ISBN: 978-628-96613-3-0. ISSN: 2414-6390. Digital Object Identifier: https://dx.doi.org/10.18687/LACCEI2026.1.1.955

Clustering Analysis for Identification of Gastrointestinal Health Risk Profiles: A Community-Based Study Using Machine Learning Approaches

Ethel Flores[1]; Rene Gonzales[2]; Yenny Velasquez[3]; Oneyda Zelaya[4]; Karen Oliva[5]; Victor Funez[6]; Jorge Urmeneta[7]
[1,2,3,5,7]Faculty of Medical Sciences, National Autonomous University of Honduras, Honduras, *ethel.flores@unah.edu.hn, rene.gonzales@unah.edu.hn, yenny.velasquez@unah.edu.hn, karen.oliva@unah.edu.hn, jorge.urmeneta@unah.edu.hn*
[4]Policia Nacional de Honduras, Honduras, *lizethzelaya2010@gmail.com, oneydaespinoza1996@gmail.com*
[6]Instituto Hondureño de Seguridad Social, Honduras, *victorhugofunezm@gmail.com*

Abstract*– Developing countries struggle with functional gastrointestinal issues, which lower quality of life and healthcare use. Traditional epidemiological methods often miss risk profile variability in afflicted groups. This study used unsupervised machine learning to determine Honduran community gastrointestinal health risk profiles for targeted intervention. From September to December 2025, 1,838 Honduran 12-49-year-olds from five departments participated in a cross-sectional survey. Data included sociodemographics, gastrointestinal symptoms, lifestyle, and medical history. K-means, DBSCAN, and hierarchical agglomerative clustering were used. Silhouette coefficient, Davies-Bouldin index, and Calinski-Harabasz index assessed cluster validity. PCA with t-SNE reduced dimension for visualization. The best partitioning was K-means clustering with K=3 (silhouette score: 0.162). There were three risk profiles: (1) High-risk lifestyle cluster (10.2%; n=187) with universal smoking (100%), elevated alcohol consumption (61%), and moderate symptom prevalence; (2) Low-risk cluster (57.3%; n=1,054) with younger individuals with healthy lifestyle habits and minimal symptoms (9.2% abdominal pain); and (3) High-symptom cluster (32.5%; n=597) predominantly female (71.9%) with significant gastrointestinal complaints (67% abdominal pain, 62.6% heartburn) and Machine learning-based clustering enabled individualized primary healthcare intervention design by stratifying the population into clinically meaningful risk profiles. These data support gastrointestinal health program resource allocation optimization.

Keywords*-- Clustering analysis, gastrointestinal symptoms, K-means, primary healthcare, risk profiles.

I. Introduction

Gut-brain Axis Disorders (GBAD) are a diverse collection of illnesses with persistent or recurrent gastrointestinal symptoms without anatomical or biochemical abnormalities [1]. According to Rome IV criteria, 11-15% of the global population has these illnesses, putting strain on healthcare systems [2]. Up to 20% of Latin American adults may have gastrointestinal symptoms, with regional differences due to socioeconomic variables, food, and healthcare access [3].

Functional gastrointestinal symptoms are caused by complex biological, psychological, and social interactions [4]. Traditional epidemiological methods may ignore synergistic effects and miss subpopulations with diverse risk profiles by analyzing risk factors separately. This hinders targeted intervention efforts and optimum resource allocation in basic healthcare [5].

Population stratification using multidimensional health data is promising with machine learning, especially unsupervised clustering techniques [6]. These methods can identify natural groupings in datasets without preset categories, exposing clinically important patient subgroups that conventional analysis may miss [7]. Clustering methods have been shown to be useful for patient phenotyping in irritable bowel syndrome and functional dyspepsia [8], [9].

The Honduran National Health Model emphasizes PHC as the foundation for universal health coverage [10]. Designing community-based therapies requires understanding population-level gastrointestinal health patterns.

The Honduran National Health Model designates primary healthcare (PHC) as the cornerstone of universal health coverage, emphasizing community-level diagnosis and prevention [10]. While gastrointestinal epidemiology has been described in Latin American populations using traditional survey-based methods [3], [11], [12]; these approaches typically characterize prevalence at the group level without identifying internally coherent risk subgroups. To our knowledge, no prior Honduran study has applied unsupervised machine learning to segment community populations by gastrointestinal risk profile. This represents a meaningful gap: without computational stratification, PHC resource allocation remains undifferentiated, and high-risk subgroups defined by co-occurring symptom-lifestyle patterns may go undetected [5], [6]. The present study directly addresses this gap by combining epidemiological data collection with clustering-based analysis in a geographically diverse Honduran sample, offering a methodological and contextual contribution that extends beyond prior descriptive work in the region.

