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.

An Explainable XGBoost Framework for Anxiety and Depression Risk Prediction with SHAP and Counterfactual Reasoning

Anima Dahal

anima.bit_J2023@padmashreecollege.edu.np

Nilai University https://orcid.org/0009-0000-2718-1524

Bibash Basnet

Nilai University https://orcid.org/0009-0006-9610-9546

Research Article

Keywords: Anxiety prediction, depression prediction, XGBoost, explainable artificial intelligence (XAI), SHAP, counterfactual explanation, probability calibration, class imbalance, fairness evaluation

Posted Date: September 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11150010/v1

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

Additional Declarations: The authors declare no competing interests.

Abstract

Anxiety and depression are prevalent mental health conditions for which early risk identification may support timely intervention. However, many machine learning approaches for mental health prediction emphasize predictive performance while providing limited insight into model decisions and often under-address class imbalance, decision thresholds, probability calibration, and subgroup performance. This study presents an explainable machine learning framework for anxiety and depression risk prediction using demographic and psychological questionnaire features. Optimized XGBoost classifiers were developed for anxiety and depression, with class imbalance incorporated during model training and hyperparameters optimized through cross-validation. Decision thresholds were selected exclusively using validation data and subsequently applied to an isolated test set to prevent test-set leakage. The framework further integrates probability calibration, SHAP-based global and local explanations, counterfactual reasoning, and subgroup performance analysis to improve model transparency and evaluation. On the final test set of 3,644 participants, the selected models achieved accuracies of 97.59% for anxiety and 94.62% for depression, with ROC-AUC values of 0.9601 and 0.9386, respectively. Using a 0.90 decision threshold, the models obtained F1-scores of 0.5319 and 0.5075 for anxiety and depression. SHAP analysis identified stress- and sleep-related questionnaire features as prominent predictors, while counterfactual analysis demonstrated how multiple feature changes could alter model decisions. The framework provides an interpretable and evaluation-focused approach for mental health risk assessment, while its predictions are not intended to replace clinical diagnosis.

I. INTRODUCTION

Artificial intelligence (AI) has increasingly become an important component of modern healthcare, enabling computational systems to analyze large and complex datasets and support clinical decision-making [1]. In particular, machine learning (ML) can identify patterns within demographic, behavioral, and questionnaire-based information that may be difficult to recognize consistently through conventional analysis. In mental healthcare, AI-assisted prediction has potential value for the early identification of individuals who may be at risk of anxiety or depression [2]. Such systems can provide rapid, data-driven risk estimates and support practitioners in prioritizing individuals for further psychological assessment. However, AI-based prediction should complement rather than replace professional clinical judgment because mental health assessment involves subjective and contextual factors that cannot be fully represented computationally.

Although the field of healthcare-related artificial intelligence has made strides, early identification of mental health risks is still a difficult task. Common methods for clinical assessment of anxiety and depression include clinical interviews, psychological assessment, and the use of standardized self-reported questionnaires, such as the Perceived Stress Scale (PSS) (Cohen et al., 1983) and the Insomnia Severity Index (ISI) (DCS). While these tests are useful, signs and symptoms can be subtle, vague or can depend on the person's situation. From a computational perspective, mental health datasets may also exhibit substantial class imbalance, causing classifiers to achieve high accuracy while poorly identifying

minority cases. Furthermore, ML models may provide predictions without adequately explaining the factors influencing individual decisions. Probability calibration and decision-threshold selection are additional concerns, as predicted probabilities may not represent reliable risk estimates and a default threshold of 0.50 may not suit every screening objective. Performance differences across demographic and behavioral subgroups may further affect model reliability and fairness.

A research gap therefore exists between developing predictive models and developing trustworthy, interpretable, and systematically evaluated mental health prediction frameworks. Existing approaches often emphasize aggregate predictive performance while giving less attention to the combined effects of class imbalance, probability calibration, threshold selection, individual-level explanations, counterfactual reasoning, and subgroup performance. In addition, using the test set during hyperparameter optimization, feature selection, threshold selection, or calibration can introduce information leakage and lead to overly optimistic performance estimates. This study addresses these issues by separating model development from final evaluation, using validation data for optimization and threshold selection while reserving an independent test set for final assessment.

The main contribution of this study is an explainable and evaluation-oriented framework based on optimized XGBoost models for separate anxiety and depression risk prediction. The framework integrates class-imbalance handling, cross-validation-based hyperparameter optimization, validation-based threshold selection, probability calibration, SHAP-based global and local explanations, counterfactual reasoning, and subgroup performance analysis. The final test set is reserved for evaluation using accuracy, precision, recall, F1-score, specificity, ROC-AUC, PR-AUC, Brier score, confusion matrices, and comparative model assessment. The framework attempts to offer a more transparent and reliable method of AI-assisted mental health risk assessment while preserving the distinction between computational risk prediction and clinical diagnosis by combining prediction with interpretability, probability assessment, and subgroup analysis.

II. RELATED WORK

Machine learning has increasingly been applied to predict anxiety and depression using demographic and psychological assessment data. Tian et al. (2023) proposed a depression-anxiety prediction method based on the XGBoost technique and employing a data set from the collection of responses from 29,841 participants during the period of the epidemic. Through a psychometric and demographic analysis, they were able to distinguish a depressed group, an anxious group or a group which presented depressive and a combination of these disorders.They tested XGBoost and obtained accuracy rates of 78% for depression, 77% for anxiety, and 89% for a combination of depression and anxiety with the corresponding AUC values were 0.86, 0.85, and 0.95 respectively. It was demonstrated by them that reducing the number of features from 56 to 19 would still give a fair level of prediction accuracy [3].

Chowdhury et al (2024) investigated anxiety, depression and insomnia in a sample of 1250 Bangladeshi university students by giving the GAD-7, PHQ-9 and ISI-7 together with demographic information. The Decision Tree, Random Forest and XGBoost model were applied and XG Boost has the best predictive

performance. The AUC the for XG Boost were 0.94, 0.88 and 0.89 for anxiety, depression and insomnia and accuracy were 86%, 84% and 81% respectively. Although this study is closely related to the present work because of its use of GAD-7, PHQ-9, ISI, and XGBoost, its population was restricted to university students and its evaluation primarily focused on conventional predictive performance measures [4].

More recently, XGBoost has also been combined with explainability techniques for mental-health prediction. A study of 496 rural hypertensive adults in Bangladesh used GAD-7 and PHQ-9 to identify anxiety and depression symptoms and compared XGBoost with Logistic Regression. SHAP was applied to rank influential variables, demonstrating the potential of combining prediction with model interpretation. However, the study focused on a specific hypertensive rural population and used only 13 predictor variables, limiting its generalizability to broader populations [5].

