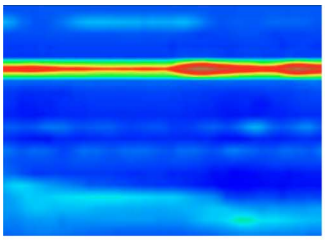
Artificial Intelligence Driven Diagnosis of Motility Patterns in High-Resolution Esophageal Manometry: A Multicentric Multidevice Study

Miguel Mascarenhas, MD, PhD[1,2,3,*], Joana Mota, MD[1,2,*], João Rala Cordeiro, PhD[4,5], Francisco Mendes, MD[1,2,3], Miguel Martins, MD[1,2,3], Pedro Cardoso, MD[1,2,3], Maria João Almeida, MD[1,2], Antonio Pinto da Costa, MD[6], Ismael Hajra Martinez, MD[6], Virginia Matallana Royo, MD[6], Benjamin Niland, MD[7], Jack Di Palma, MD[7], João Ferreira, MD[8], Guilherme Macedo, MD, PhD[1,2,3] and Cecilio Santander, MD, PhD[9]

INTRODUCTION: **Esophageal motility disorders (EMDs) are common in clinical practice, with a high symptomatic burden and significant impact on the patients' quality of life. High-resolution esophageal manometry (HREM) is the gold standard for the evaluation of functional esophageal disorders. The Chicago Classification offers a standardized approach to HREM. However, HREM remains a complex procedure, both in data analysis and in accessibility. This study aimed to develop and validate machine learning (ML) models to detect EMDs according to the Chicago Classification.**

Artificial Intelligence Driven Diagnosis of Motility Patterns in High-Resolution Esophageal Manometry

HREM is the **gold standard** for evaluating **esophageal motility disorders.**



Existing AI models are suboptimal.

618 HREM procedures
3 Hospitals
2 Devices

80% training dataset

20% testing dataset

Model performance
Accuracy
AU-ROC

Model Performance

Model	Target disorder	Accuracy	AU-ROC
Gradient Boosting Classifier	Disorders of EGJ outflow	94.2%	0.92
xGBClassifier	Disorders of peristalsis	80.9%	0.87

First worldwide multicenter and multidevice study accurately **identified disorders of the EGJ outflow and peristalsis** in HREM according to the **Chicago Classification.**

Mascarenhas et al. *Clin Trans Gastroenterol.* 2025. doi: 10.14309/ctg.0000000000000941

Clinical and Translational GASTROENTEROLOGY

[1]Precision Medicine Unit, Department of Gastroenterology, Centro Hospitalar Universitário São João, Porto, Portugal; [2]WGO Gastroenterology and Hepatology Training Center, Porto, Portugal; [3]Faculty of Medicine, University of Porto, Portugal, Porto, Portugal; [4]Instituto Telecommunications, Instituto Universitário de Lisboa, Lisbon, Portugal; [5]Department of Information Science and Technology, Instituto Universitário de Lisboa, Lisbon, Portugal; [6]Hospital Universitario Puerta de Hierro-Majadahonda, Madrid, Spain; [7]Division of Gastroenterology, University of South Alabama, College of Medicine, Mobile, Alabama, USA; [8]Faculty of Engineering of the University of Porto, Porto, Portugal; [9]Hospital Universitario La Princesa, Madrid, Spain. **Correspondence:** Miguel Mascarenhas, MD, PhD. E-mail: miguelmascarenhassaraiva@gmail.com.
*Miguel Mascarenhas and Joana Mota contributed equally to this work.
Received May 17, 2025; accepted October 9, 2025; published online October 23, 2025

METHODS: **We retrospectively analyzed 618 HREM examinations from 3 centers (Spain and the United States) using 2 recording systems. Labels were assigned by expert consensus as either disorder present or absent for 2 categories: esophagogastric junction outflow disorders and peristalsis disorders. Several ML models were trained and evaluated. ML classifiers were developed using an 80/20 patient-level stratified split for training/validation and testing. Model selection was guided by internal evaluation through repeated 10-fold cross-validation. Model performance was assessed by accuracy and area under the receiver-operating characteristic curve (AUC-ROC).**

