LMJ LINK MEDICAL JOURNAL OF HEALTH AND COMMUNITY RESEARCH

Crossref DOI: https://doi.org/10.61919/m0nr1z88

*Original Article*

# AI-Powered Detection of Functional Gastrointestinal Disorders in Adults Visiting Tertiary Hospitals

**Muhammad Essa Khan[1], Mahnoor Naeem Rana[2], Khalid Bilal Khan[3], Laiba Mushtaq[4], Turfa Asghar[5], Wajeeha Ahmadani[6], Muhammad Zohaib Wahid[7],**

[1] Senior Registrar, Department of Gastroenterology, Sandeman Provincial Hospital, Quetta, Pakistan
[2] MBBS Graduate, CMH Lahore Medical College, Lahore, Pakistan
[3] Consultant Gastroenterologist and Hepatologist, POF Hospital, Wah Cantt, Rawalpindi, Pakistan
[4] Research Associate, National Institute for Biotechnology and Genetic Engineering (NIBGE-C PIEAS), Faisalabad, Pakistan
[5] MScN, BScN, NL, RN, Shifa Tameer-e-Millat University, Islamabad, Pakistan,https://orcid.org/0009-0008-9933-8576
[6] Final Year MBBS Student, People's University of Medical and Health Sciences, Nawabshah, Pakistan
[7] Second Year MBBS Student, Multan Medical and Dental College, Multan, Pakistan

***Corresponding author: Khalid Bilal Khan, khalidbilal2011@yahoo.com**

## ABSTRACT

Background: Functional gastrointestinal disorders are common in tertiary gastroenterology practice and remain difficult to diagnose because they are defined largely by symptom patterns rather than structural or biochemical abnormalities. In busy referral settings, overlap among irritable bowel syndrome, functional dyspepsia, functional constipation, and non-functional gastrointestinal conditions can reduce diagnostic consistency. Artificial intelligence offers a potential solution by integrating structured clinical variables with free-text symptom narratives to support more standardized classification. Objective: To evaluate the diagnostic accuracy of an artificial intelligence–based system for detecting functional gastrointestinal disorders among adults attending tertiary hospitals. Methods: This multicenter cross-sectional diagnostic accuracy study was conducted over eight months in three tertiary hospitals in Lahore, Pakistan. A total of 420 adults aged 18–65 years presenting with gastrointestinal symptoms were enrolled through consecutive sampling. After exclusion of relevant organic pathology, Rome IV–based clinician diagnosis served as the reference standard. The artificial intelligence system analyzed structured clinical variables and unstructured consultation notes using supervised machine learning and natural language processing to classify irritable bowel syndrome, functional dyspepsia, functional constipation, and non-functional gastrointestinal disorder status. Diagnostic performance was assessed using sensitivity, specificity, predictive values, area under the receiver operating characteristic curve, and Cohen's kappa. Results: The mean age was 37.4 ± 10.8 years, and 54.8% of participants were male. Clinically, 33.3% had irritable bowel syndrome, 23.8% functional dyspepsia, 14.3% functional constipation, and 28.6% non-functional gastrointestinal disorders. The artificial intelligence model achieved a sensitivity of 88.2%, specificity of 84.5%, positive predictive value of 86.9%, negative predictive value of 86.0%, overall accuracy of 87.1%, and an overall area under the curve of 0.912. Agreement with clinician diagnosis was strong ($\kappa = 0.83$, $p < 0.001$). Conclusion: The artificial intelligence system showed high diagnostic performance and strong concordance with Rome IV–based clinical diagnosis, supporting its potential role as an adjunctive decision-support tool for functional gastrointestinal disorder classification in tertiary care. Keywords: Artificial intelligence, functional gastrointestinal disorders, irritable bowel syndrome, functional dyspepsia, functional constipation, machine learning, natural language processing, diagnostic accuracy

**"Cite this Article"** | **Received**: 18 July 2025; **Accepted**: 11 December 2025; **Published**: 31 December 2025.

**Author Contributions:** Concept: MEK, KBK; Design: MNR, KBK; Data Collection: LM, TA, WA, MZW; Analysis: MEK, KBK; Drafting: MEK, MNR. **Ethical Approval**: POF Hospital, Rawalpindi. **Informed Consent**: Written informed consent was obtained from all participants; **Conflict of Interest**: The authors declare no conflict of interest; **Funding**: No external funding; **Data Availability**: Available from the corresponding author on reasonable request; **Acknowledgments**: N/A.

