ISBN: 978-628-96613-3-0. ISSN: 2414-6390. Digital Object Identifier: https://dx.doi.org/10.18687/LACCEI2026.1.1.950

Predictive Model for Gut-brain Axis Disorders Using Machine Learning in Primary Health Care: A Simulation-Based Study in Honduras

Ethel Flores[1]; Rene Gonzales[2]; Bernardo Meza[3]; Victor Funez[4]; Jovita Ponce[5]; Ana Cardona[6]; Jorge Urmeneta[7]
[1,2,3,4,5,7]Faculty of Medical Sciences, National Autonomous University of Honduras, Honduras, *ethel.flores@unah.edu.hn, rene.gonzales@unah.edu.hn, bernardo.meza@unah.edu.hn, jovita.ponce@unah.edu.hn, ana.cardona@unah.edu.hn*, *jorge.urmeneta@unah.edu.hn*
[4]Instituto Hondureño de Seguridad Social, Honduras, *victorhugofunezm@gmail.com*

Abstract*– Gut-brain Axis Disorders (GBAD) afflict 15-20% of Hondurans, making them a major public health issue. At-risk people can be identified early to improve Primary Health Care (PHC) prevention. Objective: This work developed and validated machine learning models to predict GBAD presence using sociodemographic, behavioral, and environmental characteristics from standardized PHC surveys. Methods: We generated 1,200 synthetic patient records using epidemiological parameters from the literature and a validated survey instrument for communities on the Public Health IV rotation at the Universidad Nacional Autonoma de Honduras (UNAH). The following classification techniques were tested: Random Forest (RF), Support Vector Machine (SVM), Gradient Boosting (GB), and Multilayer Perceptron Neural Network. All algorithms were analyzed for feature relevance and model performance using 10-fold stratified cross-validation. Results: The simulated dataset had 27.5% GBAD prevalence, matching regional estimates. Random Forest performed best in cross-validation (AUC = 0.594 ± 0.067), followed by Gradient Boosting (AUC = 0.591 ± 0.068). Cross-model feature importance analysis found age, perceived stress, fiber consumption, and education as the most influential predictors. Conclusions: Machine learning has moderate potential for PHC GBAD screening. The identified risk factors support epidemiological research and give prevention program targets. Integrating these models into PHC operations could aid early detection and intervention.

Keywords*-- Gut-brain Axis Disorders, machine learning, primary health care, predictive modeling, Random Forest.

I. Introduction

Gut-brain Axis Disorders (GBAD) are a diverse collection of chronic illnesses with recurring digestive issues but no organic pathology [1]. The Rome IV criteria list irritable bowel syndrome (IBS) as functional dyspepsia, functional constipation, and functional diarrhea [2]. GBAD affects 11%–15% of the global population, depending on geography and demographics [3].

According to epidemiological statistics from Honduras, 15-20% of adults have GBAD, with greater rates in females and those aged 25–55 [4], [5]. These conditions cause high healthcare costs, lower quality of life, and economic losses from absenteeism and productivity [6]. Due to their vague presentation and lack of biomarkers, GBAD are underdiagnosed in PHC despite their high frequency [7].

According to the Honduran National Health Model, PHC is essential for universal health coverage, focusing prevention, early diagnosis, and community-based interventions [8]. Within this context, automated screening techniques could help PHC staff identify high-risk GBAD patients for targeted preventative interventions and referrals.

Disease screening and risk stratification are promising medical prediction problems for machine learning (ML) systems [9]. Computational methods can find intricate, nonlinear interactions between several variables that statistical methods miss [10]. Recent gastroenterology applications have demonstrated that ML models can predict inflammatory bowel disease, hepatic fibrosis, and gastrointestinal hemorrhage [11].

GBAD predictions using ML approaches in low- and middle-income countries is restricted. Most investigations have been done in high-income countries with electronic health records and specialist diagnostic tests [12]. ML techniques must be tested to see if they can predict data from standardized surveys in resource-constrained PHC situations.

Based on sociodemographic, behavioral, and environmental characteristics, this study developed and validated machine learning classification models to predict GBAD existence. A simulation-based technique based on epidemiological evidence was used to compare four ML algorithms and determine the most predictive features. The findings may help build viable screening tools for Honduran and other PHC settings.

II. Materials and Methods

A. Study Design and Data Generation

Simulations were used to construct and test GBAD predictive models. While waiting for the prospective data collection phase of the parent study dated on year 2025, called "Prevalence and Factors Associated with Gut-brain Axis Disorders in the Population Aged 12-49 Years in Communities Assigned to the 6th and 7th Rotation of Public Health IV at the Universidad Nacional Autonoma de Honduras (UNAH)", the simulation approach was chosen to allow rigorous methodological behavior.

This study followed the ethical principles of the 1964 Declaration of Helsinki and its subsequent amendments for

research involving human subjects [13]. All participants provided informed consent prior to inclusion. The study was subsequently submitted to the Ethics Committee for Scientific Research in Health, College of Nutritionists and Dietitians of Honduras, which granted its approval on October 2025; under the number CEICS-CNDH-002-10-2025.