With all these developments, the focus of most of the published works has been either the classification accuracy or the feature relevance. Only a small number of these papers consider the issue of using prediction models and assessing class imbalance in a systematic way with various decision-thresholds, well-calibrated probability, local explanations, counterfactual reasoning, subgroup performance etc. Also, such studies can be hardly compared directly due to the different populations, target definitions, datasets and protocols used for evaluation. The present study addresses this gap by developing independently optimized XGBoost models for anxiety and depression risk prediction and extending conventional evaluation with validation-based threshold selection, probability calibration, SHAP-based global and local explanations, counterfactual analysis, and subgroup evaluation. The models are assessed on an isolated test set using discrimination, classification, and calibration metrics, providing a broader evaluation of predictive performance, reliability, and interpretability.

III. MATERIALS AND METHODS

A. Study Design and Dataset

This study developed a machine-learning framework for the prediction of anxiety and depression risk using demographic and psychological questionnaire information. The objective was not to establish a clinical diagnosis, but to investigate whether routinely collected questionnaire responses could be used to identify participants belonging to positive or negative screening classes. The dataset contains 24,292 participants with information derived from the Patient Health Questionnaire-9 (PHQ-9), Generalized Anxiety Disorder-7 (GAD-7), Perceived Stress Scale (PSS), Insomnia Severity Index (ISI), and demographic and behavioral variables [6]. Anxiety and depression were formulated as two independent binary classification tasks, denoted Y_A and Y_D, respectively. Figure 1 summarizes the overall processing and evaluation pipeline used throughout this study.

As illustrated in Fig. 2, both target classes are highly imbalanced: anxiety-positive participants and depression-positive participants each represent a small minority of the overall sample. The available variables were organized into demographic, contextual, and questionnaire-based feature groups. The demographic information included variables such as age, gender, educational level, smoking status, and

drinking imbalance. Psychological/contextual information included ISI and PSS questionnaire responses, while PHQ-9 and GAD-7 responses were used in the feature-analysis experiments. The implementation explicitly constructed alternative feature sets, including demographic-only, contextual, questionnaire-based, and full feature configurations, enabling investigation of the contribution of different information sources.

B. Data Preprocessing and Feature Representation

Data preprocessing was performed to obtain a consistent machine-learning representation. The anxiety and depression targets were treated independently as binary outcomes, and exploratory analysis was conducted to examine target distributions, class imbalance, and the descriptive characteristics of the PHQ-9, GAD-7, ISI, and PSS variables. Categorical variables, including gender, education, smoking, and drinking status, were one-hot encoded, while numerical variables were retained. The final representation contained 36 features, including age, ISI and PSS questionnaire items, and encoded demographic and behavioral variables. All preprocessing transformations were fitted using training data only and then applied to the validation and test sets to prevent information leakage.

Because the positive classes were relatively underrepresented, class imbalance was considered during model training. For the Logistic Regression baseline, balanced class weights were calculated from the training labels and used to increase the contribution of minority-class observations.

C. Data Partitioning and Leakage Prevention

The dataset was divided into training, validation, and final test subsets. The training subset was used for model fitting and hyperparameter optimization, whereas the validation subset was used to construct the model and decision-thresholds. The final test set consisted of 3,644 participants and was strictly isolated from all development decisions. This separation was maintained throughout the complete experimental pipeline. In particular, the final test set was not used for hyperparameter optimization, feature selection, threshold selection, model selection, or calibration fitting; it was used only for the final evaluation of the selected models. This protocol was implemented to minimize optimistic bias and provide an independent estimate of generalization performance [7].

D. Baseline Machine-Learning Models

As first baseline methods, two traditional machine-learning techniques were used: the logistic regression (LR) and random forest (RF) classifiers. Logistic regression was selected mainly because it gives a straightforward and explainable linear reference model and is one of very few methods which are suitable for binary classification problems in general; it was separately trained for anxiety and depression with balanced class weights and a much higher maximum-iteration limit was set to guarantee convergence. RF was subsequently evaluated as a nonlinear ensemble baseline, whose tree-based structure enables modeling of nonlinear relationships and interactions between psychological and demographic variables. Both models were trained to estimate probabilities and not only make hard predictions of which class to choose and then calculated threshold independent measures like ROC-AUC, PR-AUC and Brier score. The

accuracy, precision, recall, F1-score, ROC-AUC, PR-AUC and Brier score were used to evaluate the framework.

E. Optimized XGBoost Model

To improve predictive flexibility, Extreme Gradient Boosting (XGBoost) was employed as the primary advanced model [8]. Separate XGBoost classifiers were developed for anxiety and depression because the two outcomes represent distinct prediction tasks and exhibit different class distributions. Hyperparameter optimization was performed using the training data with five-fold cross-validation and PR-AUC as the optimization objective, selected because the positive classes were highly imbalanced, making precision-recall performance particularly informative compared with accuracy alone [9]. Table I lists the optimized hyperparameters obtained for the two target-specific models. Class imbalance was additionally addressed using target-specific positive-class weighting. The resulting optimized XGBoost models were then evaluated on the validation set before any final test evaluation was performed.

Table I: OPTIMIZED XGBOOST HYPERPARAMETERS

Hyperparameter	Anxiety Model	Depression Model
Learning rate	0.03	0.03
Max tree depth	3	2
N estimators	200	500
Min child weight	10	3
Subsample	1.0	0.9
Colsample_bytree	0.9	0.8
Gamma	0.1	—
L1 (alpha)	0.1	—
L2 (lambda)	2	5

F. Decision-Threshold Optimization

Rather than assuming the conventional 0.50 probability threshold was clinically appropriate, the decision threshold was selected using the validation dataset, because the relative cost of false-negative and false-positive predictions can differ substantially in screening-oriented applications [10]. Figure 3 shows precision, recall, F1-score, and specificity as functions of the classification threshold for the anxiety model on the validation set; recall and specificity trade off directly as the threshold increases, while F1-score peaks toward higher thresholds because of the strong class imbalance. The validation-selected threshold for both anxiety and depression was 0.90. This threshold was then fixed before evaluating the final test set. The final, fine-tuned XGBoost prediction models delivered 97.59% test accuracy for anxiety and

94.62%test accuracy for depression at this threshold without a significant trade-off for precision compared to other lower(threshold) model setups.

