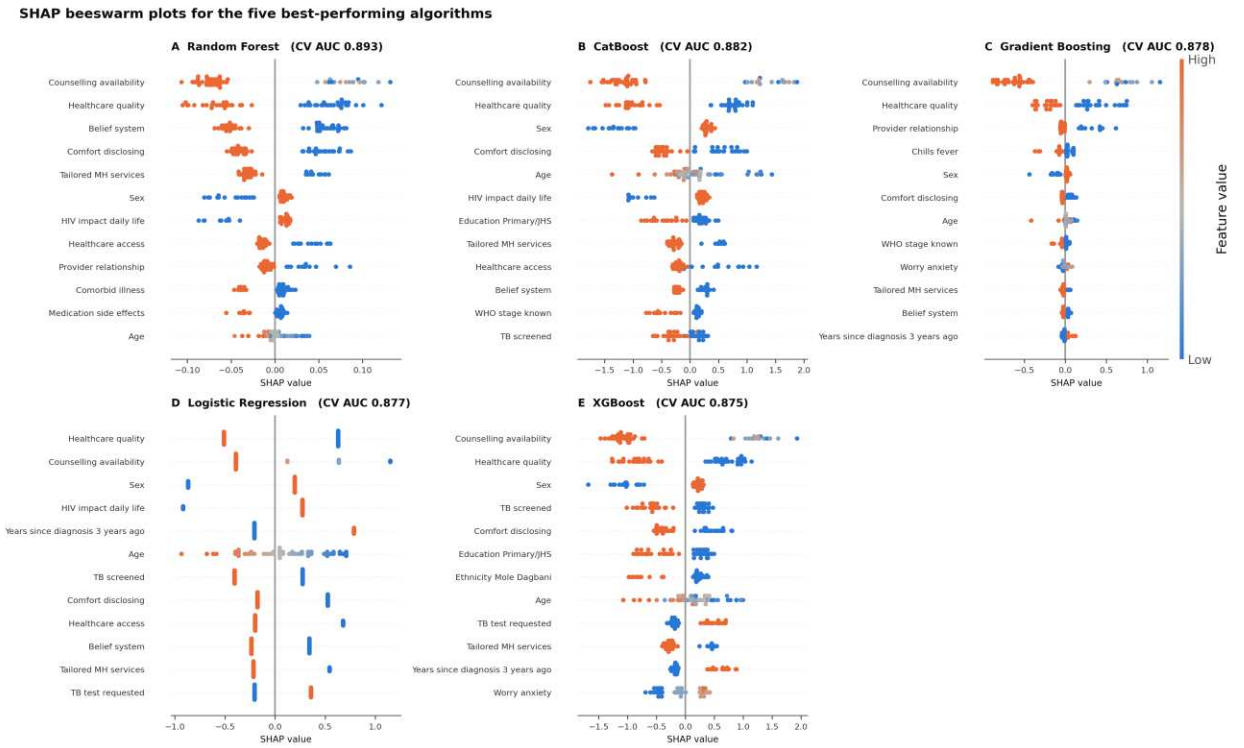


Figure 10: SHAP beeswarm plots for five leading algorithms, each showing the 12 features with the largest mean absolute SHAP value for that model. Each point represents one participant; horizontal position gives the feature contribution to the model output, and colour represents the observed feature value.

Four features appeared in the top ten of all five models: availability of counselling (mean rank 1.2), perceived healthcare quality (1.8), sex (4.0) and comfort in disclosing psychological difficulty (5.4). Age and availability of tailored mental-health services appeared in the top ten of four models, while perceived HIV impact on daily life, belief system and healthcare access appeared in three (Table S5). Penalised Logistic Regression placed healthcare quality and counselling availability first and second, respectively. Agreement across model families supports internal robustness of these feature rankings within this dataset but does not establish causal effects or transportability to other settings.