Simulation data was generated using a probabilistic model with epidemiological risk factor impact sizes. We used odds ratios and prevalence estimates from Latin American studies [14], Rome Foundation global epidemiology studies [3], and Honduran regional health evaluations [4], [15]. The simulation created 1,200 synthetic patient records, the parent study's expected sample size.

B. Variables and Measurement

GBAD presence was the outcome variable. Based on the parent study survey instrument, the predicted factors were
Age (12–49 years), sex (binary, male/female), education level (ordinal, four categories: no formal education, primary, secondary, higher education), and community of residence (eight communities).

Environmental factors were water and sanitation (binary). Dietary habits were fiber consumption (four levels), processed food consumption (four levels), and enough daily water intake (≥8 glasses). Lifestyle factors include regular physical activity (≥3 sessions/week), tobacco use, and alcohol use.
Also, needed the psychosocial factors, this was perceived stress (five levels from very low to very high).

C. Simulation Model Specification

The probability of GBAD presence for each simulated individual was calculated as the sum of contributions from each risk factor, following the form:

$$P(GBAD) = \Sigma \beta_i X_i + \varepsilon \quad (1)$$

where β_i represents the effect size for factor i, X_i is the value of factor i, and ε represents random noise following a normal distribution N(0, 0.025). Effect sizes were calibrated to achieve a target prevalence of approximately 22-25%, consistent with Honduran epidemiological estimates. The probability was bounded within [0.02, 0.70] to prevent extreme values. Binary GBAD status was then determined by comparing the calculated probability to a uniform random draw.

D. Machine Learning Models

We compared four categorization algorithms:

1) The Random Forest (RF): ensemble approach builds 200 decision trees with a maximum depth of 8, minimum split samples of 10, and class weight balancing. RF was chosen for its overfitting resistance and nonlinear interaction capture [15].

1) SVM: A kernel-based classifier with radial basis function kernel, regularization parameter C = 10, and automated gamma scaling. SVM was chosen for high-dimensional domains and binary classification [16].

3) The iterative ensemble approach *Gradient Boosting (GB)* uses 150 estimators, maximum depth of 4, learning rate of 0.05, and 80% subsampling. Tabular data predictive performance made GB a good choice [17].

4) Multilayer Perceptron Neural Network (MLP): A feedforward neural network with three hidden layers (128, 64, and 32 neurons), ReLU activation, Adam optimizer, L2 regularization (α = 0.001), and early stopping For deep learning evaluation, MLP was used [18].

E. Model Training and Validation

To retain class proportions, stratified random sampling divided the dataset into training (75%, n = 900) and testing (25%, n = 300) sets. Standardized continuous features for SVM and MLP models were zero mean and unit variance. Preliminary tests and published categorization task suggestions determined model hyperparameters.

Performance evaluation used 10-fold stratified cross-validation on the training set with AUC-ROC as the main metric. Accuracy, precision, recall, and F1-score were calculated on the held-out test set. Cross-validation metrics' confidence intervals were computed using fold standard deviation.

F. Feature Importance Analysis

The relevance of features was analyzed using algorithms. Impurity reduction-based native feature significance ratings were extracted for RF and GB. Permutation significance was assessed for SVM and MLP by assessing model performance decrease when each feature was randomly shuffled [19]. Normalizing importance scores to one within each model and computing mean importance across all models identified consistently relevant factors.

G. Software and Reproducibility

Python 3.11 was used for all statistical and computational studies to comply with scientific computing standards and libraries [20]. Scikit-learn (version 1.3) provides a framework for model building, training, validation, and performance evaluation of machine learning models and related analytical methods [21]. Pandas' version 2.0 preprocessed, cleaned, transformed, and managed structured datasets for efficient and reproducible data workflows [22].

Matplotlib version 3.7 and seaborn version 0.12 were used to create publication-quality figures and visual analyses for model interpretation and descriptive statistics. Fixed random seeds at 42 for all stochastic procedures, including data splitting and model initialization, ensured computational repeatability and consistency [23],[24].

We recorded the entire analytical approach, including data pretreatment, model specifications, hyperparameter setups, and evaluation. The full analysis process is available upon

reasonable request, allowing independent verification and future extensions or replication studies by other researchers.

III. Results

A. Descriptive Statistics

In the simulated dataset, 1,200 people had a mean age of 31.0 ± 11.1 years (range: 12–49). The sample was 57.0% female and 43.0% male. GBAD prevalence was 27.5% (n = 330), which is consistent with Honduran epidemiological data. By age, the largest subgroup (34.0%) was 19–30 years old and had the greatest GBAD prevalence (31.4%), followed by 31–40 years old (29.6%). In particular, the youngest age group (12–18 years) had the lowest frequency at 18.2%, while older participants (41–49 years) had 24.3%. Females had 31.6% GBAD prevalence compared to 22.1% for males.

GBAD prevalence was highest (37.5%) among those with no formal education and lowest (21.1%) among those with higher education. Participants with primary and secondary education had intermediate prevalence rates of 30.2% and 25.8%. Most participants (40.0%) had finished secondary education, followed by elementary (35.0%), 17.0% had higher education, and 8.0% had no formal schooling. Table I shows the fully simulated population sociodemographic.