RESULTS: **The GradientBoostingClassifier model outperformed the remaining ML models with an accuracy of 0.942 ± 0.015 and an AUC-ROC of 0.921 ± 0.041 for identifying disorders of esophagogastric junction outflow. The xGBClassifier model detected disorders of peristalsis with an accuracy of 0.809 ± 0.029 and an AUC-ROC of 0.871 ± 0.027. Performance was consistent across repeated validations, demonstrating model robustness and generalization.**

DISCUSSION: **This multicenter, multidevice study demonstrates that ML models can accurately detect EMDs in HREM. Artificial intelligence-driven HREM may improve diagnosis by standardizing interpretation and reducing interobserver variability.**

KEYWORDS: artificial intelligence; high-resolution esophageal manometry; machine learning; esophageal motility disorders

ABBREVIATIONS: AI, artificial Intelligence; EGJ, esophagogastric junction; HREM, high resolution esophageal manometry; ML, machine Learning.

Clinical and Translational Gastroenterology 2025;00:e00941. https://doi.org/10.14309/ctg.0000000000000941

INTRODUCTION

The esophagus is a muscular tube with 2 sphincters that primarily transports food and liquids into the stomach through coordinated contractions and relaxation of both sphincters, enabling digestion (1). Esophageal motility disorders (EMDs) are frequently encountered in clinical practice and are associated with a high symptomatic burden and significant impairment in quality of life. Because the diagnosis of these pathologies can be challenging, it is often delayed or even missed in the early stages, relative to the onset of symptoms. Therefore, early and accurate diagnosis is essential for improving patient outcomes and enabling effective treatment and follow-up (2).

High-resolution esophageal manometry (HREM) is currently the gold standard for evaluating patients with functional esophageal disorders. The most widely used classification system for diagnosis is the Chicago Classification (version 4.0), which provides a standardized framework based on algorithmic assignment of motility patterns (3,4). Despite significant technological advancements, HREM remains a complex procedure, with several limitations: notably, high intraobserver and interobserver variability, which hinders its availability and reproducibility, and the intrinsic difficulty of data analysis/interpretation, which can result in suboptimal diagnostic accuracy (5).

The application of artificial intelligence (AI) in gastroenterology procedures is rapidly increasing. AI models have shown potential in overcoming examination limitations, as previously demonstrated in other gastroenterology fields (6)—real-time automatic detection of colorectal polyps in colonoscopy. However, applying AI to HREM presents unique challenges.

Unlike image-based data in endoscopy or colonoscopy, HREM generates complex, high-dimensional spatiotemporal pressure signals. This requires tailored approaches for data preprocessing, feature extraction, and model design. In addition, the presence of noise and artifacts such as patient movement or sensor misalignment can affect the quality of the Data sets. The analysis process itself is time-consuming and depends on specialized clinical expertise, which limits the availability of large, high-quality Data sets for robust training and validation.

This study aimed to develop and validate an AI-based model for identifying motility disorders in HREM according to the Chicago Classification. The models were designed to detect 2 main diagnostic categories: disorders of peristalsis and esophagogastric junction (EGJ) outflow disorders, according to Chicago Classification, leveraging expert-labeled examinations from multiple international centers and devices.

METHODS

A total of 618 HREM procedures were retrospectively reviewed from 3 reference centers for functional disorders: Hospital Universitario La Princesa (Spain), Hospital Universitario Puerta del Hierro Majadahonda (Spain), and the University Hospital of South Alabama (United States).

Patients were eligible if they were over 18 years of age and underwent HREM for a clinical indication. Pediatric and pregnant patients, as well as those deemed ineligible according to each center's clinical evaluation, were excluded. Patients with active opioid use or a history of previous esophageal surgery were not excluded, to better reflect real-world practice.