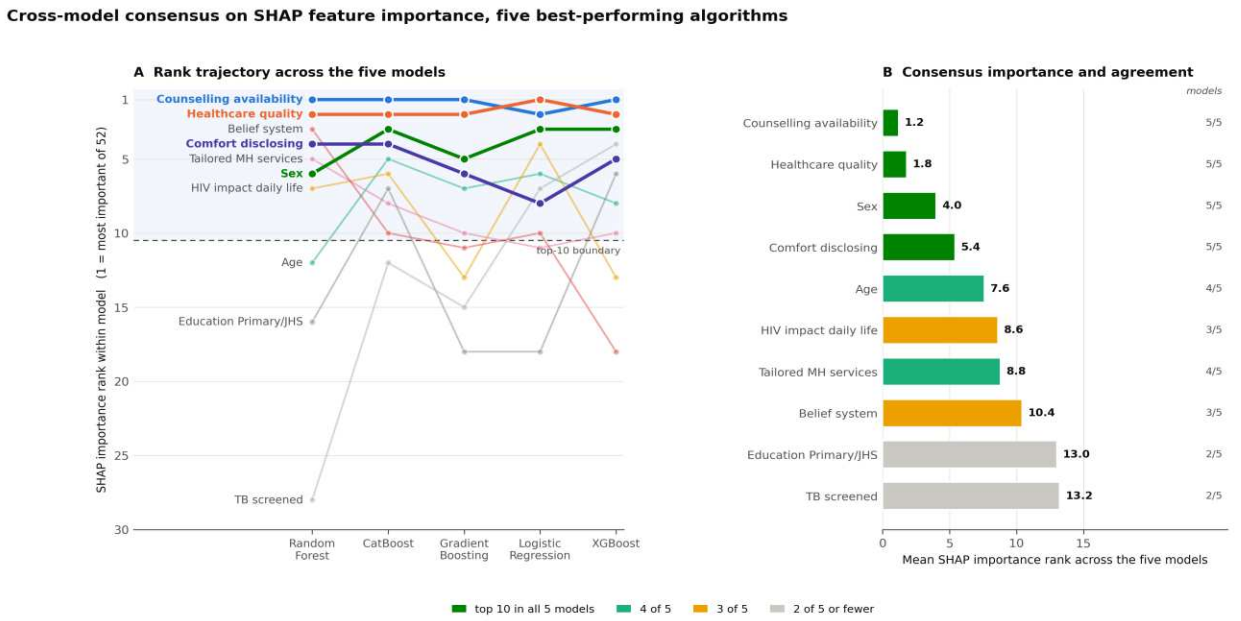


Figure 11: (A) SHAP importance rank of each feature within each of the five examined models; darker cells indicate higher importance. (B) Mean rank across the five models, with colour indicating the number of models placing that feature in the top ten. Only the top 15 features by mean rank are shown.

SHAP dependence across models

Agreement on which features matter does not by itself establish agreement on how they matter. We therefore examined SHAP dependence for the four consensus features in each of the five models (Figure 12), plotting each participant's SHAP value against their observed value of the feature and joining the mean SHAP value at each observed level.

The direction and shape of every effect was reproduced across all five algorithms. Availability of counselling showed a strongly negative dependence in every model ($\rho = -0.91$ to -1.00; mean -0.93), with a pronounced step at the highest observed level: the three lower levels carried positive SHAP values of similar magnitude, and the top level dropped sharply below zero. This threshold shape was reproduced by CatBoost, Gradient Boosting and XGBoost independently and approximated linearly by the penalised logistic regression, which is constrained to monotone linear effects on the log-odds. Perceived healthcare quality ($\rho = -0.88$ to -1.00; mean -0.95) and comfort in disclosing psychological difficulty ($\rho = -0.88$ to -1.00; mean -0.95) behaved similarly. Sex showed a consistently positive dependence ($\rho = +0.82$ to $+1.00$; mean $+0.94$) see Table S6.

The near-perfect feature-SHAP correlations for penalised Logistic Regression are a structural property of a linear model rather than an empirical finding. The tree ensembles were free to learn non-monotone patterns, but their mean-SHAP profiles were broadly monotone and similar. This supports internal consistency of the fitted models; it does not demonstrate external validity or a causal relationship between these features and depressive symptoms.

A scale difference between panels should be noted when reading Figure 12: TreeExplainer returns Random Forest SHAP values in probability units, whereas the gradient-boosting implementations and the linear model return them in log-odds units. Only the shape and sign of each profile, not the absolute magnitude, is comparable across columns.