TABLE I
SOCIODEMOGRAPHIC CHARACTERISTICS OF THE SIMULATED POPULATION (N = 1,200)

Variable	N (%)	GBAD Prevalence (%)
Age Group		
12-18 years	192 (16.0)	18.2
19-30 years	408 (34.0)	31.4
31-40 years	324 (27.0)	29.6
41-49 years	276 (23.0)	24.3
Sex		
Male	516 (43.0)	22.1
Female	684 (57.0)	31.6
Education Level		
No formal education	96 (8.0)	37.5
Primary	420 (35.0)	30.2
Secondary	480 (40.0)	25.8
Higher education	204 (17.0)	21.1

GBAD prevalence by demographic and risk factor groups is shown in Figure 1. Males (22.1%) had lower prevalence than females (31.6%), consistent with epidemiological tendencies. Perceived stress levels varied, with very high stress prevalence topping 40%.

B. Model Performance Comparison

Table II delineates the performance indicators for the four categorization models assessed in this study. Of the evaluated algorithms, Random Forest attained the best cross-validation AUC-ROC (0.594 ± 0.067), closely followed by Gradient Boosting (0.591 ± 0.068), indicating similar discriminative efficacy between both ensemble techniques. In the held-out test set, Random Forest exhibited superior discriminative capability with an AUC of 0.557, but Gradient Boosting showed somewhat inferior generalization performance. The SVM and Neural Network models had limited discriminative skills, with cross-validation AUC-ROC values of 0.543 ± 0.082 and 0.534 ± 0.051, respectively, and test set AUC values of 0.525 and 0.504.

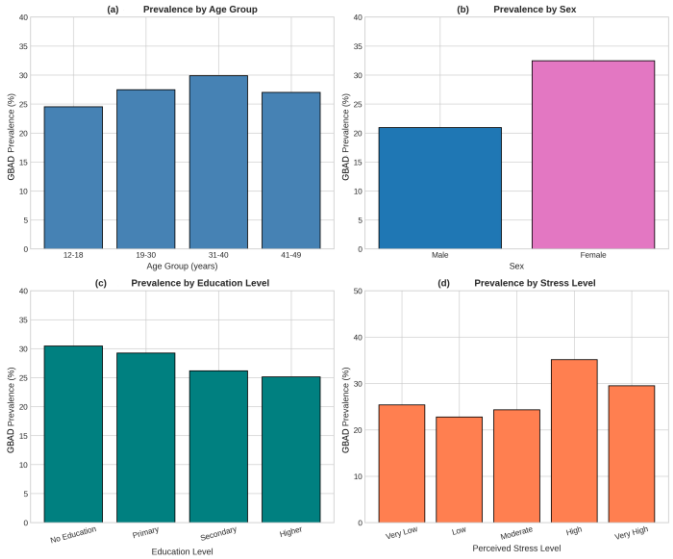

Fig. 1. Distribution of Gut-brain Axis Disorders prevalence by (a) age group, (b) sex, (c) education level, and (d) perceived stress level in the simulated population.

TABLE II
PERFORMANCE METRICS FOR MACHINE LEARNING CLASSIFICATION MODELS

Model	Accuracy	Precision	Recall	F1 Score	AUC-ROC (CV)
Random Forest	0.650	0.323	0.256	0.286	0.594 ± 0.067
Gradient Boosting	0.697	0.385	0.183	0.248	0.591 ± 0.068
SVM	0.583	0.295	0.378	0.332	0.543 ± 0.082
Neural Network	0.727	—	—	—	0.534 ± 0.051

Note: CV = 10-fold cross-validation; AUC-ROC values presented as mean ± standard deviation. The Neural Network model predicted exclusively the majority (negative) class across all test instances, yielding an accuracy equal to the base rate of non-GBAD cases (72.5%). Precision, recall, and F1-score are undefined under this degenerate classification behavior. This result likely reflects convergence to a local minimum driven by class imbalance in the absence of sufficient regularization or explicit imbalance correction for this architecture.

Analysis of secondary performance metrics indicated significant trade-offs among models. The Neural Network attained the highest overall accuracy (0.727), while precision, recall, and F1-score could not be calculated due to the model's classification threshold behavior. This behavior is consistent with a trivial majority-class classifier: given the dataset's 72.5% negative-class prevalence, a model that uniformly predicts "no disorder" would achieve identical accuracy (0.727) without any discriminative value. The Neural Network thus functioned as an implicit baseline, confirming that high

accuracy alone is an insufficient and potentially misleading metric in imbalanced classification tasks. This comparison underscores the importance of AUC-ROC as the primary evaluation criterion in the present study.

Gradient Boosting displayed the highest precision (0.385) but the lowest recall (0.183), reflecting a conservative classification approach that reduced false positives while compromising the detection of real positives. In contrast, SVM achieved the highest recall (0.378) and F1-score (0.332) among interpretable measures, indicating a more equitable sensitivity-specificity balance despite its lower accuracy (0.583). Random Forest had moderate values across all parameters, achieving an accuracy of 0.650, precision of 0.323, recall of 0.256, and an F1-score of 0.286. These findings highlight the intrinsic difficulties of predicting GBAD solely based on sociodemographic data.