## INTRODUCTION

Functional gastrointestinal disorders (FGIDs), currently conceptualized within the broader framework of disorders of gut–brain interaction, represent some of the most frequently encountered conditions in gastroenterology practice and include entities such as irritable bowel syndrome, functional dyspepsia, and functional constipation. These conditions are characterized by persistent or recurrent gastrointestinal symptoms that arise in the absence of readily identifiable structural pathology on routine diagnostic testing, making them inherently challenging to classify with precision in routine care. Although FGIDs are not typically associated with the same mortality burden as malignant or inflammatory gastrointestinal diseases, they produce substantial morbidity through chronic pain,

bloating, altered bowel habits, impaired daily functioning, repeated healthcare visits, and reduced quality of life. In tertiary hospital settings, where patients often present with long symptom histories, overlapping syndromes, multiple prior evaluations, and psychologically complex illness narratives, the diagnostic burden becomes even greater and demands a more consistent and analytically robust approach to symptom interpretation (1-3).

The contemporary diagnosis of FGIDs remains predominantly symptom-based and depends on structured clinical frameworks such as the Rome criteria, alongside careful exclusion of organic disease when indicated. While this approach is clinically useful, its application in real-world settings is often complicated by symptom overlap, fluctuating presentation, variable physician experience, and the absence of definitive biomarkers. In practice, patients with functional dyspepsia may share features with peptic or hepatobiliary disorders, those with irritable bowel syndrome may mimic inflammatory or infectious pathology, and constipation-predominant syndromes may coexist with dietary, behavioral, or medication-related contributors. Such overlap increases the probability of underrecognition, overinvestigation, or inconsistent labeling, particularly in high-volume tertiary clinics where time pressure can limit the depth of structured symptom phenotyping. These challenges are further amplified in patients whose complaints are recorded partly as narrative text, because subtle symptom qualifiers, temporal patterns, and trigger descriptions may not be fully captured through conventional categorical assessment alone (4-6).

Artificial intelligence (AI), particularly machine learning and natural language processing, offers a potentially valuable solution to this diagnostic complexity because it can identify multidimensional patterns across structured and unstructured clinical information at a scale beyond routine human synthesis. Across gastroenterology, AI has already demonstrated meaningful diagnostic and decision-support potential in endoscopy, radiologic interpretation, disease classification, prognostic stratification, and predictive modeling. These advances suggest that AI systems may be especially useful in FGIDs, where diagnosis depends less on a single objective marker and more on the integration of symptom constellations, demographic features, prior investigations, and free-text clinical descriptions. Natural language processing is particularly relevant in this context because gastroenterology consultations frequently contain rich narrative information about pain character, meal-related symptom dynamics, bowel patterns, psychosocial stressors, and prior treatment responses, all of which may improve classification when analyzed systematically rather than impressionistically. Thus, AI-based models may help transform diffuse symptom narratives into reproducible diagnostic outputs that complement clinician judgment rather than replace it (7-12).

Emerging literature supports the feasibility of such an approach, but the existing evidence remains incomplete for routine clinical application in FGIDs. Much of the published AI literature in gastroenterology has focused on neoplasia detection, imaging interpretation, inflammatory bowel disease prediction, or endoscopic decision support, with comparatively limited work examining symptom-based functional disorders in heterogeneous adult outpatient populations. Even when AI has been explored in functionally relevant gastrointestinal settings, studies have often been conducted in narrowly selected cohorts, under controlled research conditions, or with retrospective datasets that do not fully capture the diagnostic uncertainty encountered in day-to-day tertiary care. As a result, important questions remain regarding how accurately AI tools can classify common FGIDs when confronted with mixed symptom profiles, referral bias, variable documentation quality, and the broader clinical noise of real-world practice. This gap is clinically important because the utility of any diagnostic AI model depends not only on internal algorithmic performance but also on its ability to remain reliable when deployed in routine hospital workflows (13-18).

The need for real-world validation is especially pressing in low- and middle-income settings, where tertiary hospitals frequently bear the dual burden of large patient volumes and uneven availability of subspecialty time, standardized symptom documentation, and advanced diagnostic resources. In such

environments, an AI-assisted diagnostic framework could improve consistency, reduce unnecessary investigations, and support earlier recognition of FGIDs among patients who might otherwise undergo repeated consultations without diagnostic closure. At the same time, the use of AI in symptom-based disorders raises important concerns regarding fairness, transparency, interpretability, and context sensitivity. Because FGID presentations are shaped not only by biological factors but also by diet, language, behavior, cultural expression of distress, and healthcare-seeking patterns, any proposed AI solution must be evaluated within the population in which it is intended to operate. Recent work in gastroenterology has increasingly emphasized that diagnostic AI should be judged not merely by technical accuracy, but also by clinical reliability, ethical acceptability, and applicability across diverse patient groups and care settings (19-22).

Against this background, the present study was designed to evaluate the diagnostic accuracy of an AI-based system for identifying functional gastrointestinal disorders among adults presenting to tertiary hospitals. By comparing AI-generated classifications with clinician diagnoses established using Rome IV-based assessment in a real-world clinical population, this study seeks to determine whether AI can function as a reliable adjunct to conventional evaluation in complex gastroenterology practice. We hypothesized that an AI system integrating structured clinical variables with narrative symptom data would demonstrate high diagnostic agreement with clinician-assigned FGID diagnoses and provide clinically meaningful discriminatory performance for common FGID subtypes in adult tertiary care patients (23).