These centers used 2 different high-resolution manometry systems: the Medtronic ManoScan ESO High-Resolution Manometry system and the Laborie Solar GI High-Resolution Manometry, performed by highly experienced HREM gastroenterologists from those centers, ensuring variability in device source and patient demographics.

All procedures were independently reviewed and labeled by 2 independent expert gastroenterologists according to the Chicago

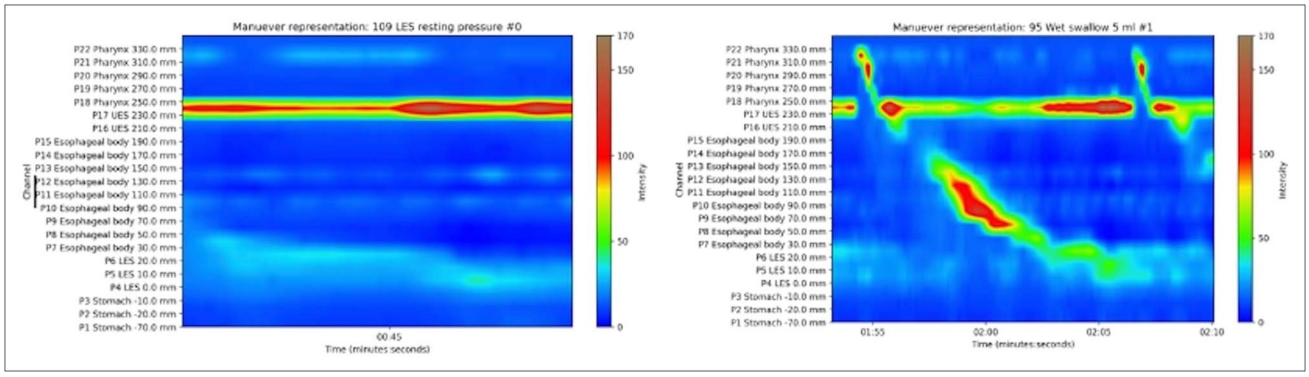


Figure 1. Examples of the data of high-resolution esophageal manometry used to train the model.

Classification version 4.0. Anatomic landmarks used to define swallow patterns were manually assessed and confirmed by the expert reviewers before inclusion in the Data set. Discrepancies were resolved through joint consensus, and only studies with complete agreement on landmark identification and swallow pattern classification were used for model training and testing. Examinations for which consensus could not be obtained were excluded. Final labels reflected 2 diagnostic categories (Figure 1 illustrates the data used in the model):

1. Disorders of EGJ outflow, including achalasia (types I, II, and III) and EGJ outflow obstruction.
2. Disorders of peristalsis, such as absent contractility, distal esophageal spasm, hypercontractile esophagus, and ineffective esophageal motility.

Patient confidentiality and anonymization were guaranteed during the analyses of the data. This study was approved by the ethics committees of the respective institutions.

HREM procedure

All procedures followed a standard HREM protocol per Chicago Classification version 4.0. Patients fasted for at least 4 hours before the procedure. Most of the examinations were initially performed in the supine position. After catheter placement, a stabilization period of at least 60 seconds took place to allow for adaptation. Subsequently, the patient was asked to do at least 3 deep inspirations to confirm catheter position. Next, a baseline period of at least 30 seconds was recorded to identify anatomic landmarks, including the upper esophageal sphincter, lower esophageal sphincter, respiratory inversion point, and basal EGJ pressure. After this, ten 5 mL wet swallows of ambient temperature water or saline were performed, ensuring a minimum interval of 30 seconds between each. A multiple rapid swallow sequence was then performed, consisting of five 2 mL wet swallows administered 2–3 seconds apart using a 10 mL syringe. After that, the patient position was changed to an upright position. A minimum of 60 seconds for adaptation was given, and catheter position was reassessed with at least 3 deep inspirations, followed by another baseline period of at least 30 seconds to identify anatomic landmarks. Subsequently, at least 5 mL wet swallows were performed with the same interval of 30 seconds between each. Finally, a rapid drink challenge was conducted, with ingestion of 200 mL of water as quickly as possible through a straw. If no major motility disorder was identified, or findings were inconsistent with clinical presentation, additional supportive maneuvers were performed (7). Each HREM procedure included at least 15 swallows (10 supine, 5 upright), generating high-resolution spatiotemporal pressure signals across 36 channels. These raw pressure signals were standardized and converted into structured feature vectors, forming the input to the machine learning (ML) models.