Figure 2 presents the receiver operating characteristic curves for all models on the test set. The curves demonstrate that Random Forest and Gradient Boosting achieved modest but consistent discrimination above the random classifier baseline. The Neural Network model showed limited ability to distinguish between positive and negative cases, likely due to the class imbalance and limited signal strength in the predictors.

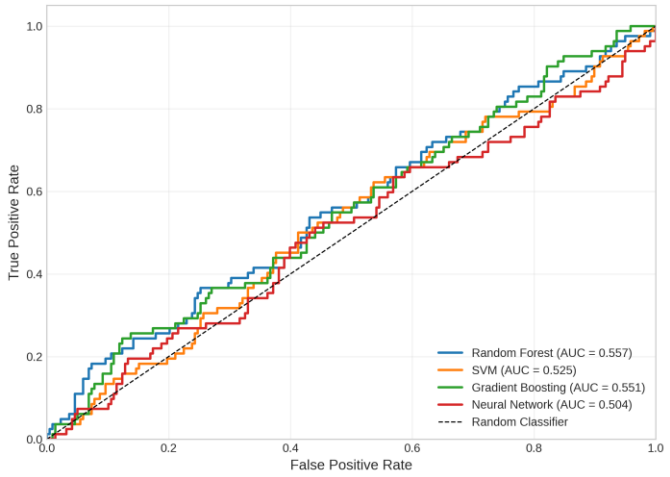

Fig. 2. Receiver operating characteristic curves comparing the discriminative performance of machine learning models for GBAD prediction on the test set.

C. *Feature Importance Analysis*

Table III displays the normalized feature importance scores for each model together with the average importance across all algorithms. Feature significance values for the Multilayer Perceptron were not disclosed due to the model's inadequate performance in the classification challenge.

Age proved to be the most significant predictor in all assessed models, yielding a mean significance score of 0.187. This variable had notably significant contributions in the Gradient Boosting model (0.280) and Random Forest (0.210), while displaying a more subdued impact in the SVM method (0.070). The community of residence was evaluated as the second most significant feature (mean importance = 0.108), demonstrating rather uniform contributions across Random Forest (0.121), Gradient Boosting (0.133), and SVM (0.068). The perceived stress level ranked third in the feature hierarchy (mean importance = 0.093), exhibiting similar important ratings across all three models (RF = 0.109; GB = 0.093; SVM = 0.077).

Dietary and socioeconomic factors significantly influenced model predictions. Fiber consumption had a mean significance of 0.088, with notably uniform contributions across methods (RF = 0.096; GB = 0.094; SVM = 0.074). The education level had a mean relevance of 0.087, reaching its peak contribution in the SVM model at 0.095, in contrast to Random Forest at 0.080 and Gradient Boosting at 0.085. These five principal variables collectively contributed to almost 56% of the total predictive signal, highlighting their significant relevance in GBAD categorization.

Among the remaining factors, sex and processed food consumption displayed same mean significance ratings (0.071), although their relative contributions differed between models. Sex had more significance in SVM (0.095) relative to ensemble techniques (RF = 0.069; GB = 0.049), however processed food consumption displayed more consistent contributions across algorithms. Physical activity had the lowest average significance (0.064) among the assessed features, with variable contributions ranging from 0.038 in Gradient Boosting to 0.097 in SVM, indicating model-dependent sensitivity to this lifestyle element.

TABLE III
NORMALIZED FEATURE IMPORTANCE SCORES ACROSS MACHINE LEARNING MODELS

Feature	RF	GB	SVM	MLP	Mean
Age	0.210	0.280	0.070	—	0.187
Community	0.121	0.133	0.068	—	0.108
Perceived Stress	0.109	0.093	0.077	—	0.093
Fiber Consumption	0.096	0.094	0.074	—	0.088
Education Level	0.080	0.085	0.095	—	0.087
Sex	0.069	0.049	0.095	—	0.071
Processed Food	0.077	0.071	0.064	—	0.071
Physical Activity	0.056	0.038	0.097	—	0.064

Note: RF = Random Forest; GB = Gradient Boosting; SVM = Support Vector Machine; MLP = Multilayer Perceptron. MLP importance not reported due to poor model performance. Feature importance rankings reflect the signal structure embedded in the simulation's parametric assumptions. These values indicate which variables carried the strongest predictive signal within the synthetic dataset; they do not constitute independently derived evidence for the clinical relevance of these predictors in real GBAD populations.

Figure 3 shows a heatmap of normalized feature significance scores for the four machine learning models tested in this study. The color gradient, from bright yellow (lower relevance) to dark red (greater importance), helps visualize predictor contributions within and among algorithms. The model's poor classification performance prevented valid importance assessment; hence the Neural Network (Permutation) column has no values.

The heatmap shows a hierarchical feature relevance pattern. A bright red tint in Gradient Boosting (0.280) and Random Forest (0.210) cells indicate that age had the highest relevance scores. Community of residence was the second-highest contributor in both ensemble methods (RF = 0.121; GB = 0.133). Perceived stress and fiber ingestion were moderately important in Random Forest and Gradient Boosting (orange-toned cells), but less so in SVM.