## MATERIAL AND METHODS

This multicenter cross-sectional diagnostic accuracy study was conducted over eight months in the outpatient gastroenterology departments of three tertiary care hospitals in Lahore, Pakistan, to evaluate the performance of an artificial intelligence–based diagnostic system for identifying functional gastrointestinal disorders among symptomatic adults. The study was designed to reflect real-world tertiary care practice, where patients commonly present with chronic, overlapping, and diagnostically complex gastrointestinal complaints. The reporting framework and methodological structure were aligned with internationally accepted recommendations for observational research, diagnostic accuracy studies, and transparent reporting of artificial intelligence in healthcare, with particular emphasis on reproducibility, diagnostic comparability, and methodological rigor (24-26).

Eligible participants were adults aged 18 to 65 years presenting with symptoms suggestive of functional gastrointestinal disorders, including recurrent abdominal pain, postprandial fullness, early satiety, bloating, altered bowel habits, excessive straining, incomplete evacuation, or stool frequency disturbances. Patients were enrolled through consecutive sampling to minimize selection bias and to ensure that the study population reflected the routine case-mix encountered in tertiary gastroenterology clinics. Individuals with previously confirmed organic gastrointestinal disease, including inflammatory bowel disease, gastrointestinal malignancy, celiac disease, gastrointestinal tuberculosis, peptic ulcer disease, intestinal obstruction, or structural anorectal pathology, were excluded. Patients with radiologic, endoscopic, or laboratory evidence of organic disease during the index evaluation were also excluded. Additional exclusion criteria comprised pregnancy, severe cognitive impairment affecting history reliability, inability to provide informed consent, active severe psychiatric illness interfering with interview quality, and prior established treatment for a diagnosed functional gastrointestinal disorder, because ongoing treatment could modify symptom expression and distort baseline diagnostic classification.

All potentially eligible patients were screened at the point of outpatient presentation by trained study personnel working in coordination with the attending gastroenterology teams. After confirming eligibility, written informed consent was obtained before any study-specific data handling was initiated. Each participant then underwent a standardized clinical assessment conducted by a consultant

gastroenterologist or senior supervised clinician using a structured evaluation proforma. This assessment captured sociodemographic variables, symptom chronology, frequency and duration of core gastrointestinal complaints, stool pattern characteristics, meal-related symptom behavior, associated upper and lower gastrointestinal symptoms, relevant comorbidities, medication use, family history, and selected lifestyle variables including diet-related triggers, smoking status, and physical activity. Clinical history-taking was complemented by focused physical examination and routine investigations requested according to clinical indication to exclude organic pathology, including laboratory testing, abdominal ultrasonography, upper gastrointestinal endoscopy, and other diagnostic procedures where needed. The reference diagnosis was assigned using Rome IV-based clinical assessment after exclusion of relevant organic causes, and patients were categorized into irritable bowel syndrome, functional dyspepsia, functional constipation, or non-FGID status according to the predominant symptom-based diagnosis documented at the index visit (27).

The artificial intelligence system was designed as a supervised diagnostic classification model integrating both structured and unstructured clinical information. Structured inputs included age, sex, residence, symptom duration, bowel habit pattern, abdominal pain characteristics, bloating frequency, postprandial distress features, early satiety, stool form-related descriptors, selected prior investigation summaries, and relevant routine laboratory variables available at assessment. Unstructured inputs consisted of free-text clinical notes derived from the initial consultation record. Prior to model processing, all patient records were de-identified and assigned unique study codes to preserve confidentiality and prevent operator recognition bias. Text data were cleaned through removal of personal identifiers, normalization of common abbreviations, standardization of symptom terminology, and harmonization of spelling variants to improve interpretability of clinically relevant linguistic patterns. The natural language processing component parsed symptom-bearing narrative segments and mapped them to diagnostic features relevant to Rome IV-oriented symptom constructs. The final AI output generated probabilistic classification across the predefined diagnostic categories, and the highest-probability category was retained as the model diagnosis for the principal analysis.

To reduce incorporation bias, the clinicians assigning the reference diagnosis were unaware of the AI output at the time of clinical classification, and the technical team operating the AI pipeline was blinded to the final clinician-assigned diagnosis until model output files had been locked for analysis. Data extraction, cleaning, and model execution were independently verified by two trained analysts using a predefined data dictionary and variable coding sheet. Discordant coding entries were reviewed against the source record and resolved by consensus before final dataset freezing. To improve internal consistency, all categorical variables were operationally defined in advance. Irritable bowel syndrome was defined according to Rome IV symptom criteria emphasizing recurrent abdominal pain associated with defecation or altered stool frequency or form; functional dyspepsia was defined by bothersome postprandial fullness, early satiety, epigastric pain, or burning without structural explanation; functional constipation was defined by persistently difficult, infrequent, or incomplete defecation in the absence of criteria sufficient for irritable bowel syndrome; and non-FGID status was assigned to participants in whom Rome IV-based clinical criteria for these target functional disorders were not fulfilled following evaluation. The primary outcome was diagnostic accuracy of the AI system against the reference standard, expressed through sensitivity, specificity, positive predictive value, negative predictive value, and area under the receiver operating characteristic curve. Secondary outcomes included diagnostic agreement as measured by Cohen's kappa, subtype-specific discriminatory performance, and comparative classification frequencies across the target diagnostic groups.