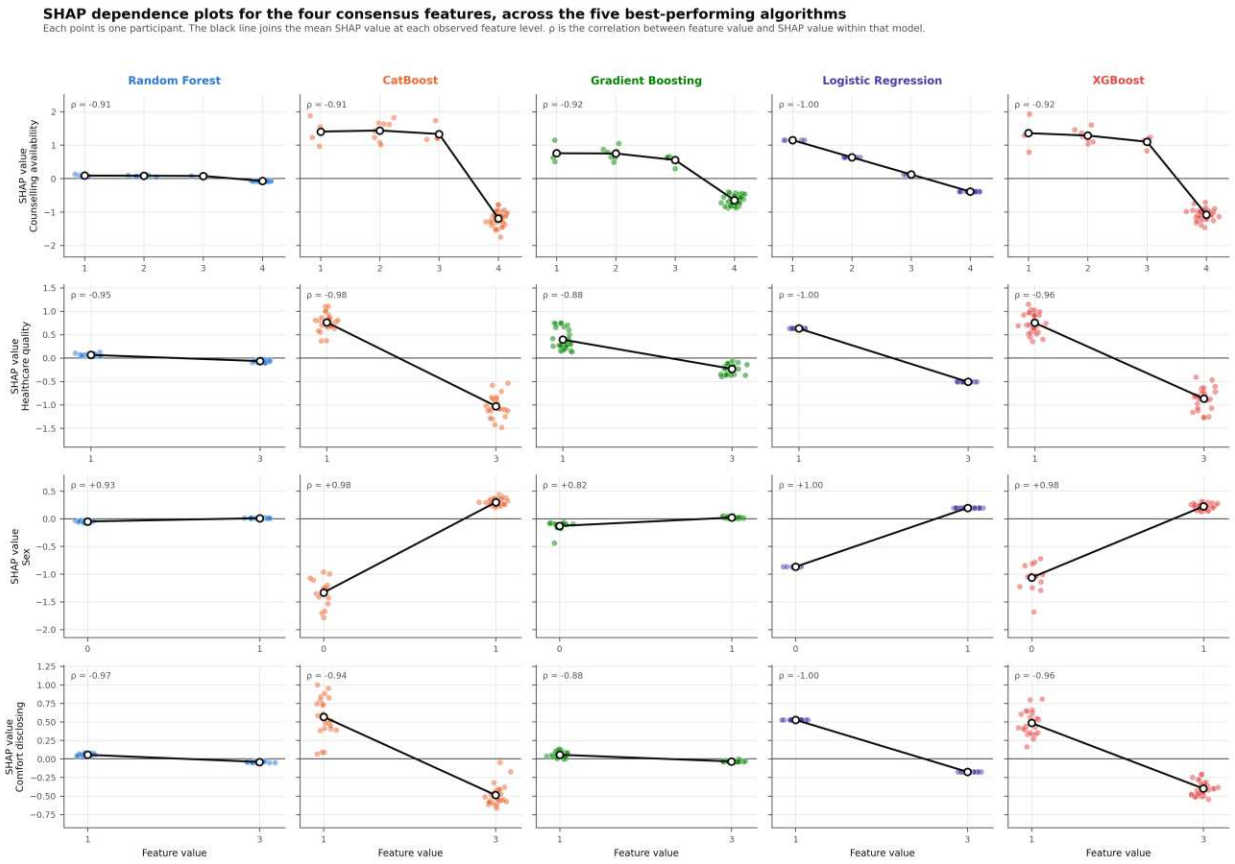
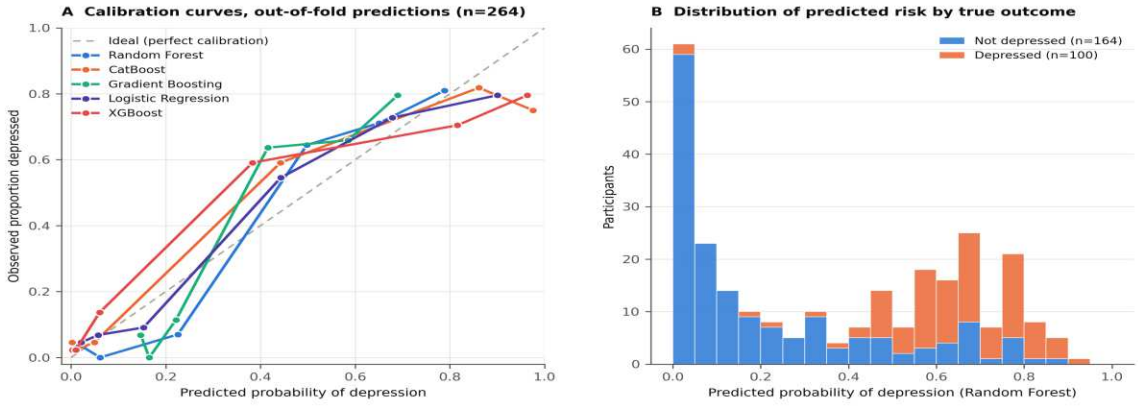


Figure 12: SHAP dependence plots for the four features ranked in the top ten by all five models. Rows are features, columns are algorithms. Each point is one participant with horizontal jitter added for legibility; the black line joins the mean SHAP value at each observed feature level. ρ is the within-model correlation between feature value and SHAP value. Random Forest values are in probability units and the remaining models in log-odds units, so magnitudes are not comparable across columns; direction and shape are.

Calibration and predicted probabilities

Exploratory out-of-fold calibration for five leading models is shown in Figure 13. For the full-feature Random Forest, the out-of-fold calibration slope was 1.26; penalised Logistic Regression had a slope of 0.84. The reduced-feature Random Forest slope of 1.047 is reported only as an exploratory result because reduced-feature selection used outcome information from the full dataset. These curves are useful for describing within-sample calibration patterns but should not be interpreted as external calibration.




```
Calibration statistics (out-of-fold):
Random Forest        slope  1.26   intercept   0.07   Brier 0.124
CatBoost             slope  0.53   intercept  -0.10   Brier 0.132
Gradient Boosting    slope  1.80   intercept   0.35   Brier 0.145
Logistic Regression  slope  0.84   intercept  -0.01   Brier 0.139
XGBoost              slope  0.55   intercept  -0.01   Brier 0.142
```

Figure 13: (A) Calibration curves for the five best-performing algorithms using out-of-fold predicted probabilities across the full dataset (n=264), with six quantile bins. The dashed diagonal is perfect calibration. (B) Distribution of Random Forest predicted probabilities by observed outcome.

The distribution of individual predicted probabilities (Figure 14) illustrates within-sample separation. Median predicted probability was 0.098 among participants below the PHQ-9 >=10 threshold and 0.653 among those meeting it (Table S7). PHQ-9-positive participants were concentrated toward the upper end of the ranked-risk distribution, although some received low predicted probabilities and would be missed by threshold-based triage. These plots describe the present dataset and do not substitute for external validation.