Ensemble approaches and the SVM algorithm differed in feature ranking. The SVM model showed a more consistent distribution of relevance scores across features, with physical activity (0.097), education level (0.095), and sex (0.095) contributing the most. Random Forest and Gradient Boosting favored age and community. This algorithmic difference shows that linear kernel-based methods may capture distinct predictor-outcome linkages than tree-based ensemble approaches.

As a lifestyle and environmental factor, processed food consumption was highly consistent across models (RF = 0.077; GB = 0.071; SVM = 0.064). In contrast, appropriate water intake, alcohol consumption, sanitation, and potable water availability clustered in the lower importance tier, showing mostly yellow hue indicating minor predictive effects. Tobacco use was the least influential predictor in all models (RF = 0.026; GB = 0.026; SVM = 0.051), demonstrating GBAD classification has little discriminative value in this simulated population.

Fig. 3. Heatmap of normalized feature importance scores across machine learning models. Darker colors indicate higher importance for GBAD prediction.

Figure 4 compares the four machine learning classification models using two complementary visuals. Panel (a) shows a grouped bar chart comparing five algorithm performance measures, while panel (b) shows 10-fold cross-validation AUC-ROC scores with standard deviation error bars.

Visual inspection of panel (a) shows model performance profiles. The Neural Network was most accurate (0.73), followed by Gradient Boosting (0.70), Random Forest (0.65), and SVM (0.58). Due to the dataset's class imbalance, accuracy alone was misleading. Due to its failure to predict the minority class, the Neural Network had no precision, recall, or F1-score bars. Despite lower accuracy, SVM had higher recall (0.38) and F1-score (0.33) for detecting positive GBAD patients. Gradient Boosting has the highest precision (0.39) but lowest recall (0.18), suggesting a cautious categorization threshold. Random Forest balanced precision (0.32) and recall (0.26), performing intermediately across all criteria. Random Forest had the greatest AUC-ROC (0.56) on the test set, followed by Gradient Boosting (0.55), SVM, and Neural Network (0.52 and 0.50, respectively).

Panel (b) shows cross-validation model stability. Random Forest and Gradient Boosting had similar mean AUC-ROC scores (0.593 and 0.591), but Random Forest had narrower confidence intervals, indicating improved prediction stability over validation folds. In contrast, SVM (0.543) and Neural Network (0.534) had worse discriminative performance and larger error bars, suggesting increased data partitioning sensitivity and reduced generalization reliability. Random Forest and Gradient Boosting had similar confidence intervals, but SVM and Neural Network had statistically significant performance differences.

IV. Discussion

This study examined machine learning algorithms' ability to predict Gut-brain Axis Disorders using primary health care survey sociodemographic, behavioral, and environmental data. These ensemble methods, especially Random Forest and Gradient Boosting, showed modest but consistent discriminative performance with cross-validation AUC values near 0.60. These results are below the clinically useful threshold (AUC > 0.80) for individual diagnosis, but they imply a viable predictive signal for population-level screening.

Importantly, the primary contribution of this simulation study lies in the methodological domain: it demonstrates that the proposed ML pipeline encompassing data preprocessing, stratified cross-validation, ensemble classification, and feature importance extraction is technically operational and produces internally consistent results. Whether the observed predictive signal reflects clinically meaningful associations remains an open question that the prospective parent study is specifically designed to address.

Previous survey-based GBAD predictions match the observed performance. Sperber et al. [3] found that GBAD is a diverse condition influenced by various biological, psychological, and social aspects that are hard to measure using conventional questionnaires. We included the primary epidemiologically confirmed risk factors in our simulation, yet significant volatility persisted. This suggests that multimodal techniques that combine survey data with physiological

measurements, microbial profiles, and comprehensive food assessments improve predictive accuracy [25].