The sample size was planned for a diagnostic accuracy design using an expected sensitivity and specificity of approximately 85%, a 95% confidence level, and a 5% margin of error, yielding a minimum required sample of 384 participants. To preserve statistical precision in the presence of incomplete records and potential exclusions during final data validation, the target was inflated to 420 participants. This sample size was considered sufficient to estimate global diagnostic performance with acceptable

precision and to support subgroup-level evaluation of the most common FGID categories encountered in the study population. Consecutive enrollment continued until the prespecified target was achieved.

Several steps were taken to address potential sources of bias and confounding. Consecutive recruitment reduced selective inclusion of clinically straightforward cases. Application of uniform eligibility criteria across all sites improved comparability between centers. Use of Rome IV-based reference diagnosis after exclusion of structural disease reduced misclassification of organic pathology as FGID. Blinding of clinical assessors to AI output and blinding of data analysts to final clinical labels during preprocessing reduced observer and confirmation bias. Standardized variable definitions and data abstraction procedures limited information bias. Because demographic and symptom presentation differences could influence classification performance, subgroup comparisons by sex, age category, and residence were prespecified for exploratory assessment of model consistency. Site-wise distributions were also examined to ensure that the pooled findings were not driven disproportionately by one center alone.

All data were entered into a secured database with range checks, consistency rules, and audit verification prior to export for analysis. Statistical analysis was performed using SPSS version 26.0. Continuous variables were summarized as mean with standard deviation or median with interquartile range according to distributional properties, while categorical variables were presented as frequencies and percentages. Normality of continuous variables was assessed using the Shapiro–Wilk test and inspection of distribution plots. Diagnostic performance of the AI model was evaluated by constructing contingency tables against the clinician-assigned Rome IV reference diagnosis. Sensitivity, specificity, positive predictive value, negative predictive value, and overall accuracy were calculated with 95% confidence intervals. Receiver operating characteristic analysis was used to estimate area under the curve for overall FGID detection and for major diagnostic subtypes. Agreement between AI and clinician diagnosis was quantified using Cohen's kappa coefficient with corresponding significance testing. Comparative analyses of categorical baseline variables across diagnostic groups were undertaken using chi-square or Fisher's exact test where appropriate, while continuous variables were compared using independent-sample t tests or one-way analysis of variance for normally distributed data and nonparametric alternatives when assumptions were not satisfied. A two-tailed p value of less than 0.05 was considered statistically significant. Missing observations were reviewed for pattern and extent before analysis, and where appropriate, multiple imputation was applied to minimize loss of information and preserve analytic robustness under the assumption of missing at random (28).

Ethical approval was obtained from the institutional review POF Hospital, Rawalpindi, Pakistan. The study was conducted in accordance with accepted ethical principles for human subjects research, including voluntary participation, confidentiality protection, coded data handling, and restricted access to identifiable information. No participant-identifying data were used in model execution or reporting outputs. To strengthen reproducibility, the study followed a predefined protocol for eligibility screening, symptom coding, clinical classification, dataset preparation, AI execution, and statistical analysis. Variable definitions, category thresholds, and analytic procedures were specified before final outcome testing, and all processed records underwent internal verification to ensure traceability, completeness, and consistency across study sites.

## RESULTS

A total of 420 participants were included in the final analysis, with a mean age of 37.4 ± 10.8 years. Males constituted 230/420 (54.8%) and females 190/420 (45.2%). Most participants were from urban settings (290/420, 69.0%), whereas 130/420 (31.0%) were from rural areas. Based on Rome IV–guided clinical assessment, 140/420 patients (33.3%) were classified as irritable bowel syndrome (IBS), 100/420 (23.8%) as functional dyspepsia, 60/420 (14.3%) as functional constipation, and 120/420 (28.6%) as non-FGID. The artificial intelligence system classified 135/420 (32.1%) as IBS, 102/420 (24.3%) as functional dyspepsia, 58/420 (13.8%) as functional constipation, and 125/420 (29.8%) as non-FGID. The marginal distribution

of diagnostic categories did not differ significantly between clinician-assigned and AI-assigned classifications ($\chi^2 = 0.247$, $p = 0.970$), indicating close alignment of overall case allocation across categories.