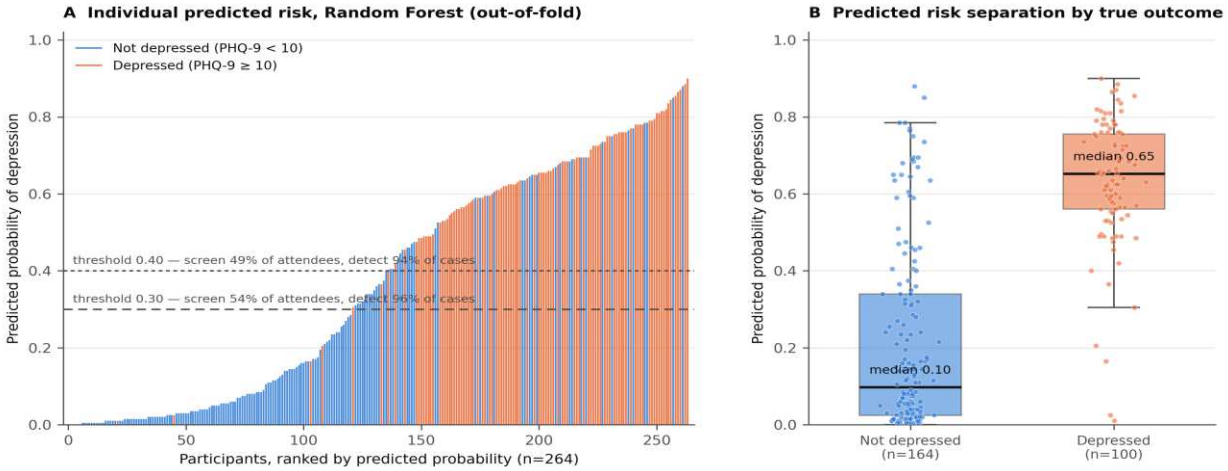


Figure 14: (A) Individual out-of-fold predicted probabilities from the Random Forest model for all 264 participants, ranked in ascending order and coloured by observed PHQ-9-defined status; horizontal lines mark two candidate operating thresholds. (B) Distribution of predicted probability by observed PHQ-9-defined status, with individual participants overlaid.

Exploratory decision-curve analysis

Exploratory decision-curve analysis (Figure 15A) suggested higher model-based net benefit than the screen-all or screen-none strategies across much of the examined threshold range. Beyond a threshold of approximately 0.38, the modeled net benefit of the screen-all strategy fell below zero while the model-guided strategy remained positive. These are statistical decision-curve results, not evidence that model-guided screening improves patient outcomes or reduces total clinic workload.

At a risk threshold of 0.40, the model would designate 48.9% of attendees for PHQ-9 administration and identify 94% of participants meeting the PHQ-9 >=10 threshold in the out-of-fold predictions (Figure 15B). A threshold of 0.30 would designate 54.2% and identify 96%, whereas 0.50 would designate 40.9% and identify 83%. These figures count PHQ-9 administrations only; they do not include the time and burden of collecting the predictors needed to generate the risk score.

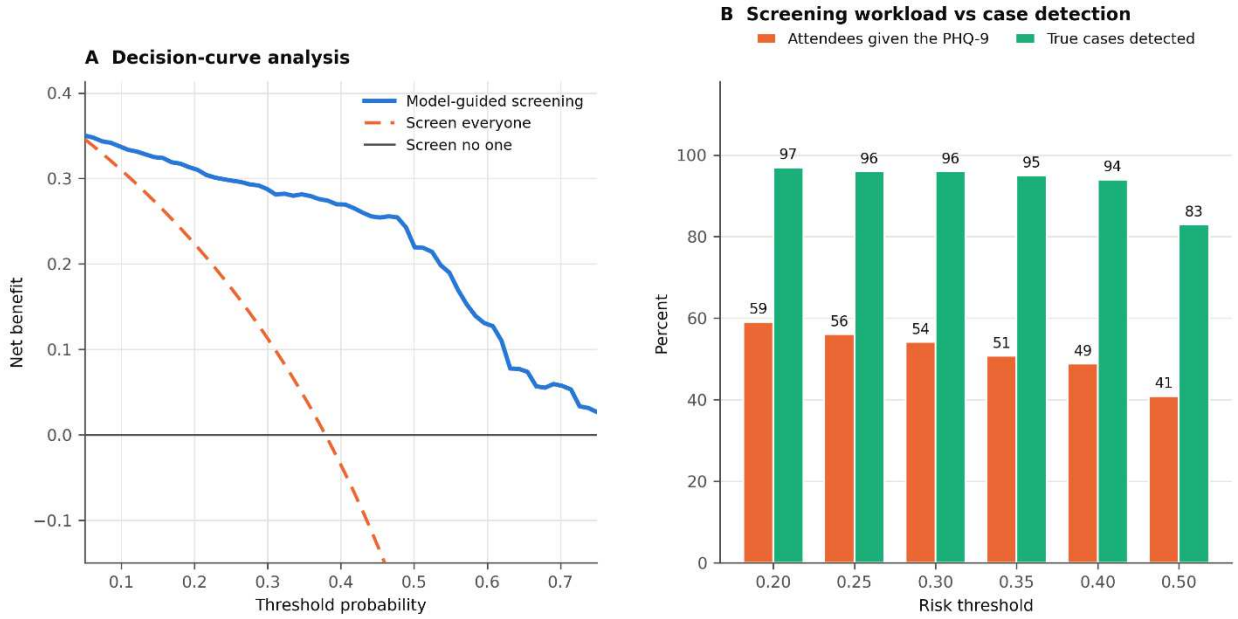


Figure 15: (A) Exploratory decision-curve analysis comparing model-guided PHQ-9 administration with screen-all and screen-none strategies using out-of-fold predicted probabilities. (B) Hypothetical proportion receiving the PHQ-9 and proportion of PHQ-9-positive cases identified at candidate thresholds. Predictor-collection burden is not included.

Sensitivity analysis

Repeating the analysis with a PHQ-9 threshold of >=5, where 181 of 264 participants (68.6%) had at least mild depressive symptoms, produced similar model rankings with higher discrimination. Random Forest again had the highest secondary cross-validated AUC (0.911, SD 0.055) and a held-out test AUC of 0.928, with accuracy 84.9%, sensitivity 83.3%, specificity 88.2%, F1 0.882 and kappa 0.674 (Table S8). These findings are a sensitivity analysis and are subject to the same internal-validation limitations as the primary analysis.