Model selection and tuning and model performance

When identifying procedural disorders, it is common to encounter a significant imbalance between the number of procedures without disorders and those where disorders are verified. To maximize labeling effectiveness, a human-in-the-loop/active learning strategy was adopted, prioritizing procedures that provided the most informative data for ML training.

To achieve this goal, we developed 2 independent machine-learning models—one for each disorder. Each model was trained and validated on 80% of the available procedures, with the remaining 20% held out for testing. The split was patient-level and stratified so that all sets shared similar characteristics while ensuring that any given patient appeared in only one set. We applied cross-validation within the training phase.

The models ingest a comprehensive set of features extracted from the manometry time-series recordings, covering both the resting phase and every swallowing maneuver. These features range from basic statistics such as mean and SD to more advanced descriptors like rolling-window measures and wavelet-based metrics, resulting in more than 1.600 input variables and a binary output (yes or no). Inclusive examinations were excluded in the models' training.

In the initial stage, a base model was developed using the DecisionTreeClassifier algorithm to establish baseline discrimination capabilities. The ability to distinguish between procedures with and without disorders was then significantly enhanced through an automated ML architecture optimization process. This optimization evaluated various algorithms, including XGBoost (XGBClassifier), LightGBM (LGBMClassifier), AdaBoostClassifier, Gaussian Naive Bayes (GaussianNB), Gradient Boosting Classifier, and CatBoostClassifier.

After this optimization, the final model architecture was trained and evaluated over 10 repeated runs using different random seeds to assess performance stability. Mean accuracy and area under the receiver-operating characteristic curve (AUC-ROC) values, along with SDs, were calculated.

All analyses were conducted on a system equipped with a 2.1 GHz Intel Xeon Gold 6130 processor (Intel, Santa Clara, CA) and