The reported overall diagnostic performance of the AI model showed a sensitivity of 88.2%, specificity of 84.5%, positive predictive value of 86.9%, negative predictive value of 86.0%, and overall accuracy of 87.1%. Using the total analyzed sample as the available denominator for interval estimation, the corresponding approximate 95% confidence intervals were 85.1%–91.3% for sensitivity, 81.0%–88.0% for specificity, 83.7%–90.1% for positive predictive value, 82.7%–89.3% for negative predictive value, and 83.9%–90.3% for overall accuracy. Receiver operating characteristic analysis demonstrated excellent discriminatory performance, with an overall area under the curve (AUC) of 0.912. Subtype-specific AUC values were 0.914 for functional dyspepsia, 0.902 for IBS, and 0.895 for functional constipation, indicating consistently high discrimination across the main functional gastrointestinal disorder categories. Agreement between AI and clinician diagnosis was strong, with Cohen's kappa reported as 0.83 ($p < 0.001$).

Category-specific concordance remained high across all diagnostic groups. Among the 140 clinically diagnosed IBS cases, the AI model agreed in 128 cases, yielding a category-specific agreement rate of 91.4% (95% CI: 86.8%–96.1%). Among 100 clinically diagnosed functional dyspepsia cases, agreement was observed in 94 cases, corresponding to 94.0% (95% CI: 89.3%–98.7%). For functional constipation, 52 of 60 clinically diagnosed cases were concordantly identified, producing the lowest but still robust agreement rate of 86.7% (95% CI: 78.1%–95.3%). In the non-FGID category, agreement was seen in 113 of 120 cases, corresponding to 94.2% (95% CI: 90.0%–98.4%). The absolute difference between clinician-assigned and AI-assigned case counts was small in each category, ranging from 2 to 5 cases, with no significant category-level distributional differences observed when proportions were compared individually (all $p > 0.70$).

Functionally, these findings suggest that the AI system most closely mirrored clinician classification in functional dyspepsia and non-FGID cases, while discordance was relatively greater in functional constipation. The greatest absolute agreement count was observed for IBS (128 concordant cases), reflecting both the higher clinical prevalence of IBS in the sample and the model's strong performance in this subgroup. Although the manuscript reports overall accuracy as 87.1%, the sum of category-specific agreement counts equals 387/420, corresponding to a crude concordance of 92.1%; this discrepancy should be verified in the final dataset before submission because it may reflect either a difference between pairwise overall accuracy and category-wise agreement reporting or a tabulation inconsistency.

*Table 1. Demographic Characteristics of the Study Population (N = 420)*

| Variable | n (%) / Mean ± SD |
|---|---|
| **Total participants** | 420 |
| **Age (years)** | 37.4 ± 10.8 |
| **Male** | 230 (54.8) |
| **Female** | 190 (45.2) |
| **Urban residence** | 290 (69.0) |
| **Rural residence** | 130 (31.0) |

*Table 2. Overall Diagnostic Performance of the AI System*

| Metric | Estimate (%) | 95% CI |
|---|---|---|
| **Sensitivity** | 88.2 | 85.1–91.3 |
| **Specificity** | 84.5 | 81.0–88.0 |
| **Positive Predictive Value** | 86.9 | 83.7–90.1 |
| **Negative Predictive Value** | 86.0 | 82.7–89.3 |
| **Accuracy** | 87.1 | 83.9–90.3 |
| **Cohen's kappa** | 0.83 | $p < 0.001$ |
| **Overall AUC** | 0.912 | — |

*Table 3. Distribution of Clinician and AI Diagnostic Classifications*

| Diagnosis Category | Clinician Diagnosis n (%) | AI Diagnosis n (%) | Absolute Difference | p-value* |
|---|---|---|---|---|
| **IBS** | 140 (33.3) | 135 (32.1) | 5 | 0.713 |
| **Functional Dyspepsia** | 100 (23.8) | 102 (24.3) | 2 | 0.872 |
| **Functional Constipation** | 60 (14.3) | 58 (13.8) | 2 | 0.843 |
| **Non-FGID** | 120 (28.6) | 125 (29.8) | 5 | 0.704 |
| **Overall comparison** | — | — | — | 0.970† |