Discussion

This study developed and internally evaluated ten supervised machine-learning models for identifying PHQ-9-defined probable depression among PLHIV attending rural Ghanaian ART clinics using variables obtainable without laboratory investigation. Random Forest was selected during training-stage comparison and achieved a held-out test-set AUC of 0.870, accuracy of 77.4%, sensitivity of 80.0% and Cohen's kappa of 0.536. A secondary repeated-cross-validation analysis gave a mean AUC of 0.893. Permutation importance and SHAP highlighted relational and service-related features, while exploratory decision-curve analysis suggested potential statistical net benefit for risk-guided PHQ-9 administration. These findings should be interpreted as proof-of-concept rather than as evidence for immediate clinical deployment.

The discrimination observed here is broadly comparable with or higher than estimates reported in some ML studies that motivated this work. Alemayehu (2024) reported a best AUC of 0.813 for measles-vaccination dropout prediction using Ethiopian DHS data [8], while Vu et al. (2025) reported a best AUC of 0.69 for depressive-disorder prediction using NHANES data [10]. Demsash et al. (2023) reported higher raw accuracy for childhood-vaccination prediction in Ethiopia [9]. Direct comparisons across these studies are limited by differences in populations, outcomes, predictors and validation designs. The relatively strong discrimination in the present dataset may reflect signal in the psychosocial and health-system variables, but the small sample and internal evaluation preclude strong comparative claims.

The interpretability analyses are informative but should not be read causally. Comfort in disclosing psychological difficulty, belief system, counselling and tailored mental-health service availability, and perceived healthcare access and quality ranked above many HIV-clinical variables. Depression may alter a participant's willingness to disclose distress or perceptions of services, while service environments may also influence mental health; a cross-sectional design cannot separate these directions. In addition, many of the influential predictors require additional questioning. Consequently, the present 52-feature model should be viewed as a development model for studying predictive structure, not yet as a low-burden screening tool. A clinically practical model would need a substantially more parsimonious predictor set based primarily on routinely available ART data or questions already embedded in counselling workflows.

Female sex emerged as an important sociodemographic contributor, consistent with the pooled African evidence of higher odds of depressive symptoms among women living with HIV [4]. The convergence between established epidemiological patterns and model-based feature attribution provides a useful face-validity check, but does not establish causality.

The results also illustrate why algorithm comparison should not rely on AUC alone. Differences in secondary cross-validated AUC among the leading models were small relative to fold-to-fold variability, and regularised Logistic Regression performed within the same range as several ensemble methods. No formal paired comparison established statistical superiority among the leading algorithms. On the held-out test set, the Support Vector Machine had the highest AUC but detected only 4 of 20 PHQ-9-positive participants at the default 0.5 threshold and had poor calibration. Reporting discrimination together with sensitivity, agreement, Brier score and calibration therefore materially changes interpretation of model performance [13].

The workflow implications should be interpreted cautiously. The threshold analysis suggests that fewer PHQ-9 instruments might be administered at selected model thresholds, but the full model contains numerous psychosocial, health-system and psychological predictors, including approximately 26 items that are not simply routine register fields. Therefore, fewer PHQ-9 administrations do not necessarily mean less total questioning or lower staff workload. Before implementation, a parsimonious model based on routinely available or already-collected variables should be developed and externally validated, and its total consultation time, acceptability, equity and safety should be tested prospectively. In particular, model-based triage should not be used to withhold assessment of suicidal ideation or other urgent mental-health needs.

Strengths and limitations

Strengths include comparison of diverse algorithm families under a common analysis framework; preservation of an untouched held-out test set for primary split-sample evaluation; reporting of

calibration, Brier score and chance-corrected agreement alongside discrimination; complementary permutation and SHAP interpretation; exploratory decision-curve analysis; and use of TRIPOD+AI to guide reporting. The secondary repeated cross-validation analysis additionally describes the stability of model rankings within the training data.

Several limitations constrain the conclusions. First, the original sample size was determined for the parent cross-sectional survey using a finite-population survey formula rather than contemporary prediction-model sample-size criteria; 264 participants and 100 PHQ-9-positive events are limited for a 52-feature development model [26]. Second, the held-out test set contained only 53 participants and 20 events, so its performance and calibration estimates are imprecise. Third, although the test set remained untouched, preprocessing parameters and SMOTE were estimated before the internal tuning folds rather than within each fold; training-set cross-validation results may therefore be optimistic and are reported only as model-selection/stability measures. Fourth, the reduced-feature sensitivity analysis used outcome information from the full dataset to choose cluster representatives and is consequently exploratory. Fifth, no external or temporal validation was available, so transportability to other districts or facility types is unknown. Sixth, the cross-sectional design means that several strong predictors may be concurrent manifestations of depressive symptoms rather than antecedents, limiting causal interpretation. Seventh, clinical variables were self-reported or unknown for many participants, and the response-option codebook for several psychosocial variables was unavailable for this analysis. Eighth, the PHQ-9 outcome is a screening definition of probable depression rather than a diagnostic interview-confirmed disorder; any outcome misclassification propagates into the models. Ninth, SMOTE can distort predicted probabilities [19], and the exploratory decision-curve analysis did not account for predictor-collection burden or establish the safety of withholding full PHQ-9 assessment. Finally, this was a two-facility study in one rural district, and the results should not be extrapolated to urban or other Ghanaian ART populations without further validation.