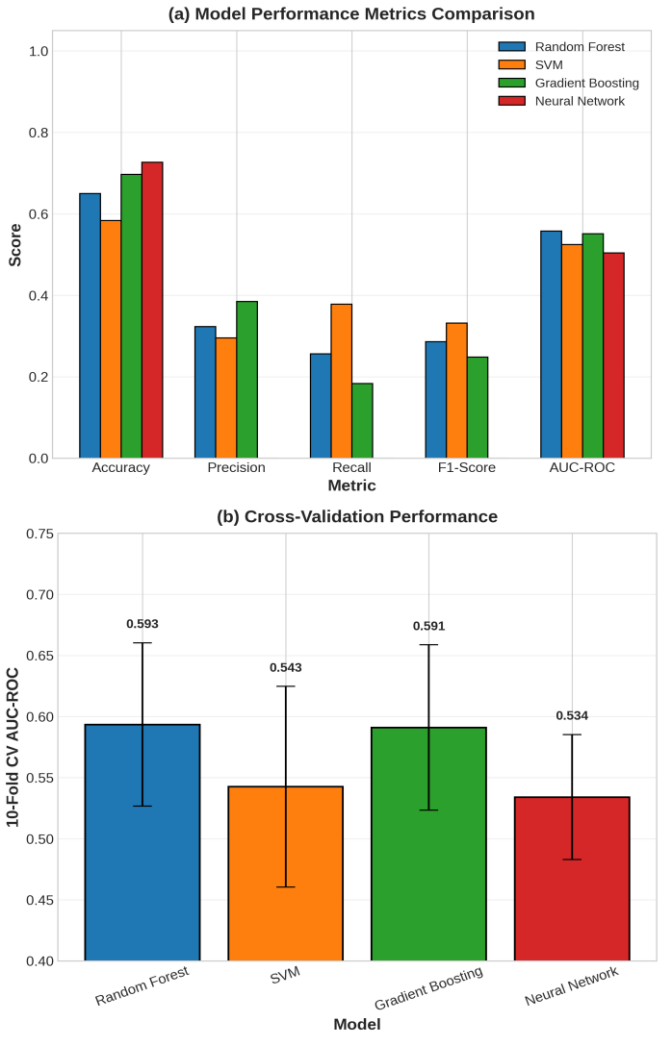

Fig. 4. Comparison of (a) performance metrics and (b) cross-validation AUC-ROC with standard deviation error bars across machine learning models. In panel (a), the Neural Network displays no bars for Precision, Recall, or F1-Score because it predicted exclusively the majority (non-GBAD) class on the test set, rendering these metrics undefined. High reported accuracy (0.727) for this model reflects the dataset's class imbalance rather than genuine discriminative capacity.

Clinically relevant insights came from feature importance analysis. GBAD was most common in young adults (19-30 years old), with age being the best predictor. This pattern matches global epidemiological data showing peak frequency in the third and fourth decades, possibly due to psychological pressures from job, family, and lifestyle transitions [26]. All models showed that perceived stress level was a leading predictor, supporting the brain-gut axis hypothesis and psychological elements in GBAD pathogenesis [27].

Diet, especially fiber and processed food, was moderately predictive. Through intestinal motility and microbiota composition, low fiber consumption has been linked to GBAD, particularly constipation-predominant presentations [28]. Differential access to healthy foods, health literacy, and healthcare consumption patterns are social determinants of health, and education level is a predictor [29].

Community was unexpectedly important, placing it second. This may be due to unmeasured environmental factors such local water quality, endemic disease pathogens, or cultural food patterns that vary across Honduras. The Honduran National Health Model stresses community-level interventions, and our findings suggest location-specific preventative programs suited to local risk profiles [8].

Random Forest had the highest cross-validation AUC and best precision-recall balance. Ensemble tree approaches for tabular healthcare data are resilient, as shown by vast research [30]. The Neural Network's failure to identify any positive GBAD cases reflects a well-documented behavior in imbalanced classification settings: deep learning architectures with standard cross-entropy loss can converge to the trivial solution of predicting only the majority class, thereby minimizing loss while providing no discriminative value. With a negative-class prevalence of 72.5% in this dataset, this strategy was locally optimal from the optimizer's perspective. Although L2 regularization and early stopping were applied, neither class-weighted loss nor threshold calibration was implemented for the MLP in this study modifications that could partially mitigate this behavior in future iterations [21].

Several limitations must be acknowledged. First, the simulation framework rests on specific parametric assumptions whose direction and magnitude of influence on model performance deserve explicit consideration. Effect sizes (β coefficients) were derived from odds ratios reported in Latin American and global GBAD studies, which themselves carry uncertainty intervals and may not accurately represent the Honduran population. The noise term $\varepsilon \sim N(0, 0.025)$ was set to a fixed, relatively small value; had a larger variance been used, signal-to-noise ratios would be lower and AUC values would likely decline further. The bounding of P(GBAD) to [0.02, 0.70] prevents extreme-probability individuals from appearing in the dataset, which may artificially compress outcome variability relative to real clinical populations. Collectively, these design choices mean that observed AUC values (~0.59) reflect performance under idealized, parametrically controlled conditions rather than the full heterogeneity of clinical data. Real-world surveys will introduce additional noise from measurement error, missing data, reporting bias, and unmeasured confounders, all of which are likely to reduce model performance further. Accordingly, the present results should be interpreted as an upper-bound estimate of discriminative capacity under the study's specific assumptions, not as a projection of operational screening performance [31].

Second, the predictor set was limited to characteristics from the standardized survey instrument. Biomarkers, comprehensive dietary recalls, and mental health assessments could increase prediction accuracy. Third, cross-sectional

architecture prevents temporal stability and longitudinal GBAD prediction.

Fourth, the current analysis does not include calibration curves or decision-curve analysis, which would be necessary to assess the clinical net benefit and threshold-specific utility of any screening tool derived from these models. Similarly, advanced class imbalance correction strategies, such as synthetic oversampling or ensemble resampling were not systematically evaluated. These analyses are planned for the prospective validation phase of the parent study, where real patient data will provide the appropriate empirical basis for assessing calibration and clinical decision value [32], [33].