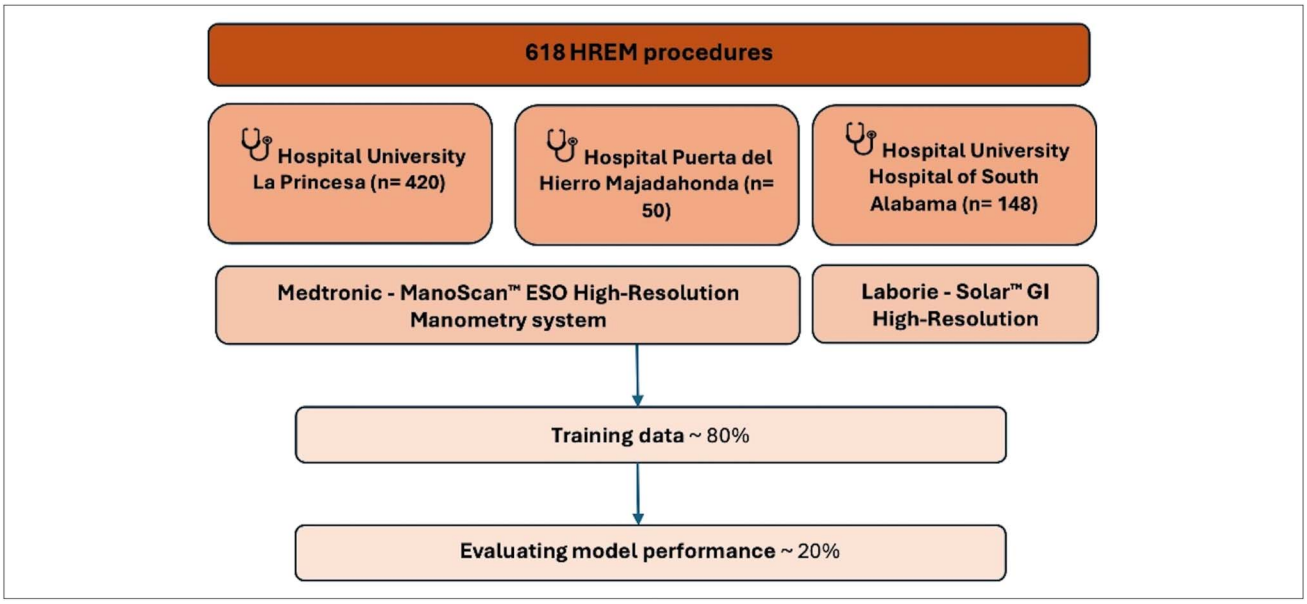


Figure 2. Graphical flowchart of the study. HREM, high-resolution esophageal manometry.

dual NVIDIA Quadro RTX 8000 GPUs (NVIDIA, Santa Clara, CA), ensuring efficient processing and model training.

A graphical representation of the study design is shown in Figure 2.

Statistical analysis

The performance of the ML models was evaluated through their accuracy and AUC-ROC in differentiating disorders of peristalsis and disorders of EGJ outflow. Performance metrics were compared against the expert consensus labels.

RESULTS

A total of 618 HREM examinations were used for the Data set. The mean age was 58 ± 14 years, and 56% of patients were women. Among these, 54 were classified as EGJ outflow disorders, and 187 as peristalsis disorders, based on the final consensus. In total, the Data set comprised 420 examinations from Hospital Universitario La Princesa, 50 from Hospital Puerta del Hierro Majadahonda, and 148 from Health University Hospital of South Alabama, ensuring diversity in patient origin and device usage.

Table 1 summarizes the agreement between the ML model predictions and expert-labeled final diagnosis. The GradientBoostingClassifier achieved an accuracy of 0.942 ± 0.015 and an AUC-ROC of 0.921 ± 0.041 for the identification of EGJ outflow disorders. The xGBClassifier model detected peristalsis disorders with an accuracy of 0.809 ± 0.029 and an AUC-ROC of 0.871 ± 0.027. Additional cross-validation runs confirmed consistent model performance, and further validation per center is planned to assess generalizability. Table 2 displays the SD obtained across 10 independent training iterations for the LGBMClassifier and CatBoostClassifier models.

Table 1. The agreement between the machine learning model predictions and the expert-labeled final diagnosis

	Final diagnosis		
Model classification	Yes	No	Inconclusive
Disorders of EGJ	54	500	64
Disorders of peristalsis	187	373	58

EGJ, esophagogastric junction.

DISCUSSION

This study was conducted in a multicenter, multidevice setting, demonstrating the feasibility of creating AI models that can be applied globally, with the potential to improve diagnostic accuracy and reduce variability in the identification of EMDs in HREM according to the Chicago Classification.

HREM is the current gold standard for diagnosing EMDs. Its current classification system, used worldwide, is the Chicago classification (version 4.0), which provides a widely accepted algorithmic scheme using metrics from HREM to achieve the diagnosis (7). However, its application remains expert and manual-dependent, which introduces substantial intraobserver and interobserver variability that limits reproducibility and accessibility (4). This reliance on manual interpretation may also reduce accuracy and hinder the widespread implementation of HREM in clinical practice.

Table 2. The SD values from the 10 model training iterations for LGBMClassifier and CatBoostClassifier models

Model	Disorder type	Accuracy Mean ± SD	AUC-ROC Mean ± SD
LGBMClassifier	EGJ outflow disorder	0.940 ± 0.014	0.931 ± 0.038
CatBoostClassifier	Peristalsis	0.797 ± 0.033	0.839 ± 0.033

AUC-ROC, area under the receiver-operating characteristic curve; EGJ disorder, esophagogastric junction outflow disorders.