To address this gap, the present study aimed to identify gastrointestinal health risk profiles in a community-based Honduran population by applying and comparing multiple unsupervised machine learning clustering algorithms. Specifically, the study sought to: (1) stratify participants into clinically meaningful subgroups based on symptom burden, lifestyle exposures, and sociodemographic characteristics; (2)

evaluate cluster validity using internal validation metrics; and (3) generate evidence to support targeted intervention design within the Honduran primary healthcare system.

II. Materials and Methods

A. Study Design and Population

This cross-sectional epidemiological study was undertaken in communities assigned to the sixth and seventh rotations of Public Health IV at the National Autonomous University of Honduras from September to December 2025. The study included 12-49-year-olds from Intibuca, Colon, Comayagua, Olancho, and El Paraiso. A convenience selection technique yielded 1,838 participants who met inclusion criteria: 6-month community residence and informed consent.

B. Data Collection Instrument

A structured questionnaire was developed based on validated instruments for gastrointestinal symptom assessment, incorporating elements from the Rome IV diagnostic criteria and lifestyle assessment tools [1], [13]. The instrument captured: (1) sociodemographic variables including age, sex, educational level, and community of residence; (2) gastrointestinal symptoms including abdominal pain (past 30 days), heartburn (past 2 weeks), gastric reflux frequency, and post-prandial abdominal inflammation; (3) lifestyle factors including dietary fiber consumption, water intake, physical activity frequency, tobacco use, and alcohol consumption; and (4) medical history including previous gastrointestinal diagnoses and current health perception.

C. Data Preprocessing

Data preprocessing was performed using Python 3.11 with pandas (version 2.0) and scikit-learn (version 1.3) libraries. The 14-variable feature set included a mix of ordinal and nominal categorical variables alongside continuous measures. Ordinal categorical variables (e.g., educational level, reflux frequency, symptom self-rating) were encoded using integer label encoding to preserve their inherent rank structure, as these sequences carry meaningful distance information relevant to distance-based clustering. Nominal binary variables (e.g., sex, presence/absence of symptoms, smoking status, alcohol use) were encoded as binary indicators (0/1), which are equivalent to one-hot encoding for two-category variables and do not introduce artificial inter-category distances. No imputation was required, as missing data accounted for fewer than 0.5% of entries and were handled by listwise deletion. Feature standardization was applied using z-score normalization (mean = 0, standard deviation = 1), which is the appropriate scaling strategy for K-means clustering because the algorithm relies on Euclidean distance: without standardization, variables with larger absolute ranges would disproportionately dominate centroid calculation and distort cluster boundaries [14], [15].

D. Clustering Algorithms

For robust population stratification, three complementing clustering methods were used:

1) K-means clustering: A centroid-based technique minimizes within-cluster sum of squares to cluster observations into K [16]. The elbow approach and silhouette analysis identified the ideal number of clusters for K values from 2 to 10. K-means++ initialization enhanced convergence (10 initializations, maximum 300 iterations).

2) Cluster solution consistency: it was assessed across all ten independent initializations: the final partition showed identical cluster assignments regardless of random seed, indicating that the three-cluster solution represents a stable local optimum under the K-means++ criterion rather than an initialization-dependent artifact. Formal bootstrap-based stability analysis (e.g., Jaccard similarity across resampled partitions) was not conducted and is acknowledged as a limitation of this study.

3) The density-based spatial clustering of applications with noise (DBSCAN): with this the density-based approach classifies clusters as high-density regions divided by low-density areas [17]. To balance cluster granularity with noise point identification, parameter optimization examined epsilon values (1.5-3.0) and minimum samples (5-15).

4) Hierarchical Agglomerative Clustering: Under Ward's minimal variance criterion, this bottom-up method merges clusters to reduce within-cluster variance [18]. Dendrogram visualization guided cluster number selection.

E. Cluster Validation Metrics

Cluster quality was assessed using three internal validation indices:

1) Silhouette coefficient: This metric ranges from -1 to 1, measuring how similar an object is to its own cluster compared to other clusters. Values approaching 1 indicate well-separated clusters [19].

2) Davies-Bouldin index: This ratio of within-cluster distances to between-cluster distances yields lower values for better-defined clusters [20].

3) Calinski-Harabasz index: Also known as the variance ratio criterion, higher values indicate denser and better-separated clusters [21].

F. Dimensionality Reduction and Visualization

Two complementing cluster visualization methods were used:

1) PCA: Principal component analysis This linear method projects high-dimensional data onto orthogonal axes to capture maximum variance and understand main components [22].

2) T-Distributed Stochastic Neighbor Embedding (t-SNE): This non-linear method preserves local structure and shows cluster separability in two dimensions [23]. The parameters were 30 perplexity and 1000 iterations.