G. Probability Calibration

Because classification probabilities should ideally reflect the observed likelihood of the corresponding outcomes, probability calibration was incorporated into the framework [11]. Calibration was performed after model development without using the final test data for fitting. The calibrated models were assessed using the Brier score, while ROC-AUC and PR-AUC were retained to evaluate ranking-based discrimination. The Brier score provides a measure of the agreement between predicted probabilities and observed outcomes, with lower values indicating better probabilistic performance [11].

H. Explainable AI Using SHAP

To address the black-box nature of the optimized XGBoost models, SHapley Additive exPlanations (SHAP) were used for global and local interpretability. SHAP is widely used in healthcare machine-learning applications, particularly for explaining predictions from tabular models [12]. SHAP values were calculated for all 36 input features. For global interpretation, features were ranked using their mean absolute SHAP values, with higher values indicating greater overall influence on model predictions. For local interpretation, SHAP explanations were generated for representative negative, positive, and borderline participants. Signed SHAP values were used to indicate whether individual features increased or decreased the model prediction toward the positive class.

I. Counterfactual Explanations

Counterfactual analysis was incorporated to provide an additional form of individual-level interpretability by examining how hypothetical changes in input features could affect the model's prediction. Counterfactual explanations provide alternative feature configurations that can lead a fixed predictive model to produce a different outcome, with minimality and plausibility being important considerations [13]. For each selected participant, the analysis started from the original feature values and evaluated alternative feature combinations while keeping the trained XGBoost model fixed. Candidate changes were progressively expanded from smaller to larger feature combinations to identify feasible counterfactuals capable of moving the predicted probability across the predefined decision threshold of 0.90. The analysis therefore focused on identifying the minimum or relatively small set of feature modifications required to change the model's predicted class. Counterfactual explanations were generated for representative predictions, including positive and borderline cases, to examine the sensitivity of individual model decisions to changes in questionnaire responses. The resulting counterfactuals were interpreted strictly as hypothetical model scenarios and were not considered causal explanations, treatment recommendations, or evidence that modifying a questionnaire response would clinically reduce anxiety or depression risk.

J. Subgroup and Fairness Analysis

To assess whether model performance remained reasonably consistent across different participant groups, a subgroup analysis was conducted using demographic and behavioral characteristics available in the dataset. Gender, education level, drinking and smoking habits, and predetermined age groupings were all taken into account. Accuracy, precision, recall, F1-score, and ROC-AUC were utilized to evaluate each subgroup's performance using the same final model and decision threshold as the overall test evaluation. In addition, the observed positive rate and predicted positive rate were compared to identify potential differences in outcome prevalence and model predictions across groups. No subgroup was used for model training, hyperparameter optimization, threshold selection, or calibration. Subgroup analysis was treated as an additional fairness-oriented evaluation rather than as a formal determination of algorithmic fairness [14]. Particular caution was applied to groups with small sample sizes, for which performance estimates may be unstable and ROC-AUC may be undefined. This analysis was therefore intended to identify potential subgroup differences that may warrant further investigation rather than to establish definitive fairness conclusions.

Overall, the methodology establishes a leakage-controlled, optimized, calibrated, explainable, and subgroup-aware machine-learning pipeline in which the final test set remains isolated until final evaluation. This design enables the proposed framework to be assessed not only in terms of predictive performance, but also in terms of probability reliability, model interpretability, individual-level decision sensitivity, and potential subgroup differences.

IV. EXPERIMENTAL RESULTS

A. Performance Metrics

The final performance evaluation was conducted on an isolated test set containing 3,644 participants. The optimized XGBoost models for anxiety and depression risk prediction were evaluated using a decision threshold of 0.90, which was selected exclusively from the validation set. For anxiety risk prediction, the model achieved an accuracy of 97.59%, precision of 47.17%, recall of 60.98%, and F1-score of 53.19%. The model also obtained a specificity of 98.43%, ROC-AUC of 0.9601, and PR-AUC of 0.5650. For depression risk prediction, the model achieved an accuracy of 94.62%, precision of 50.25%, recall of 51.27%, and F1-score of 50.75%, with a specificity of 97.10%, ROC-AUC of 0.9386, and PR-AUC of 0.5571.

The ROC curves and confusion matrices are presented in Figs. 4. The ROC-AUC values indicate strong discrimination, while the PR-AUC values provide additional evaluation under class imbalance. The confusion matrices show 3,506 true negatives and 50 true positives for anxiety, and 3,452 true negatives and 101 true positives for depression, with the remaining cases representing false-positive and false-negative predictions. The high accuracy therefore reflects, in part, the correct classification of the majority negative class, supporting the use of precision, recall, F1-score, and PR-AUC alongside accuracy.

B. Model Comparison

The results obtained from the optimized XGBoost models were compared with that of Logistic Regression, Random Forest and default XGBoost models as presented in Table IV. The best models with

the highest ROC-AUCs were the optimized models with values of 0.9601 for anxiety and 0.9386 for depression. Anxiety's precision dropped from 22.15% to 47.17% with the increase in threshold from 0.50 to 0.90, and similarly for depression (27.29% to 50.25%). This shows the typical precision–recall relationship that comes with a more conservative threshold. Since ROC-AUC and PR-AUC are threshold-independent, they did not change.

TABLE II: COMPARISON OF BASELINE AND OPTIMIZED MODELS FOR ANXIETY AND DEPRESSION RISK PREDICTION

Model	Target	Thr.	Accuracy	Precision	Recall	F1	ROC-AUC	PR-AUC
Logistic Regression	Anxiety	0.50	91.58%	19.35%	86.59%	31.63%	0.9572	0.5483
Logistic Regression	Depression	0.50	86.99%	27.55%	86.29%	41.77%	0.9374	0.5663
Random Forest (A)	Anxiety	0.50	98.05%	58.21%	47.56%	52.35%	0.9567	0.5295
Random Forest (A)	Depression	0.50	94.68%	50.97%	40.10%	44.89%	0.9256	0.5235
Random Forest (B)	Anxiety	0.50	95.25%	29.22%	78.05%	42.52%	0.9499	0.5630
Random Forest (B)	Depression	0.50	88.94%	30.12%	79.19%	43.64%	0.9282	0.5277
Default XGBoost	Anxiety	0.50	95.28%	29.17%	76.83%	42.28%	0.9514	0.5578
Default XGBoost	Depression	0.50	88.89%	29.92%	78.68%	43.36%	0.9294	0.5308
Optimized XGBoost (Final)	**Anxiety**	**0.90**	**97.59%**	**47.17%**	**60.98%**	**53.19%**	**0.9601**	**0.5650**
Optimized XGBoost (Final)	**Depression**	**0.90**	**94.62%**	**50.25%**	**51.27%**	**50.75%**	**0.9386**	**0.5571**

RF (A) and RF (B) denote the two evaluated Random Forest class-weighting configurations. The final optimized XGBoost results are shown at the validation-selected threshold of 0.90.