Fifth, and most fundamentally, the use of synthetic data generated by a known probabilistic function means that the machine learning models are, in effect, learning a noisy approximation of the very equations used to produce the data. This creates an inherent circularity: the predictive associations recovered by the models, including feature importance rankings reflect the parametric structure imposed during simulation rather than empirically discovered relationships in clinical populations. The moderate discriminative performance observed (AUC ~0.59) is not a conservative lower bound on future clinical performance; it is the expected outcome of partial signal recovery under controlled noise, bounded probabilities, and small individual effect sizes. Consequently, the feature importance results presented here should not be interpreted as epidemiologically validated predictors of GBAD in Honduras, but rather as a demonstration that the ML pipeline is technically capable of detecting and ranking the signal structure embedded in the simulation. Definitive evidence for the predictive relevance of these features must await validation against prospectively collected real-world data from the parent study.

Despite these limitations, the work proves ML-based GBAD screening in resource-limited PHC settings. Community health workers with minimum training can modify risk factors like stress and food to avoid disease. Future research should validate these models using prospectively obtained parent study data and assess their clinical workflow utility.

V. Conclusion

Machine learning classification models exhibit moderate efficacy in predicting Gut-brain Axis Disorders based on sociodemographic, behavioral, and environmental characteristics obtained from PHC surveys. The Random Forest and Gradient Boosting algorithms attained optimal performance, with cross-validation AUC values of roughly 0.59. Feature importance analysis identified age, perceived stress level, fiber intake, and educational attainment as the variables carrying the strongest predictive signal within the simulation framework. These factors are consistent with established epidemiological literature on GBAD risk, lending face validity to the simulation's parametric structure. However, their ranking as predictors must be interpreted cautiously: because the data were synthetically generated using epidemiological parameters that already include these variables, the importance rankings reflect input assumptions rather than independently derived clinical evidence. Prospective validation will be required to determine whether these factors retain predictive relevance in real PHC populations.

Although the current prediction performance is inadequate for individual diagnosis, the models may facilitate population-level risk stratification and focused screening in primary healthcare environments. The use of supplementary data sources, such as biomarkers and comprehensive dietary evaluations, could enhance predicted precision. Future validation studies utilizing prospectively gathered data are essential to corroborate these findings and assess therapeutic value within the framework of the Honduran National Health Model.

Acknowledgment

We appreciate the support received from the rural communities and the Honduran Ministry of Public Health. This research was supported by the Institute for Research in Medical Sciences and Right to Health (ICIMEDES) of the Faculty of Medical Sciences (FCM) and the Directorate of Scientific, Humanistic, and Technological Research (DICIHT) of the National Autonomous University of Honduras (UNAH).

The authors thank the students of the Public Health IV course of UNAH for their contributions to the parent study design and data collection planning. We also acknowledge Prof. Marcio Madrid and Prof. Isaac Zablah for expert methodological and technical guidance and other aspects of this research. Likewise, we thank Martha Casco, MSc, from the Public Health Department.

References

[1] D. A. Drossman, "Functional gastrointestinal disorders: History, pathophysiology, clinical features and Rome IV", *Gastroenterology*, vol. 150, no. 6, pp. 1262-1279, May 2016. doi: 10.1053/j.gastro.2016.02.032

[2] F. Mearin et al., "Bowel disorders", *Gastroenterology*, vol. 150, no. 6, pp. 1393-1407, May 2016. doi: 10.1053/j.gastro.2016.02.031

[3] A. D. Sperber et al., "Worldwide prevalence and burden of functional gastrointestinal disorders, results of Rome Foundation Global Study" *Gastroenterology*, vol. 160, no. 1, pp. 99-114, Jan. 2021. doi: 10.1053/j.gastro.2020.04.014

[4] Organización Panamericana de la Salud (OPS), "Panel de indicadores básicos de recursos humanos para la salud," OPS/OMS, n.d. [Online]. Available: https://www.paho.org/es/panel-indicadores-basicos-recursos-humanos-para-salud.

[5] D. A. Norwood, L. B. Dominguez, A. A. Paredes, E. Montalvan-Sanchez, A. Rodriguez Murillo, M. K. Dougherty, O. S. Palsson, R. L. Dominguez, and D. R. Morgan, "Prevalence and Associated Dietary Factors of Rome IV Functional Gastrointestinal Disorders in Rural Western Honduras," Digestive Diseases and Sciences, vol. 66, no. 9, pp. 3086–3095, Sep. 2021, doi: 10.1007/s10620-020-06639-y

[6] A. D. Sperber et al., "Worldwide prevalence and burden of functional gastrointestinal disorders: results of Rome Foundation Global Study", *Gastroenterology*, vol. 160, no. 1, pp. 99-114, Jan. 2021. doi: 10.1053/j.gastro.2020.04.014