G. *Statistical Analysis*

Cluster profiles were described using means, standard deviations, and percentages for continuous and categorical variables. Heatmaps and radar plots showed cluster centroids. Python plus NumPy [24], SciPy [25], and Matplotlib [26] were used for all analyses.

H. *Ethical Considerations*

This study followed the ethical principles of the 1964 Declaration of Helsinki and its subsequent amendments for research involving human subjects [27]. 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; a legal requirement due the use of human data.

III. Results

A. *Study Population Characteristics*

The study included 1,838 eligible participants. The sample was 65.2% female (n=1,198) and had a mean age of 28.3 years (SD=10.9, range: 12-49). 42 (2.3%) were illiterate, 408 (22.2%) had incomplete primary education, 376 (20.5%) completed primary school, 467 (25.4%) had incomplete secondary education, 374 (20.4%) completed secondary school, and 171 (9.3%) had some university education. Participant locations included Intibuca (31.8%), Colon (29.9%), Comayagua (26.7%), Olancho (8.7%), and El Paraiso (2.9%).

A 32.4% (n=596) reported abdominal pain in the previous 30 days, 31.0% (n=570) heartburn in two weeks, and 31.9% (n=587) post-prandial abdominal inflammation. Gastric reflux occurred seldom (24.9%), occasionally (19.9%), frequently (8.1%), and always (2.5%). Lifestyle factors showed that 64.8% ate fiber-rich foods consistently, 42.8% drank 8 glasses of water daily, 47.8% exercised three or more times a week, 10.3% smoked, and 16.3% drank alcohol regularly. 21.1% (n=388) had gastrointestinal diagnoses.

B. *Optimal Cluster Number Determination*

The elbow method analysis revealed diminishing returns in inertia reduction beyond K=3 clusters (Fig. 1, left panel). Silhouette coefficient analysis confirmed K=3 as optimal, yielding the second-highest silhouette score (0.162) after K=2 (0.169), while providing greater clinical interpretability. The Davies-Bouldin index showed consistent values across cluster numbers, with K=3 producing an acceptable value of 2.33.

C. *Clustering Algorithm Comparison*

Table I compares three clustering techniques' performance metrics. K-means with K=3 had balanced silhouette scores of 0.162, Davies-Bouldin index of 2.33, and Calinski-Harabasz index of 232.59. DBSCAN (epsilon=1.5, minimum samples=15) found 8 clusters with 1,424 noise points (77.5%), limiting its practicality despite higher silhouette scores on clustered points. Hierarchical clustering using Ward linkage yielded fewer compact clusters with lower silhouette scores (0.084) and higher Davies-Bouldin index (2.79).

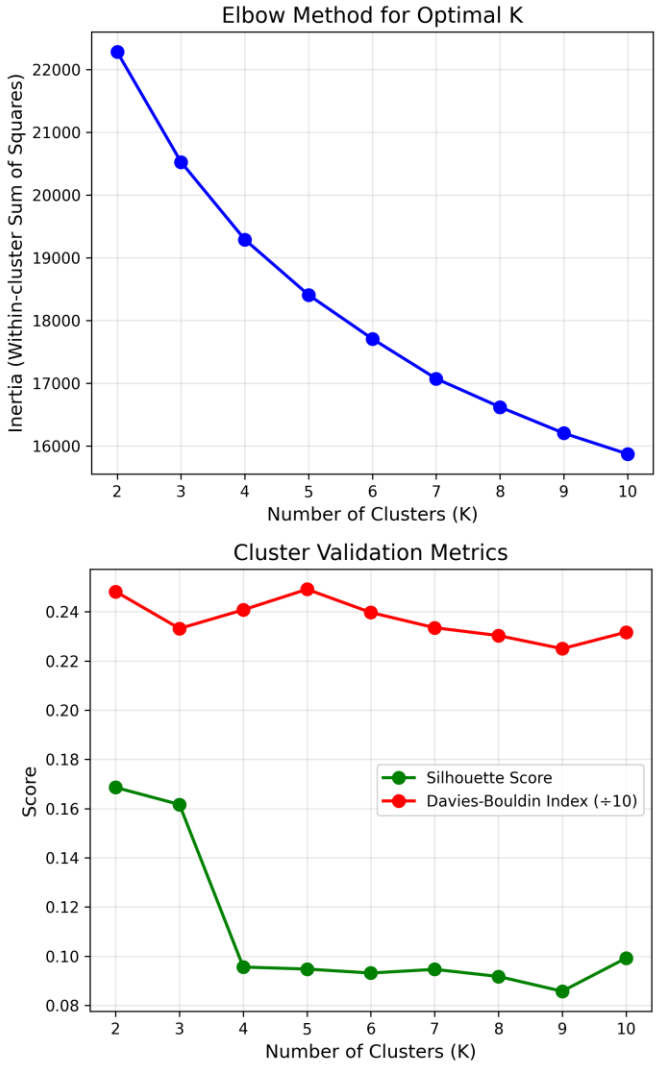

Fig. 1. Optimal cluster number determination. Left: Elbow method showing within-cluster sum of squares (inertia) versus cluster number. Right: Cluster validation metrics including silhouette score and Davies-Bouldin index (scaled by 1/10).