C. Probability Calibration

Probability calibration was applied to improve the reliability of the predicted probabilities. The Brier score decreased from 0.0542 to 0.0151 for anxiety and from 0.0921 to 0.0355 for depression after calibration, indicating improved agreement between predicted probabilities and observed outcomes. The ROC-AUC

and PR-AUC values remained unchanged after calibration (anxiety: 0.9601 and 0.5650; depression: 0.9386 and 0.5571), indicating that calibration improved probability reliability without changing the models' ranking-based discrimination.

D. Explainable Model Predictions

SHAP analysis was used to examine the contribution of the input features to the optimized XGBoost predictions. For anxiety risk prediction, PSS Q3, PSS Q2, and PSS Q1 were the most influential features, with mean absolute SHAP values of 0.8300, 0.7511, and 0.7124, respectively. For depression risk prediction, PSS Q3, PSS Q2, and PSS Q7 were the most influential features, with mean absolute SHAP values of 0.5440, 0.5440, and 0.5043, respectively. Overall, stress-related PSS features were among the strongest contributors to both predictions, while ISI features also showed notable contributions, particularly for depression.

The SHAP beeswarm plots further illustrate the magnitude and direction of feature contributions across participants. The distributions show that both PSS and ISI features contributed differently across individual predictions.

For local interpretability, waterfall plots were generated for the same representative participant using the anxiety and depression models. The plots show how individual features contributed toward or away from the positive class. As shown in Fig. 8, the anxiety and depression predictions were influenced by different combinations and magnitudes of PSS and ISI features.

E. Counterfactual Analysis

Counterfactual analysis showed that changing multiple input features could move both model predictions below the selected 0.90 threshold. For anxiety, the probability decreased from 0.9954 to 0.8709 after changing PSS Q3 and PSS Q1. For depression, it decreased from 0.9994 to 0.8771 after changing PSS Q3, PSS Q2, PSS Q7, and ISI Q1. These results demonstrate the sensitivity of model decisions to combinations of input features

TABLE III: REPRESENTATIVE COUNTERFACTUAL EXPLANATIONS (DECISION THRESHOLD = 0.90)

Target	Features Changed	Prob. Before	Prob. After
Anxiety	pss_q3: 5→2; pss_q1: 5→1	0.9954	0.8709
Depression	pss_q3, pss_q2, pss_q7, isi_q1	0.9994	0.8771

F. Subgroup Analysis

Subgroup analysis was performed to examine model performance across participant groups. For gender, the anxiety model achieved ROC-AUC values of 0.9594 for females and 0.9612 for males, while the depression model achieved 0.9302 and 0.9505, respectively. The results indicate broadly comparable

discrimination across the two gender groups, although precision, recall, and F1-score varied between groups. Small subgroup sizes limited reliable interpretation for some demographic categories.

TABLE IV: SUBGROUP PERFORMANCE FOR ANXIETY PREDICTION, STRATIFIED BY GENDER

Group	N	Pos. Rate	Pred. Pos. Rate	Accuracy	Precision	Recall	F1	ROC-AUC
Female	2301	2.04%	3.13%	97.35%	40.28%	61.70%	48.74%	0.9594
Male	1343	2.61%	2.53%	97.99%	61.76%	60.00%	60.87%	0.9612

V. DISCUSSION

The proposed framework demonstrated strong discrimination for both anxiety and depression risk prediction, while the lower precision and F1-scores relative to accuracy highlighted the effect of class imbalance and the need for multiple evaluation metrics. Probability calibration substantially improved Brier scores without changing ROC-AUC or PR-AUC, indicating improved probability reliability while preserving discrimination. SHAP analysis showed that stress-related PSS features were consistently important, while ISI features contributed particularly to depression predictions, although these feature contributions do not imply causality. Counterfactual analysis further demonstrated that combinations of input features could change model predictions across the selected threshold, providing insight into model sensitivity. Subgroup analysis showed broadly comparable discrimination between female and male participants, although some differences were observed in other metrics. Overall, integrating optimized XGBoost, calibration, SHAP, counterfactual reasoning, and subgroup evaluation provides a more comprehensive approach to mental-health risk prediction than predictive performance alone. The framework should be regarded as a risk-screening and decision-support system rather than a clinical diagnostic tool.

VI. CONCLUSION

This study presented an explainable machine-learning framework for anxiety and depression prediction using optimized XGBoost models combined with threshold optimization, probability calibration, SHAP-based explanations, and counterfactual reasoning. On the independent test set of 3,644 participants, the final models achieved ROC-AUC values of 0.9601 for anxiety and 0.9386 for depression, with accuracies of 97.59% and 94.62%, respectively. The framework also provided global and local explanations, identifying stress- and sleep-related questionnaire features as important contributors to predictions. These findings demonstrate the potential of combining predictive performance with interpretability to support transparent AI-assisted mental-health screening. Nevertheless, the system is intended as a decision-support tool rather than a clinical diagnostic system, and its predictions should be interpreted alongside appropriate professional assessment.

VII. FUTURE WORK

Future research is needed to validate the framework in larger and more diverse samples from other populations and health care environments, in order to evaluate its generalizability and robustness. Further calibration assessment, external validation, and prospective evaluation would provide greater confidence in use in the real world. For future research, fairness-aware learning and a more extensive subgroup analysis should be explored, especially for underrepresented demographic subgroups. This counterfactual part can be extended by incorporating clinically validated constraints to guarantee that the generated explanations are realistic and feasible, but not just mathematically possible. Finally, by adding temporal patient information, uncertainty estimation, and clinician-in-the-loop evaluation, the reliability, transparency, and utility of the framework for mental health decision support in real-world settings could be further enhanced.