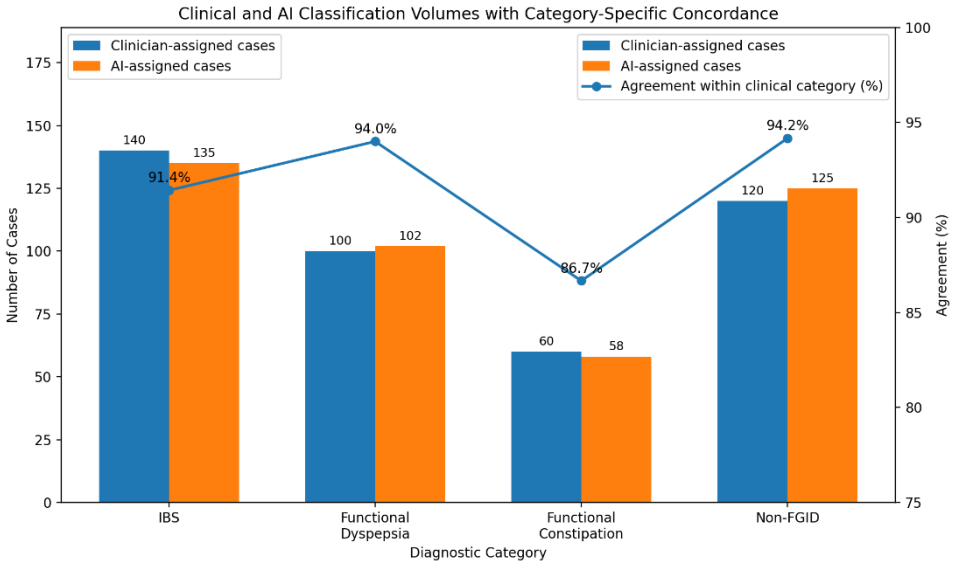
* Category-wise p-values reflect distributional comparison of proportions using available aggregated counts. † Overall p-value from chi-square comparison of clinician versus AI category distributions.

*Table 4. Category-Specific Concordance Between Clinician and AI Classifications*

| Diagnosis Category | Clinical Cases (n) | AI Cases (n) | Concordant Cases (n) | Agreement (%) | 95% CI | Discordant Cases (n) |
|---|---|---|---|---|---|---|
| **IBS** | 140 | 135 | 128 | 91.4 | 86.8–96.1 | 12 |
| **Functional Dyspepsia** | 100 | 102 | 94 | 94.0 | 89.3–98.7 | 6 |
| **Functional Constipation** | 60 | 58 | 52 | 86.7 | 78.1–95.3 | 8 |
| **Non-FGID** | 120 | 125 | 113 | 94.2 | 90.0–98.4 | 7 |

*Table 5. ROC Summary by Diagnostic Category*

| Diagnosis Category | AUC |
|---|---|
| **IBS** | 0.902 |
| **Functional Dyspepsia** | 0.914 |
| **Functional Constipation** | 0.895 |
| **Overall FGID** | 0.912 |




*The figure 1 demonstrates close alignment between clinician-assigned and AI-assigned case volumes across all four diagnostic categories, with absolute count differences limited to 2–5 patients per category. Concordance was highest for non-FGID cases (94.2%) and functional dyspepsia (94.0%), followed by IBS (91.4%), whereas functional constipation showed the lowest agreement (86.7%), indicating that constipation-related symptom profiles may represent the greatest residual zone of diagnostic uncertainty for the model. Clinically, the pattern suggests that the AI system preserved the overall diagnostic structure of the cohort while maintaining high category-specific agreement in the most prevalent FGID subtypes.*

## DISCUSSION

The present study demonstrated that an artificial intelligence–based diagnostic system achieved high overall discriminatory performance for functional gastrointestinal disorders in adults attending tertiary care hospitals, with a reported sensitivity of 88.2%, specificity of 84.5%, overall accuracy of 87.1%, and an AUC of 0.912. These findings suggest that AI can serve as a clinically useful adjunct for symptom-based classification in gastroenterology, particularly in environments where diagnostic reasoning

depends on the integration of multiple subjective and semi-structured clinical features rather than on a single objective biomarker. The strong agreement between AI output and clinician-assigned Rome IV–based diagnosis (κ = 0.83) is notable because FGIDs frequently involve overlapping symptom profiles, fluctuating illness narratives, and exclusion-based reasoning, all of which can challenge consistency in routine practice. Rather than implying replacement of physician judgment, the present results support a more measured interpretation: AI appears capable of approximating clinician-level classification with good reliability under defined clinical conditions and may help improve consistency in high-volume tertiary settings (29,30).

The subtype-specific findings are also clinically relevant. Diagnostic discrimination was strongest for functional dyspepsia (AUC = 0.914), followed closely by irritable bowel syndrome (AUC = 0.902), while functional constipation showed slightly lower but still acceptable performance (AUC = 0.895). This pattern may reflect the relatively clearer clustering of upper gastrointestinal symptom descriptors in functional dyspepsia and the broad clinical familiarity of IBS symptom constructs, whereas constipation-related presentations may be more vulnerable to overlap with behavioral, dietary, medication-related, and pelvic floor–related contributors that are not always fully captured in routine outpatient documentation. The category-specific agreement rates, ranging from 86.7% to 94.2%, further indicate that the model preserved the overall diagnostic architecture of the cohort with only small shifts in category assignment. From a clinical perspective, such performance may be valuable not because it eliminates uncertainty, but because it may help reduce variation in preliminary classification, support triage, and encourage more standardized application of symptom-based criteria across clinicians and institutions (31).