TABLE I
COMPARATIVE PERFORMANCE METRICS OF CLUSTERING ALGORITHMS

Algorithm	Clusters	Silhouette	Davies-Bouldin	Calinski-H.
K-means (K=3)	3	0.162	2.33	232.59
DBSCAN*	8	0.373	1.19	125.71
Hierarchical (Ward)	3	0.084	2.79	185.04

* DBSCAN parameters: epsilon=1.5, min_samples=15; 77.5% noise points excluded from metrics calculation.

D. *Dimensionality Reduction Visualization*

Principal Component Analysis revealed that the first two components explained 29.6% of total variance (PC1: 18.8%, PC2: 10.8%). Figure 2 presents cluster visualizations using both PCA and t-SNE projections across the three algorithms. K-means clustering demonstrated clearer separation between clusters in both representations, with Cluster 0 (high-risk lifestyle) forming a distinct group in the t-SNE projection. The t-SNE visualization particularly highlighted the overlap between Clusters 1 and 2 while maintaining clear separation of Cluster 0.

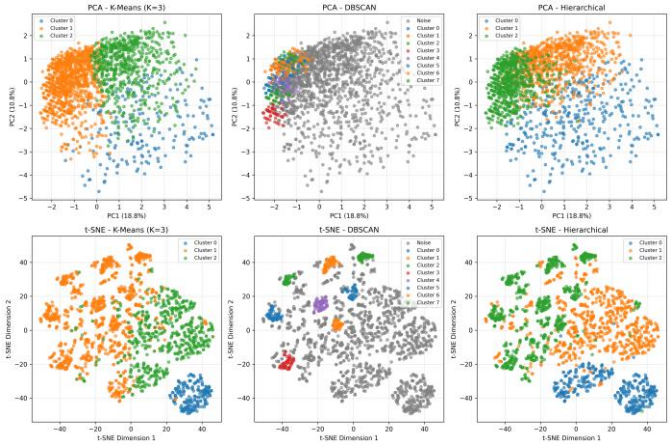

Fig. 2. Optimal cluster number determination. Left: Elbow method showing within-cluster sum of squares (inertia) versus cluster number. Right: Cluster validation metrics including silhouette score and Davies-Bouldin index (scaled by 1/10).

E. *Cluster Profile Characterization*

K-means clustering found and described three gastrointestinal health risk profiles (Table II) due to its excellent performance:

1) Cluster 0 (187, 10.2%): High-Risk Lifestyle Profile High alcohol use (61.0%) and ubiquitous smoking (100%) defined this cluster. Most participants were male (60.4%) and averaged 33.7 years old. 52.9% experienced belly pain, 52.4% heartburn, and 50.3% abdominal inflammation. With 33.7% regular exercise, healthy lifestyle behaviors were lower.

2) Cluster 1: Low-Risk Profile (n=1,054, 57.3%): This largest cluster was younger (mean age 25.5 years) and mostly female (65.9%). Abdominal pain, heartburn, and inflammation were rare: 9.2%, 9.3%, and 12.8%. High fiber consumption (72.3%), regular physical exercise (56.2%), and low substance usage (0.2% smoking, 7.1% drinking) were common.

3) Cluster 2: High-Symptom Profile (n=597, 32.5%): This cluster was mostly female (71.9%) and averaged 31.5 years old. Significantly higher gastrointestinal symptoms were 67.0% abdominal discomfort, 62.6% heartburn, and 60.0% abdominal inflammation. Intermediate alcohol use (18.6%) and worse healthy lifestyle adoption were identified despite no smoking. Medical diagnosis rates peaked at 42.5%.

G. *Risk Profile Visualization*

A radar graphic of cluster risk factor prevalence is shown in Figure 4. The visualization distinguishes profiles: Cluster 0 had high substance use, Cluster 1 had protective factors across most dimensions, while Cluster 2 had focused gastrointestinal symptom axes and low substance use.