References

1. Topol EJ (2019) High-performance medicine: the convergence of human and artificial intelligence. Nat Med 25:44–56
2. Thieme A, Belgrave D, Doherty G (2020) Machine Learning in Mental Health: A Systematic Review of the HCI Literature to Support the Development of Effective and Implementable ML Systems, Anja; Belgrave, Danielle; Doherty, Gavin. ACM Trans Computer-Human Interact, 27, 5
3. Tian Z, Qu W, Zhao Y, Zhu X, Wang Z, Tan Y, Jiang R, Tan S (2023) Predicting depression and anxiety of Chinese population during COVID-19 in psychological evaluation data by XGBoost. J Affect Disord 323:417–425
4. Chowdhury AH, Rad D, Rahman MS (2024) Predicting anxiety, depression, and insomnia among Bangladeshi university students using tree-based machine learning models. Health Sci Rep, 7, 4
5. Raihana Z, Kader M, Islam MZ, Bornee FA, Mondal MNI, Chowdhury MRK, Billah B (2025) Factors associated with the presence of anxiety and depression symptoms in rural hypertensive adults in Bangladesh: leveraging extreme gradient booster machine learning algorithm. Front Psychol 16:1650667
6. Su Z (2023) *Data from: Temporal dynamics in psychological assessments: a novel dataset with scales and response times*, Zenodo
7. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, Ghassemi M, Liu X, Reitsma JB, van Smeden M, Boulesteix A-L, Camaradou JC, Celi LA (2024) TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods, *BMJ*, vol. 385
8. Chen T, Guestrin C (2016) XGBoost: A Scalable Tree Boosting System,
9. Saito T, Rehmsmeier M (2015) The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. PLoS ONE, 10, 3
10. Wynants L, van Smeden M, McLernon DJ, Timmerman D, Steyerberg EW, Van Calster B (2019) Three myths about risk thresholds for prediction models. BMC Med, 17
11. Lane RO (2026) A Comprehensive Review of Classifier Probability Calibration Metrics. Data Sci, 9, 1

12. Allgaier J, Mulansky L, Draelos RL, Pryss R (2023) How does the model make predictions? A systematic literature review on the explainability power of machine learning in healthcare. Artif Intell Med, 143
13. Verma S, Boonsanong V, Hoang M, Hines K, Dickerson J, Shah C (2024) Counterfactual Explanations and Algorithmic Recourses for Machine Learning: A Review. ACM-CSUR, 56, 12
14. Chen RJ, Wang JJ, Williamson DFK, Chen TY, Lipkova J, Lu MY, Sahai S, Mahmood F (2023) Algorithmic fairness in artificial intelligence for medicine and healthcare. Nat Biomedical Eng 7:719–742
15. Cohen S, Kamarck T, Mermelstein R (1983) A Global Measure of Perceived Stress. J Health Soc Behav 24(4):385–396

Figures

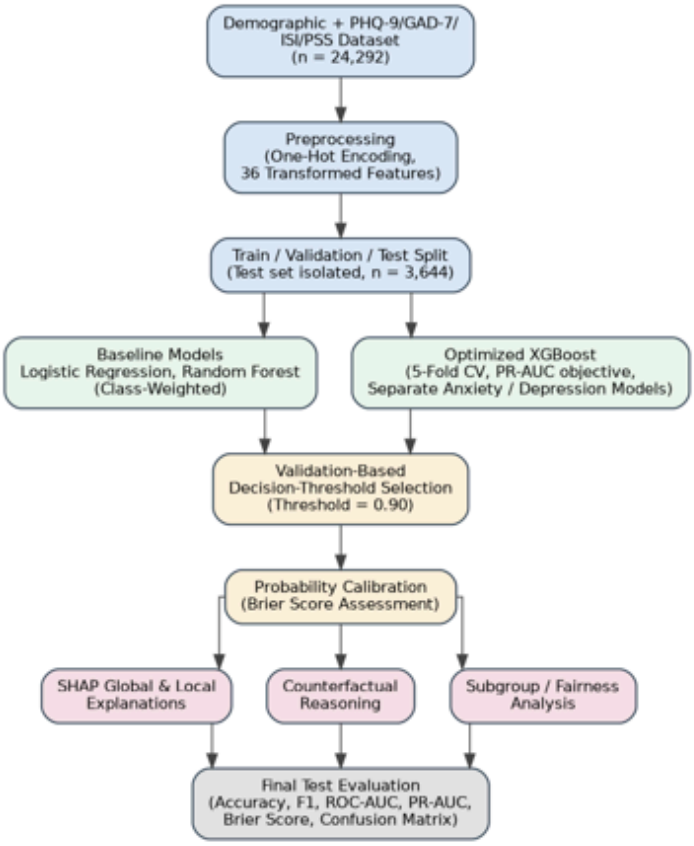


Figure 1

Overall framework pipeline, from raw questionnaire data to leakage-controlled, calibrated, and explainable final evaluation.

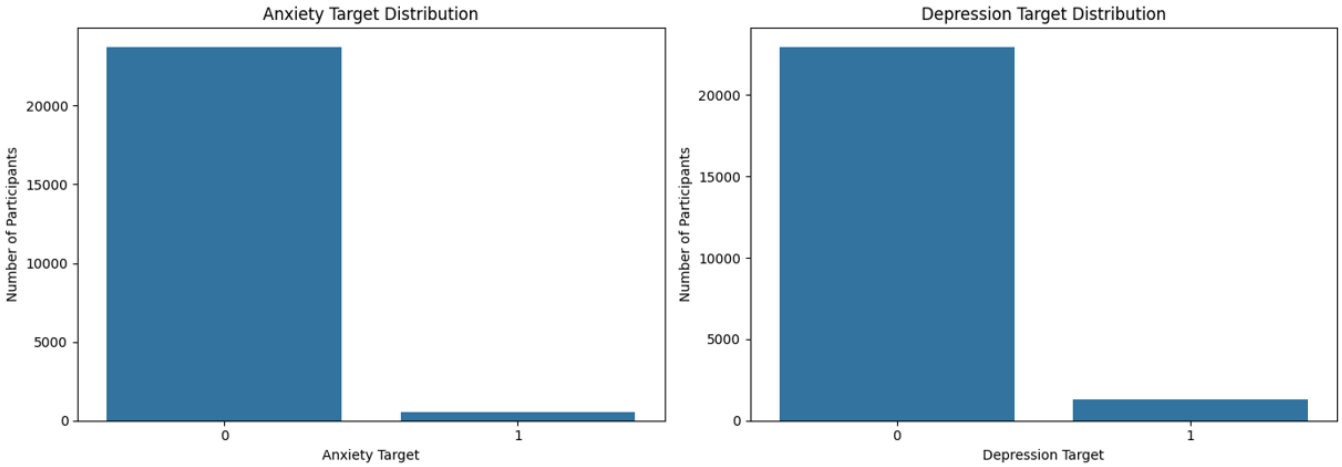


Figure 2

Class distribution of the anxiety and depression target variables, illustrating pronounced class imbalance for both outcomes.