Conclusion

Ten supervised machine-learning algorithms were compared for prediction of PHQ-9-defined probable depression among PLHIV in a rural Ghanaian ART programme. Random Forest was selected during training-stage comparison and achieved a held-out test-set AUC of 0.870, with a secondary repeated-cross-validation AUC of 0.893. Differences among the leading algorithms were small relative to resampling variability, and no formal analysis established superiority of one leading algorithm over the others. Permutation importance and SHAP consistently highlighted relational and service-related features. Exploratory decision-curve analysis suggested that risk-guided assessment could reduce the number of PHQ-9 administrations at selected thresholds, but this does not demonstrate a reduction in total workflow burden or establish safety. The model is not ready to replace routine symptom or suicide-risk assessment. External validation, development of a parsimonious workflow-compatible model, and prospective evaluation are required before implementation.

Declarations

Ethical considerations and consent to participate

This study was conducted in accordance with the Declaration of Helsinki. Ethical approval was obtained from the University of Health and Allied Sciences Research Ethics Committee (UHAS-

REC A.2[6] 23-24), and administrative permission was obtained from the Oti Regional Health Directorate and Kadjebi District Health Directorate. The study was explained to participants, and written informed consent was obtained before interview; the consent process included permission to use anonymised quotations in publications.

Contributors

VUG and DN conceived the study and designed the analysis approach. VUG, GOY, PED, DN, AZA and AKM contributed to the study methodology and implementation. PED and DN supervised data collection. VUG EKK and AZA developed and validated the models and had access to and verified the underlying data. VUG wrote the first draft. All authors contributed to interpretation, critically revised the manuscript, approved the final version and accepted responsibility for the work submitted for publication.

Declaration of interests

We declare no competing interests.

Funding

No specific funding was received for this study.

Data and code availability

The analysis code is publicly available at git repository. De-identified individual participant data, the data dictionary and study instruments are available from the corresponding author on reasonable request, subject to approval by the University of Health and Allied Sciences Research Ethics Committee and completion of an appropriate data-transfer agreement. Participant-level data are not deposited in the public code repository because they include sensitive HIV and mental-health information.

Acknowledgements

We thank the study participants for their time and contribution. We also acknowledge the Kadjebi District Health Directorate and the staff of Mary Theresa Hospital, Papase, and Kadjebi Health Centre for supporting the conduct of the study.

References

[1] B. Huang, A. Younger, M. P. Gallant, and T. J. O'Grady, "Depressive Symptoms and HIV Viral Suppression: A Systematic Review and Meta-analysis," *AIDS and Behavior 2024 29:3*, vol. 29, no. 3, pp. 870–883, Dec. 2024, doi: 10.1007/S10461-024-04571-0.

[2] N. L. Sin and M. R. DiMatteo, "Depression Treatment Enhances Adherence to Antiretroviral Therapy: a Meta-Analysis," *Annals of Behavioral Medicine 2013 47:3*, vol. 47, no. 3, pp. 259–269, Nov. 2013, doi: 10.1007/S12160-013-9559-6.

[3] C. Bernard, F. Dabis, and N. De Rekeneire, "Prevalence and factors associated with depression in people living with HIV in sub-Saharan Africa: A systematic review and meta-analysis," *PLoS One*, vol. 12, no. 8, p. e0181960, Aug. 2017, doi: 10.1371/JOURNAL.PONE.0181960.