TABLE II
CLUSTER PROFILE CHARACTERISTICS (K-MEANS, K=3)

Characteristic	Cluster 0	Cluster 1	Cluster 2
n (%)	187 (10.2)	1,054 (57.3)	597 (32.5)
Age, mean (SD)	33.7 (9.0)	25.5 (10.3)	31.5 (10.9)
Female, %	39.6	65.9	71.9
Gastrointestinal Symptoms, %			
Abdominal pain	52.9	9.2	67.0
Heartburn	52.4	9.3	62.6
Abdominal inflammation	50.3	12.8	60.0
Lifestyle Factors, %			
Fiber consumption	52.9	72.3	55.3
Physical activity (≥3x/week)	33.7	56.2	37.4
Current smoking	100.0	0.2	0.0
Alcohol consumption	61.0	7.1	18.6
Medical diagnosis, %	36.9	6.2	42.5

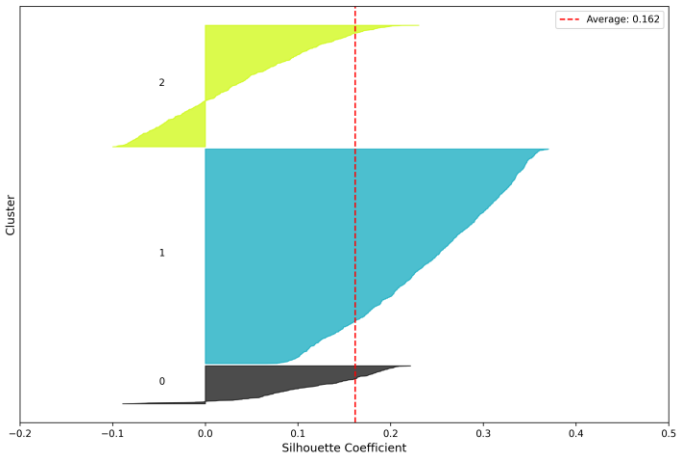

Fig. 3. Silhouette analysis for K-means clustering (K=3). Each horizontal bar represents an individual observation, with bar length indicating silhouette coefficient. Dashed vertical line indicates average silhouette score (0.162).

The normalized feature heatmap (Figure 5) shows cluster centroids and mathematical group separation. Red values indicate above-average scores, while blue values imply below-average. Cluster 0 had strong positive deviations for smoking and alcohol, Cluster 1 had negative deviations for symptom and substance variables and positive deviations for healthy lifestyle indicators, and Cluster 2 had strong positive deviations for all gastrointestinal symptom variables.

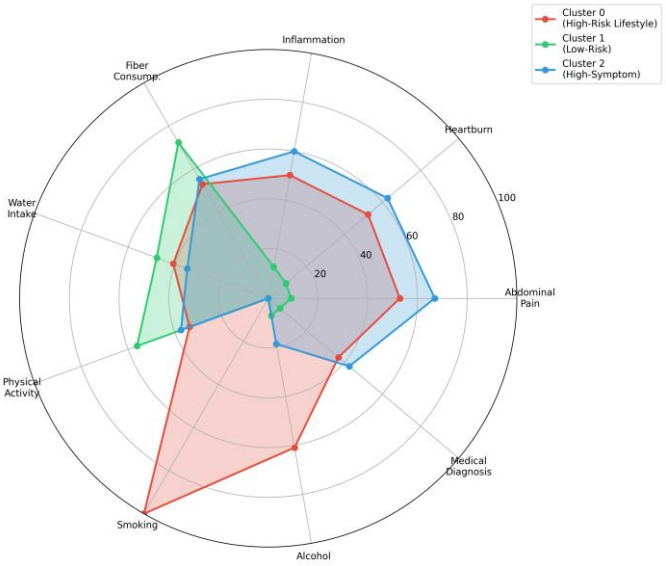

Fig. 4. Radar chart visualization of risk profile characteristics across identified clusters. Values represent percentage prevalence for each variable. Note the distinct pattern of Cluster 0 (substance use), Cluster 1 (protective factors), and Cluster 2 (gastrointestinal symptoms).

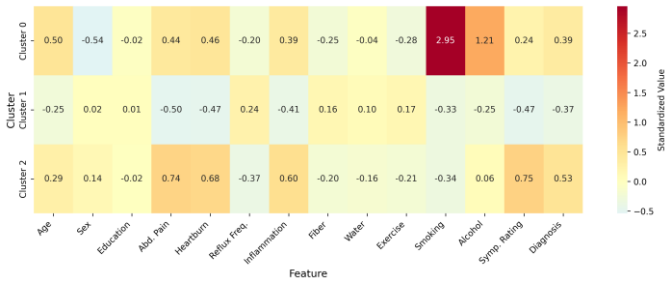

Fig. 5. Heatmap of cluster centroids showing standardized feature values. Positive values indicate above-average scores (red), negative values indicate below-average scores (blue). Features are arranged along the x-axis with clusters on the y-axis.