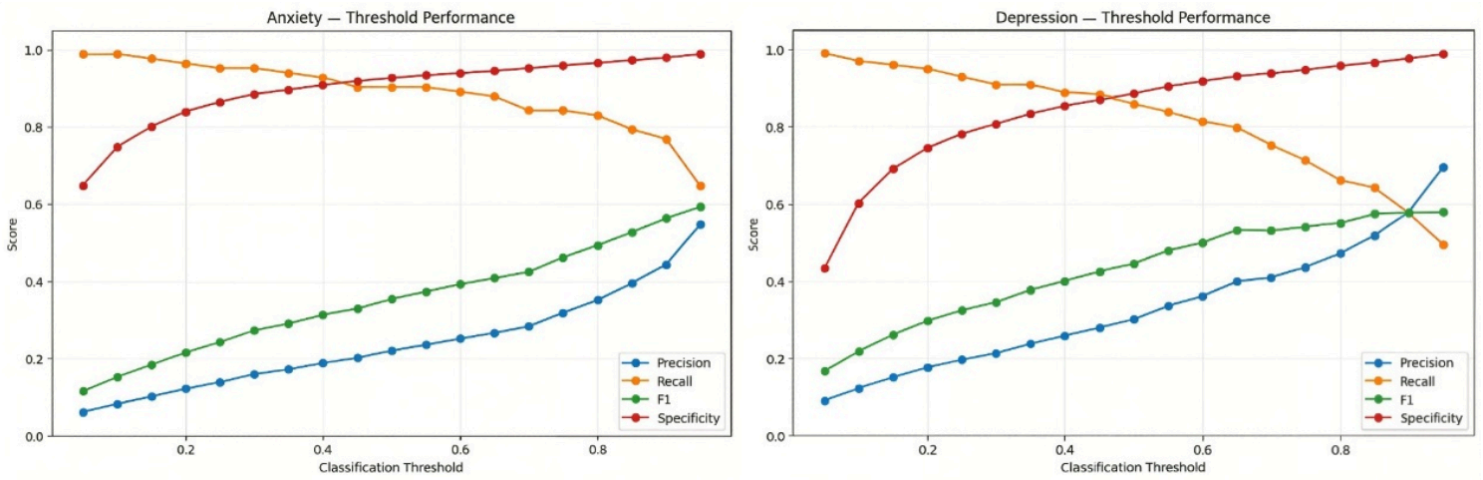


Figure 3

Precision, recall, F1-score, and specificity of the anxiety XGBoost model across candidate classification thresholds on the validation set.

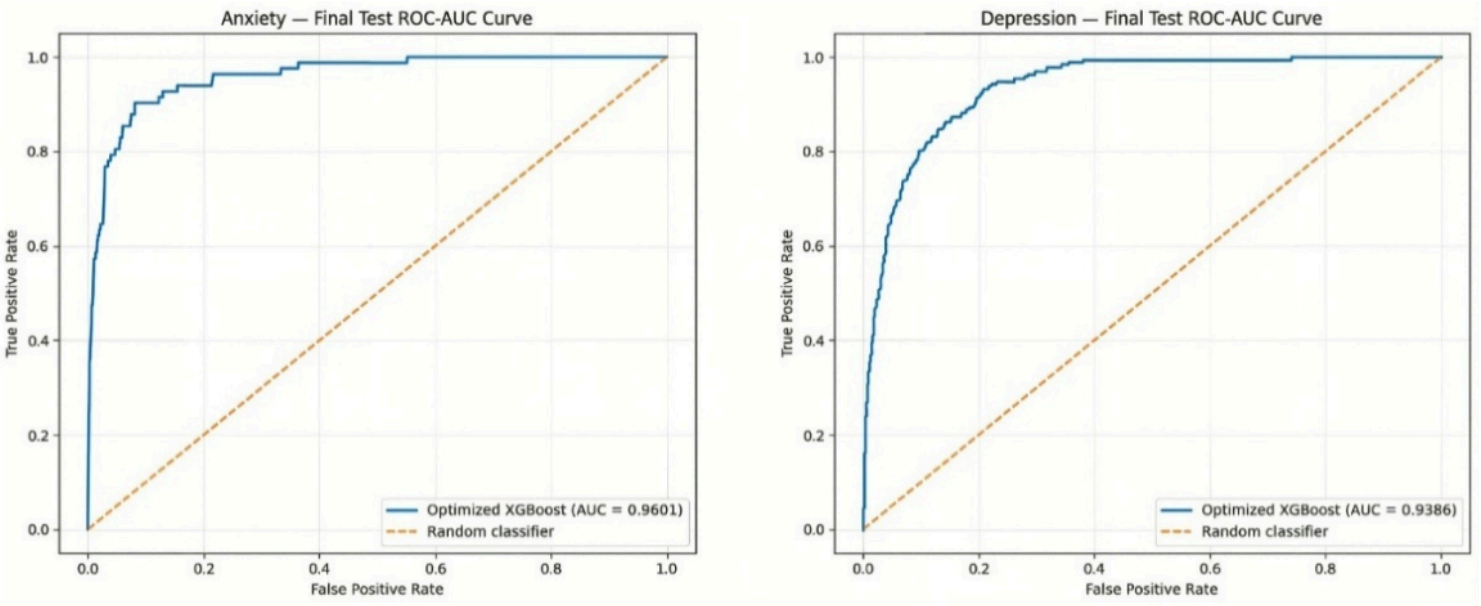


Figure 4

ROC curves of the optimized XGBoost models for anxiety and depression risk prediction.