AI use in medical fields has grown exponentially in recent years, with gastroenterology being no exception. Several commercially available AI systems are now routinely used, such as those used for real-time detection of polyps during endoscopy. In the motility field, AI holds great promise but also faces specific challenges. Currently, only a few studies have investigated this area, revealing a significant gap in the existing literature. Indeed, some pilot studies have demonstrated the utility of ML models in identifying normal motility patterns and swallowing types (8–10). More recently, models have been developed to differentiate EMDs according to the Chicago Classification (9). However, these studies were mostly limited to single centers, relied on a single device manufacturer, and were based on the outdated Chicago Classification (version 3.0). Our study addresses these limitations by developing and validating a ML model using a multicenter, multidevice Data set aligned with the most recent Chicago Classification. This approach enhances the model's potential for interoperability and external validity. To our knowledge, this is the first study to evaluate the performance of ML models in differentiating disorders of peristalsis and disorders of EGJ outflow in HREM data across multiple institutions and devices.

This study has several strengths that deserve acknowledgment, notably the inclusion of 618 HREM examinations (performed based on the Chicago Classification 4.0) and had a Data set that included a diverse population from 2 different continents, 3 different medical institutions, and 2 different medical devices. The diversity of this Data set enhances reproducibility and supports potential global application. We used 80% of the procedures for model training, whereas the remaining 20% were used for testing, ensuring class stratification to maintain similar distributions between sets. The optimized model was trained and evaluated 10 times to ensure robustness and achieved excellent results.

Despite these contributions, several limitations must be acknowledged.

First, the absence of a fully independent external validation cohort remains a major limitation. Although the Data set included examinations from various institutions across different countries and device platforms, all participating centers contributed data to both the training and testing phases. This may introduce bias by potentially overestimate the model's generalizability. Future research should focus in obtaining an independent Data set from centers not involved in model development to more accurately assess real-world performance before broader clinical implementation.

Second, the class imbalance, particularly in EGJ outflow disorders, which represented a smaller proportion of the Data set, may limit sensitivity and generalizability. To mitigate this potential bias, we applied patient-level splitting and repeated cross-validation. Nevertheless, the absence of a formal sample size calculation further reinforces the need for larger, balanced, and external cohorts to confirm the reproducibility of our findings.

Third, the Data set is inherently affected by variability in patient cooperation and anatomical characteristics, which can influence test quality and, in such cases, may necessitate substantial reliance on expert interpretation. Moreover, only examinations with full expert consensus were included in training, which may reduce generalizability in ambiguous or borderline real-world scenarios. Another limitation is the exclusive reliance on pressure-based metrics. EGJ distensibility (measured by FLIP) and impedance, were not incorporated, despite their recognized importance for certain diagnoses. Future studies should explore the integration of multimodal data to enhance diagnostic accuracy and clinical applicability.

From a technical standpoint, the lack of transparency of ML models often lead them to be perceived as black-box systems, which poses challenges in critical areas such as medicine. The lack of interpretability makes it difficult for humans to fully trust the results (11). Although explainable AI offers potential solutions, its application in motility disorders is hindered by the inherent subjectivity of certain examination findings. Improving model interpretability will be essential for future clinical deployment. In subsequent iterations, we plan to incorporate established explainability techniques such as SHapley Additive exPlanations, to quantify the contribution of individual features to the ML prediction, together with visual representation methods (e.g., heatmaps and temporal importance plots) to highlight the most relevant segments of the manometry signal (12).

Our proof-of-concept was designed to evaluate the feasibility and generalizability of our models across centers and devices. The aim is to provide clinicians with transparent, reproducible, and actionable insights that complement their expertise, and ultimately promote the broader adoption of AI in esophageal motility diagnostics. The potential integration of these models into clinical workflows also warrants consideration. One feasible application would be as a second-reader tool within existing HREM analysis platforms, designed to flag cases with uncertain or borderline findings for additional expert review. In resource-limited settings, such models could serve as an accessible decision-support tool to enhance the diagnostic capabilities of less experienced practitioners. Nevertheless, effective clinical deployment will require rigorous prospective validation, the development of intuitive user interfaces, and seamless interoperability with existing reporting systems.