[4] G. Tadesse *et al.*, "Depressive symptoms and its determinants among people living with HIV in Africa: systematic review and meta-analysis," *BMC Psychiatry*, vol. 25, no. 1, Dec. 2025, doi: 10.1186/S12888-025-06766-8.
[5] J. John *et al.*, "Depression and its associated factors among people living with HIV in the Volta region of Ghana," *PLOS Mental Health*, vol. 1, no. 1, p. e0000035, Jun. 2024, doi: 10.1371/JOURNAL.PMEN.0000035.
[6] WHO, "World Health Organization mhGAP Intervention Guide for mental, neurological and substance use disorders in non-specialized health settings. Version 2.0," *World Health Organization*, vol. 2.0, pp. 1–173, 2016, Accessed: Sep. 23, 2026. [Online]. Available: https://www.who.int/publications/i/item/9789241549790
[7] D. Chibanda *et al.*, "Effect of a Primary Care-Based Psychological Intervention on Symptoms of Common Mental Disorders in Zimbabwe: A Randomized Clinical Trial," *JAMA*, vol. 316, no. 24, pp. 2618–2626, Dec. 2016, doi: 10.1001/JAMA.2016.19102.
[8] M. A. Alemayehu, "Machine learning algorithms for prediction of measles one vaccination dropout among 12–23 months children in Ethiopia," *BMJ Open*, vol. 14, no. 11, p. e089764, Nov. 2024, doi: 10.1136/BMJOPEN-2024-089764.
[9] A. W. Demsash, A. A. Chereka, A. D. Walle, S. Y. Kassie, F. Bekele, and T. Bekana, "Machine learning algorithms' application to predict childhood vaccination among children aged 12–23 months in Ethiopia: Evidence 2016 Ethiopian Demographic and Health Survey dataset," *PLoS One*, vol. 18, no. 10, p. e0288867, Oct. 2023, doi: 10.1371/JOURNAL.PONE.0288867.
[10] T. Vu *et al.*, "Prediction of depressive disorder using machine learning approaches: findings from the NHANES," *BMC Med. Inform. Decis. Mak.*, vol. 25, no. 1, p. 83, Dec. 2025, doi: 10.1186/S12911-025-02903-1.
[11] B. Van Calster *et al.*, "Calibration: the Achilles heel of predictive analytics," *BMC Medicine 2019 17:1*, vol. 17, no. 1, pp. 230-, Dec. 2019, doi: 10.1186/S12916-019-1466-7.
[12] A. J. Vickers and E. B. Elkin, "Decision curve analysis: a novel method for evaluating prediction models," *Med. Decis. Making*, vol. 26, no. 6, p. 565, Nov. 2006, doi: 10.1177/0272989X06295361.
[13] E. Christodoulou, J. Ma, G. S. Collins, E. W. Steyerberg, J. Y. Verbakel, and B. Van Calster, "A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models," *J. Clin. Epidemiol.*, vol. 110, pp. 12–22, Jun. 2019, doi: 10.1016/J.JCLINEPI.2019.02.004.
[14] Ghana Statistical Service., "2021 PHC General Report Vol 3A_Population of Regions and Districts_181121.pdf.," *Ghana Statistical Service*, 2021.
[15] G. S. Collins *et al.*, "TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods," *The BMJ*, vol. 385, p. e078378, 2024, doi: 10.1136/BMJ-2023-078378.
[16] K. Kroenke, R. L. Spitzer, and J. B. W. Williams, "The PHQ-9: validity of a brief depression severity measure," J. Gen. Intern. Med., vol. 16, no. 9, pp. 606–613, Sep. 2001, doi: 10.1046/j.1525-1497.2001.016009606.x.
[17] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," Jan. 2012, Accessed: Mar. 25, 2025. [Online]. Available: http://arxiv.org/abs/1201.0490

[18] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/JAIR.953.

[19] R. Van Den Goorbergh, M. Van Smeden, D. Timmerman, and Ben Van Calster, "The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression," *Journal of the American Medical Informatics Association*, vol. 29, no. 9, pp. 1525–1534, Aug. 2022, doi: 10.1093/JAMIA/OCAC093.

[20] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324/METRICS.

[21] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, vol. 13-17-August-2016, pp. 785–794, Aug. 2016, doi: 10.1145/2939672.2939785/SUPPL_FILE/KDD2016_CHEN_BOOSTING_SYSTEM_01-ACM.MP4.

[22] G. Ke *et al.*, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, Accessed: Sep. 22, 2026. [Online]. Available: https://github.com/Microsoft/LightGBM.

[23] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: unbiased boosting with categorical features," *Adv. Neural Inf. Process. Syst.*, vol. 2018-December, pp. 6638–6648, Jun. 2017, Accessed: Sep. 23, 2026. [Online]. Available: https://arxiv.org/pdf/1706.09516

[24] C. Cortes, V. Vapnik, and L. Saitta, "Support-vector networks," *Machine Learning 1995 20:3*, vol. 20, no. 3, pp. 273–297, Sep. 1995, doi: 10.1007/BF00994018.

[25] S. M. Lundberg and S. I. Lee, "A Unified Approach to Interpreting Model Predictions," *Adv. Neural Inf. Process. Syst.*, vol. 2017-December, pp. 4766–4775, May 2017, Accessed: May 14, 2025. [Online]. Available: https://arxiv.org/pdf/1705.07874

[26] R. D. Riley, K. I. E. Snell, J. Ensor, D. L. Burke, F. E. Harrell Jr, K. G. M. Moons, and G. S. Collins, "Minimum sample size for developing a multivariable prediction model: Part II - binary and time-to-event outcomes," Stat. Med., vol. 38, no. 7, pp. 1276–1296, Mar. 2019, doi: 10.1002/sim.7992.

Tables

Table 1: Feature dictionary - the 52 predictors entered into the models

Domain	Features	n
Sociodemographic	Age (continuous); Sex; Ethnicity (6 levels, one-hot); Religion (3 levels); Marital status; Educational attainment (4 levels); Occupation (3 levels); Time since HIV diagnosis (4 levels)	18
HIV clinical	Duration on ART (3 levels); WHO clinical stage documented (yes/no); CD4 count known (yes/no)	4
Tuberculosis services	Ever screened for TB symptoms; ever asked to undergo TB test; currently on TB preventive treatment; chills and fever	4

Domain	Features	n
Physical health and behaviour	Comorbid physical illness; medication side-effects; HIV impact on daily life; specific physical limitation; overall HIV impact; substance use; tobacco use	7
Health system and services	Health information received; counselling attended; healthcare access; healthcare quality; counselling availability; tailored mental health services; provider-patient relationship; comfort disclosing psychological difficulty	8
Cultural and stigma	Cultural beliefs influence; belief system; cultural norms; perceived stigma; cultural attitudes; community acceptance	6
Psychological (non-PHQ)	Worry/anxiety; feelings of distress; low energy; difficulty expressing; recurrent thoughts	5
EXCLUDED by design	The nine PHQ-9 items, the PHQ-9 total and the derived severity band	-

ART=antiretroviral therapy. TB=tuberculosis. Counts are post-encoding feature columns. Nominal variables were one-hot encoded with the first level dropped; binary and ordered variables retained as integers.