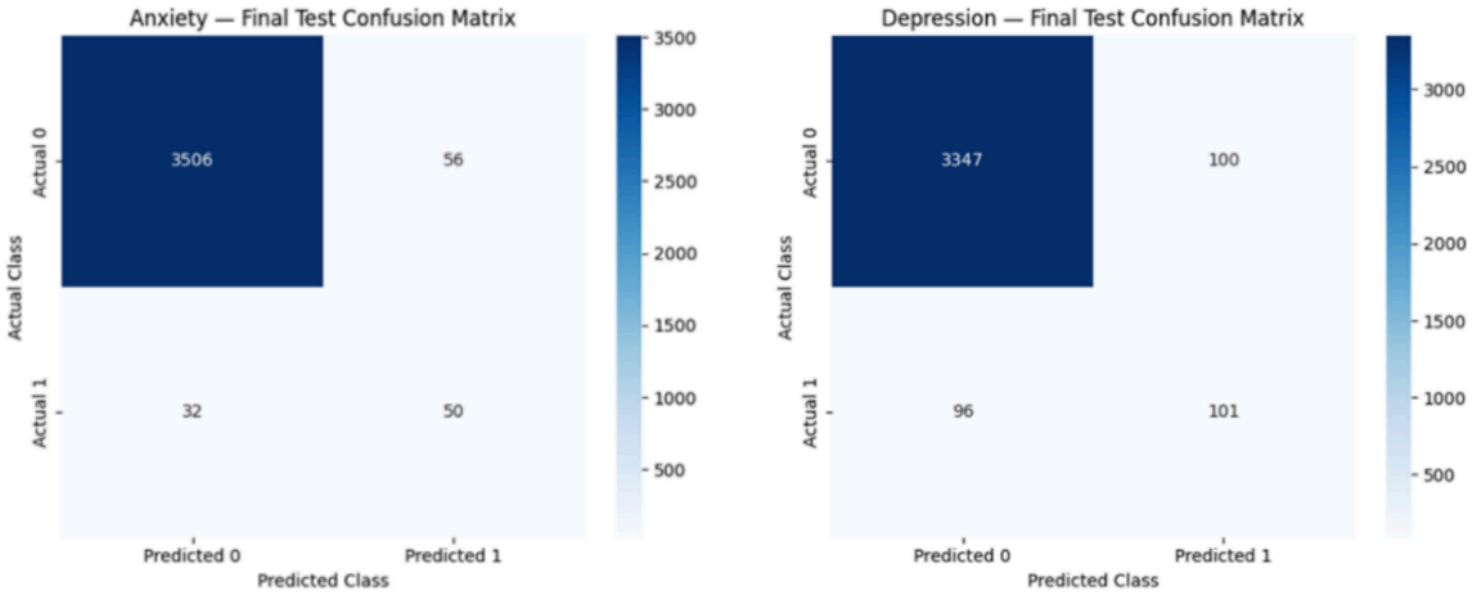


Figure 5

Confusion matrices of the optimized XGBoost models for (a) anxiety and (b) depression risk prediction.

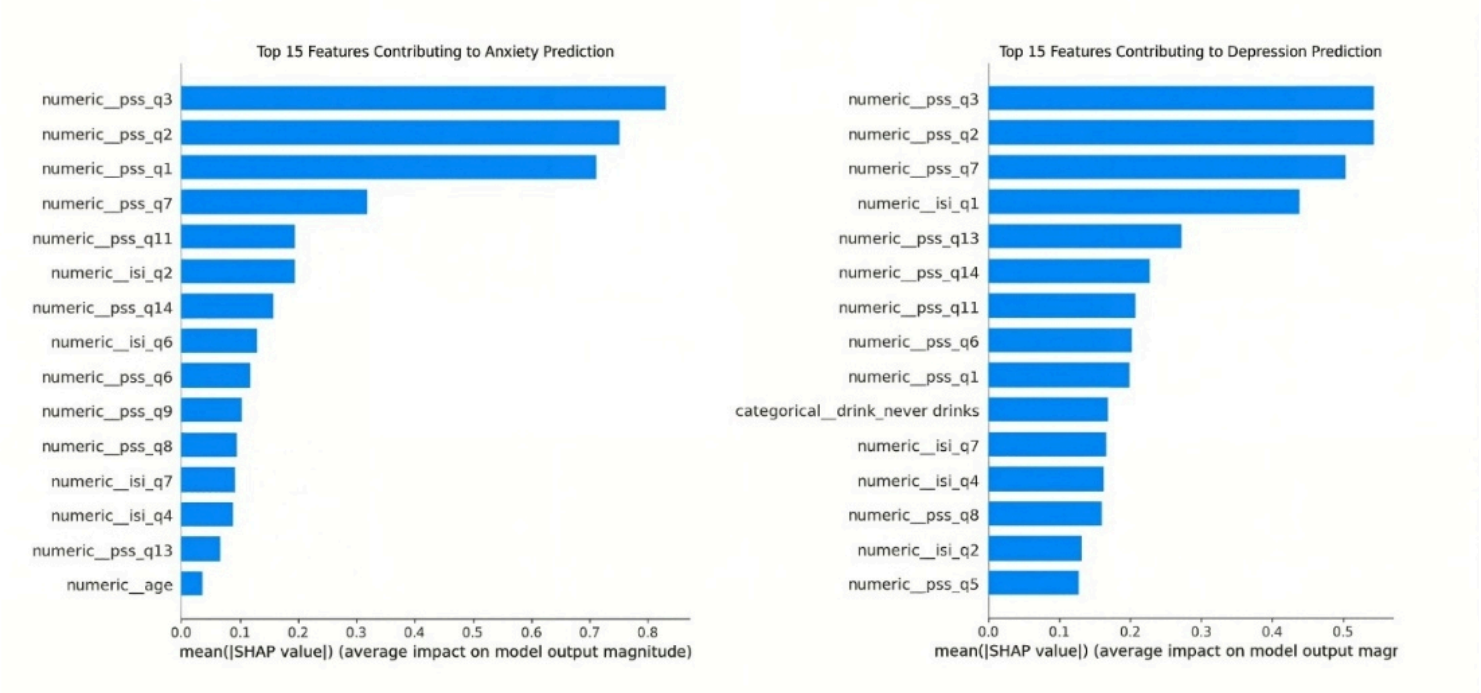


Figure 6

Global SHAP feature-importance ranking for anxiety and depression risk prediction.