In conclusion, this is, to our knowledge, the first international proof-of-concept study to develop and validate a ML model for the identification and classification of peristalsis and EGJ outflow disorders in HREM, based on the Chicago classification (version 4.0). AI-assisted interpretation of HREM has the potential to significantly enhance diagnostic standardization, reduce interobserver variability, and ultimately facilitate broader implementation and access to high-quality motility assessment across different healthcare settings.

CONFLICTS OF INTEREST

Guarantor of the article: Miguel José da Quinta e Costa de Mascarenhas Saraiva, MD, PhD

Specific author contributions: M.M. (Miguel Mascarenhas) and J.M.: equal contribution in study design, image extraction, drafting of the manuscript, and critical revision of the manuscript. F.M., T.R., P.C., M.M. (Miguel Martins), M.J.A.: bibliographic review, image extraction, critical revision of the manuscript. J.R.C.: construction and development of the machine learning models, statistical analysis, critical revision of the manuscript. I.H.M., V.M.R., B.N., G.M., C.S., and J.D.P.: study design, critical revision of the manuscript. All authors approved the final version of the manuscript.

Financial support: None to report.

Potential competing interests: None to report.

Study Highlights

WHAT IS KNOWN

- ✓ High-resolution esophageal manometry (HREM) is the gold standard for evaluating esophageal motility disorders.
- ✓ HREM interpretation is complex and depends on expert manual review.
- ✓ Existing artificial intelligence models are suboptimal.

WHAT IS NEW HERE

- ✓ Our group developed a multicenter and multidevice machine learning model based on the Chicago Classification version 4.0.
- ✓ The model's diagnostic yield may contribute to standardized, reproducible, and automated HREM analysis.

REFERENCES

1. Margolis KG, Picoraro JA. Development of gastrointestinal motility. In: Polin RA, Abman SH, Rowitch DH, et al, eds. Fetal and Neonatal Physiology. 5th edn. Amsterdam, Netherlands: Elsevier; 2017:881–8.e2.
2. Patel DA, Yadlapati R, Vaezi MF. Esophageal motility disorders: Current approach to diagnostics and therapeutics. Gastroenterology 2022;162(6): 1617–34.
3. Kröner PT, Engels MM, Glicksberg BS, et al. Artificial intelligence in gastroenterology: A state-of-the-art review. World J Gastroenterol 2021; 27(40):6794–824.
4. Zifan A, Lin J, Peng Z, et al. Unraveling functional dysphagia: A game-changing automated machine-learning diagnostic approach. Appl Sci 2023;13(18):10116.
5. Kou W, Carlson DA, Baumann AJ, et al. A multistage machine learning model for diagnosis of esophageal manometry. Artif Intell Med 2022;124:102233.
6. Kou W, Galal GO, Klug MW, et al. Deep learning-based artificial intelligence model for identifying swallow types in esophageal high-resolution manometry. Neurogastroenterol Motil 2022;34(7):e14290.
7. Yadlapati R, Kahrilas PJ, Fox MR, et al. Esophageal motility disorders on high-resolution manometry: Chicago Classification version 4.0. Neurogastroenterol Motil 2021;33(1):e14058.
8. Lundager FH, Tack J, Blondeau K, et al. Patients with esophageal motility disorders show distinct patterns based on axial force measurements. Dig Dis Sci 2012;57(11):2929–35.
9. SurdeaBlaga T, Sebestyen G, Czako Z, et al. Automated Chicago Classification for esophageal motility disorder diagnosis using machine learning. Sensors (Basel) 2022;22(14):5227.
10. Wang Z, Hou M, Yan L, et al. Deep learning for tracing esophageal motility function over time. Comput Methods Programs Biomed 2021; 207:106212.
11. Hassija V, Chamola V, Mahapatra A, et al. Interpreting black-box models: A review on explainable artificial intelligence. Cogn Comput 2024;16(1):45–74.
12. Lundberg SM, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell 2020;2(1): 56–67.