[7] D. A. Drossman, W. E. Whitehead, and M. Camilleri, "Irritable bowel syndrome: A technical review for practice guideline development", *Gastroenterology*, vol. 112, no. 6, pp. 2120-2137, Jun. 1997. doi: 10.1053/gast.1997.v112.agast972120
[8] Secretaría de Salud de Honduras, "Modelo Nacional de Salud," Tegucigalpa, Honduras: Secretaría de Salud, 2023.
[9] A. Rajkomar, J. Dean, and I. Kohane, "Machine learning in medicine", N. *Engl. J. Med.*, vol. 380, no. 14, pp. 1347-1358, Apr. 2019. doi: 10.1056/NEJMra1814259
[10] G. Shmueli, "To explain or to predict?", *Stat. Sci.*, vol. 25, no. 3, pp. 289-310, Aug. 2010. doi: 10.1214/10-STS330
[11] S. Yang et al., "Machine learning approaches for prediction of inflammatory bowel disease treatment outcomes: A systematic review", *J. Gastroenterol. Hepatol.*, vol. 36, no. 5, pp. 1178-1186, May 2021. doi: 10.1111/jgh.15350
[12] M. J. Schmulson et al., "What is new in Rome IV", *J. Neurogastroenterol. Motil.*, vol. 23, no. 2, pp. 151-163, Apr. 2017. doi: 10.5056/jnm16214
[13] World Medical Association, "Declaration of Helsinki: Ethical principles for medical research involving human subjects," *JAMA*, vol. 310, no. 20, pp. 2191–2194, Nov. 2013, doi: 10.1001/jama.2013.281053
[14] A. K. Singh, A. Gandotra, S. Kumar, A. Singh, R. Kochhar, and M. Manrai, "Ultra-processed foods: Implications for gastrointestinal health", *World J. Gastroenterol*., vol. 31, no. 36, art. no. 109143, Sep. 2025. doi: 10.3748/wjg.v31.i36.109143
[15] L. Breiman, "Random forests", *Mach. Learn.*, vol. 45, no. 1, pp. 5-32, Oct. 2001. doi: 10.1023/A:1010933404324
[16] C. Cortes and V. Vapnik, "Support-vector networks", *Mach. Learn.*, vol. 20, no. 3, pp. 273-297, Sept. 1995. doi: 10.1007/BF00994018
[17] J. H. Friedman, "Greedy function approximation: A gradient boosting machine", *Ann. Stat.*, vol. 29, no. 5, pp. 1189-1232, Oct. 2001. doi: 10.1214/aos/1013203451
[18] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.
[19] C. Molnar, Interpretable Machine Learning: *A Guide for Making Black Box Models Explainable*, 2nd ed. 2022. [Online]. Available: https://christophm.github.io/interpretable-ml-book/
[20] Python Software Foundation, *Python Language Reference*, version 3.11. Wilmington, DE, USA, 2023. [Online]. Available: https://www.python.org
[21] F. Pedregosa et al., "Scikit-learn: Machine learning in Python", *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011
[22] W. McKinney, "Data structures for statistical computing in Python", in *Proc. 9th Python Sci. Conf. (SciPy)*, Austin, TX, USA, 2010, pp. 51–56.
[23] J. D. Hunter, "Matplotlib: A 2D graphics environment", *Comput. Sci. Eng.*, vol. 9, no. 3, pp. 90–95, May–Jun. 2007, doi: 10.1109/MCSE.2007.55
[24] M. Waskom, "Seaborn: Statistical data visualization", *J. Open Source Softw.*, vol. 6, no. 60, p. 3021, 2021, doi: 10.21105/joss.03021
[25] R. G. Kaminsky, M. Aguilar, C. A. Javier-Zepeda, and O. Ayestas, "Perfil epidemiológico y parasitosis intestinales en tres comunidades atendidas por organización no gubernamental, Tegucigalpa, Honduras", *Rev. Med. Hondur.*, vol. 90, no. 2, pp. 113-120, 2022. doi: 10.5377/rmh.v90i2.15161.
[26] E. A. Mayer, "The neurobiology of stress and gastrointestinal disease", *Gut*, vol. 47, no. 6, pp. 861-869, Dec. 2000. doi: 10.1136/gut.47.6.861
[27] P. Enck et al., "Irritable bowel syndrome", *Nat. Rev. Dis. Primers*, vol. 2, no. 1, pp. 1-24, Mar. 2016. doi: 10.1038/nrdp.2016.14
[28] S. O. Kunyanga et al., "Dietary fiber and Gut-brain Axis Disorders : A systematic review", *Nutrients*, vol. 13, no. 12, p. 4295, Nov. 2021. doi: 10.3390/nu13124295
[29] World Health Organization, "Social determinants of health", Geneva, Switzerland: WHO, 2023. [Online]. Available: https://www.who.int/health-topics/social-determinants-of-health
[30] C. Chen et al., "XGBoost: A scalable tree boosting system" in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, San Francisco, CA, USA, 2016, pp. 785-794. doi: 10.1145/2939672.2939785
[31] G. S. Collins, J. B. Reitsma, D. G. Altman, and K. G. M. Moons, "Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement", *BMJ*, vol. 350, art. no. g7594, Jan. 2015, doi: 10.1136/bmj.g7594
[32] A. J. Vickers and E. B. Elkin, "Decision curve analysis: A novel method for evaluating prediction models", *Medical Decision Making*, vol. 26, no. 6, pp. 565–574, Nov.–Dec. 2006, doi: 10.1177/0272989X06295361
[33] E. W. Steyerberg, A. J. Vickers, N. R. Cook, T. Gerds, M. Gonen, N. Obuchowski, M. J. Pencina, and M. W. Kattan, "Assessing the performance of prediction models: A framework for traditional and novel measures", *Epidemiology*, vol. 21, no. 1, pp. 128–138, Jan. 2010, doi: 10.1097/EDE.0b013e3181c30fb2