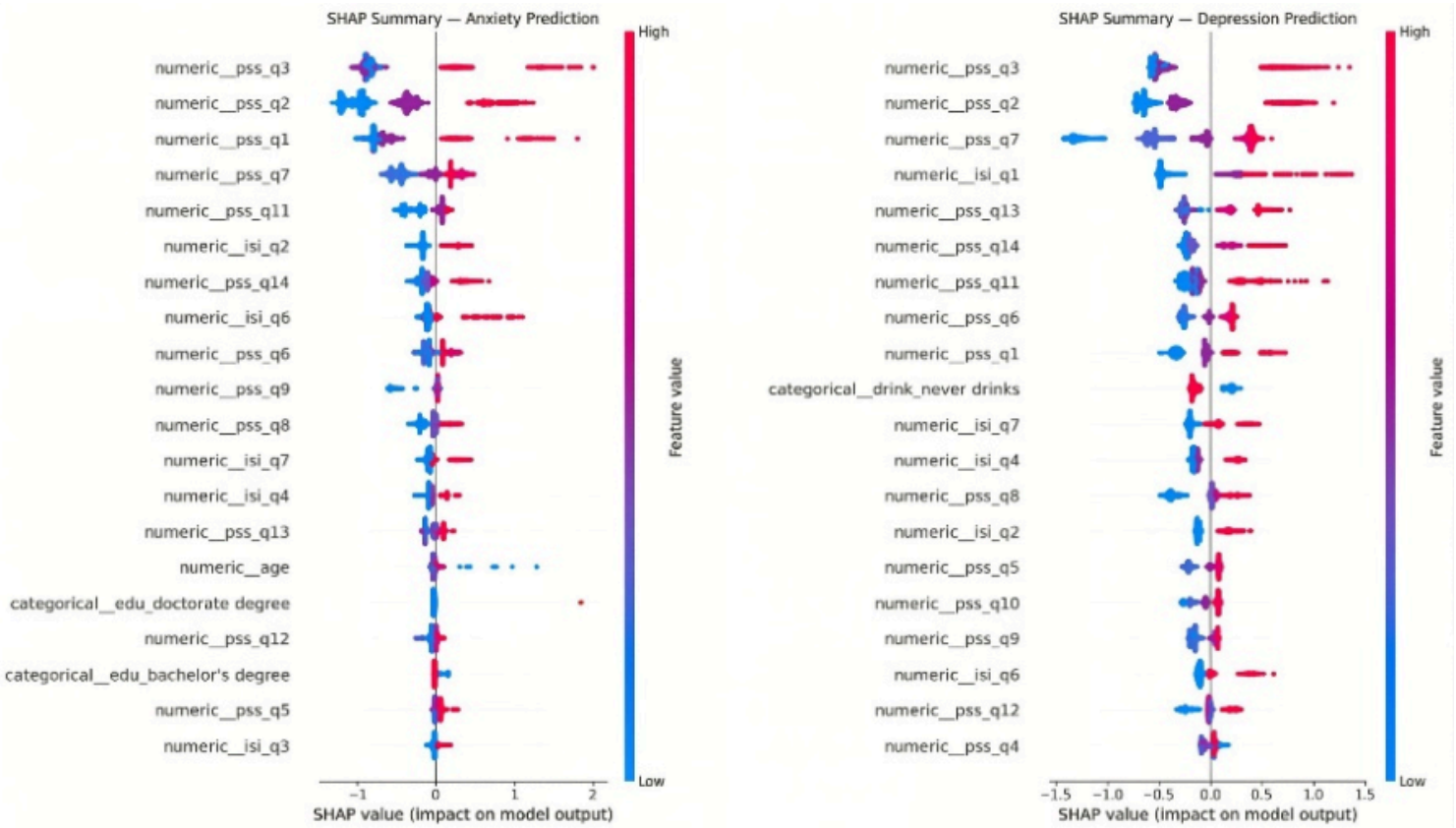


Figure 7

SHAP beeswarm plot for anxiety and depression risk prediction.

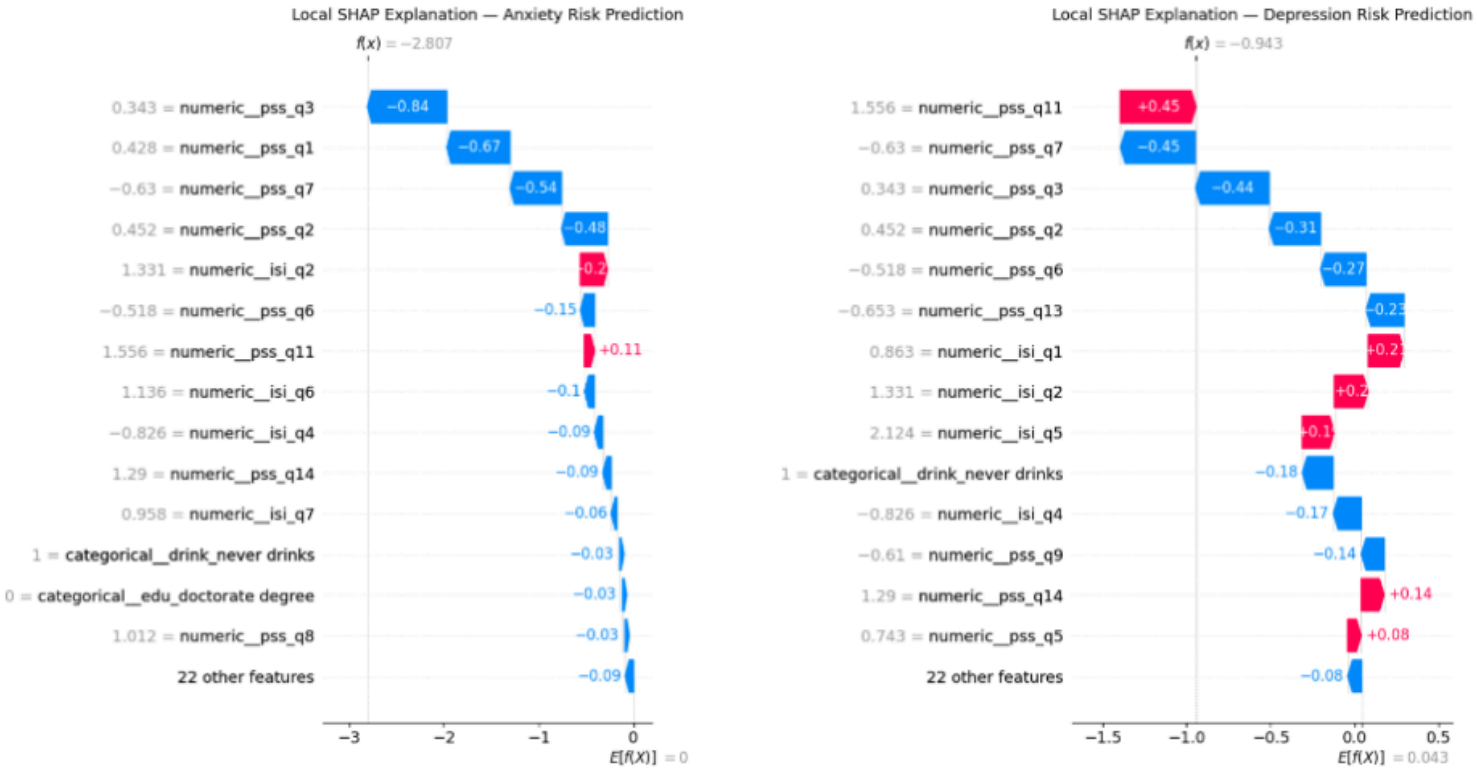


Figure 8

Local SHAP explanation for a representative participant's anxiety and depression risk predictions.