IV. Discussion

This work used unsupervised machine learning to stratify a community population into clinically significant gastrointestinal health risk profiles. Three separate clusters give actionable findings for targeted intervention development in Honduran primary healthcare settings.

Cluster 0 (the High-Risk Lifestyle cluster) is susceptible and requires integrative intervention. Universal smoking and high alcohol intake in this primarily male cluster are consistent with research relating substance use to gastrointestinal symptoms [28], [29]. The modest symptom prevalence (about 50%) may indicate early gastrointestinal injury, offering secondary prophylaxis. Interventions for this profile should include gastroenterology, substance use therapy, and behavioral modification.

Cluster 2 (High-Symptom; 71.9% female) confirms significant literature on sex-based functional gastrointestinal disorder prevalence [30], [31]. This gap is caused by hormonal effects on visceral sensitivity, gut motility, and pain perception, as well as psychosocial factors such stress response patterns and healthcare-seeking [32]. The high medical diagnosis rate (42.5%) implies this cluster is actively seeking care, indicating more severe or disabling symptoms requiring gastroenterological assessment.

The Low-Risk cluster (Cluster 1), with 57.3% of the population, has preventive lifestyle habits such high fiber eating, regular exercise, and little substance use. This suggests that lifestyle-change-focused health promotion works [33]. The lower symptom prevalence and younger age profile (mean 25.5 years) reflect healthy behavior adoption or reduced cumulative risk factor exposure. Maintenance of protective habits and prevention of higher-risk profiles as people age are primary preventative goals.

In this dataset, K-means clustering outperformed DBSCAN and hierarchical methods. DBSCAN had higher silhouette scores on clustered observations, however it identified 77.5% of participants as noise points, limiting clinical usefulness. The overall silhouette coefficient of 0.162 warrants critical interpretation. By conventional thresholds, values below 0.25 indicate weak cluster structure, and values below 0.50 suggest substantial overlap between adjacent clusters [19]. This score reflects the inherent difficulty of partitioning multidimensional health data: gastrointestinal symptom burden, lifestyle exposures, and sociodemographic characteristics form overlapping, continuous distributions rather than discrete natural groupings [34].

This behavior is consistent with reported silhouette scores in comparable clinical clustering studies, where values in the 0.10–0.25 range are common when working with heterogeneous, mixed-type survey data [35]-[37]. The observed Cluster 2 heterogeneity (mean silhouette = 0.036) is particularly noteworthy and suggests that the High-Symptom profile may encompass clinically distinct sub-phenotypes that a finer-grained analysis could separate. Critically, the decision to retain K=3 was based on a convergence of evidence interpretable elbow inflection, comparative validation metrics (Davies-Bouldin index, Calinski-Harabasz index), and clinical coherence, rather than on silhouette alone. Clinicians and public health practitioners using these clusters for intervention design should interpret cluster boundaries as probabilistic tendency zones rather than rigid classifications, acknowledging that borderline individuals may transition between profiles over time.

PCA caught only 29.6% of variance in the first two components, showing complex variable linkages, according to dimensionality reduction analysis. The t-SNE visualization improves local structure preservation, distinguishing Cluster 0 and showing overlap between Clusters 1 and 2. This overlap may reflect the shifting nature of gastrointestinal health, where Low-Risk individuals may experience High-Symptom symptoms.

These findings affect the Honduran National Health Model's primary care focus [10]. Risk profiles can guide community health worker training for targeted screening and individualized health education. Communities with higher Cluster 0 representation should prioritize smoking cessation and alcohol reduction programs, while Cluster 2 communities may benefit from gastroenterological referral pathways and stress management interventions.

Several limitations should be noted. The cross-sectional design precludes causal inference regarding risk factor-symptom relationships. The convenience sampling strategy may limit generalizability to the broader Honduran population, although the geographic diversity across five departments strengthens external validity. Self-reported symptom data are subject to recall bias, and clinical validation without objective diagnostic confirmation remains a constraint. Regarding the clustering solution itself, formal bootstrap-based stability analysis and permutation testing were not performed; while the consistency of cluster assignments across all ten K-means++ initializations provides indirect evidence of solution robustness, future work should include rigorous stability metrics such as Jaccard similarity coefficients or Fowlkes-Mallows indices [38], [39]. Future studies should incorporate longitudinal follow-up, biological markers, and validated diagnostic instruments to refine cluster characterization.

V. Conclusion

This study shows that machine learning-based clustering can identify community gastrointestinal health risk profiles. The clinically significant three-cluster solution of High-Risk Lifestyle, Low-Risk, and High-Symptom profiles allows for tailored intervention design. In this epidemiological context, K-means clustering with K=3 balanced statistical validity and practical interpretability better than density-based and hierarchical methods.