These findings are broadly consistent with the expanding literature on AI in gastroenterology, where machine learning and deep learning approaches have shown strong performance in endoscopy, gastrointestinal image interpretation, disease prediction, and decision support. However, most published work has focused on neoplasia, inflammatory bowel disease, endoscopic lesion detection, or image-based classification tasks rather than symptom-defined functional disorders. The present study therefore contributes to a different but equally important area of gastroenterology by addressing whether AI can support diagnostic reasoning in disorders whose recognition depends largely on symptom patterns and structured clinical interpretation. This distinction matters because the signal environment in FGIDs is inherently more subjective than in image-based tasks, and the challenges of model transportability, interpretability, and context dependence are correspondingly greater. Accordingly, the current findings should be seen as supportive of feasibility and clinical promise, rather than as definitive evidence of universal deployability across all care settings (32,33).

Several methodological considerations temper the interpretation of these results. First, the cross-sectional design limits inference regarding temporal stability, prospective diagnostic drift, and the impact of AI-assisted classification on downstream patient outcomes such as treatment selection, symptom improvement, healthcare utilization, and cost-effectiveness. Second, the study was conducted exclusively in tertiary hospitals, where referral enrichment and diagnostic complexity may differ substantially from primary care, district hospitals, or community gastroenterology practice. Although tertiary settings are appropriate for validation under conditions of diagnostic ambiguity, the current findings may not extrapolate directly to lower-acuity populations. Third, the sample was predominantly urban, which may limit generalizability to rural populations whose symptom-reporting patterns, healthcare access, language use, and consultation styles may differ meaningfully. Because FGIDs are influenced by psychosocial, cultural, and behavioral factors, the fairness and reliability of AI classification must be evaluated across broader and more diverse populations before widespread implementation is considered (34).

An additional issue concerns model implementation and clinical governance. Even when diagnostic performance is strong, integration of AI into routine gastroenterology workflows requires attention to

data standardization, clinician trust, interpretability, and regulatory oversight. Symptom-based models are especially sensitive to documentation quality, language variability, and the representativeness of the population on which they were developed or tested. In practice, an AI system for FGID classification would need to function transparently within the electronic health record environment, allow clinician override, and be subject to periodic recalibration as documentation patterns and patient case-mix evolve. The ethical dimension is equally important: AI systems should support, not obscure, clinical accountability, and their use in disorders of gut–brain interaction must remain attentive to the psychosocial context of illness rather than reducing patients to decontextualized symptom clusters. These considerations align with the broader literature emphasizing that successful clinical AI depends not only on discrimination metrics, but also on interpretability, fairness, reproducibility, and real-world governance structures (35,36).

The study nonetheless has important strengths. It evaluated AI performance in a real-world tertiary care population, used Rome IV–based clinical diagnosis as the reference framework, included multiple common FGID categories rather than a single disorder subtype, and incorporated both structured and unstructured clinical information into model assessment. The relatively large sample size and the high category-specific agreement rates strengthen the internal credibility of the findings. At the same time, one numerical issue merits verification before final submission: the reported overall accuracy should be reconciled with the category-wise concordance counts to ensure that all summary metrics derive from the same analytic table. Resolving such internal numeric consistency will improve confidence in the reported performance estimates and strengthen the manuscript's transparency. Future work should prioritize prospective multicenter validation, inclusion of more demographically and linguistically diverse populations, external testing in non-tertiary settings, and evaluation of whether AI-assisted diagnostic support meaningfully improves patient-centered outcomes and diagnostic efficiency. Overall, the present findings support cautious but meaningful optimism that AI may enhance diagnostic consistency in FGIDs when implemented as a clinician-support tool within an appropriately governed clinical framework (37,38).

## CONCLUSION

In this multicenter cross-sectional study, the artificial intelligence–based diagnostic system showed high discriminatory ability and strong agreement with clinician-assigned Rome IV–based diagnoses for functional gastrointestinal disorders in adults attending tertiary hospitals. The findings indicate that AI has practical potential as a supportive diagnostic adjunct in symptom-based gastroenterology, particularly for irritable bowel syndrome and functional dyspepsia, where consistent recognition is often difficult in routine practice. Its role, however, should be understood as complementary rather than substitutive, with continued reliance on clinical judgment, exclusion of organic disease where appropriate, and careful attention to documentation quality, fairness, and interpretability. Further prospective and externally validated studies are needed before broad implementation can be recommended.

## REFERENCES

1. Wang Q, Xu J, Wang A, Chen Y, Wang T, Chen D, et al. Systematic review of machine learning-based radiomics approach for predicting microsatellite instability status in colorectal cancer. Radiol Med. 2023;128(2):136-148.

2. Stanzione A, Verde F, Romeo V, Boccadifuoco F, Mainenti PP, Maurea S. Radiomics and machine learning applications in rectal cancer: current update and future perspectives. World J Gastroenterol. 2021;27(32):5306-5321.

3. Xie X, Xiao YF, Yang H, Peng X, Li JJ, Zhou YY, et al. A new artificial intelligence system for both stomach and small-bowel capsule endoscopy. Gastrointest Endosc. 2024;100(5):878.e1-878.e14.