Table 2: Characteristics of the 264 participants

Characteristic	Category	n (%)
Sex	Female	200 (75.8)
	Male	64 (24.2)
Age (years)	Median (IQR)	39 (31.5-49)
	18-29	45 (17.1)
	30-39	75 (28.4)
	40-49	77 (29.2)
	≥50	67 (25.4)
Education	No formal education	116 (43.9)
	Primary / junior high	104 (39.4)
	Senior high	30 (11.4)
	Tertiary	14 (5.3)
Marital status	Unmarried	165 (62.5)
	Married	99 (37.5)
Occupation	Self-employed	194 (73.5)
	Unemployed	62 (23.5)

Characteristic	Category	n (%)
	Formal employment	8 (3.0)
Religion	Christian	135 (51.1)
	Muslim	107 (40.5)
	Traditionalist	22 (8.3)
HIV clinical	WHO stage not known	201 (76.1)
	CD4 count not known	243 (92.0)
	ART duration not known	118 (44.7)
TB services	Ever TB-screened	108 (40.9)
	On TB preventive treatment	68 (25.8)
Target variable	Probable depression (PHQ-9 >=10)	100 (37.9)
	Below probable-depression threshold	164 (62.1)

839 IQR=interquartile range. ART=antiretroviral therapy. TB=tuberculosis.

840 *Table 3: Comparative performance of ten algorithms for predicting PHQ-9-defined probable depression (PHQ-9 >=10)*

Algorithm	CV AUC (SD)	AUC	Acc	Prec	Sens	Spec	F1	κ	Brier	Cal. slope
Random Forest	0.893 (0.041)	0.870	0.774	0.667	0.800	0.758	0.727	0.536	0.137	1.38
k-Nearest Neighbours	0.890 (0.045)	0.871	0.755	0.640	0.800	0.727	0.711	0.503	0.148	0.16
Support Vector Machine	0.888 (0.039)	0.894	0.679	0.800	0.200	0.970	0.320	0.199	0.173	2.95
CatBoost	0.882 (0.036)	0.845	0.755	0.652	0.750	0.758	0.698	0.493	0.185	0.41
Gradient Boosting	0.878 (0.046)	0.832	0.755	0.630	0.850	0.697	0.723	0.512	0.166	1.57
Logistic Regression	0.877 (0.041)	0.827	0.717	0.593	0.800	0.667	0.681	0.437	0.192	0.64
XGBoost	0.875 (0.046)	0.832	0.774	0.682	0.750	0.788	0.714	0.527	0.178	0.51
Naive Bayes	0.873 (0.040)	0.802	0.698	0.600	0.600	0.758	0.600	0.358	0.279	0.09
LightGBM	0.865 (0.042)	0.858	0.755	0.640	0.800	0.727	0.711	0.503	0.172	0.48

Algorithm	CV AUC (SD)	AUC	Acc	Prec	Sens	Spec	F1	κ	Brier	Cal. slope
Decision Tree	0.825 (0.054)	0.814	0.679	0.571	0.600	0.727	0.585	0.324	0.196	0.18

CV AUC = mean AUC from the secondary five-fold x four-repeat cross-validation stability analysis in the training partition, with standard deviation across folds. All remaining columns are metrics from the untouched held-out test set (n=53, 20 PHQ-9-positive events) at a classification threshold of 0.5. The held-out test-set results are the primary split-sample internal-validation estimates. Acc=accuracy; Prec=precision; Sens=sensitivity; Spec=specificity; kappa=Cohen's kappa; Cal.=calibration.

Table 4: Exploratory PHQ-9 administration and case-detection characteristics of the Random Forest model at candidate risk thresholds

Risk threshold	% designated for PHQ-9 administration	% of PHQ-9-positive cases identified	PHQ-9 administrations per 264 attendees
0.20	59.1%	97%	156
0.25	56.1%	96%	148
0.30	54.2%	96%	143
0.35	50.8%	95%	134
0.40	48.9%	94%	129
0.50	40.9%	83%	108
PHQ-9 for all	100%	100%	264

Computed from out-of-fold predicted probabilities over the full dataset (n=264, 100 PHQ-9-positive cases). Percentages refer to PHQ-9 administrations and detected PHQ-9-positive cases only; they do not include the burden of collecting model predictors and should not be interpreted as demonstrated workload reduction.

Table 5: Top ten features by permutation importance and by SHAP

Rank	Permutation importance (Δ AUC)	SHAP (mean |value| ranking)
1	Comfort disclosing (0.0167)	Counselling availability
2	Belief system (0.0143)	Healthcare quality
3	Tailored mental health services (0.0102)	Belief system
4	Counselling availability (0.0089)	Comfort disclosing
5	Healthcare access (0.0081)	Tailored mental health services
6	Education: senior high (0.0073)	Sex

Rank	Permutation importance (Δ AUC)	SHAP (mean |value| ranking)
7	Cultural beliefs influence (0.0065)	HIV impact on daily life
8	Medication side-effects (0.0058)	Healthcare access
9	Overall HIV impact (0.0058)	Provider relationship
10	Comorbid illness (0.0056)	Comorbid illness

Both methods applied to the Random Forest model. Permutation importance computed on the held-out test set over 50 repeats; SHAP computed with TreeExplainer. Seven of ten features appear in both lists.

Supplementary materials

S1: Supplementary Tables and Figures

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- S1.docx