Profiles support focused primary healthcare resource allocation: substance use therapies for Cluster 0, protective behavior maintenance for Cluster 1, and improved gastroenterological services for Cluster 2. These findings provide evidence-based health policy that supports the Honduran National Health Model's equitable and effective community-based care.

Future work should address two complementary directions. First, external validation of the identified profiles is needed: a predictive modelling approach — such as a supervised classifier trained on cluster-derived labels and evaluated against prospective health outcomes (e.g., incident gastrointestinal diagnoses, healthcare utilization, symptom progression) — would provide evidence that the clusters capture clinically meaningful variation beyond internal metric performance. Second, integration of these clustering algorithms into routine health information systems could enable real-time population stratification and adaptive primary care targeting. The methodological framework described here is transferable to other primary healthcare contexts where multidimensional risk assessment and resource allocation are priorities.

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 as 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] M. J. Schmulson and D. A. Drossman, "What Is New in Rome IV", *J. Neurogastroenterol. Motil.*, vol. 23, no. 2, pp. 151-163, Apr. 2017. doi: 10.5056/jnm16214

[4] D. A. Drossman and M. Camilleri, "Irritable bowel syndrome: a technical review for practice guideline development", *Gastroenterology*, vol. 123, no. 6, pp. 2108-2131, Dec. 2002. doi: 10.1053/gast.2002.37095

[5] A. C. Ford et al., "Global prevalence of, and risk factors for, uninvestigated dyspepsia: a meta-analysis", *Gut*, vol. 64, no. 7, pp. 1049-1057, Jul. 2015. doi: 10.1136/gutjnl-2014-307843

[6] A. Esteva et al., "A guide to deep learning in healthcare", *Nat. Med.*, vol. 25, no. 1, pp. 24-29, Jan. 2019. doi: 10.1038/s41591-018-0316-z

[7] B. J. Frey and D. Dueck, "Clustering by passing messages between data points", *Science*, vol. 315, no. 5814, pp. 972-976, Feb. 2007. doi: 10.1126/science.1136800

[8] O. Palsson et al., "Development and validation of the Rome IV diagnostic questionnaire for adults", *Gastroenterology*, vol. 150, no. 6, pp. 1481-1491, May 2016. doi: 10.1053/j.gastro.2016.02.014

[9] Y. Koloski et al., "The brain-gut pathway in functional gastrointestinal disorders is bidirectional: a 12-year prospective population-based study", *Gut*, vol. 61, no. 9, pp. 1284-1290, Sep. 2012. doi: 10.1136/gutjnl-2011-300474

[10] Secretaria de Salud de Honduras, "Modelo Nacional de Salud", Tegucigalpa: Secretaria de Salud, 2023.

[11] 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, 2021, doi: 10.1007/s10620-020-06639-y

[12] M. P. Curado, M. M. de Oliveira, and M. de Araújo Fagundes, "Prevalence of *Helicobacter pylori* infection in Latin America and the Caribbean populations: A systematic review and meta-analysis", *Cancer Epidemiology*, vol. 60, pp. 141–148, 2019, doi: 10.1016/j.canep.2019.04.003

[13] D. A. Drossman and W.L. Hasler, "Rome IV - Functional GI Disorders: Disorders of Gut-Brain Interaction", *Gastroenterology*, vol. 150, no. 6, pp. 1257-1261, May 2016. doi: 10.1053/j.gastro.2016.03.035