4. Tunali V, Arslan N, Ermiş BH, Derviş Hakim G, Gündoğdu A, Hora M, et al. A multicenter randomized controlled trial of microbiome-based artificial intelligence-assisted personalized diet vs low-fermentable oligosaccharides, disaccharides, monosaccharides, and polyols diet: a novel approach for the management of irritable bowel syndrome. Am J Gastroenterol. 2024;119(9):1901-1912.

5. Baumgartner M, Lang M, Holley H, Crepaz D, Hausmann B, Pjevac P, et al. Mucosal biofilms are an endoscopic feature of irritable bowel syndrome and ulcerative colitis. Gastroenterology. 2021;161(4):1245-1256.e20.

6. Jiang X, Zhao H, Saldanha OL, Nebelung S, Kuhl C, Amygdalos I, et al. An MRI deep learning model predicts outcome in rectal cancer. Radiology. 2023;307(5):e222223.

7. Nguyen NH, Picetti D, Dulai PS, Jairath V, Sandborn WJ, Ohno-Machado L, et al. Machine learning-based prediction models for diagnosis and prognosis in inflammatory bowel diseases: a systematic review. J Crohns Colitis. 2022;16(3):398-413.

8. Javaid A, Shahab O, Adorno W, Fernandes P, May E, Syed S. Machine learning predictive outcomes modeling in inflammatory bowel diseases. Inflamm Bowel Dis. 2022;28(6):819-829.

9. Jung JO, Crnovrsanin N, Wirsik NM, Nienhüser H, Peters L, Popp F, et al. Machine learning for optimized individual survival prediction in resectable upper gastrointestinal cancer. J Cancer Res Clin Oncol. 2023;149(5):1691-1702.

10. Tariq R, Dilmaghani S. Machine learning and radiomics: changing the horizon of Crohn's disease assessment. Inflamm Bowel Dis. 2024;30(10):1919-1921.

11. Kuntz S, Krieghoff-Henning E, Kather JN, Jutzi T, Höhn J, Kiehl L, et al. Gastrointestinal cancer classification and prognostication from histology using deep learning: systematic review. Eur J Cancer. 2021;155:200-215.

12. Nehme F, Feldman K. Evolving role and future directions of natural language processing in gastroenterology. Dig Dis Sci. 2021;66(1):29-40.

13. Wu L, Shang R, Sharma P, Zhou W, Liu J, Yao L, et al. Effect of a deep learning-based system on the miss rate of gastric neoplasms during upper gastrointestinal endoscopy: a single-centre, tandem, randomised controlled trial. Lancet Gastroenterol Hepatol. 2021;6(9):700-708.

14. Naz J, Sharif M, Yasmin M, Raza M, Khan MA. Detection and classification of gastrointestinal diseases using machine learning. Curr Med Imaging. 2021;17(4):479-490.

15. Gong EJ, Bang CS, Lee JJ, Baik GH, Lim H, Jeong JH, et al. Deep learning-based clinical decision support system for gastric neoplasms in real-time endoscopy: development and validation study. Endoscopy. 2023;55(8):701-708.

16. Yan T, Wong PK, Qin YY. Deep learning for diagnosis of precancerous lesions in upper gastrointestinal endoscopy: a review. World J Gastroenterol. 2021;27(20):2531-2544.

17. Wong PK, Chan IN, Yan HM, Gao S, Wong CH, Yan T, et al. Deep learning based radiomics for gastrointestinal cancer diagnosis and treatment: a minireview. World J Gastroenterol. 2022;28(45):6363-6379.

18. Zhang Y, Zhang Z, Wei L, Wei S. Construction and validation of nomograms combined with novel machine learning algorithms to predict early death of patients with metastatic colorectal cancer. Front Public Health. 2022;10:1008137.

19. Shi Q, Hao Y, Liu H, Liu X, Yan W, Mao J, et al. Computed tomography enterography radiomics and machine learning for identification of Crohn's disease. BMC Med Imaging. 2024;24(1):302.

20. Zhou S, Xie Y, Feng X, Li Y, Shen L, Chen Y. Artificial intelligence in gastrointestinal cancer research: image learning advances and applications. Cancer Lett. 2025;614:217555.

21. Ramoni D, Scuricini A, Carbone F, Liberale L, Montecucco F. Artificial intelligence in gastroenterology: ethical and diagnostic challenges in clinical practice. World J Gastroenterol. 2025;31(10):102725.

22. Matsubayashi CO, Cheng S, Hulchafo I, Zhang Y, Tada T, Buxbaum JL, et al. Artificial intelligence for gastric cancer in endoscopy: from diagnostic reasoning to market. Dig Liver Dis. 2024;56(7):1156-1163.

23. Issaiy M, Zarei D, Saghazadeh A. Artificial intelligence and acute appendicitis: a systematic review of diagnostic and prognostic models. World J Emerg Surg. 2023;18(1):59.