[14]F. Pedregosa et al., "Scikit-learn: Machine learning in Python", *J. Mach. Learn. Res.*, vol. 12, pp. 2825-2830, Oct. 2011.
[15]A. K. Jain, "Data clustering: 50 years beyond K-means", *Pattern Recognit. Lett.*, vol. 31, no. 8, pp. 651-666, Jun. 2010. doi: 10.1016/j.patrec.2009.09.011
[16]D. Arthur and S. Vassilvitskii, "K-means++: the advantages of careful seeding", in *Proc. 18th Annu. ACM-SIAM Symp. Discret. Algorithms*, New Orleans, LA, USA, 2007, pp. 1027-1035.
[17]M. Ester et al., "A density-based algorithm for discovering clusters in large spatial databases with noise", in *Proc. 2nd Int. Conf. Knowl. Discov. Data Min.*, Portland, OR, USA, 1996, pp. 226-231.
[18]J. H. Ward Jr., "Hierarchical grouping to optimize an objective function", *J. Am. Stat. Assoc.*, vol. 58, no. 301, pp. 236-244, Mar. 1963. doi: 10.1080/01621459.1963.10500845
[19]P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis", *J. Comput. Appl. Math.*, vol. 20, pp. 53-65, Nov. 1987. doi: 10.1016/0377-0427(87)90125-7
[20]D. L. Davies and D. W. Bouldin, "A cluster separation measure", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-1, no. 2, pp. 224-227, Apr. 1979. doi: 10.1109/TPAMI.1979.4766909
[21]T. Calinski and J. Harabasz, "A dendrite method for cluster analysis", *Commun. Stat.*, vol. 3, no. 1, pp. 1-27, 1974. doi: 10.1080/03610927408827101
[22]I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments", *Philos. Trans. R. Soc. A*, vol. 374, no. 2065, p. 20150202, Apr. 2016. doi: 10.1098/rsta.2015.0202
[23]L. van der Maaten and G. Hinton, "Visualizing data using t-SNE", *J. Mach. Learn. Res.*, vol. 9, pp. 2579-2605, Nov. 2008.
[24]C. R. Harris et al., "Array programming with NumPy", *Nature*, vol. 585, no. 7825, pp. 357-362, Sep. 2020. doi: 10.1038/s41586-020-2649-2
[25]P. Virtanen et al., "SciPy 1.0: fundamental algorithms for scientific computing in Python", *Nat. Methods*, vol. 17, no. 3, pp. 261-272, Mar. 2020. doi: 10.1038/s41592-019-0686-2
[26]J. D. Hunter, "Matplotlib: A 2D graphics environment", *Comput. Sci. Eng.*, vol. 9, no. 3, pp. 90-95, May 2007. doi: 10.1109/MCSE.2007.55
[27]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
[28]R. Bode and J. C. Bode, "Activation of the innate immune system and alcoholic liver disease: effects of ethanol per se or enhanced intestinal translocation of bacterial toxins induced by ethanol?", *Alcohol. Clin. Exp. Res.*, vol. 29, no. 11 Suppl., pp. 166S-171S, Nov. 2005. doi: 10.1097/01.alc.0000189280.19073.28
[29]T. Bjorkholm et al., "Tobacco smoking and nicotine-mediated changes in the gastrointestinal tract", *World J. Gastroenterol.*, vol. 20, no. 12, pp. 3058-3067, Mar. 2014. doi: 10.3748/wjg.v20.i12.3058
[30]Y. Kim et al., "Sex-gender differences in irritable bowel syndrome", *J. Neurogastroenterol. Motil.*, vol. 24, no. 4, pp. 544-558, Oct. 2018. doi: 10.5056/jnm18082
[31]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
[32]L. A. Houghton et al., "Age, gender, and women's health and the patient", *Gastroenterology*, vol. 150, no. 6, pp. 1332-1343, May 2016. doi: 10.1053/j.gastro.2016.02.017
[33]World Health Organization, "Global Status Report on Noncommunicable Diseases 2021", Geneva: WHO, 2021.
[34]D. D. Penrice, P. Rattan, and D. A. Simonetto, "Artificial Intelligence and the Future of Gastroenterology and Hepatology", *Gastro Hep Adv*., vol. 1, no. 5, pp. 546-556, Sep. 2022. doi: 10.1016/j.gastha.2022.02.025
[35]C. E. Coombes, X. Liu, Z. B. Abrams, K. R. Coombes, and G. Brock, "Simulation-derived best practices for clustering clinical data," *Journal of Biomedical Informatics*, vol. 118, art. no. 103788, Jun. 2021, doi: 10.1016/j.jbi.2021.103788
[36]C. Ciucurel and E. I. Iconaru, "An Epidemiological Study on the Prevalence of the Clinical Features of SARS-CoV-2 Infection in Romanian People", *Int. J. Environ. Res. Public Health*, vol. 17, no. 14, art. no. 5082, Jul. 2020. doi: 10.3390/ijerph17145082
[37]E. Yamga, L. Mullie, M. Durand, A. Cadrin-Chenevert, A. Tang, E. Montagnon, C. Chartrand-Lefebvre, and M. Chassé, "Interpretable clinical phenotypes among patients hospitalized with COVID-19 using cluster analysis," *Frontiers in Digital Health*, vol. 5, art. no. 1142822, 2023, doi: 10.3389/fdgth.2023.1142822
[38]A. Ben-Hur, A. Elisseeff, and I. Guyon, "A stability based method for discovering structure in clustered data", *Pacific Symposium on Biocomputing*, vol. 7, pp. 6–17, 2002, doi: 10.1142/9789812799623_0002
[39]C. Hennig, "Cluster-wise assessment of cluster stability", *Computational Statistics & Data Analysis*, vol. 52, no. 1, pp. 258–271, Sep. 2007, doi: 10.1016/j.csda.2006.11.025.